A Cross-Cultural Analysis of Music Structure

Mi Tian

A thesis submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London
United Kingdom

July 2016

“Te ofrezco la lealtad de un hombre que nunca ha sido leal.”

Abstract

Music signal analysis is a research field concerning the extraction of meaningful information from musical audio signals. This thesis analyses the music signals from the note-level to the song-level in a bottom-up manner and situates the research in two Music information retrieval (MIR) problems: audio onset detection (AOD) and music structural segmentation (MSS).

Most MIR tools are developed for and evaluated on Western music with specific musical knowledge encoded. This thesis approaches the investigated tasks from a cross-cultural perspective by developing audio features and algorithms applicable for both Western and non-Western genres. Two Chinese Jingju databases are collected to facilitate respectively the AOD and MSS tasks investigated.

New features and algorithms for AOD are presented relying on fusion techniques. We show that fusion can significantly improve the performance of the constituent baseline AOD algorithms. A large-scale parameter analysis is carried out to identify the relations between system configurations and the musical properties of different music types.

Novel audio features are developed to summarise music timbre, harmony and rhythm for its structural description. The new features serve as effective alternatives to commonly used ones, showing comparable performance on existing datasets, and surpass them on the Jingju dataset. A new segmentation algorithm is presented which effectively captures the structural characteristics of Jingju. By evaluating the presented audio features and different segmentation algorithms incorporating different structural principles for the investigated music types, this thesis also identifies the underlying relations between audio features, segmentation methods and music genres in the scenario of music structural analysis.

Statement of Originality

I, Mi Tian, confirm that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. I acknowledge the helpful guidance from my supervisor, Prof. Mark Sandler.

Acknowledgments

First and foremost, I would like to thank my supervisor, Prof. Mark Sandler for the tremendous support throughout the four years. Thank you for giving me freedom to develop my own motivation and in the meantime, always kindly reminding me to think over the most detailed questions which otherwise I would have taken for granted such as "why using a Hann window and why a 1024-point FFT".

I would like to thank Dr. Dawn. A. A. Black, for introducing me to such an amazing research field and for being an awesome advisor. Thanks to Dr. Haojun Ai, who supervised my undergraduate final project, for teaching me from scratch how to do research.

Thanks to Prof. Xavier Serra and your awesome team, Ajay, Kainan, Rafael for hosting my visit in Music Technology Group. Special thanks to Xavier and Rafael for the inspiring discussions. Thanks to Prof. Roger Dean for inviting me to visit your lab. Thanks everyone in MARCS especially Yvonne for making it such an enjoyable visit. Special thanks to the anonymous reviewers of the papers I have submitted for your great input and the most helpful comments.

Thanks to my main funding party, China Scholarship Council, to support this research. I also gratefully acknowledge financial support for my research travels and publications from various funding parties (ordered alphabetically): EPSRC C4DM Travel Funding, EPSRC Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1), EPSRC Platform Grant on Digital Music (EP/K009559/1), European Research Council project CompMusic, International Society for Music Information Retrieval Student Grant, QMUL Postgraduate Research Fund, QMUL-BUPT Joint Programme Funding and Women in Music Information Retrieval Grant.

Thanks to everyone in C4DM: Beici, Bob, Chunyang, Chris, Chris, Dan, Dave, Di, Elio,

Emmanuel, Fran, George, Holger, Janis, Jiajie, Jordan, Julie, Katerina, Ken, Luwei, Maria, Matthias, Matthieu, Mike, Pablo, Peter, Sebastian, Sid, Simin, Siying, Steve, Tian, Thomas, Veronica, Yading... I would like to thank Dr. George Fazekas for the help and advice during my tough starting days. Thanks to Prof. Geraint Wiggins, Prof. Mark Plumbley and Prof. Elaine Chew for the occasional chats, advisory and friendly. Thanks to Chris, Dan, Ken, Matthias, Matthieu and Sebastian for being the nicest friends and seniors. Thanks to Chunyang, Siying, Tian and Yading for so much lovely time spent together in and outside the lab. Thanks to Beici, Fran, Jiajie, Jordan, Manos, Maria, Matthias, Matthieu, Pablo, Sid, Siying, Veronica and Yading for the pub nights, board game nights, cooking parties, ping pongs... Thanks to Nela and the fabulous G.Hack team, for your dedication in the exchange of knowledge and joy for female engineers. I feel proud to be one of you.

Thanks to Jieyu, Lei, Lifeng, Shan, Tian, Yansha, Yanru and Zhen for being the most awesome flatmates and for being family. Thanks to Valeria for so many lovely hanging-outs in England and in Italy and for being such a precious friend. Thanks to my old buddies Cheng, Jing, Qiao, Xiaoting, Shanhui, Yaling, Yuanhua, Yujie, Yuling, for keeping me company across thousands of kilometres. I would like to thank Daniel for all the love and support. Thank you for playing me the funniest songs on those dreadfully long revision evenings, for the continuous encouragement, for putting up with my bad mood and for everything not named here.

Finally, thank my parents for the unconditional love and for always supporting me to pursue what I want. I always feel proud to be your daughter. Hope one day not too far from now, I can use “Dr” to title my name, just like you guys do.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3D	Three-Dimensional
ACF	Autocorrelation function
ASR	Automatic speech recognition
AOD	Audio onset detection
AW	Adaptive whitening
BER	Broadband energy rise
BPM	Beats per minute
BPO	Bins per octave
CB	Critical band
CC	Constrained clustering
CD	Complex domain
$CDSF_L$	Linear fusion of onset detection algorithm Complex domain and SuperFlux
CJ	Chinese Jingju song database
CM	Complex mixture
CNMF	convex non-negative matrix factorisation
CNN	Convolutional neural network
CV	Cross validation
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
EVD	Eigen value decomposition

ERB	Equivalent rectangular bandwidth
F0	Fundamental frequency
FDLP	Frequency domain linear prediction
FFT	Fast Fourier transform
FN	False negative
FP	False positive
FT	Fourier transform
GMM	Gaussian mixture model
GC	Gammatone contrast
GCCs	Gammatone cepstral coefficients
GF	Gammatone featureset
GT	Ground truth
HFC	High frequency content
HPSS	Harmonic-percussive source separation
HMM	Hidden Markov model
IAA	Inter-annotator agreements
IQRs	interquartile ranges
JP	Jingju percussion database
JPB	Onset detection based presented by Juan P. Bello et al.
k-NN	K-nearest neighbours
LDA	Linear discriminant analysis
LP	Low-pass filtering (boolean parameter)
LPC	Linear predictive coding
LPCCs	Linear predictive coding coefficients
LSTM	Long Short-Term Memory
MIR	Music information retieval
MF	Median filter

MFCCs	Mel-frequency cepstral coefficients
MSA	Music strctural analysis
MSS	Music structural segmentation
NC	Novelty curve
NMF	Non-negative matrix factorisation
NN	Neural network
NPP	Non-pitched percussive
ODF	Onset detection function
OWL	Web Ontology Language
PNP	Pitched non-percussive
PP	Pitched percussive
PCA	Principal component analysis
PD	Phase deviation
PF	Polynomial fitting
QMVP	Queen Mary Vamp Plugins
QN	Quadratic novelty
RDF	Resource Description Framework
RMS	Root mean square
RNN	Recurrent neural network
RP	Recurrence plot
RT	tempogram featureset
RWC	Real world computing
SB	Onset detection database presented by Sebastian Böck et al.
SC	Spectral constrast
SD	Spectral difference
SF	SuperFlux
S-IA	SALAMI Internet Archive

SSM	Self-similarity matrix
std	Standard deviation
STFT	Short-time Fourier transform
SW	Semantic Web
TCCs	Tempogram cepstral coefficients
TDLP	Time domain linear prediction
TF	Time-frequency
TFR	Time-frequency representation
TI	Tempo intensity
TIR	Tempo intensity ratio
TP	True positive
TPCs	Tempogram principal coefficients
WT	Adaptive whitening

List of Symbols

A_g	Gain in a transform of an electronic signal
B	Number of bins in the chromatic scale
bt	backtracking threshold
C	Number of subbands in the Gammatonegram bandwise processing
$C_b(z)$	Critical bandwidth of the zth band
C_c	Number of clusters used in a clustering algorithm
$E(z)$	Energy in the zth band
$ERB(fc)$	Equivalent Rectangular Bandwidth of the band with center frequency fc
f	Frequency in Hz
fc	Centre frequency of the filter bank
$F(k,q)$	A triangle filterbank used in the SuperFlux onset detector
f_c	Cutoff frequency for low-pass filtering
$\mathscr{F}(f)$	Fourier transform of a time signal x(t)
G(n)	Gammatonegram
H	The activation matrix after NMF decomposition with size (R,P)
H_G	Hop size between consecutive summation windows (in frame)
k	Frequency index of the SFTF
K	Number of frequency bins after the STFT
k_G	Size of Gaussian Kernels used for calculating novelty curves from SSMs

l	Combination weight for linear fusion
L	Number of subbands in the tempogram bandwise processing
l_h	Maximum filter size used to separate the harmonic component in HPSS
l_p	Maximum filter size used to separate the percussve component in HPSS
M	Number of individual filters in a filterbank
minBW, e_q	Galsberg and Moore parameters used to calculate $ERB(f_c)$
m_d	The embedding dimension in a multi-dimensional time series
m_f	Memory coefficient for the adaptive whitening
m_h	Median filter size used to separate the harmonic component in HPSS
m_p	Median filter size used to separate the percussive component in HPSS
N_G	Length of the summation window (in frame)
N_F	Number of filterbanks divided into each subband
nfft	Number of FFTs
p	Order of the Gammatone filter
princarg	A function mapping the phase deviation to $[-\pi, \pi]$ range
r	Floor value for the adaptive whitening
R	Number of ranks in a NMF decomposition
$S(i, j)$	The self-similary square matrix
s_r	Feature rate used to resample tempogram features
sens	Sensitivity for peak picking
th_a	Threshold for the a parameter of the quadratic model in polynomial fitting
th_c	Threshold for the c parameter of the quadratic model in polynomial fitting
thresh	Local threshold yielded by adaptive thresholding
V	A feature matrix with size (N,P)
W	The basis matrix after NMF decomposition with size (N,R)
$\mathscr{W}$	A Window functioin
W_a	Rectangular window used to calculate the autocorrelation based tempogram

W_f	Hann window used to calculate the Fourier transform based tempogram
W_H	Hop size between consecutive windows (in frame)
W_N	Length of the window (in frame)
x(n)	A finite duration sequence where $0 \leq n \leq N-1$
x(t)	A time domain signal
X(n,k)	Spectrogram of x(t)
$\widehat{x}_i$	A multi-dimensional time series
$X_h(n,k)$	Harmonic component after HPSS
$X_p(n,k)$	Percussive component after HPSS
$X_r(n,k)$	Residue component after HPSS
$X_w(n,k)$	X(n,k) after adaptive whitening
Y(n,k)	Magnitide spectrogram of x(t)
α	Normalisation coefficient for exponential weighting
β	Separation coefficient for HPSS
δ	Constant offset for adaptive thresholding
ϵ	A smal constant introduced to avoid zero division
Λ	Time lag (in second)
μ	Neighbourhood size of STFT frames for calculating deviations
$\phi(n,k)$	Phase of X(n, k)
$\phi'(n,k)$	Rate of phase change at the nth frame
ψ	Median window size
τ	Tolerance window length for late fusion (in second)
τ_t	Tempo (in BPM)
θ	Energy threshold used in the Broadband energy rise onest detector
ϱ	Exponential fraction root used to calculate the *tempogram intensity* feature
ζ	Time delay in the multi-dimensional time series

Chapter 1

Introduction

1.1 Scope and Motivation

Music plays an indispensable role in our everyday lives. *Music information retrieval* (MIR) is an interdisciplinary research field concerning the extraction of meaningful information from music, including the musical content itself as well as its ancillary knowledge sources related to such as the artist details and the performance venues. It has many real-world purposes such as music retrieval, indexing and browsing, recommendation and audio production. While early MIR research focused on working with symbolic representations such as MIDI [Swi97], progress in signal processing and data mining in recent decades has enabled the processing and analysis of the digitised acoustical sound signal [SGU14]. The scope of this thesis is the acoustic aspect, where we will analyse the music content from its audio-based representations.

Music is primarily an event-based phenomenon. It comprises a series of musical elements that unfold in time such as the melodic stanzas and harmonic passages. These elements then collectively composite the overall aesthetics of music with vivid structural characteristics. In both listening and analysis activities, one can give music boundaries to pinpoint specific musical units and to facilitate the music content discovery with certain

within-piece sectional information.

Music structure can be rendered at different hierarchies with different levels of abstractions. It can be broken all the way down from large compositional parts, such as sections, into smaller elements such as bars and beats whose basic units includes the music notes. What occurs at low-level hierarchies is embodied in high-level ones [Nar91]. While the lower levels of the hierarchy relate closely to the physical properties of the sound signal, the higher levels can reflect how humans perceive and interpret the music contents. The overarching goal of this thesis is to analyse the music structure at various abstractions from polyphonic audio recordings in a *bottom-up* manner.

Previous works have shown that MIR systems rely on musical knowledge to yield meaningful results [Bel03]. The term "knowledge-based" originates from the field of artificial intelligence (AI) [CF81]. It refers to the ability to represent, manage and use knowledge within a given problem domain. *Source knowledge* characterising the music itself can be used to inform the design of the analysis methodologies of the MIR system. For example, Eggink and Brown introduced a knowledge-based fundamental frequency (F0) detection system where source knowledge about the instrument properties and tone durations was used to select the correct F0 from the candidates [EB04]. Another type of knowledge is the *contextual knowledge*. It can assist putting the analysis in specific contexts so that algorithms are able to make informed decisions in specific scenarios or for required applications. For example, musical phenomena observed in the recent past can be used as contextual information to assist the prediction of current events in a probabilistic manner [LS08; MND09]. This thesis proposes to use knowledge-based feature extraction techniques to enhance the *contextual awareness* to solve MIR problems. Specifically, we situate this goal in two tasks: *audio onset detection* and *music structural segmentation*.

Musical notes describe the music content at the lowest meaningful level. The melody and harmony are largely determined by the sequential occurrence of the pitched notes; the music beat, tempo and rhythm are set by the temporal recurrence patterns of note

onsets. Much effort has been made by the MIR community to approach the automatic detection of note onsets [Bel+05; Dix06; Mar+14]. Many methods detect the presence of note onsets based on changes of spectral properties. Machine learning techniques are also employed by more recent methods. The majority of these methods, however, either employ only a single feature to encode the entire onset properties, or require an expensive training process for the relevant information to be learnt from the audio signals.

A promising solution to onset detection problems hence lies in the *fusion* of different onset-related features or detection rules for a more comprehensive representation of onset characteristics and a more precise standard to select onsets [Deg+09; ZG11]. A widely accepted definition for data fusion is provided by the Joint Directors of Laboratories (JDL) [Fra87]: "A multi-level process dealing with the association, correlation, combination of data and information from single and multiple sources to achieve refined position, identify estimates and complete and timely assessments of situations, threats and their significance." This thesis will investigate a fusion-based approach for improved onset detection. We will present new onset detection algorithms by combining different features to summarise the music signal, as well as by combining the results from multiple detections. The underlying hypothesis is that the inferences made by one constituent feature or method can provide complementary contextual information to those by another.

The structure of a music piece typically refers to its *song structure* or *musical form*. The foundation of the structure of popular music is the "verse" and "chorus" form. Although much work has been presented to discover music structure in recent years, the focus of the majority of the existing research is Western pop music [Smi14; Jia15; Ong06; Nie15; Gro12; MND09; ME14a]. Most state-of-the-art methods attempt to analyse the music structure by finding repetitive patterns at a segment level. Such strategy, however, does not address the form of the so-called *through-composed* music (a term derived from the German word "durchkomponiert") which is relatively continuous and non-repetitive. The features employed describe attributes observed mainly for Western pop music such

as the chord progressions therefore may be less effective beyond the Western scenario. Meanwhile, this thesis aims to analyse music structure setting aside the assumption that sectional repetitions must present.

There has been a dedication of recent research on the algorithms for music structural analysis to improve the precision of structural discoveries. To this end, we propose to improve the representation and interpretation of the underlying structural information by investigating new feature descriptors and segmentation methods. In this process, the source knowledge related to the music will be used to develop novel audio features, and the properties of such features can be used as the contextual knowledge to assist the design of segmentation algorithms.

The current MIR research is predominantly Western-centric [Dow03b; SGU14]. Understanding of the music characteristics can be genre-dependent and an MIR algorithm designed for one corpus may have weak assumptions for another. The acquisition of non-Western datasets and methodologies is particularly valuable to combat the bias towards Western popular music genres within current MIR research [Dow03b]. In the recent decade, a few non-Western traditional music corpora have been included in the MIR research, such as Turkish Makam, Indian Carnatic, Arabian Andalusian and Chinese Jingju music. Among these music cultures, the Chinese music is the one that has had the least contact with the Western academic school and most existing studies have been published only in Mandarin and available in China [Ser11; Tia+13]. Jingju, also referred to as Beijing Opera or Peking Opera, is one of the most representative Chinese traditional music genres combining singing, dance and theatre art. 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 recent years [Rep+14].

Jingju orchestra uses very characteristic instruments that are traditional to the Chinese music repertoire. This hence broadens the variety of existing research subject and confronts the retrieval of the music content with new challenges. Unlike Western classical or popular music, the bar-beat structure is rather loose for Jingju and the music

does not follow a "chorus-verse" or "ABBA" structure. The music form of Jingju can have distinctively different characteristics from Western music in the sense that harmony or chordal structure is hardly present in Jingju songs at the segment-level. The music form is composed, however, mainly of its sung sections with narrative lyrics and vivid melodies [Wic91]. It is what musicians call "through-composed". This music melody is characterised by a *heterophonic* texture, i.e., variations introduced by different instruments exist in the unitary basic melody. One of the main characteristics of Jingju aesthetics is the rhythmicity that directs the overall performance [Rep+14]. With an absence of bars or measures as defined for Western music, Jingju uses very specific percussion and metrical patterns to set its timing. Therefore, as a genre newly brought under the MIR spotlight, Jingju has many unique musical properties and can offer intriguing research questions to complete the existing methods and paradigms.

So far we have highlighted the importance of knowledge awareness in a music analysis system. One aspect of the musical knowledge we are particularly interested in is the music genre. We hereby propose a *cross-genre* approach for the research carried out of this thesis. By introducing Jingju into our analysis framework, we do not intend to investigate into methods suitable for this genre alone. Because in this way, we would only be introducing a new bias to replace the existing Western bias instead of confronting the semantic gap between different music types. Rather, this thesis will be focused on investigating novel features and algorithms to address the overarching goal of music structure analysis for different music genres. To evaluate the analysis beyond the Western context, new databases of Jingju music have been constructed featuring its instrument ensembles for an onset detection study and its songs for a structural segmentation study. Besides assessing the presented features and algorithms, we also aim to identify the new challenges for existing MIR paradigms by evaluating our experiments on both existing datasets and the new Jingju ones.

1.2 Outline of the Thesis

This thesis is outlined as follows:

Chapter 2 reviews the background of the computational music content analysis and the signal processing methods used in this thesis. It starts by introducing the time-frequency analysis of music signals and proceeds to introduce the MIR tasks focusing on the two main targets of this thesis: audio onset detection and music structural segmentation. It also surveys common methods for evaluating MIR tasks.

Chapter 3 is devoted to presenting the musical background and the design of two evaluation datasets for this thesis. A survey of the current MIR research for Jingju are firstly presented, by which we are aiming to give an essential background of our work and to identify the research goal of this thesis. We will introduce the basic musical elements of Jingju and identify how they collectively constituent the music structure – from the note level to the piece level. Two annotated Jingju datasets are presented along with a survey of existing Western ones from related work. The first dataset is designed for music onset detection tasks consisting of ensembles of the four major Jingju percussion instruments. The second one is a music structural segmentation dataset with audio samples collected from commercial CDs and annotations made by professional listeners. This chapter hereby identifies the source knowledge of two different roles. The music database holds knowledge related to problem solving, from which observations can be made. The annotations are explicitly related to the questions the algorithms attempt to address and will be used as the ground truth to evaluate their behaviours.

Chapter 4 presents original work on audio onset detection using fusion strategies and evaluates the new onset detectors in the scenario of Western and Chinese music. By fusion, distinct knowledge sources, in our case different kinds of spectral information or features are merged together to yield a more comprehensive representation of the onset properties. An exhaustive parameter analysis is carried out aiming to assess the effects of signal processing methods in an AOD framework as well as to discover the optimal

uses of signal processing methods in the scenario of different music types.

Chapter 5 is focused on feature design for music structural analysis. This chapter starts with revisiting commonly used features designed for Western pop music in the scenario of Jingju and proposes using chroma features with different *bins-per-octave* (BPO) settings. It then proposes *harmonic-percussive source separation* (HPSS) as a feature enhancement technique for music structural segmentation and analyses its effect in the scenario of structural description. Two new sets of features are presented to summarise the rhythmic and auditory aspects of the music content. The presented features are extracted in a *Vamp plugin* environment with the *Resource description framework* (RDF) web ontology tools [Onl09] and evaluated in a segmentation framework. We also analyse the effects of different audio features and signal processing methods for the investigated music genres.

Chapter 6 aims to investigate different segmentation algorithms in an integrated music structural analysis system. Five state-of-the-art methods are included covering different structural hypotheses. These methods are evaluated using features introduced in Chapter 5 on three datasets consisting Western and Chinese music. This chapter hereby attempts to answer the following question: how do audio features capture the structure of different music types and how do different segmentation methods interpret the reflected structural patterns? This chapter will also investigate the effects of signal processing parameters involved in the segmentation algorithms.

Finally, Chapter 7 concludes this thesis and identifies some directions for future work.

1.3 Main Contributions

This thesis describes signal processing methods for music structure analysis, where the use of music knowledge plays a key role. The main contributions of this thesis is summarised as below:

Chapter 3 presents two Jingju datasets to replenish the existing Western-centric evaluation corpora. The first one consists of studio recordings of Jingju percussion instruments. It is used to evaluate the onset detection study carried out in this thesis. We also identify that it can be used to evaluate other MIR tasks such as instrument recognition and percussion transcription. The second dataset is designed for MSS with audio samples collected from commercial CDs. The structural annotations are created by professional musicians with the associated metadata recorded.

New onset detection methods based on fusion are presented in Chapter 4 outperforming the state-of-the-art algorithms. By fusion, multiple knowledge sources are brought together to complement the inference each other makes. A thorough evaluation of involved signal processing components is carried out with their effects extensively tested. We highlight that parameter configuration for the underlying signal processing system can be a significant factor for a successful onset detection and that some default parameter configurations in existing systems have adverse effects. In this experiment, various processes responsible for the fusion, onset detection, parameter selection as well as result evaluation are conceptualised and fully automated in an integrated system relying on Audio Features Ontology and Vamp Plugin Ontology. We demonstrate that Semantic Web technologies can be useful tools for the analysis of large-scale audio signals.

Novel audio features are developed for music structural description in Chapter 5. The new rhythmic features demonstrate different temporal patterns for different genres. We also indicate further applications of this feature set. Features extracted from the Gammatone auditory representation are proved to be at least as effective as the commonly used structural descriptors such as MFCCs and chroma features. The advantage of this feature set is more pronounced for music characterising singing voice. A critical analysis of existing chroma features for Jingju is carried out. We find that it is an effective feature descriptor for Jingju too although different bins per octave settings should be used.

In Chapter 6, We contribute a cross-cultural evaluation of the state of the art in MSS algorithms with various audio features. We show that although repetition-based

segmentation methods are very successful for Western pop music, they are less effective for Jingju with its lack of chordal structures. Novelty- and homogeneity-based methods (introduction of these methods will be given in Section 2.5.3) can be more genre-invariant when used to summarise the similarity-level music structure. The outcomes of this chapter 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 signals such as the genre and the level of music structure to analyse.

1.4 Associated Publications

This thesis covers the work carried out by the author between September 2012 and June 2016 at Queen Mary University of London. The majority of the work presented in this thesis has been published in peer-reviewed conferences and journals:

- M. Tian and M. B. Sandler, 2016, "Towards music structural segmentation across genres: features, structural hypotheses and annotation principles", Intelligent Systems and Technology, Special Issue on Intelligent Music Systems and Applications, ACM Transactions on, Volume 8, Number 2, pp. 23-41.

- M. Tian and M. B. Sandler, 2016, "Music structural segmentation across genres with Gammatone features, in Proceedings of International Society on Music Information Retrieval Conference", pp. 561-567, New York, USA.

- M. Tian, G. Fazekas, D. A. A. Black and M. B. 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, pp. 419-423, Brisbane, Australia.

- L. Yang, M. Tian and E. Chew, 2015, "Vibrato characteristics and frequency histogram envelopes in Beijing Opera singing", in Proceedings of the 5th International Workshop on Folk Music Analysis, pp. 139-140, Paris, France.

- M. Tian, G. Fazekas, D. A. A. Black and M. B. Sandler, "Design and evaluation of onset

detectors using different fusion policies", in Proceedings of International Society on Music Information Retrieval Conference, pp. 631-636, Taipei, Taiwan, 2014.

- D. A. A. Black, L. Ma and M. 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, pp. 250-255, Seoul, South Korea.

- M. Tian, A. Srinivasamurthy, M. Sandler and X. Serra, 2014, "Study of instrument-wise onset detection in Beijing Opera percussion ensembles", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2159-2163, Florence, Italy.

- M. Tian, G. Fazekas, D. A. A. Black and M. B. Sandler, 2013, "Towards the representation of Chinese traditional music: A state of the art review of music metadata standards", in Proceedings of International Conference on Dublin Core and Metadata Applications, pp. 71-81, Lisbon, Portugal.

Chapter 2

Signal Analysis for Music Information Retrieval

2.1 Introduction

This chapter reviews the background, state-of-the-art methods, applications and evaluation approaches for *Music information retrieval* (MIR) research. We will firstly introduce the mathematical background for the time-frequency representations of audio signals in Section 2.2. This is followed by an introduction of several popular audio features in Section 2.3. As the main focuses of this thesis, *audio onset detection* and *music structural segmentation* will be introduced in Section 2.4 and Section 2.5. The background and research methodologies for these tasks are also surveyed. We introduce the Vamp Plugin audio processing framework as will be used in this thesis in Section 2.6. Section 2.7 introduces the analysis of signal processing parameters in audio-based MIR systems. Finally, evaluation methods for MIR tasks will be introduced in Section 2.8.

2.2 Audio Signal Representations

2.2.1 Human Listening and Music Perception

Sounds can be described in terms of their perceptual attributes such as pitch, loudness, subjective duration and timbre. The human auditory system is capable of analysing complex sounds by performing a spectrographic analysis of any auditory stimulus.

Hearing perception has been studied by researchers since the 1870s [Hel63]. Fletcher suggested that the peripheral auditory system behaves as if it contains a bank of band-pass filters with overlapping passband [Ros07]. Computational models for sounds require the input signal to be subject to a band-pass filterbank which approximates the frequency selectivity function of the inner ear. These filters are denoted auditory filters [Fle40].

Besides the semitone or the logarithmic scale which is typically used to visualise the pitch contours for music, different scales have also been established from auditory observations. The *Mel* scale is introduced from psychoacoustic experiments with simple tones (sinusoids). In the experiment, subjects were asked to divide the frequency ranges given into four perceptually equal intervals or to adjust the frequency of a tone to be half as high as that of a reference tone [SVN37]. The name "Mel" indicates that the scale is based on *pitch* to measure the music melody. One Mel is defined as one thousandth of the pitch of a 1-kHz tone. Increasingly large intervals are indicated by listeners to produce equal pitch increments when the frequency goes beyond 500 Hz.

Despite its popularity in MIR research, some surveys have pointed out that the Mel scale is often used mainly for the reason of its historical priority [Tra90]. Since the hearing system also performs a temporal analysis that contributes to frequency resolution for low frequencies, some work argues that the auditory frequency resolution cannot be fully represented by the Mel scale, which does not provide sufficient resolutions for the low-frequency sound sources [Tra97; MG83].

Auditory frequency resolution can also be rendered by the *equivalent rectangular bandwidth* (ERB) [MG83]. It is the bandwidth of a rectangular filter which approximates human hearing [Ben08]. ERB is proportional to the centre frequency fc when it is above 500 Hz. For lower frequencies, it decreases with dropping fc because the fine temporal structure of the signal contributes substantially to frequency resolution [MG83]. Glasberg and Moore [GM90] express the ERB as a function of fc of each filter as,

$$ERB(fc) = minBW + fc * 10^{-3}/e_q, \tag{2.1}$$

where $minBW$ and e_q are the Glasberg and Moore parameters respectively set to 24.7 and 9.26449 derived from psychoacoustic experiments.

The effect of bandwidth on loudness is essential in human perception of music. For example, for a given amount of energy, a complex sound is louder if its bandwidth exceeds one ERB than if its bandwidth is less than one ERB [Moo12]. Altogether, *Mel* and ERB are both very effective scales for sound signal analysis.

2.2.2 Time-frequency Representations of Music Signals

The human auditory perception discussed so far starts with the frequency analysis of the sound in human perception. The musical sounds must use time-dependent variables to expose the dynamics of the sounds and represent its spectral content evolving over time. There are two main domains that music signals can be represented by: *time* and *frequency* [KD06]. Dennis Gabor wrote a fundamental article in which he explicitly defined a TF representation of a signal [Gab46]. Gabor demonstrated that, in addition to the time representation (the signal itself) and the frequency representation, one can construct a two-dimensional representation, where each point corresponds to both a limited interval of time and a limited interval of frequency.

The *time-frequency (TF) representation* of a music signal mapping its time represen-

tation to the frequency representation is thus an intuitive starting point for a computational music analysis. There exist different forms of TF representations. In this section we review two commonly used TF representations for audio analysis and visualisation: *spectrogram* and *Gammatonegram*. The first characterises the *Fourier transform* (FT) and the latter incorporates human auditory perception cues and processes the signal in the time domain using the *Gammatone* filters on an ERB scale.

The theorem of FT is that any complex waveform can be decomposed into a series of sinusoids with specific frequencies. In the context of audio or music analysis, the FT is normally applied to a finite series of signal. At this point it is usually called a *short-time Fourier transform* (STFT). At the vicinity of a time instant t_0, we extract a portion of the signal $x(t)$, denoted a *frame*. The corner of this portion is rounded after multiplied with a window [KD06]. More precisely, the frame at time t_0 with the window function $\mathscr{W}$ is defined as:

$$x_{t_0}^{\mathscr{W}}(t) = x(t)\mathscr{W}(t_0 - t). \tag{2.2}$$

From frames, it is easy to build the STFT as the FT of successive frames in a short-time window:

$$STFT_x^{\mathscr{W}}(t, f) = \int_{-\infty}^{+\infty} x(\varsigma)\mathscr{W}(t - \varsigma)e^{-2jw\pi f\varsigma}d\varsigma. \tag{2.3}$$

Spectrograms are the energy of the STFT mathematically defined as its squared modulus:

$$X(t, f) = |STFT_x^{\mathscr{W}}(t, f)|^2, \tag{2.4}$$

where the spectrogram of a short-time discrete signal $x(n)$ is expressed as $X(n, f)$.

The length and shape of the selected window is a measure of the bandwidth discriminating the analysis. The main parameters that influence the TF analysis are the *window size* W_N and the *step size* (also called the *hop size*) between adjacent windows W_H. The

frequency resolution, i.e., the distance between consecutive frequency bins, is defined by $\frac{f_s}{W_N}$, where f_s is the sample rate with larger W_N leads to better frequency resolution. For an effective analysis, a compromise must be reached between having a good temporal resolution by using shorter windows and having a good frequency resolution by using longer windows. Experiments by Rasch indicate that the perception of tones in ensemble music is accurate to only 30 to 50 ms [Ras79]. A window size of 5 to 50 ms is typically used for STFT calculations in the MIR scenario [SGU14]. Standard window functions include *Hann*[1], *Hamming* and *Blackman* and the choice is generally flexible.

One way to address the conflict in retaining high temporal resolution and high frequency resolution in audio signal analysis is to resort to auditory filters. An auditory representation is typically obtained as the output of a computational model approximating the auditory processing with parameters of the model derived from psychoacoustic experiments [PP07]. As opposed to the spectrogram which displays the TF components according to the physical intensity levels of the audio signals, auditory representations attempt to emphasise their perceptually salient aspects [PP07]. This is mainly aimed to counter the effect of *auditory masking*, which means that the perception of one sound source can be affected by the presence of another in its vicinity.

In the Patterson model, the cochlea processing is simulated by *Gammatone auditory filters* [Pat+87]. The Gammatone function is defined in the time domain by the impulse response of the audio signal:

$$G(t, fc) = at^{(p-1)}e^{(-2\pi bt)}cos(2\pi(fc)t + \phi), \tag{2.5}$$

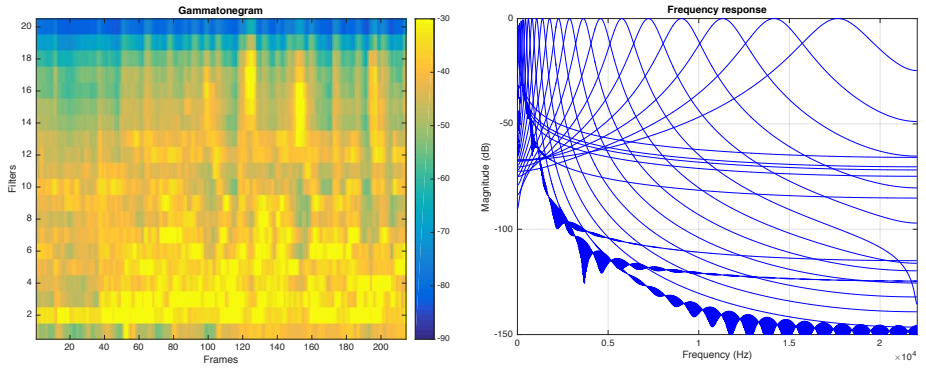
where a is the amplitude factor, p is the order of the filter, b is the frequency bandwidth of the filter in Hz which largely determines the duration of the impulse response, fc in Hz is the central frequency of the filterbank and t is the time in seconds.

[1] "Hann" is the original name of this window function. However, the name "Hanning" is also used occasionally derived from the expression "hanning a signal" which had originally been used to mean "to apply a Hann window to the signal".

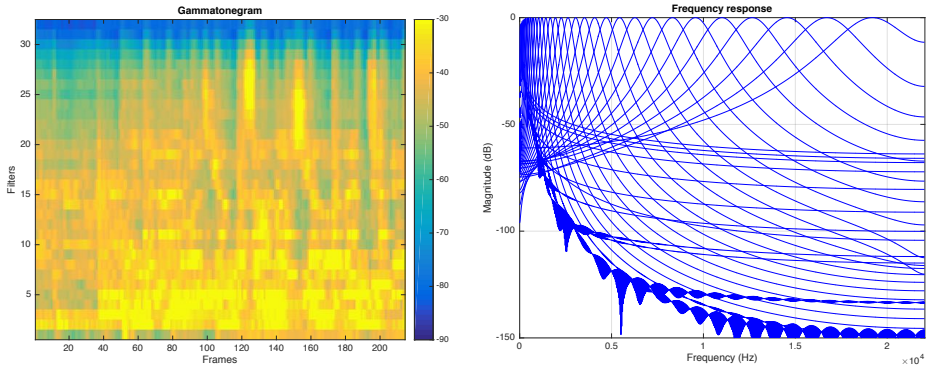
Given the number of filters M, the lower and upper frequency bounds, centre frequencies of the filter banks are uniformly spaced between the lower and upper frequency bound on the ERB scale. The impulse response of the Gammatone filter can be seen as a burst of fc enclosed in an envelope defined by the Gamma function, Here, the name "Gammatone" refers to the fact that the exponential expression in Equation 2.5 is the gamma function from mathematics, and that the cosine term is a tone for an auditory frequency range [Pat+87]. Fourth-order Gammatone filters form frequency responses that can closely approximate auditory filter shapes measured from psychoacoustic experiments [PP07]. An efficient implementation is provided by Slaney [Sla93]. The filter outputs may also be subject to rectification, compression and low-pass filtering as well as adaptive thresholding, which approximate the functions of the neural transduction. More details for the calculation of and the applications for Gammatone filters will be given in Section 5.5.

The Gammatone filters process a audio signal to yield $GF(t, fc)$ which keeps the original sample frequency f_s. In order to convert it into a spectrogram-like TF representation, noted a *Gammatonegram* or a *Cochleagram*, it is necessary to sum up the energy over fixed time windows. Although Gammatone filters have been successfully applied for TF descriptions in speech processing [Sch+07; Qi+13], relatively little work has attempted to use it to address MIR problems [Cas+08].

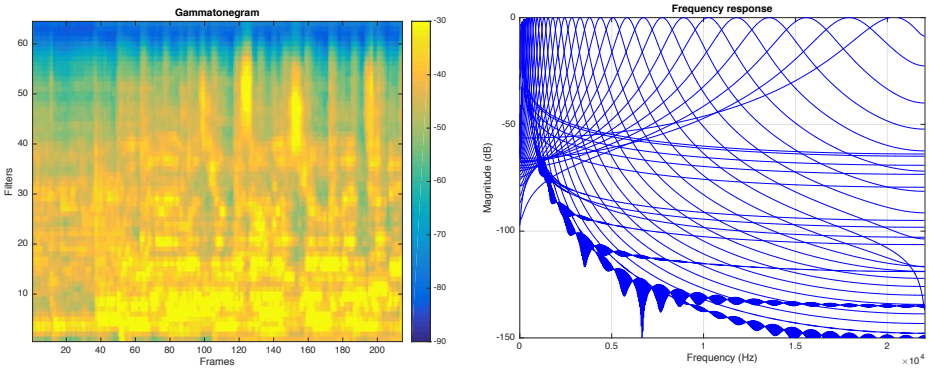
The main parameters influencing the calculation of a Gammatonegram are hence the number of filters m, the summation window N_G as well as the hop size between consecutive frames H_G. M = 20, 32, 48, 64 and 128 are typically used following practices in speech and audio processing where many works have reported that 32 - 64 channels are appropriate for a robust speech recognition or music spectral mapping [VA12; HWW14; Qi+13; Sch+07]. McKinney and Breebart used 18-channel Gammatone features for music classification [MB03]. Newton and Smith also showed that a 15-channel Gammatone filterbank with a frequency range of 200 Hz to 5 KHz can yields an abstract yet effective tone descriptor for musical instrument recognition [NS12]. The selection of



(a) 20 filters.



(b) 32 filters.



(c) 64 filters.

Figure 2.1: Gammatonegram computed with 20, 32 and 64 filters and the corresponding filter frequency responses for a 5-second excerpt (10.0 s - 15.0 s) of song "Help" by The Beatles. The lower and upper frequency bounds are respectively 50 Hz and 22.1 KHz. Only every 2nd channel is shown when using 64 filters for visualisation purposes.

N_G and H_G are typically similar to the window size and hop size chosen for the STFTs.

Figure 2.1 illustrates the Gammatonegrams and the filter bank frequency responses calculated using the implementation provided by Slaney [Sla93] on a 5-second excerpt

of the Beatles song "Help" (10.0 s - 15.0 s). M is set respectively to 20, 32 and 64 and N_G and H_G are set to 46 ms and 23 ms. The lower and upper frequency bounds are 50 Hz and 22.1 KHz (half the sample rate) for all. We can observe narrower frequency bins and increasing frequency resolutions with increasing number of channels. Unlike the Fourier-based TFRs, however, this does not lead to any loss of temporal resolution. We also observe that the "brightest" areas become more compressed and that the spectral centroids of individual frames appear to be shifting downwards with increasing M, although a constant frequency range is used, due to the non-linearity of the ERB scale. A promising approach for distinguishing the frequency or temporal events hence lies in the replacement of the standard spectrogram with the output of auditory filters such as the Gammatone filters for the audio signal representation. We will investigate the feature extraction with Gammatone filters in Chapter 5.

2.3 Audio Feature Extraction

Feature extraction constitutes the core process for audio content description. Though the digital signal contains all the information of the sound, it must be summarised somehow to facilitate the subsequent information retrieval. The goal of feature extraction in our scenario is to find descriptors of the audio content that are invariant to irrelevant transformations and having good discriminative power across different classes. Note that in this thesis we refer "audio features" to descriptors calculated from the audio signal in a derived way, i.e., from its TFRs, to summarise specific aspects of the audio content, since one may argue that the more general TFRs such as the spectrogram are features of the signals too [SGU14].

Audio features can be categorised based on the abstraction level of the information they summarise. The *low-level* features are considered to be related to the physical properties or musical attributes related to the signal itself, such as the *root mean square* (RMS) *energy*. The so-called *mid-level* features are commonly derived from the statistical

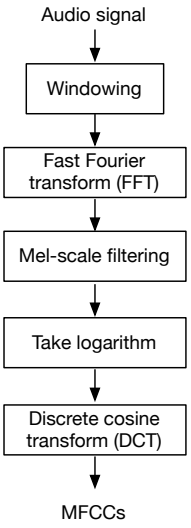


Figure 2.2: Calculation of the MFCCs feature.

descriptions or summaries of the low-level features, for example, the *tempogram*, as will be introduced shortly. The *high-level* features, sometimes also denoted *semantic* features, expose aspects that are typically approximating how humans would perceive the music.

A popular approach for feature extraction is to use explicit domain knowledge about the underlying source signal to derive abstract and salient signal representations. In MIR, the selection of audio features contributes crucially to the acquirement of a representation of specific aspects of the music content, such as the timbre, loudness, key, harmony, melody and rhythm [Cas+08]. Well-known feature descriptors include the *Mel-frequency cepstral coefficients* (MFCCs) which capture the music timbre, *chroma* features which describe the harmony, perceptual descriptors such as *loudness* and *sharpness* and rhythmic features characterising the metre, timing, tempo as well as their grouping rules [Pee04]. A popular method is to extract features from consecutive frames of the TFRs, typically the spectrogram, of the input audio signal. Here we introduce three features, MFCCs, chromagram and tempogram, describing respectively the timbral, harmonic and rhythmic contents of the music content.

2.3.1 Mel-frequency Cepstral Coefficients

Timbre refers to the quality of the sound. The American National Standards Institute defines timbre as "the attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". It is the third attribute of the subjective experience of musical tones after *pitch* and *loudness*. Timbre-related features are of general importance in describing audio content [Deu12].

The use of MFCCs in MIR studies was first proposed by Logan [Log00]. MFCCs provide a compact way to model the signal spectra. Figure 2.2 shows the process of calculating the MFCCs. First, the input signal is divided into short frames. Then the framed signal is subject to a fast Fourier transform (FFT) (see Section 2.2.2). The magnitude spectrum of the transformed signal is passed through the Mel filterbank (see Section 2.2.1) and then converted to a logarithmic scale. The purpose of the Mel warping is to convert the representation into a perceptually meaningful scale where higher resolution is assigned to lower frequency components. As the final step, a *discrete cosine transform* (DCT) [ANR74] is taken to derive a compact feature representation. DCT is a Fourier-based transform similar to the discrete Fourier transform (DFT) but operates only on real data hence leading to an even symmetry. The "log-DCT" analysis, also referred to as the "cepstral" analysis, is commonly used to decorrelate convolved data for feature representation (for example, human speech can be modelled as the convolution of an excitation and a vocal tract).

Despite its wide application in many MIR tasks [APS05; Jen+06; Rum+10; LS06; RBH13; TSB05], little work has attempted to justify the choice of the Mel filters used in MFCCs after Logan, whose conclusion is only "using the Mel cepstrum to model music in the speech and music discrimination problem is not harmful" [Log00]. It hence remains unaddressed to quantify the suitability of the Mel scale in MIR scenarios. In this thesis, we will investigate the use of different TF representations to characterise the

timbre description.

2.3.2 Chromagram

Chroma features, which closely correlate to the aspect of tonality, are well-established tools in processing and analysing pitched music data [Góm06; BW05; EP07]. A *chromagram*, for the definition of which readers are often directed to Fujishima [Fuj99], also called *Harmonic pitch class profile* (HPCP), is a B-dimensional vector representation of a chroma, where B is the *bins per octave* (BPO) in the scale. A chromagram hence represents the relative intensity of each bin in a chromatic scale with equal temperament in a tuning-independent manner. Since the type of a chord rests on the position of its constituent tones in a chroma, a chromagram would consist the whole information to represent a chord. Besides used to capture the harmonic aspect of music, chromagram features have also been proven useful as timbre descriptors [MBH12].

One popular implementation of this feature is based on the constant-Q transform (CQT) as described in Brown [Bro91]. The framework is shown in Figure 2.3. First, the framed input audio signal is subject to a CQT and passed through a logarithmic frequency filterbank with centre frequencies decided by Equation 2.6:

$$f(k_{lf}) = f_{min} * 2^{\frac{k_f}{B}}, \tag{2.6}$$

where f_{min} is the minimum frequency of the analysis in Hz, $k_f \in [0, B * N_o]$ is the filter index with B as the number of semitones in the scale and N_o the number of octaves. The outcome of this step is a log-frequency magnitude spectrum. The chromagram is then derived from binning the spectrum across octaves into a B-dimensional vector representation.

Alternatively, the CQT can be replaced by a DFT followed by a log-frequency warping. Whereas the CQT-based method provides a constant ratio of frequency resolution

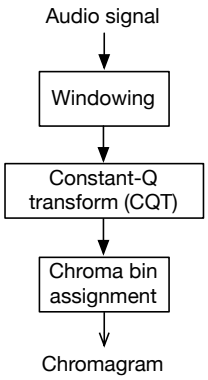


Figure 2.3: Computation of the chromagram feature.

by adjusting the window size for each frequency bin, the DFT-based approach would lead to unfixed resolutions – the resolution is lower at low frequencies and higher at high frequencies.

While $B = 12$ is typically used for the Western equal temperament, some work has proposed to use different BPO settings for specific tasks or music genres [HSG06; Góm06]. In [HSG06], Harte and Sandler proposed a 36-bin semitone-quantised chromagram to improve the accuracy of locating the boundaries between semitones for chord recognition. We will investigate the effect of BPO settings for different music types in Chapter 5.

Different variants of the chroma features have been proposed to improve its timbre- and orchestration-invariance such as Chroma Energy Normalized Statistics (CENS) [MEK09], Chroma DCT-Reduced log Pitch (CRT) [ME10a] and the Complexity Features [WM15]. Different approaches have also been presented to counter the interference for chroma bin assignment introduced by overtones. Mauch and Dixon introduced the Non-negative Least Squares (NNLS) Chroma with improved note dictionary [MD10]. Jiang et al. introduced a filterbank-based chroma feature with a logarithmic compression before the octave mapping for chord recognition [Jia+11].

2.3.3 Tempogram

Rhythmic features describe the temporal events and their organisations in music. Most rhythmic descriptors are based on locating the note events, typically note onset, as will be introduced in Section 2.4. By measuring their periodicity of note onsets, higher level rhythmic descriptions can be obtained such as the beat and metre estimations [GD05].

One rhythmic descriptor is the *tempogram*. A tempogram is a time-pulse representation of an audio signal indicating the variation of pulse strength over time given a specific time lag l_t or a BPM value τ_t. This property makes it an appropriate base for higher level feature representations to be extracted from incorporating tempo and rhythm information. By *pulse*, also called *beat*, or more technically *tactus*, we mean the basic unit of time in music which can be categorised according to where it falls in the bar: as "weak" (the second and fourth in a four-beat bar, the second and third in a three-beat bar, or the second in a two-beat bar), or "strong" (the first and, to a lesser degree, the third in a four-beat bar, and the first in a three-beat or a two-beat bar)[2]. The tempo of a piece is defined as the speed of the pulse. We hence note the periodicity of a pulse with a tempo value in *beats per minute* (BPM).

The calculation of a tempogram is based on a one-dimensional temporal signal indicating the presence of the beginning of notes, denoted the *onset detection function* (ODF) as will soon be detailed in Section 2.4. The two main methods presented in the literature employ respectively the Fourier transform (FT) and the autocorrelation method. While the first method yields tempograms with harmonics, the latter yields tempograms emphasising the subharmonics [Pee05]. Here we introduce these two approaches.

In the first approach, the ODF is analysed using an FT. We note $F(t, \tau_t)$ the amplitude spectrum of $ODF(n)$ for a tempo τ_t and at time t where τ_t in BPM can be mapped to the frequency f in Hz by $\tau_t = 60 * f$. A Hann window with window length W_f is used.

[2]Besides being used as synonyms, a distinction is occasionally made between "pulse" and "beat": for example, 6/8 time may be said to have six "pulses" but only two "beats" [SW70]

This process is shown in Equation 2.7.

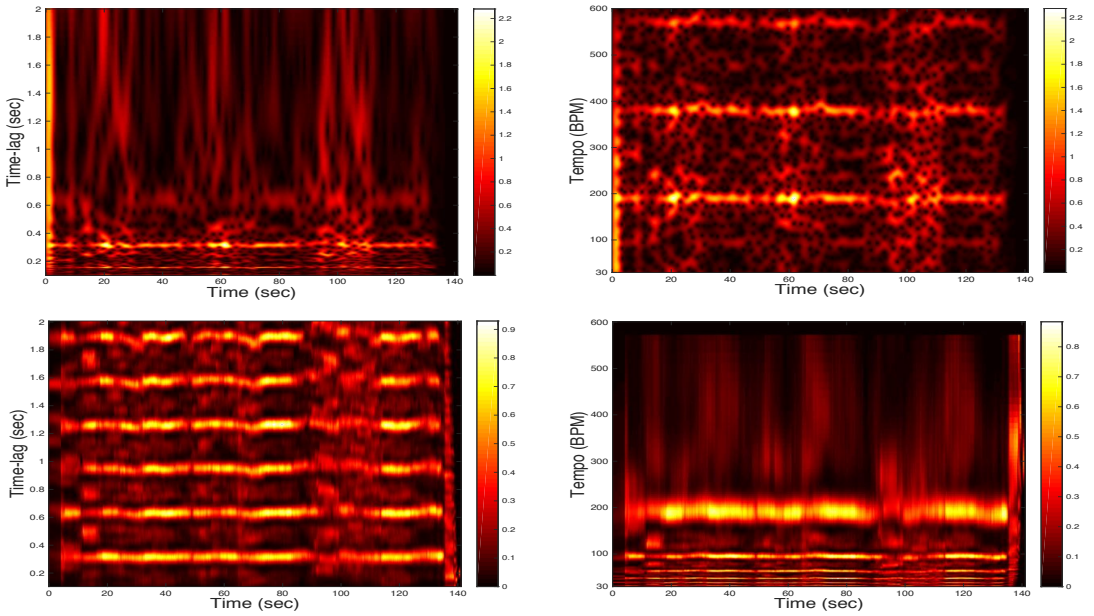
$$F(t,f) = |\sum_{n\in\mathbb{Z}} ODF(n) \cdot W_f(n-t) \cdot e^{-2\pi i f n}|, \tag{2.7}$$

These periodic patterns may also be characterised by peaks in the autocorrelation function (ACF) of the ODF at certain time lags. In the autocorrelation-based method, the local ACF of the ODF is calculated using a rectangular window W_a as shown in Equation 2.8. We note $A(t,\Lambda)$ the ACF of the ODF for a time lag Λ and time t. $A(t,\Lambda)$ will hence represent the level of periodicity at the lag Λ/fs where fs is the sample rate and the time lag Λ corresponds to the tempo $\tau_t = 60/(s_r \cdot \Lambda)$.

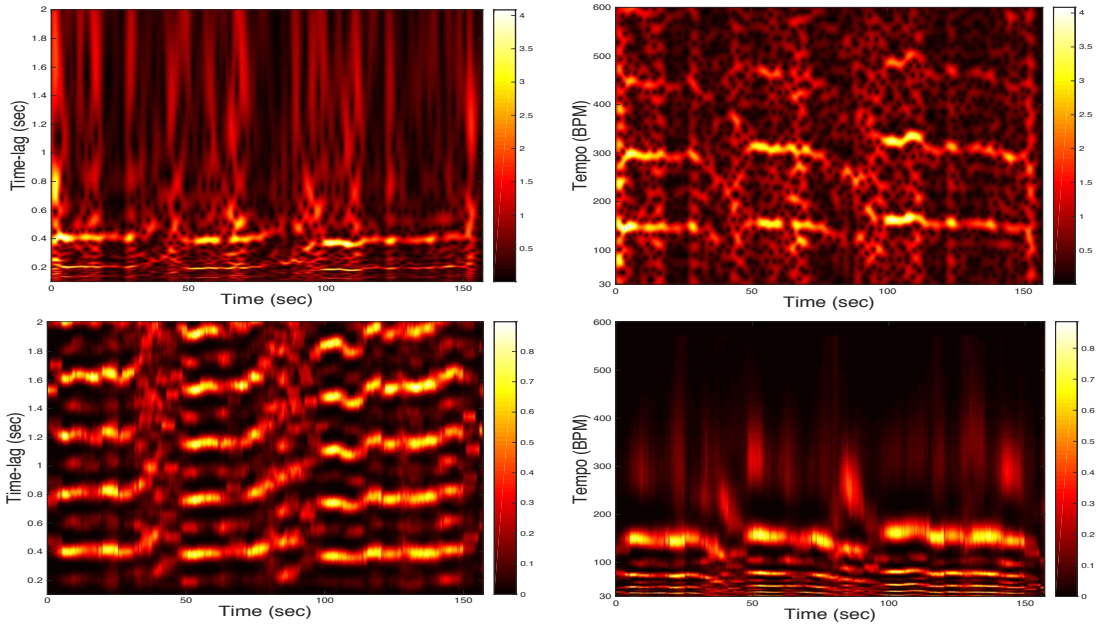
$$A(t,\Lambda) = \frac{\sum_{n\in\mathbb{Z}} ODF(n)ODF(n+\Lambda) \cdot W_a(n-t)}{2N+1-\Lambda}, \tag{2.8}$$

for time $t \in \mathbb{Z}$, time lag $\Lambda \in [0:N]$ and s_r is the feature rate.

Figure 2.4 demonstrates the tempogram computed using both approaches of the pop music "Help" (Figure 2.4a) and Jingju music "Jin yu nu" (Figure 2.4b) from our evaluation dataset BeatlesTUT and CJ (details of the datasets will be given in Section 3.4). The average tempo for the two tracks are respectively 95 BPM and 78 BPM. The FFT window size and step size are respectively 0.046 s and 0.023 s to derive the ODF, and the time window W_f and W_a are set to 6 s. Note that in the FT-based method, the STFT has been carried out twice, once to derive the spectrogram where the ODF is calculated and once to derive the tempogram. We use the implementation from *Tempogram Toolbox* by Grosche and Müller [GM11b]. The upper and bottom pane show the tempogram calculated using the FT- and the ACF-based method respectively and the left and right pane show the tempogram in the tempo (in BPM) and lag (in second) domain. It is noticeable that while the two approaches both present the dominant tempo, the subharmonics are delineated in the ACF-based tempogram but suppressed in the FT-based tempogram.



(a) Tempograms for the audio example "Help" using the Fourier Transform based method (left) and the autocorrelation based method (right).



(b) Tempograms for the audio example "Jin yu nu" using the Fourier Transform based method (left) and the autocorrelation based method (right).

Figure 2.4: Tempogram for Western pop music "Help" and an excerpt of Jingju music "Jin yu nu" from dataset BeatlesTUT and CJ (see Section 3.4). The left and the right pane show respectively tempograms derived using the autocorrelation and the Fourier transform based method.

Some works also propose to use the tempogram as a basis to extract rhythmic descriptions at a higher abstraction. Grosche and Müller introduced the Cyclic tempogram and the derived *Predominant local pulse* (PLP) curve reflecting the predominant pulse at each time position [GM11a]. The main idea of the PLP curve is to derive for each time position a sinusoidal kernel for the tempogram indicating the local periodicity of the novelty curve obtained prior to the tempogram. The kernels are then accumulated over time using the overlap-add technique such that the derived signal feature can reveal the PLP. This feature is proved an effective tool for beat tracking [GM11a].

As can be seen from the above examples, specific audio features normally describe specific aspects of the music content and convey specific semantic meanings. The design and the selection of features in the MIR scenario hence rest with the tasks they are used for. An overview of the extraction of commonly used audio features and their applications can be found in [Pee04; BDC15]. In the next two sections we introduce in detail two MIR tasks investigated in this thesis: *audio onset detection* (AOD) and *music structural segmentation* (MSS).

2.4 Audio Onset Detection

2.4.1 Definition and Applications

Musical notes are discrete events that underpin the music composition. The succession of pitched notes comprises the melody of a piece of music, and the starting times of notes, percussion notes in particular, decide its metre and tempo. The primary goal for automatic AOD is to parse the note events and analyse their inherent patterns and periodicities to extract more abstract representations such as tempo, timing and metrical structure [GD05].

The establishment of a *note onset* requires the perception of sound event usually with short intervals where the sound signal undergoes rapid changes, defined as *tran-*

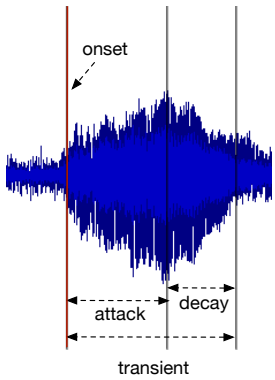


Figure 2.5: A violin note onset. The red vertical line indicates onset location annotated by listeners.

sients [SW70; Hon02]. The note onset in MIR research is commonly defined as the start of a *transient*, which marks the time interval during which the amplitude envelope of the signal increases. The detection of a note onset can hence be approached by pinpointing the start of the *transient*.

A transient is normally defined as a high-amplitude, short-duration sound at the beginning of a waveform [Cro98]. For many musical instruments, the transient is often associated with a sudden increase of the signal energy after an external excitation, such as a pluck on the string or a hammer strike on the membrane, is applied and then damped, followed by a decay at the resonance frequencies of the body [Bel+05]. In case of instruments with longer transient times and without sharp bursts of energy, the interpretation of the typical ODF may become ambiguous. Based on this rationale, music onsets can be categorised into hard and soft onsets. Figure 2.5 illustrates the onset of a single note produced by a violin. The audio track and onset annotation are provided in [Bel+05]. We will provide more details on how to create onset annotations in Chapter 3. It can be observed that the appearance of an onset is accompanied by the amplitude increase from a low level to a peak followed by a decay.

The automatic detection of note onsets is an essential part in many music and audio signal analysis schemes. It has various applications in content-based music processing

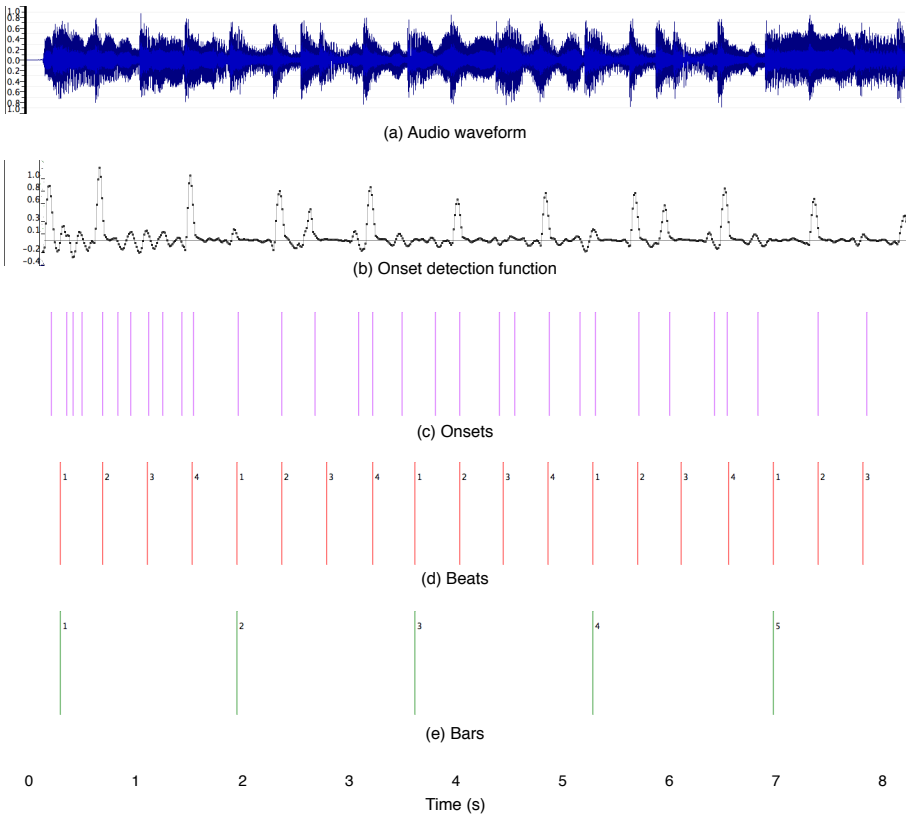


Figure 2.6: The waveform, onset detection function, estimated onsets, beats and bar positions for an excerpt of Beatles song "Love me do".

and is also considered as the first step of many other MIR tasks such as tempo tracking and the extraction of beat and bar positions. Figure 2.6 shows the onset detection function (ODF) and onset positions obtained using [Dux+03] as well as beat and bar detected following [DP07; SDP09]. The local beat is estimated based on the periodicity in the ODFs and the bar positions are derived from the downbeats given the music metre. Both implementations are provided by Queen Mary Vamp Plugins [Plu06].

2.4.2 Related Work

Many methods for AOD have been presented in related works [Ros+99; Bel+05; Col05a; Col05b; Dix06; Deg+09; Deg+10; Eyb+10; BKS12; Böc+12; BW13a; BW13b; HSL13; HS08; Hol+10; Zha+13; Wan+06; TZW08; SP07; RRM12; SB13; Mar+14; LD04; LK06;

LE06; Kla99; KP04]. Energy-based onset detection methods were primarily developed to detect sudden changes or a burst of signal energy introduced by the emergence of a new note. Common techniques include *High frequency content* (HFC) and *Spectral difference* (SD) [Mas96].

However, one of the drawbacks of many energy-based methods is that they are less effective in detecting soft onsets and partially masked onsets in complex mixtures of musical sounds. To compensate for this, phase information can be incorporated into the analysis, which relies on the fact that phase change is less predictable during transients than steady-state. This leads to phase-based and complex-domain techniques introduces in Duxbury et al. [Dux+03] with further improvements by Dixon [Dix06].

Pitch contours and harmonicity information can also be indicators for onset events. Collins introduced the use of pitch change to detect note onsets [Col05b]. Heo et al. introduced a method where the ODF is built using the summation of *harmonic cepstral coefficients* with their "quefrency" indices derived from previous frames [HSL13]. These methods show some advantages for detecting soft onsets. Stylistic musical elements such as vibrato and tremolo can lead to the detection of onsets not corresponding to actual notes, referred to as false positives. Techniques for vibrato suppression have been presented in recent works. *SuperFlux* (SF) is introduced as a modified version of the spectral difference method which reduces false positive detections considerably by tracking spectral trajectories of partials [BW13b]. In the next section, we will review the composition of an AOD system and introduce some popular onset detection algorithms.

2.4.3 Onset Detection System

An onset detection framework generally comprises three main components: *pre-processing* of the input audio signal, calculation of the *onset detection function* and *post-processing and peak picking*. The framework for onset detection provided by the Queen Mary Vamp Plugins (QMVP) [Plu06] is illustrated in Figure 2.7. Although more advanced algorithm

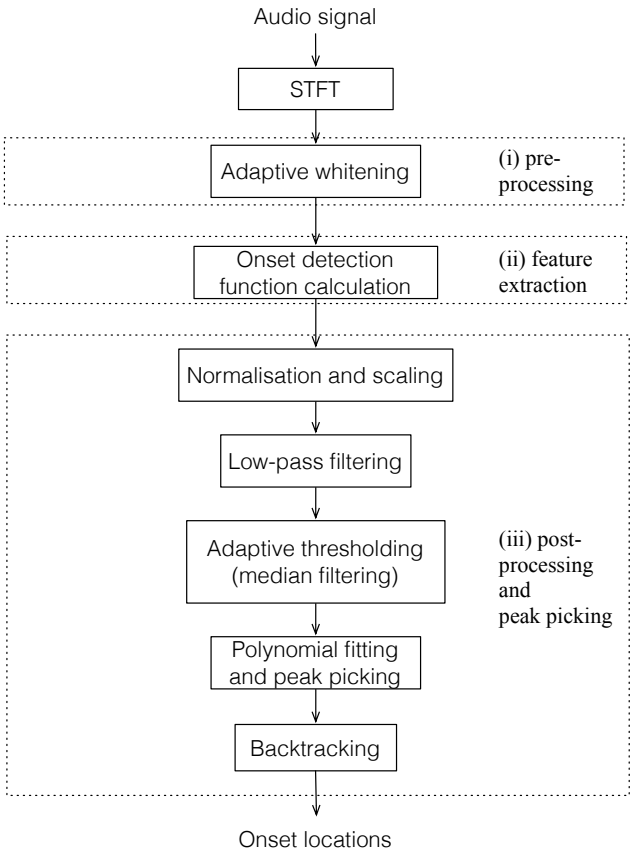


Figure 2.7: Flowchart of the onset detection framework.

have been presented in recent works, we take QMVP as a representative of the standard spectral feature based onset detection workflow.

Pre-processing of Input Signal

Pre-processing concerns modifying the input audio signal, for instance using whitening techniques, to facilitate the subsequent extraction of onset-related features. The spectra of musical pieces exhibit significant temporal variability due to dynamics as well as prominent variability across frequency bins. This may lead to over-emphasising note onsets in the lower bands or other louder parts of an audio signal.

Adaptive whitening [SP07] is designed to mitigate these issues. It is implemented in our system as a pre-processing step, where the magnitude of each STFT bin is normalised

using the recent maxima. This will bring the magnitudes of different frequency bins to similar dynamic ranges. The whitened magnitude spectrogram $X_w(n,k)$ is expressed by Equation 2.10 such that for the magnitude spectrogram $X(n,k)$,

$$P(n,k) = \begin{cases} \mathsf{max}(|X(n,k)|, r, m_f P(n-1,k)), \text{if } n > 0 \\ \mathsf{max}(|X(n,k)|, r), \quad \text{if } n = 0 \end{cases} \tag{2.9}$$

and then,

$$X_w(n,k) = \frac{X(n,k)}{P(n,k)}, \tag{2.10}$$

in which $n \in [0 : N-1]$ and $k \in [0 : K-1]$ where N is the number of frames and k is the frequency index of the STFT of the input audio signal with a window length of $2K$. m_f is the memory coefficient allowing $P(n,k)$ to decay over time exponentially such that past peak values will be "forgotten" over time such that the whitening is applied "adaptively". To prevent $P(n,k)$ from falling so low that noise gets overly amplified, r is a parameter to set the floor condition.

Onset Detection Function

An onset detection algorithm concerns transforming the input audio signal into a discrete time varying feature exhibiting the occurrence of transients indicating potential onset locations. This feature is denoted *onset detection function* (ODF). This section introduces a few well-established methods in the literature based on spectral information.

Five of these are outlined in Bello et al. including: *High frequency content* (HFC), *Spectral difference* (SD) *Complex domain* (CD), *Broadband energy rise* (BER) and *Phase deviation* (PD) [Bel+05]. These methods have been published as open source software within the Queen Mary Vamp Plugins [Plu06]. Finally, the sixth method, SuperFlux (SF) is also included in our study representing the more recent improvement on spectral-based techniques [BW13a]. Although more recent methods have been presented [Eyb+10; BW13a; BW13b; HSL13; Zha+13; RRM12; SB13; Mar+14], we introduce these methods

here because they are commonly employed in related MIR tasks using onset detection as the fundamental step such as beat and tempo tracking [GMK10; DP04; Tia+15].

In the spectral domain, energy increases related to transients tend to appear as wide-band noise. This causes more energy to present in the higher frequency bands compared to harmonic sounds. Measuring the *High frequency content* (HFC) [MB96] of a signal is a commonly used method with the hypothesis that a note contains high frequency energy at its onset times. Consider the magnitude spectrum $Y(n,k) = |X(n,k)|$. The ODF is constructed from the weighted local energy:

$$HFC(n) = \frac{1}{K} \sum_{k=1}^{K-1} k * Y(n,k), \tag{2.11}$$

This function has a few variations. For example, Jensen and Anderson rewrote it as the sum of the amplitudes weighted by the square of the frequency value [JA04].

The *Spectral difference* (SD) method looks for sudden changes in the energy spectrum. The ODF expresses the first-order difference between successive STFT frames, given by:

$$SD(n) = \sum_{k=0}^{K-1} |Y(n,k) - Y(n-1,k)|. \tag{2.12}$$

As opposed to instantaneous measures such as HFC, this method takes into consideration the relative amount of energy changes between frames.

Various vector norms have been proposed in the literature to produce the distance function, including the L^2 norm (Euclidean distance) and the L_1 norm (Manhattan distance) [KK83]. HFC and SD are successful in detecting highly percussive onsets. However, they tend to be less effective in detecting softer onsets such as those of bowed string instruments, or in cases when the energy profiles of the weaker notes are masked by those of stronger notes.

SuperFlux (SF) is a modification of SD where spectral trajectory-tracking is applied to suppress the effects of vibrato. The magnitude of the STFT is first processed through a triangular filterbank and then converted to the logarithmic scale. This process is described by the following formula:

$$X_{log,filt}(n,q) = log_{10}(|X(n,k)| * F(k,q) + 1), \tag{2.13}$$

where $F(k,q)$ is the filterbank and q is the frequency bin index on the quarter-tone frequency scale. The constant one is added to ensure the positivity of $X_{log,filt}(n,l)$. The maximum filtered spectrogram with a size $2m$ is then given by:

$$X^{max}_{log,filt}(n,q) = max(X_{log,filt}(n, q - W_m : q + W_m)). \tag{2.14}$$

SF is then defined as the spectral difference between the current frame of the maximum-filtered spectrogram and another which is μ frames apart, as expressed in Equation 2.15:

$$SF(n) = \sum_{k=0}^{K-1} H(X_{log,filt}(n,k) - X^{max}_{log,filt}(n-\mu,k)), \tag{2.15}$$

with $H(x) = \frac{x+|x|}{2}$ as the half wave rectification. In [BW13a], W_m and μ are set respectively to 1 and 2 where the sample rate and the window size are 44.1 KHz and 5 ms.

To better characterise soft onsets exhibiting longer but less pronounced transients, Duxbury et al. introduced the *Phase deviation* (PD) method based on quantifying the phase deviation between the target and current frame regardless of their energy intensity [Dux+03]. When $\phi(n,k)$ is the phase of $X(n,k)$, the ODF is given by

$$\begin{aligned} PD(n) = \frac{1}{K} \sum_{k=0}^{K-1} & princarg(\phi(n,k) - \phi(n-1,k)) \\ & - princarg(\phi(n-1,k) - \phi(n-2,k))|, \end{aligned} \tag{2.16}$$

where the *princarg* function maps the phase deviation to $[-\pi, \pi]$ range. As a modification, Dixon introduced a half-wave rectification to preserve only the increases of energy in spectral bins to rule out the offsets from the onsets [Dix06].

An alternative approach is to combine phase and magnitude and to measure the distance between a target and the current frame in the complex domain. The Complex domain (*CD*) detection function is calculated by summing the measured stationarity across all bins [Bel+04]:

$$CD(n) = \sum_{k=0}^{K-1} \{X(n,k) - |X(n-1,k)|e^{\phi(n-1,k)+\phi'(n-1,k)}\}^2 , \tag{2.17}$$

where $\phi'(n-1,k)$ is the rate of phase change at the $(n-1)$th frame.

As opposed to the above introduced methods, the *Broadband energy rise* (BER) method [Bar+05] does not consider the total energy change in the signal. The aim is to measure how *broadband* or *percussive* the onset is. This approach is successful in detecting drum onsets which can be characterised by a rapid broadband rise in energy followed by a fast decay. The ODF is obtained by summing the number of bins in which the log difference between consecutive STFT frames exceeding a certain threshold θ, as shown in:

$$diff(n,k) = 10log_{10}\frac{Y(n-1,k)}{Y(n,k)}, \tag{2.18}$$

and

$$BER(n) = \sum_{k=0}^{K-1} \begin{cases} 1 & \text{if } diff(n,k) > \theta \\ 0 & \text{otherwise.} \end{cases} \tag{2.19}$$

Post-processing and Peak Picking

In an onset detection system, *post-processing* is an optional step applied to the ODF to remove the non-onset related noises and increase the detectability of the transients thus facilitating the subsequent *peak picking*.

Onset detection functions may exhibit a very large number of low magnitude peaks due to small, non-onset related local variations in the audio signal, as can be seen in Figure 2.6b. These would then be unnecessarily selected as peaks in the subsequent stages. It is therefore beneficial to smooth the detection function and reduce the number of peaks that are not related to onsets. A popular approach is to use the *low-pass filter* to attenuate the high-frequency noises by passing the signals with a frequency lower than a selected *cut-off* value. In this system, a zero-phase *Butterworth* low-pass filter is applied for this purpose. Standard filters can introduce a time delay in the detected peaks. This issue can be addressed by applying the same filter twice, once forwards and once backwards. Therefore, the combined filter will have a zero phase.

For the purposes of peak picking from the ODF, choosing a fixed threshold leads to difficulties as music signals exhibit a great extent of local dynamics and intensity variations throughout a piece. To counter this, many works propose to use a *median filter* to generate the threshold adaptively [Bel+05]. A median filter has the property of preserving edges or stepwise discontinuities in the signal while eliminating small local variations by replacing signal samples with the median values computed over short-duration sliding windows [Vas08]. A median filter based adaptive thresholding process is given in Equation 2.20,

$$\begin{aligned} th_{ad}(n) = \delta + \mathsf{median}(ODF[n-\psi], ..., \\ ODF[n], ..., ODF[n+\psi+1]), \end{aligned} \tag{2.20}$$

where δ is a constant threshold and ψ is the median window length. The resulting signal will then be used as the adaptive threshold to be subtracted from the ODF. Onsets can

subsequently be identified by searching for local maxima that exceed a defined threshold. This method is adopted by a majority of related works [Bel+05; BW13a].

Besides the median filter based adaptive thresholding which selects peaks with substantial magnitude, an alternative approach selects peaks by assessing their shapes. A polynomial fitting based onset detection is used in the QM Vamp Plugin Onset Detectors [Plu06] which fits each peak in the ODF into a quadratic function to represent its property hence measuring its probability as a potential onset. First, initial onset candidates are selected using the median filter based adaptive thresholding. In order to assess both the spikiness and the amplitude of each peak, a second-degree *polynomial* is fitted on samples around each peak. For each local maximum in the smoothed ODF, five discrete samples centred around it from its neighbouring frames are used as the input for the polynomial fitting implemented with the least squares method [Wil12]. This estimates the coefficients of the second-degree quadratic function given in Equation 2.21,

$$y = ax^2 + bx + c, \tag{2.21}$$

where coefficient a and c correspond respectively to the sharpness and the amplitude of each peak.

A peak will only be accepted as an onset when the following conditions are satisfied: $a > th_a$ and $c > th_c$. The coefficient b is not assessed as the acceptance of a peak depends on but the position but only the shape of the parabola. Lower th_a and th_c indicate loose conditions a peak in the ODF can be selected as an onset, where th_a and th_c are obtained from a single sensitivity parameter $sens$ and two experimentally defined values using $th_a = (100 - sens)/1000$ and $th_c = (100 - sens)/1500$. By feeding $sens$ a value within the range of $[0, 100]$, th_a and th_c will fall into the range of $[0, 1]$ and $[0, 0.67]$ respectively with experimentally defined floor values[3]. Hence the higher the sensitivity is, the more onsets will be retrieved from the ODF.

[3] `https://code.soundsoftware.ac.uk/hg/qm-dsp`

However, in this system (see Figure 2.7) the option is given of using an alternative simple peak picker using the same parameter *sens*. A unified peak picking threshold *thresh* will be calculated for the post-processed ODF as shown in Equation 2.22. In the case, the median filter is no longer used to provide the onset candidates, but to smooth the ODF by removing the local medians.

$$thresh = \mathsf{min}(ODF) + sens * (\mathsf{max}(ODF) - \mathsf{min}(ODF)), \tag{2.22}$$

where $\mathsf{max}(ODF)$ and $\mathsf{min}(ODF)$ are the maximum and minimum value of the post-processed ODF. A local maximum will be picked as a peak when its value exceeds *thresh*.

Backtracking for Onset Relocalisation

In case of many musical instruments such as the bowed string types, onsets have long transients without a sharp burst of energy. This may cause detection functions to exhibit peaks after the onset locations listeners perceive, as can be noticed from Figure 2.5. The *backtracking* is applied to counter this effect [Tia+14a]. We trace the detected onset locations from the peak position in the ODF to a hypothesised earlier "perceived" location. The *backtracking* procedure is illustrated in Algorithm 1. In this expression, *bt* is used as a stopping condition to measure the relative differences of adjacent samples in the ODF. In this way, it monitors the steepness of a peak in the ODF hence controls how "far" the onsets should be traced back from the initially detected location. A experimentally defined value of 0.9 was used in Queen Mary Vamp Plugins [Plu06]. We will investigate the effect of backtracking in Section 4.3.

2.4.4 Discussion

The techniques discussed so far rely on encoding some assumptions on the signal's properties into the design of onset detection algorithms. Alternatively, methods using prob-

Algorithm 1 Backtracking of onset locations.

Require: *i: index of a peak location in the ODF, bt: threshold*

```
1: procedure BACKTRACKING(i, ODF)
2:     Δ, Γ ← 0
3:     while i > 1 do
4:         Δ ← ODF[i] − ODF[i − 1]
5:         if Δ < Γ ∗ bt then
6:             break
7:         i ← i − 1
8:         Γ ← Δ
9:     return i
```

abilistic or machine learning models attempt to provide more generic solutions by allowing these properties to be learnt from the signal itself. Up to date the state-of-the-art methods are based on the neural networks (NNs) [Eyb+10; SB13; Mar+14]. Eyben et al. adopted a bidirectional Long Short-Term Memory recurrent neural network (LSTM RNN) trained on auditory spectral features and relative spectral differences [Eyb+10]. Schlüter and Böck trained a convolutional neural network (CNN) to find onsets in spectrogram excerpts [SB13]. The latter CNN-based method achieved comparable detection rate with the RNNs [Eyb+10], but required less manual pre-processing despite with higher computational costs. Marchi et al. revisited the LSTM RNN approach using the wavelet coefficients and multi-resolution linear prediction errors combined with auditory spectral features [Mar+14], which substantially improved the results of Eyben and his colleagues [Eyb+10]. Due to the data-driven nature of these methods, however, a computationally expensive training process is normally required.

Most existing methods are focused on characterising note onsets by capturing a single physical property of the underlying audio signal. A viable and promising approach for onset detection lies in the combination of various detection methods using fusion techniques aiming at bringing together the strengths of individual methods and overcoming their drawbacks [ZMZ08; Deg+09; Hol+10; Zha+13; Tia+14a]. Zhou et al. proposed a system integrating two detection methods selected according to properties of the target onsets [ZMZ08]. In Holzapfel et al. [Hol+10], pitch, energy and phase information are considered in parallel for the detection of pitched onsets. Another fusion

strategy is to combine peak candidates detected by different methods to form new onset estimations [Deg+09; Zha+13]. Although the concept of fusion has been adopted in related works, a thorough investigation of different fusion policies has not previously been undertaken.

Several databases have been designed to evaluate audio onset detection works in recent works [Bel+05; Hol+10; BKS12]. Most of the publically available ones are focused on the Western instruments and music types [Bel+05; BKS12], leaving the non-Western music relatively unexploited.

2.5 Music Structural Segmentation

2.5.1 Definition and Applications

In *The Oxford Companion to Music*, music structure is defined as "a series of strategies designed to find a successful mean between the opposite extremes of unrelieved repetition and unrelieved alteration" [SW70]. The main goal of *music structural analysis* (MSA) is to generate high-level structural descriptions for music. MSA is one of the most investigated topics in MIR. It has broad applications such as informed listening [Got03], thumbnailing [LC00; LSC06] and audio tagging [LWW10].

Music structural segmentation (MSS) concerns dividing an input music signal into various structural sections and deals with the structure of the entire music piece. The definition for the MSS task then depends on the understanding we have of what defines the structural sections of a music piece. Hence, a precise and consensual definition for the *section* may be hard to apply to different music genres or for individual listeners. However, what can be generally agreed on is that most musical compositions can be divided into sectional units or elements whose combination and repetitions define the global structure of the music piece. These elements may be chord progressions, melodic phrases, rhythmic patterns, motifs or other musically discriminative entities at various

hierarchical levels.

2.5.2 Audio Features for Music Structural Description

Bruderer showed that the strongest cues for the presence of sectional boundaries are *harmonic progression*, *rhythm changes*, *timbre changes*, and *tempo novelty* [Bru08]. He also suggests that the harmony or the timbre feature aspect alone carries all the information needed for the structural segments to be discerned. However, the observations made are mainly based on Western pop music and the evaluation corpora used are collected on a basis of general structural coherence. We will investigate different audio features as structural descriptors for different music genres in Chapter 5.

The most popular features for music structural analysis include chroma features and MFCCs [PMK10]. Some works also combine the two to obtain a composite description [Ero07; LNS07]. Although rhythmic content is one of the most important factors to influence human perception of the music structure [Coo63], features characterising this aspect are much less employed for analysing music structure compared to timbre and harmonic features and are hardly used alone. A few works however indicated the potential of rhythmic features in music segmentation such as the *cyclic tempogram* [GMK10] and the *beat spectrum* [FU01].

Besides extracting features synchronised at a frame level, a popular approach is to extract features at a scale divided based on tracked beat events, i.e., to *beat-synchronise* the features extracted. This can be done by setting the hop size equal to the beat-length of the music (typically 300 - 400 ms), or by synchronising the feature fragments according to the beat position using statistical aggregation functions [LS08]. This is enabled by a reliable *beat tracking*. Figure 2.8 shows the beat-synchronised chroma feature for the song "Love me do" by The Beatles. The beat-tracking algorithm used is presented by Ellis [Ell07]. Higher temporal coherence is presented as a result of frame-to-beat aggregation, despite degradation in the resolution. For music with structured beat

patterns, beat synchronisation allows for extracting the features at musically meaningful scales and is proven generally beneficial for the quality of audio summarisation [Pee03].

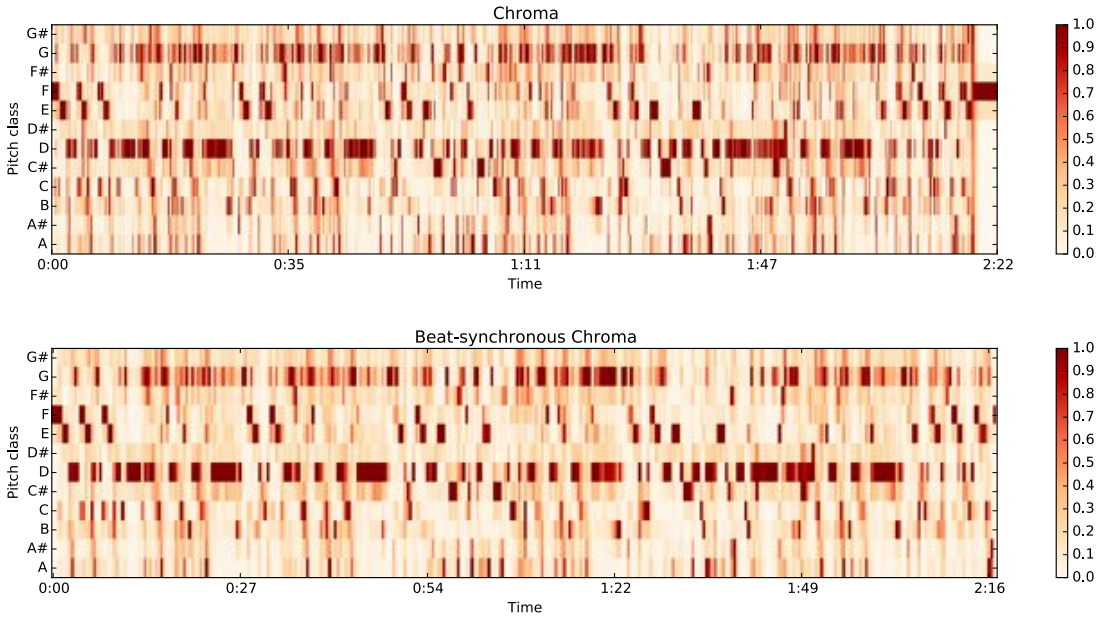


Figure 2.8: Beat-synchronised chroma feature for audio example "Love me do".

2.5.3 Methods for Music Segmentation

The extraction of relevant features underpins the structural information of the piece of music. The analysis typically proceeds with the generation of a representation indicating the structural characteristics more intuitively to enable the recognition of the sectional units. In this section, we present a survey of MSS methods presented in recent works with the evaluation methodologies introduced shortly in Section 2.8.

One common approach operates by measuring the similarity within vectors of the feature matrix of a song, which compares all pairwise combinations of feature matrices using a quantitative similarity measure. We note the ith vector of a D-dimensional feature matrix v_i, where $i \in [1, 2, .., N]$ denoting the index of the frame. The resulted similarity data is embedded in an $N \times N$ square matrix S derived using a generic similarity measure. The elements of S are:

$$S(i,j) = 1 - d(v_i, v_j), \tag{2.23}$$

where $i, j = 1, 2, ..., N$ and $d(v_i, v_j) \in [0, 1]$.

Distance metrics $d(v_i, v_j)$ commonly used include the Euclidean distance, cosine distance, and the exponential distance [FC03]. Different distance metrics also include the squared Euclidean, Manhattan and correlation distance [KK83].

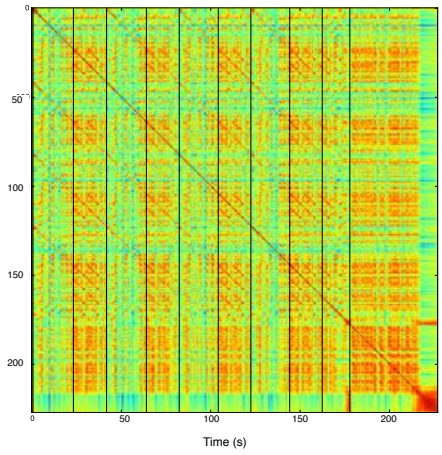


Figure 2.9: Self-similarity matrix calculated for the Beatles song "Hello goodbye" using the chromagram feature with the Euclidean distance. Black vertical lines indicate segment boundaries.

The output of this step is a symmetric matrix indicating the pairwise resemblance of each entry of the feature matrices, denoted the *self-similarity matrix* (SSM) [Foo00]. Figure 2.9 shows the SSM computed for the Beatles song "Hello goodbye" (see Section 3.4.1 for dataset description) using the chromagram feature with the Euclidean distance. The black vertical lines represent annotated segment boundaries. We can observe brighter pixels indicating higher similarities aggregate into stripe and block patterns with their borders coinciding with boundaries in the ground truth annotation. Such patterns exhibited in the SSMs are typically used as indicators of the structure of the music in MSS works.

Alternatively, one can recognise the structural patterns from the feature representations directly using classification methods. For example, features can be represented by statistical or probabilistic models, such as the *hidden Markov models* (HMMs), where segment types are represented by HMM states [AS01]. In recent years, methods relying on supervised learning, particularly, convolutional neural networks (CNNs), have been successfully applied and leading the state-of-the-art for MSS tasks [USG14; GS15a; GS15b]. In the majority of these works, CNNs are trained using the *logarithmic Mel spectrogram* feature with manual annotations. The network output is typically an activation curve indicating boundary probabilities, similarly to the novelty curves.

An overview of recent work for MSS is summarised in Table 2-A, where the segmentation methods are categorised into three types: the *novelty*-based, *homogeneity*-based, and the *repetition*-based. In the remainder of this section, we introduce a few popular segmentation algorithms relying on these three strategies which will later be discussed in this thesis.

Novelty-based Method

The novelty-based method employs the hypothesis that segment boundaries are characterised by prominent changes in audio features. One classical example is introduced by Foote [Foo00]. Firstly, a sliding Gaussian-tapered kernel is correlated along the main diagonal of the SSM. A time-aligned signal indicating the "novelty" of the current frame in its vicinity, denoted the *novelty curve* (NC), can subsequently be obtained by summarising the element-wise multiplication of the kernel and the SSM diagonal stripe. In this process, the size of the kernel k_G directly affects the properties of NC. A small kernel detects novelty on a finer time scale while a large one produces much smoother NC hence detects longer temporal structure. Then, segment boundaries are extracted when the local novelty scores exceed certain thresholds, typically the median-filter derived adaptive thresholds (see Section 2.5.3). The segment boundaries can subsequently be detected corresponding to peaks in the NC. Besides the Foote approach, many recently

Name	Features	Principle
Foote [Foo00]	MFCCs	Novelty
Ong and Herrera [OH05]	Combined features (MFCCs, subband energy and spectral features)	Novelty
Tian el al. [Tia+14a]	Tempogram	Novelty
Ullrich et al. [USG14]	Log Mel spectrogram	Novelty
Grill and Shlüter [GS15a]	Log Mel spectrogram	Novelty
Grill and Shlüter [GS15b]	Log Mel spectrogram, self-similarity lag matrices	Novelty
Aucouturier et al. [APS05]	Spectral envelope features	Homogeneity
Barrington el al. [BCL10]	Chromagram and MFCCs	Homogeneity
Levy and Sandler [LS08]	timbre features	Homogeneity
Nieto and Jehan [NJ13]	Chromagram	Homogeneity
Peeters and Rodet [Pee03]	Dynamic features	Homogeneity
Rhodes et el. [Rho+06]	Constant-Q envelope features	Homogeneity
Chai [Cha05]	Chromagram	Repetition
Mauch et al. [MND09]	Chromagram	Repetition
McFee and Ellis [ME14b]	Chromagram and MFCCs	Repetition
McFee and Ellis [ME14a]	Constant-Q with time delay embeddings	Repetition
Nieto and Bello [NB14]	2D-Fourier Magnitude Coefficients (2D-FMCs)	Repetition
Ong et al. [OGS06]	Pitch class distribution features	Repetition
Paulus and Klapuri [PK06]	Chromagram and MFCCs	Repetition
Rhodes and Casey [RC07]	Timbral features	Repetition
Shiu [SJK05]	Chromagram	Repetition
Paulus and Klapuri [PK09]	Chromagram, MFCCs and rhythmogram	Combined
Serrà et al. [Ser+12]	Structural features	Combined
Xu et al. [XMK06]	Cepstral and chroma features	Combined

Table 2-A: A summary of music structural segmentation methods.

emerged CNN-based approaches can also be categorised to the novelty category [GS15a; GS15b; USG14].

However, one limitation of many novelty-based methods is that they can only locate boundary positions of the underlying segments, leaving the types of these segments undetermined. Many works then resort to a clustering algorithm to address the segment type labelling [Foo00], or to use novelty-based methods as part of or as the first step of more complex methods [Ser12b; PP13].

Homogeneity-based Method

The *homogeneity*-based approach, also referred to as the *state* approach, assumes stationarity in local statistical properties of features in structural segments. It however can be considered as the other side of the same coin as the novelty-based methods. The concept of *state* is taken rather explicitly in methods employing HMMs for the analysis [AS01; LS08]. Here, the basic assumption is that musical sections can be represented by different HMM states which produce observations from the underlying probability distribution. The states retrieved will then be classified or clustered into labelled structural sections.

One popular method is *Constrained clustering* (CC) which attempts to find structural segments by clustering audio frames into different types of sections [LS08]. First, a *hidden Markov model* (HMM) with a relatively large number of states is trained on an entire track using timbre features, with one Gaussian output distribution for each state and a single covariance matrix tied across all states. The features are then Viterbi-decoded using trained models to yield state labels for each analysis frame representing section types.

State labels within a fixed temporal window are then histogrammed where one histogram is considered as one sample of the observation. Subsequently, histograms of neighbouring frames are clustered into C_c clusters, each denoting a segment type. In the clustering process, "must-link" constraints modelled by the hidden Markov random field are set to enforce the *temporal continuity*, i.e., observations within a regulated vicinity must have come from the same section. The two main parameters involved in this

process is the number of clusters C_c which will decide the number of section types and the neighbourhood size S_n in frame to set the temporal constraints. Finally, segment boundaries are detected by locating changes in the sequence of the cluster assignments.

Another popular approach is to deploy the Non-negative matrix factorisation (NMF) algorithm [KS10; WB10; Gro+13; NJ13]. NMF has been successfully applied in various MIR tasks for parts-based decomposition since first used by Smaragdis and Brown for music transcription [SB03]. In the context of music segmentation, NMF decomposes the non-negative input feature matrix or SSM and structural sections can be discriminated over the dimensions of the decomposition matrices obtained.

Mathematically, NMF factorises a matrix $V \in \mathbb{R}^{N \times P}$ with non-negative entries into two non-negative matrices $W \in \mathbb{R}^{N \times R}$ and $H \in \mathbb{R}^{R \times P}$ such that $V \approx WH$, where R is the rank of decomposition. With this decomposition, V can be approximated as the product of two non-negative matrices W and H, where W contains the basis vectors, and the $R \times P$ matrix H supplies in its columns the coefficients to approximate each column of V as the linear combination of the columns of W. The activation functions obtained can be used to estimate events corresponding to specific basis function.

Based on the standard NMF segmentation algorithms [KS10; Gro+13], Nieto and Bello introduced a *convex constraint* to W such that it becomes the convex combinations of V, expressed as $W = VC$ ($C \in \mathbb{R}^{P \times R}$). Here the feature matrix is used as the input V. To ensure the convexity of the combination, all coefficients of C must be non-negative and the sum of each column must be 1, i.e, $c_{ij} \geq 0, \sum_j c_{ij} = 1$. The effect of this operation is that observation frames of W become the weighted cluster centroids representing potential sections of the music piece.

To detect segment boundaries from the decomposition matrices, a k-means clustering [Mac03] is carried out with the number of clusters set to *two* where each class represents respectively if there is a boundary or not. Finally, boundaries detected from each decomposition matrix are grouped and those emerged within a tolerance window are

merged into their average locations and accepted as the final boundaries. This method is denoted Convex NMF (CNMF).

Repetition-based Method

In principle, *repetition*-based approaches make structural discoveries by finding the temporally ordered repetitions which can typically be observed in the SSMs or state sequences. Most algorithms in this category operates by searching for stripes or line structures parallel to the main diagonal in an SSM (see Figure 2.9). Although normally appear recognisable to humans, the automatic detection for such patterns faces much difficulty especially with the presence of various musical dynamics [PMK10]. Many works have been presented to enhance the SSMs to make the repetitions more discernable, such as to enhance the contrast of the stripes to the neighbourhood areas and to combat the tempo variance which may lead a stripe to "bend" [Got03; Ong06; Pee07; Gro+13].

Mauch el al. presented a method, *Segmentino*, which finds structural segments relying on the repetitive harmonic properties and heuristics of beat lengths [MND09]. As the first step, a *correlation matrix* (CM) is calculated using the Pearson correlation score between each pair of beat synchronised chromagram feature vectors of the whole song. Here a CM is used as a replacement of an SSM. These two share the property in the sense that when two sequences starting at i and j with the length l are exactly the same, then the diagonals $Diag_{i,j,l}$ in both CM and SSM will be vectors of all ones. Next, an exhaustive search is carried out along all diagonals of the CM to find all repeated chroma sequences as segment candidates of reasonable beat length. A temporal constraint is introduced based on musical heuristics that the sequence is assumed to be within 12 - 128 beat length to be qualified as a potential segment.

Then, structural segments are retrieved from the candidate sequences using a greedy algorithm based on a score indicating the possibilities that a given sequence can represent a structural segment. Specifically, if we note S_l the length of a chord sequence which has

repeated S_n times, the candidate which maximises the score $S_r = S_l * (S_n - 1)$ will be selected to represent a segment type and any other sequence overlapping with the selected one will be removed from the candidates set. If a top score is obtained by multiple chord sequences, the one with the highest mean of the within-sequence correlation quartile is chosen. This selection step is repeated for the remaining sequences of the song until none is left in the set [Mau10]. Structural boundaries will then be detected where one type of segment changes to another.

Combined Method

Instead of relying on a single strategy, an alternative approach is to focus on modelling the properties of a good structural description, and in doing so, to combine different segmentation principles. In Sargebt et al., the segmental Viterbi algorithm was formalised as a cost optimisation problem which combined the SSM-based approach and the clustering-based approach [SBV11]. Peeters performed the SSM-based segmentation as an initial step and then used the average feature value over each individual segment as the initial cluster centroids in a k-means clustering to produce pruned segmentation [PLR02]. The obtained cluster centroids were then used to initialise the training of an HMM to obtain the final clustering result as the segmentation outcome. In [PK09], a cost function for structural descriptions of a piece that considers all the desired properties is employed, the final structure decision is then made by minimising the cost function.

Serrà et al. presented an unsupervised segmentation algorithm based on a feature called *structure features* (SF) [Ser+12]. SF is able to incorporate global properties to account for structural information in the recent past. To start with, a multi-dimensional time series is obtained by accumulating vectors of a standard chroma feature ranging across a span with time embedding, i.e., the ith sample in a multi-dimensional series $\widehat{x}_i$ is expressed as,

$$\widehat{x}_i = \left[x_i^T, x_{i-\zeta}^T, \ldots, x_{i-(m_d-1)\zeta}^T\right]^T, \tag{2.24}$$

where ζ is the time delay and m_d is the embedding dimension. A *recurrence plot RP* is then computed from the pairwise resemblance between vectors of time series. An element $RP(i,j)$ 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. It has to be noted that RP differs from an self-similarity matrix (SSM) in the sense that $RP(i,j)$ is calculated between elements embedded with time-shifts, $\widehat{x}_i$ and $\widehat{x}_j$, instead of static feature vectors x_i and x_j. 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 [EKR87].

Subsequently, RP is circularly shifted to derive the lag matrix RP_L such that $RP_L(i,j) = RP(i,(i+j-2)\%N+1)$ for $i, \in= 1,2,\ldots,N$ where % stands for a modulo function. RP_L is then multiplied with a bivariate rectangular Gaussian kernel, yielding a kernel density P_k which can be seen as another multi-dimensional time series where $P = [p_1, \ldots, p_k]^T$. The structure features SF are then defined as the row vectors of P, namely $SF(i) = P_i$ where $i = 1,2,\ldots,N$. Finally, a novelty curve is calculated from SF from which segment boundaries are detected using a standard thresholding mechanism following Foote [Foo00]. In this way, all the three segmentation mechanisms (novelty, homogeneity and repetition) are combined in the segmentation process. This algorithm is denoted *SERRÀ* and will be investigated later in this thesis.

The concept of *fusion* has also been employed in recent work to combine different segmentation methods. While the CNN-based methods commonly find boundaries according to the novelty cues exhibited from the trained networks hence can miss segment boundaries defined by global repetitions [USG14], Grill and Shlüter fused self-similarity lag matrices representing long-term properties with the log Mel spectrogram [USG14]. However, the authors noted that the network cannot retrieve the longer-term structural information contained within the lag matrices, which provide only additional novelty and homogeneity cues. One explanation of this observation is that not enough anno-

tated boundaries characterising the repetitions emerge in the analysis frame to train the network to develop the corresponding memory [GS15b]. Peeters and Bisot combined the segmentation frameworks of Goto [Got03] and Serrà et al. [Ser+12], where the former is used locally to derive the lag priors, the global structure feature is then calculated following the latter on the lag dimension to derive the final novelty curve where segment boundaries are located [PB14].

Wang and Mysore introduced an iterative boundary adjustment algorithm which can be incorporated into other baseline segmentation algorithms to refine the boundary decisions they have made [WM16]. This boundary adjustment algorithm operates by maximising the Kullback-Leibler (KL) divergence of two adjacent detected segments modelled by multinomial distributions. However, an improvement is hardly guaranteed by this boundary adjustment. For the two baseline methods tested in Wang and Mysore [WM16], it brought marginal improvement to one and rather degraded the performance of the other due to dubious boundary decisions newly introduced.

2.5.4 Discussion

To date, most MSS works are focused on Western pop music types, typically the Beatles songs [Got06; PK06; Mau+09; Smi+11; MJG14]. Although the RWC dataset also consists of also jazz and classical music, as well as a few pieces of world music, many of them contain only the "chorus" part [Got06]. Smith studied several segmentation algorithms and suggested that algorithms designed originally for the structural analysis of Western popular music are widely applicable [Smi10]. Nonetheless, the corpora used in Smith [Smi10] were still collected on the basis of general structural coherence. A result of this fact could be that the selection of audio features and the design of segmentation methods exhibit Western bias, i.e., musical knowledge observed for specific music types is encoded into the design of a segmentation algorithm, making it less suitable for music from different genres or with diverse properties.

Another dataset is created to address this problem consisting a large diversity of music genres and sources of recordings [Smi+11]. This dataset is then included in the MIREX evaluation together with the pop music datasets. Except for the machine learning based methods which are predominantly data driven, a large discrepancy in terms of segmentation performance can be found between the pop music datasets characterising the chorus-verse structure (MIREX09 and RWC) and dataset with more genre diversities (SALAMI) [MIR; MIR14; MIR15b]. This hence confirms the existence of the Western pop bias in the current MSS paradigms.

While the neural network based methods effectively counter such limitations by requiring few understandings and making few assumptions of the characteristics of the music signals to design the segmentation algorithm, their success is highly dependent on the training data supplied. The input feature is typically the Mel spectrogram and its variants [GS15b; GS15a; USG14]. However, the focus of this thesis is to extract more domain-specific audio features to summarise the music structure. By doing this, we are aiming at an interpretable representation to model the underlying structural properties of the investigated music types. Therefore, despite its recent popularity, we attempt little investigation into the black-box approaches based on neural networks for the research carried out in this thesis.

We also notice that fusing different feature aspects does guarantee stable improvements over the baselines in some works [GS15b; WM16]. This can mainly be due to the fact that segmentation algorithms used in these methods fail to recognise structural patterns reflected in newly merged features, and that different audio features indicate different boundaries hence cannot be simultaneously retrieved agreeably [Smi14]. To this end, instead of fusing multiple features or boundary decisions from multiple methods, we propose to develop new audio features to convey the structural information with genre-invariance, as well as to investigate segmentation algorithms to interpret the encoded structural patterns from feature descriptions.

2.6 Vamp Environment for Semantic Audio Processing

Although various software packages developed for MIR exist, there are issues related to the lack of standardisation of feature extraction components and input/output data formats. This leads to problems hindering reproducibility, data exchange, the use of multiple tools or scaling experiments to problems involving *big data*. A solution to these problems is presented by the *Vamp* audio analysis framework [Can09], which provides an *application programming interface* (API) for feature extractor plugins, as well as file formats for algorithm configuration and output relying on *Semantic Web* technologies [BHL01].

Vamp is an audio processing plugin system capable of extracting various features from audio data [Plu07]. A *Vamp plugin* is a binary module that can be loaded by a host application to execute feature extraction tasks. The input to a Vamp plugin can be audio data both in the frequency domain and the time domain. This simplifies the plugins enabling a straightforward frequency-domain processing and permits the host to cache frequency-domain data when necessary.

The host takes the responsibility for converting the input data using FFT of windowed frames under user configuration. After initialisation, to supply audio data and run the plugin, the host calls the plugin's feature extraction function repeatedly. The plugin then receives a set of input pointers and a timestamp. While Vamp plugins receive their data block-by-block, they are not required to return output immediately on receiving the input. They may store up data based on its input until the end of a processing run and then return all results at once.

Vamp plugins can be executed using *Sonic Annotator*, a command line Vamp plugin host capable of applying plugins to audio input and producing structured output[4]. Sonic Annotator accepts plugin descriptors detailing the parameters of the algorithms, as well as the transform descriptors which prescribe how the algorithms should be run. These

[4]Sonic Annotator: `http://vamp-plugins.org/sonic-annotator/`

descriptors are provided in *N3* format [BC11], a specific syntax to encode data using the *Resource Descriptor Framework* (RDF) [LS09]. Sonic Annotator supports audio files with common formats including *MP3*, *OGG*, and a number of PCM formats such as *WAV* and *AIFF*. It can process a list of audio files on the command line, making it a useful tool for batch processing in audio feature extraction.

Unlike an audio effects plugin [Ste99a], a Vamp plugin generates not more audio, but rather some symbolic information descriptive of certain information or pattern within it, i.e., the audio features. Typical things that a Vamp plugin might calculate include the locations of temporal instants such as note onsets, visualisable representations of the audio such as spectrograms, or two-dimensional data with numeric feature value and the timestamps such as fundamental frequency (F0). The description of output feature values is provided by another ontology, the *Audio Features Ontology* [Rai07], which conceptualises the common elements in different kinds of features of audio signals. The Audio Features Ontology can be produced by RDF-capable vamp host such as Sonic Annotator. In one word, one can monitor or launch the feature extraction through the host in an offline mode, i.e., the Vamp Plugin Ontology describes and prepares what the host will take as input for the actual feature extraction instances, and the Audio Features ontology describes its output to format the results.

One primary advantage of using Vamp plugins is that it enables flexible handling of the feature data and the feature extraction process within the context of Semantic Web. This is supported by the Vamp Plugin Ontology, which conceptualises the feature extraction process by encoding the complete configuration and execution information into the metadata associated with the plugin [Onl09]. Using this metadata, one can set up the feature extraction specifications without having to query the plugins themselves.

Another advantage the Vamp architecture may bring is that it facilitates the visualisation of the feature extraction process. Apart from Sonic Annotator, another Vamp host is *Sonic Visualiser*. It offers a graphical user interface (GUI) to visualise the audio content and the extracted features for wide audiences beyond computer scientists such

as musicologists and archivists. Vamp plugins will be used for feature extraction as will be introduced in the remainder of this thesis.

2.7 Parameter Analysis for Music Information Retrieval Systems

Parameter analysis and optimisation is an indispensable part for most music signal analysis systems. Especially in methods involving data decomposition or factorisation [KS10; SB03], metric learning [SWW08] or machine learning [Mar+14; Böc+12], parameter optimisation can have a significant influence on the system performance. Noland and Sandler investigated the low-level signal processing parameters for several tonality estimation algorithms and concluded that these parameters have a significant effect on the results [NS07]. Weihs and Ligges [WL06] compared different methods for automatic singing transcription with the parameter optimisation carried out for each singer from the database individually. The purpose of their work is to automate the transcription for unknown songs by unknown singers. McKay and his colleagues presented a framework for using and optimising a variety of classifiers implemented in the WEKA machine learning toolbox [SF16] for music classification [McK+05]. This work highlighted the limitations of traditional MIR pattern recognition tools in the sense that they are not easily usable by users with a variety of skill levels or for different research tasks. We will investigate the signal processing methods and the involved parameters in this thesis, and by doing this, to assess the applicability of MIR tools.

2.8 Evaluation of Music Information Retrieval Tasks

2.8.1 Evaluation Metrics

Audio-based MIR research is driven by various real-world applications. Core applications include retrieval tasks such as music indexing and identification, audio alignment, as well as many high-level or semantic tasks such as playlist generation, recommendation and music mood estimation [Dow03a; Cas+08; SGU14]. The evaluation of an MIR system is a complex task. MIR research is cross-disciplinary involving substantial physical and perceptual aspects of music and is also related to information science. This property may subject the evaluation of individual MIR tasks to multiple guidelines and criteria.

There has been a wealth of research investigating how to build and to improve the evaluation frameworks, mainly focusing on the development of evaluation datasets, algorithms and metrics [Dow03b; SGU14]. Many commonly investigated MIR tasks are evaluated in the annually held Music Information Retrieval Evaluation eXchange (MIREX), an international community-based evaluation campaign for various MIR tasks held annually[5]. Byrd et al. pointed out that the standard evaluation models for information retrieval systems might not be necessarily valid for music [BC02]. Therefore, evaluation of MIR systems is normally specifically designed for the tasks being assessed and is based on the test datasets [San10]. In this section, we provide a survey of the evaluation methods used for AOD and MSS as will be investigated in this thesis. We first introduce the metrics shared by these two tasks.

The quality of the onset detection or segment boundary retrieval can be assessed with the universally used metrics for event retrieval in pattern recognition with binary classification: *precision* (P), *recall* (R) and *F-measure* (F) [BR99]. Given a set of estimations $\mathbb{E}$ and a set of annotations $\mathbb{A}$, the set of correct retrieval is indicated by their intersection $\mathbb{E} \cap \mathbb{A}$, denoted *true positive* (TP). The retrieval errors fall into two categories: *type I*

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

and *type II* errors. From a statistic perspective, these two types of errors correspond to respectively the incorrect rejection of a true null hypothesis, called a *false positive* (FP), and the failure to reject a false null hypothesis, called a *false negative* (FN). In the MIR context, an FP typically refers to an estimation which is retrieved but does not belong to the annotation set; an FN refers to a target which the retrieval system has failed to detect that actually exists in the ground truth annotation set. Therefore, $\mathbb{E}$ and $\mathbb{A}$ can be seen respectively as the collection of the TP set and the FP set, and the collection of the TP set and the FN set. The precision and recall can be expressed by Equation 2.25 and 2.26,

$$P = \frac{\mathbb{A} \cap \mathbb{E}}{\mathbb{A}} = \frac{TP}{TP + FP}, \tag{2.25}$$

and,

$$R = \frac{\mathbb{A} \cap \mathbb{E}}{\mathbb{E}} = \frac{TP}{TP + FN}, \tag{2.26}$$

The F-measure is defined as the harmonic mean of the two,

$$F = \frac{2PR}{P + R} \tag{2.27}$$

Each of the above three metrics takes a value between 0 and 1, with 1 representing flawless result. To this end, an aspect that is crucial for the evaluation is, how to decide whether an estimation is correct, i.e., a TP.

For onset detection, a correct match indicates that a detected onset is within a reasonably small temporal window to the closest annotated onset. This window is typically defined as 0.05 s (±0.025 s) [MIR15a]. For MSS, a detected boundary is accepted to be correct if within 0.5 s (±0.25 s) [Tur+07] or 3 s (±1.5 s) [LS08] from an annotated boundary in the ground truth [MIR15c]. The time proximity is introduced mainly to

combat the limited precision in the human annotation. Smith and Chew however argue that locating structural boundaries at 3 s and 0.5 s are two distinct tasks, suggested by the weak correlations between the two [SC13b]

Besides the standard P, R and F, a specific set of metrics is also used for the evaluation of MSS tasks. Two median deviation measures between boundaries in the result and ground truth are calculated. *Median true-to-guess* measures the median time from boundaries in ground truth to the closest boundaries in the result, and the *median guess-to-true* calculates the median time from boundaries in the result to the closest boundaries in ground truth [Tur+07; MIR15c]. These metrics reflect how close the detected boundaries and the annotated boundaries actually are to each other.

The metrics discussed above handle the evaluation in real time. The evaluation of the MSS boundary retrieval can also operate on a frame basis, i.e., both the result and the ground truth are described and handled in short frames describing musical entities (e.g., a beat) [LS08; MIR15c].

Additional evaluation methods for music structural segmentation have also been proposed in recent works. Lukashevich proposed the *over-segmentation* and the *under-segmentation* measure relying on the information-theoretic conditional entropies [Luk08]. The effectiveness of the F-measure has also been argued against in the scenario of evaluation of segment boundary retrieval [Nie+14]. McFee et al. proposed metrics capable of handling multiple annotations at musical hierarchies [MNB15]. However, the precision, recall and F-measure are still the most popular evaluation metrics in this scenario.

2.8.2 Statistical Tests

Statistics are helpful in analysing most collections of data. A statistical test can be used to assess the significance of experiment observations before any conclusion can be drawn. The necessity of statistical evaluation of MIR experiments has been identified in several surveys [Urb+12; Fle06].

There exist different statistical tests, all of which are valid only under certain conditions. These tests can be categorised as *independent* and *dependent* (also called *paired* and *correlated*) based on whether the samples tested are independent and identically distributed or not. For example, the *Student's t-test* is a classical test used to determine if two sets of data are significantly different from each other given a pre-selected significance level. The independent and dependent Student's t-test test are used respectively, for example, when there are two samples independent from each other or when there is only one population which has been tested twice.

Tests can also be divided into *parametric* and *non-parametric*. The first category has stricter assumptions on the test data than the latter, for example, that it should be of normal distribution. Most practical non-parametric statistical tests rely on a qualitative rather than quantitative analysis. This is often achieved by ranking the data. That is, we compare the ranks of observations, rather than the direct data distributions. The *Wilcoxon signed-rank test* is a non-parametric alternative to the paired or dependent t-test.

However, a limitation of the t-test and its variant methods is that they can only assess the significance of varying at most two conditions or a single factor. The *Analysis of variance* (ANOVA) test offers a solution to multiple comparison problems, with two or more groups and one or more independent variables (factors) to test. There are a number of different types of ANOVA. The simplest is the One-way ANOVA which is the randomised experiment with one independent variable (a single factor) and three or more samples (or groups) of measurements (or subjects). *Repeated measures ANOVA* is used when the same subjects are used for each treatment, i.e. correlated samples. We may also be interested in whether different independent variables interact with each other. Factorial ANOVA is the most common way of analysing the results of a factorial experiment.

Table 2-B lists the different types of statistical tests commonly used to evaluate the results of scientific experiments. We will use them in the experiments carried out in this

thesis.

Number of Measurements	Parametric		Non-parametric	
	Correlated	Independent	Correlated	Independent
Two	Paired t-test	Independent t-test	Wilcoxon signed rank test	Mann-Whitney U-test
Three or more	Repeated Measures ANOVA	One-Way ANOVA for independent samples	Friedman test	Kruskal-Wallis H-test

Table 2-B: Some popular statistical tests.

2.9 Summary

This chapter presented the essential background for this thesis with an emphasis on the signal processing and feature extraction methods. It first introduced some commonly used auditory scales associated with human perception and the time-frequency representations for audio signals. A survey was then presented to introduce some popular audio features used for MIR. Two MIR tasks investigated in thesis, audio onset detection and music structural segmentation, were introduced in detail. We summarised the general background, the state-of-the-art methods and their applications. The Vamp Plugin framework was introduced as a tool for audio processing and feature extraction using Semantic Web technologies. This chapter is concluded by introducing the scientific evaluation of MIR tasks including the metrics and statistical evaluation tests.

Chapter 3

Music Corpora

3.1 Introduction

This chapter is devoted to presenting the evaluation databases used in this thesis. We will first introduce the musical background for Jingju as well as some related work from the MIR perspective. Several datasets have been published to facilitate the evaluation of audio onset detection (AOD) and music structural segmentation (MSS) tasks [Mau+09; Smi+11; PK09; Got+02; BKS12; Bel+05]. However, the majority of them are Western-centric consisting mainly of Western popular music [SGU14; Dow03b]. Hence they are considered less comprehensive for the research purpose of this thesis. After a survey of existing databases, this chapter proceeds to introduce the construction of two new Jingju databases designed for the AOD and MSS research respectively.

In the annotation process of an AOD dataset, annotators are responsible to spot the presence of any note onset, and to record the temporal location of its occurrence. The annotation inaccuracy can be introduced mainly by the perception of onset positions and manual labelling. For many other tasks that are considered as *high-level* such as MSS, the annotation work may involve more empirical or cognitive decision making from the annotators. For example, a segment boundary is not a physical entity that can be

"heard" from the music itself. Instead, annotators need to search for "indicators" of structural changes based on their own understanding of the music and the instructions they are given. Under such circumstances, the rules defined to produce the ground truths for a dataset can be highly influential for the evaluation result to be meaningfully interpreted.

This chapter is organised as below. We introduce the music background for Jingju as well as related work in MIR in Section 3.2. Section 3.3 discusses the audio onset detection databases that will be used in this thesis, including two from the literature and a new one comprising Jingju percussion music. Here we are concerned with the percussion instruments because we intend to analyse the rhythmic aspect as the music, will introduced shortly in the following chapters. Section 3.4 is devoted firstly to presenting a survey of the publicly available datasets for music structural analysis with their individual annotation principles, and then, to introducing a new dataset with Jingju songs. Finally, we summarise this chapter in Section 3.5.

3.2 Aural Dimensions and Music Information Retrieval for Jingju

3.2.1 Listening to Jingju

The musical system used in Jingju is known as *pihuang* (皮黄). The pihuang system is characterised by three major elements: *melodic-phrases*[1] (腔 "qiang"), *metrical patterns* (板式 "banshi") and *modes (*调式 *"diaoshi") and modal systems (*声腔系统 *"shengqiang xitong")* [Sto99; Jid81; Jia95; Liu89; Wic91]. When composing a Jingju play, specific modal systems and modes are firstly chosen to set the overall atmosphere of that play and the fundamental psychology of its major characters [Wic91]. The metrical types and melodic lines are then arranged accordingly to expressively interpret the content of each

[1] A *melodic-phrase* in Jingju 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".

passage of lyrics.

Jingju employs very characteristic tuning system. In contrast to the Western tonal music styles, the concept of "semitone" is missing in Jingju theory [Che13] (although it is sometimes used in practice as noted recently [YTC15]). An *anhemitonic pentatonic* scale is used for its main melody structure. The two additional notes do exist, they however adopt a different tuning scale [Wic91; Che13]. In the numbered notation Jingju uses, when C is the keynote, the seven notes 1, 2, 3, …, 7 correspond to C, D, E, …, B. The five notes except for the 4th and 7th adopt an equal temperament. The 4th and 7th, however, use a different tuning scale and convey more musical expressiveness. These two however are responsible for modulating between keys in Jingju performance hence are indispensable for the analysis of its pitched content [Wic91].

Liu and his colleagues demonstrated 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 a Western classical music, with around 90% of the total energies distributed in frequency components correspond to the five notes (C, D, E, G, A) [Liu+09]. Chen analysed the pitch histogram for a Jingju collection and confirmed the use of pentatonicism with small energy distribution also presented for the 4th and the 7th degree [Che13].

The melodic lines or melodic phrases corresponding to a couplet are considered the smallest meaningful musical units. A passage of melodic lines expressing specific music ideas can be grouped into a *melodic section* (腔节 "qiangjie") which can play a relatively independent role in the overall musical form. The song lyrics are organised in a *couplet* structure which lays the basis of its structural framework. A couplet is comprised of two *melodic phrases* sung with tendencies toward certain melodic patterns, and 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. The progression of melodic phrases can be grouped into a *melodic section* (腔节 "qiangjie") which can express specific music ideas and play a rather integrate role in the overall musical form.

The *metrical pattern*, meaning "accented beat style", is the most expressive characteristic element of Jingju [Wic91; Rep+14]. There are fixed number of metrical patterns, each has predefined organisation of accented beats (板 "ban") and unaccented beats (眼 "yan"). These metrical patterns are responsible for setting the arias and signalling the structural points from the play to the aria levels [Rep+14]. There are fixed types of metrical patterns, each is associated with certain melodic tendencies and dramatic contexts. Metrical patterns can be classified into two categories: *metred* metrical patterns (beat styles that use accented beats) and *free* metrical patterns (beat styles free of accented beats). While the former have specific tempi, the latter have no rhythmic regulation and the duration of every pitched note is unfixed.

Mode and modal systems are the most encompassing elements of Jingju [Wic91]. There are two major modal systems: *xipi* (西皮) and *erhuang* (二黄)– hence the name of the musical system "pihuang"– each has a principal mode under the same name of the modal system. A principal mode can be identified by its unique modal identify: patterns of its modal rhythm, lyric structure, melodic contours and key characteristics. When creating a Jingju play, the mode and modal systems are the first to be chosen to settle the overall musical atmosphere. Subsequently, melodic lines, metrical patterns and other related musical elements are selectively set and organised.

Jingju songs also have *instrumental connectives* (过门 "guomen", which literally means "through the door") with percussion ensembles. These connectives consist of cymbals and gongs and rarely overlap with the sung part. Melodic passages of a song are introduced by instrumental connectives which serve as preludes. Interlude instrumental connectives link melodic lines and clarify the structure of the song. By introducing, bridging or finalising the melodic sections, instrumental connectives are integral to the structure of Jingju songs.

3.2.2 Related Work in Music Information Retrieval

To date the majority of existing MIR research is focusing on Western music categories [SGU14; Ser12a]. Jingju is one of the most representative genres of Chinese traditional music [Onl16]. In modern times, the art form of Jingju has undergone changes with newly introduced popular and regional characteristics. This thesis only investigates the classical repertoire with traditional music pieces. It should be also noted that a Jingju song has to be differentiated from the full Jingju play, where the former excludes the theatre performance, mime, dance, and acrobatics aspects of the latter. Similar to most works in the MIR context, this thesis only concerns the analysis of Jingju songs.

From an MIR perspective, Jingju music offers interesting research topics which can challenge the current MIR tools and frameworks. Despite its rich musical heritage and the sheer size of its audience, little work has been done to analyse the music content of Jingju from an MIR perspective until very recently. It has been included as a target in a few genre classification works [ZZ03] and the acoustical properties of its singing has been studied [Sun+12]. Several works have been presented addressing its melody and pitch analysis from an acoustic [ZB14; ZCS15] or linguistic [ZRS14] perspective. Its singing performance has been studied in recent works [YTC15; Rep+15]. Black et al. investigated the potential of some low-level features as emotion cues for Jingju songs [BLT14]. These works mainly target the melodious content of the music. Featuring vivid metrical and rhythmic patterns (see Section 3.2.1), Jingju has also appealed to researchers to analyse its rhythm aspect focusing on the onset detection [Tia+14b] and percussion transcription [Sri+14].

So far the research questions for Jingju in an MIR scenario are mainly targeting its melodic and rhythmic discoveries [Rep+14]. A higher level analysis of its musical structure, however, is left largely unexploited. Jingju is initially improvised at its birth. An analytical discovery of its structure will largely assist its standardisation and popularisation as well as subsequent applications in areas such as music production and education.

In this thesis, we include this music genre in our study along with commonly studied Western pop music, aiming to bridge the gap between different music cultures.

Despite of the existence of research work, only three Jingju corpora have been presented to the best of our knowledge. The first has been used to evaluate a mood estimation study [BLT14] and a melody extraction research [ZB14]. The second has identified the possibilities of the presented corpus for melody analysis [RS14]. The third one is introduced for the recognition of percussion patterns in Jingju [Sri+14]. However, they mainly feature the singing and percussion properties of Jingju hence are considered less relevant to the research presented in this thesis.

3.3 Music Onset Detection Datasets

3.3.1 Existing Collections

Several databases have been presented to evaluate the AOD works. Two are included in this thesis. The first one comes from Bello et al. [Bel+05] containing 23 audio tracks with a total duration of 190 seconds and has 1058 onsets[2]. A small subsection of this database is MIDI-generated which removes the error introduced by manual labelling. The annotations for the rest of the database are created human labelling. We denote this dataset *JPB* in this thesis.

The second dataset is composed of 30 samples[3] of 10-second ballroom dance music, containing 1559 onsets in total, presented by Böck et al. [BKS12]. Onsets were annotated manually during slowed down playback. The annotators were given multi-resolution spectrograms of the audio to obtain good resolutions both in time and frequency. The annotations were then manually verified and corrected. All onsets within a 30-ms window

[2]A 7-onset discrepancy (1058 instead of 1065) from the reference paper is reported by the original author due to revisions of annotations. See Appendix A for details.

[3]Only a subset of this dataset presented in the original paper is received from the author for the evaluation in this thesis.

were combined into a single one with its position taking the arithmetic mean of the positions of the individual onsets. This dataset is denoted *SB* in this thesis.

Based on their instrumentation characteristics, samples in these two databases can be divided into four groups: pitched non-percussive (PNP), e.g. bowed strings; pitched percussive (PP), e.g. piano; non-pitched percussive (NPP), e.g. drums; and complex mixtures (CM), e.g. ballroom dance music. Table 3-A lists number of onset of each category for the two datasets JPB and SB. Audio samples in the *JPB* dataset are in *wav* format while in *SB* they are in *flac* lossless format. Metadata for the tracks in these datasets is given in Appendix A.

Dataset	PP	PNP	NPP	CM
JPB	482	93	212	271
SB	152	233	115	1059

Table 3-A: Number of onsets in each category in dataset JPB and SB.

3.3.2 Jingju Percussion Ensemble Corpus

As introduced in Section 3.2.1, Jingju is characterised by its vivid metrical patterns. There are four major kinds of percussion instruments in Jingju as shown in Figure 3.1: bangu (clapper drum), naobo (cymbal), daluo (big gong) and xiaoluo (small gong). Examples of annotated audio examples of these instruments can be found online[4].

Bangu is the composite form of two single instruments, a *ban* (a wooden clapper) and a *danpigu* (a wooden drum struck by two wooden sticks). It is the principal instrument in Jingju percussion and is played by the conductor to direct the whole orchestra. *Danao* and *qibo* are two cymbals that are collectively called "*naobo*". *Daluo* and *xiaoluo* are two gong instruments with different shapes and pitch ranges. Daluo normally delivers a deep, solemn sound. The sound is generated by hitting a wooden stick with its tip wrapped in a piece of cotton cloth against the gong. Xiaoluo is higher in pitch and more refined in timbre. It is played by hitting a slice-shaped stick made of bamboo or wood against

[4] `http://compmusic.upf.edu/examples-percussion-bo`

Figure 3.1: Jingju percussion instruments.

it. The Both daluo and xiaoluo carry specific connotations and are used for specific role types. Since the sounds generated by ban and danpigu, and those of danao and qibo are very similar to each other and always happen at the same time, for this thesis, we follow the established practice to group *ban* and *danpigu* into a general class *bangu*, and group *danao* and *qibo* into a general class *naobo* [LS99; Wic91].

This dataset consists of 10 30-second ensemble recordings of the four instruments introduced above. Unlike pitched instruments, most idiophones cannot be tuned. These percussion instruments are made from metal casting or wood carving therefore very subtle differences might exist between the acoustical properties of individual instruments. For each of the above kinds of instruments, we recorded audio samples with single strokes with 2 - 4 individual instruments to widen the timbre coverage. The instruments were played by a professional musician. Each instrument was played with different playing styles commonly used in Jingju performances. This would enable us to obtain a large diversity of playing techniques to approximate real Jingju music.

The recording was carried out in studio conditions at the Centre for Digital Music,

Queen Mary University of London. The audio was recorded in monophonic using an AKG C414 microphone with a sample rate of 44.1 KHz.

To generate the onset detection database with mixed instrument types, we manually mixed the individually recorded instrument examples together using Audacity[5] into 30-second long tracks, with possibly simultaneous onsets to closely reproduce the real world conditions having 732 onsets in total, all belong to the *non-pitched percussive* (NPP) category.

Manual labelling of onset locations is tedious and time consuming, especially for complex ensemble music consisting of instruments with diverse properties. The onset ground truth was constructed by taking the average onset locations marked by three participants without any Jingju background. Annotators were asked to mark the onset locations in each recording using the audio analysis tool Sonic Visualiser [Can+06] displaying the waveform and corresponding spectrogram with slow-down playback. This dataset is denoted *JP* in this thesis.

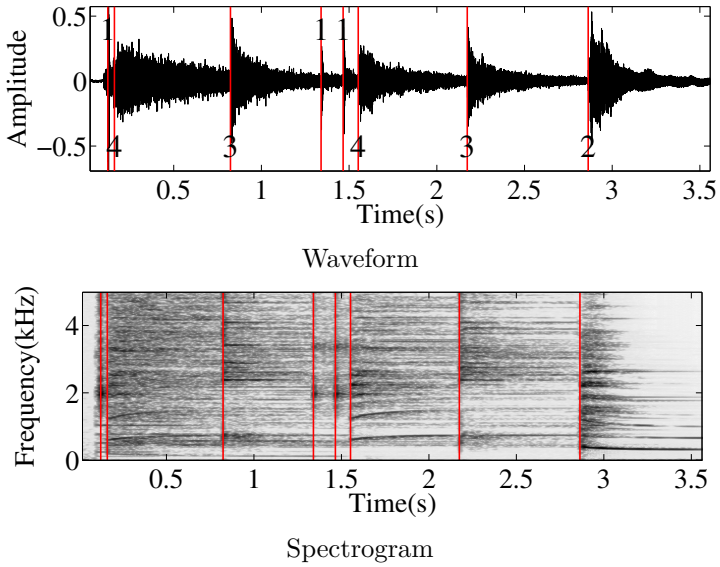


Figure 3.2: An audio example containing all four considered instruments.

Onsets generated by bangu have in general much lower amplitudes and shorter tran-

[5] `http://audacity.sourceforge.net`

sients and happen in higher densities than those generated by the cymbal instruments in an ensemble. Therefore, bangu onsets can be easily masked by cymbals and gongs. Figure 3.2 shows an audio example with all the four instruments. The top panel shows the waveform and the bottom panel is the spectrogram, the x-axis for both panels is time (in seconds). Note onsets are marked by the read vertical lines. The onsets are labelled to indicate the specific instrument onset: bangu-1, daluo-2, naobo-3, xiaoluo-4. We can see the amplitude dynamics and spectral shapes for each instrument. It can also be noted how the bangu stroke is masked by an adjoining xiaoluo stroke (0 - 0.5 s in Figure 3.2) [Tia+14b].

3.4 Music Structural Analysis Datasets

3.4.1 Existing Corpora

There are a few available collections for MSA research released over the last years. The Beatles Dataset [Pol00; PK08], The Real World Computing (RWC) Popular Music Database (PMD) [Got06], Structural Analysis of Large Amounts of Music Information (SALAMI) Internet Archive (S-IA) [Smi+11] and the Isophonics music collection by Queen Mary University of London (QMUL) [Mau+09] are among the most used ones in related work. The first three are also used for the structural segmentation evaluation in MIREX.

The *Beatles* set consists of 174 songs from The Beatles. It was manually annotated first at Universitat Pompeu Fabra (UPF) and corrected at Tampere University of Technology (TUT). However, the original annotation data provided by TUT consists of 175 instead of 174 songs[6]. In this thesis, one song is removed from the database because we have found its annotation by TUT to be dubious[7].

[6] `http://www.cs.tut.fi/sgn/arg/paulus/structure/dataset.html`

[7] This song is "Helter Skelter" from the album "The Beatles (White Album)" (CD 2). Its annotation by TUT has only one segment which starts at 0.00 s and ends at 0.11 s and labelled at "E". Additionally, this annotation file is in ".bak" format. We hence suspect that this file is a backup file and that the

We note this dataset *BeatlesTUT* in the remainder of this thesis. Due to the popularity of The Beatles music, the availability of these annotations and its formed contemporary popular music structure, it has become one of the most widely used corpora for the evaluation of structural analysis algorithms.

RWC is a large music database designed for research purpose [Got+02]. It is composed approximately half of pop music, half of jazz and classical, with a few additional pieces of world music, having 285 annotated pieces in total. However, many of the jazz and the classical pieces included have only chorus parts. The PMD is a subset of the RWC database commonly used for the evaluation of music structural analysis. RWC PMD is comprised of 80 Japanese popular music pieces and 20 Western popular music pieces. Its structural annotation is created by a graduate student majored in music.

The *SALAMI Internet Archive* (S-IA) is a publicly available subset of the full database collected in the SALAMI project[8] comprising 272 pieces. The main design consideration of the SALAMI dataset is to cover a wide variety of musical genres, mainly including Western classical music, popular music, jazz as well as world music. The SALAMI dataset has an overlap with RWC PMD and BeatlesTUT and share with the latter two 97 and 35 recordings respectively. This dataset also features a diversity of audio qualities by including a large set of live recordings. The metadata of this dataset is provided[9] by Smith with information of the track IDs, names, artists and album. The annotation data with the revision history is held on Github[10].

Finally, the *Isophonics* dataset contains 300 Western popular music pieces [Mau+09]. Around half of them are from The Beatles, the rest are from Michael Jackson, Carole King, Queen and Zweieck. It has to be noted that the annotation of The Beatles songs from this dataset is independent from TUT and MTG, although the constituent recordings are shared and they are both based on the work of Pollack [Pol00].

actual annotation file has accidentally been replaced by it.

[8] `http://ddmal.music.mcgill.ca/salami`

[9] `https://github.com/DDMAL/salami-data-public/blob/master/metadata/id_index_internetarchive.csv`

[10] `https://github.com/DDMAL/salami-data-public`

3.4.2 Annotation Principles

Different principles are employed to annotate the music structural segmentation databases including the *music similarity level*, *music function level* and *lead instrument level* annotations.

BeatlesTUT and Isophonics are annotated with section labels mainly including: "*intro*", "*verse*", "*chorus*", "*bridge*", "*refrain*" and "*outro*" with their variations such as "*verseA*" and "*verseB*", as well as a few others such as "*break*" and "*silence*". RWC PMD is annotated with the same principle. However, vocabulary for the variations of the basic terms and those belong to the "*others*" category are comparably more constrained, containing a total of only 14 unique labels. The annotations for these three datasets are made on a *music function* level, i.e., the music is segmented into structural parts expressing specific musical functions or semantic meanings. Well recognised as this principle is, it may introduce its own problems. The use of function labels conflate the notion of musical similarity with musical function, which may cause uncertainties in annotation decisions [PD09; Smi10]. Meanwhile, there constantly exists a dilemma between the completion of section types and the maintaining of precision of section nomenclature.

As opposed to these three datasets which use single annotation principles, S-IA is annotated on multiple scales incorporating the approach proposed in [PD09]. In the lowest *music similarity* level, the segments are identified to address similarities in "music ideas". The *function* level annotation is rather identical to that of BeatlesTUT and RWC 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*". Although it is beyond the scope of this thesis to analyse the labelled section types (see Section 2.8), a brief overview of the vocabulary used for labelling may give us an intuitive grasp of the underlying annotation ideas.

Figure 3.3 shows the annotations made on different levels for the song "Yellow submarine" from the above two discussed datasets with individual cover versions. The two

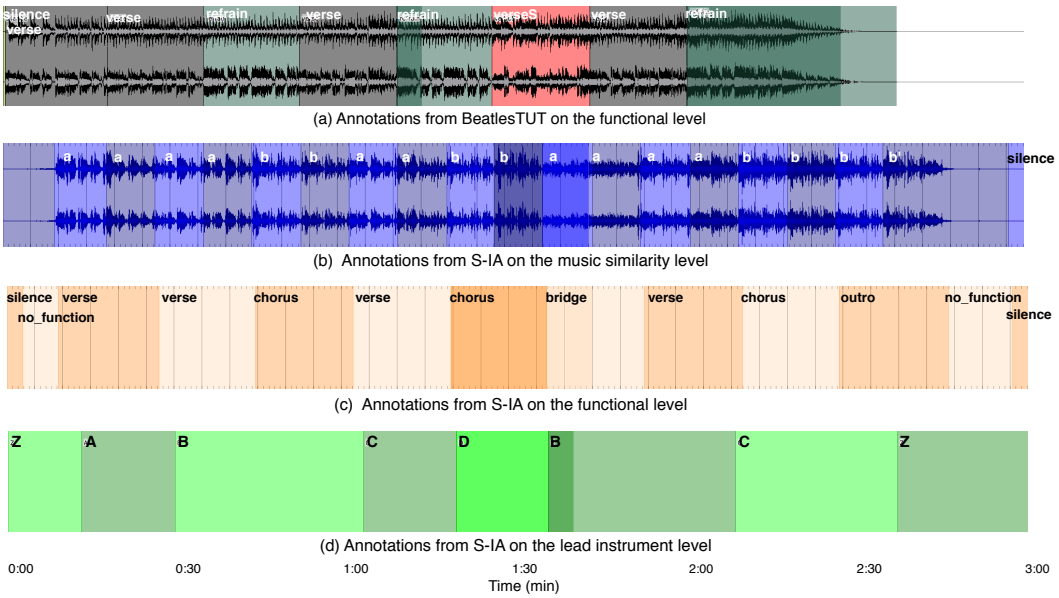


Figure 3.3: Annotations with boundary locations and segment types for "Yellow submarine" from Beatles S-IA and BeatlesTUT. Each pane from top to bottom shows respectively the music function level annotation from BeatlesTUT, music similarity level, function level and lead instrument level annotations from S-IA.

audio samples have different lengths because while the first is the studio recoding from the album of The Beatles, the second is the live recording which also contains sections with non-music background sounds. While the lowest music similarity level annotation (Figure 3.3b) can be seen as a subset of the function level annotation (Figure 3.3c) for S-IA, the lead instrument level annotation (Figure 3.3d) appears to be more independent from the two. Annotations at the same level from the BeatlesTUT (Figure 3.3a) and S-IA (Figure 3.3c) share much similarity at the chorus (refrain) and verse sections. However, annotations for the non-functional parts such as "silence" show more dependency on individual covers.

In our work, instead of investigating all the above datasets, we include only BeatlesTUT and S-IA with the lowest similarity-level annotation besides our own dataset introduced shortly. This is partly for the interest of brevity and partly for experiment design considerations. Note that although S-IA has an overlap with BeatlesTUT of 35 songs, the actual recording conditions and cover versions differ from the latter. Here we

use only the similarity-level annotation for S-IA. This is because firstly, the functional level annotations for the same songs from BeatlesTUT and S-IA differ from each other hence may conflate the evaluation of segmentation algorithms, as can be seen from Figure 3.3a and Figure 3.3c. Secondly, the highest lead instrument level corresponds only to the instrumentation variations therefore may provide limited perspectives when used to evaluate different audio features.

These datasets cover two different segmentation principles hence can serve as a comprehensive testbed for the segmentation algorithms. Besides, each of them respectively provides a textbook example of Western pop music and a large coverage of different music styles. The inclusion of these two datasets will offer us meaningful reference for the analysis of Jingju.

3.4.3 Dataset Collection

The Jingju corpus used in this thesis is composed of 30 excerpts from commercial CDs [CMG10], 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 composed of 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.

3.4.4 Annotation Process

In this work, the structural annotations are created prior to the design of audio features and segmentation algorithms. Annotations are arranged to describe the *musical novelty* or *similarity* within a piece setting aside the musical functions of segments, similar to the lowest level of the annotations for S-IA. This is mainly 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. However, low-level music similarity is a phenomenon that can be observed across different genres [Deu12]. Assessing the structure on a music similarity level provides a fair comparison between genres and datasets. Second, the melodic sections are never repeated as the chorus and verse sections would do in Western pop music. There also exists 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 thesis (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 [chu92]. The software used for annotation is Sonic Visualiser which displays the waveform and the corresponding spectrogram of the music[11]. The metadata and additional annotation information for this dataset is published online[12]. This dataset is denoted *CJ* in the remainder of this thesis.

[11] `http://www.sonicvisualiser.org/`

[12] `http://isophonics.net/content/jingju-structural-segmentation-dataset`

In this process, A1, A2 and A3 first 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 homogeneities break. Within a section, high similarity should present with a single musical idea or subject expressed. We denote annotators "in agreement" if they have independently identified the same segment boundary within a 1.0 s tolerance window. An annotated boundary will be accepted if there are at least two annotators "in agreement" with each other. The final position of this boundary will be the average of the positions that the annotators who are in agreement noted individually.

For boundaries noted by only one of A1, A2 and A3, hence the existence of disagreement, we proceed with further verifications from V1 and V2. Each would individually examine such a boundary and then decide whether it should be discarded or accepted. Should disagreement still exist for a boundary, V1 and V2 will have a discussion and decide the acceptance of it as well as its position consciously.

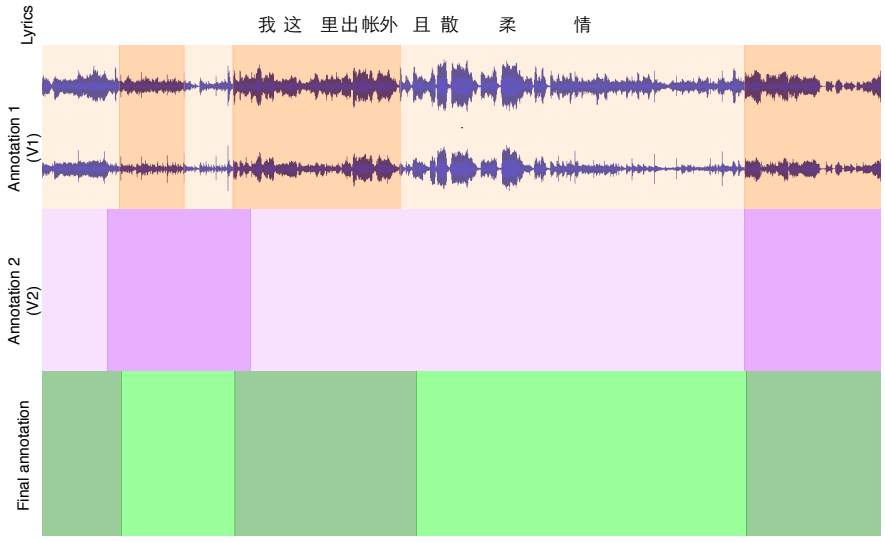


Figure 3.4: Two annotations with boundary locations and segment types for a 60-second excerpt of the recording "Ba wang bie ji" from Dataset *CJ*. Panes from top to bottom pane show respectively the lyrics of the singing (in Chinese), annotation from annotator V1 and annotator V2 and the final annotation.

Figure 3.4 shows respectively the annotations by V1 and V2 individually 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 comprises 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 their 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 find 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 Figure 3.4. 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 notes of gongs and cymbal and the fade-out effect of the vocals. 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, a detected segment boundary is accepted to be correct if within a 3-second tolerance window from an annotated one in the ground truth (see Section 2.8.1).

3.4.5 Statistics of Dataset Annotations

From the variety of existing measures commonly used to compare multiple annotations [Smi10], 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 0.5 s ($F_{0.5}$) and 3 s (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 3-B, the agreement between the annotators measured at 0.5 s ($F_{0.5} = 0.693$) is reasonably close to that measured at 3.0 s ($F_3 = 0.743$). 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 another. This is mainly because the two annotators noted different numbers of segment boundaries, as shown in Figure 3.4. This indicates that the structural annotations do depend on the annotators' individual understanding of the music as observed for Western music [SSC14].

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

Table 3-B: 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 0.5 s and 3.0 s; M_{ad} and M_{da}: median of the distances between each annotated segment boundary to its closest detected segment boundary (in second); S_{ad} and S_{da}: standard deviation of the distances between each annotated segment boundary to its closest detected segment boundary (in second).

Here we discuss some statistics of the segmentation datasets used in this thesis, including CJ, BeatlesTUT and S-IA. Table 3-C reports the number and average lengths of songs in each dataset, as well as the average number of segments and segment lengths of songs in a dataset. The distribution of segment lengths for each dataset is shown in Figure 3.5. It can also be noticed that many annotated segments in S-IA are comparably

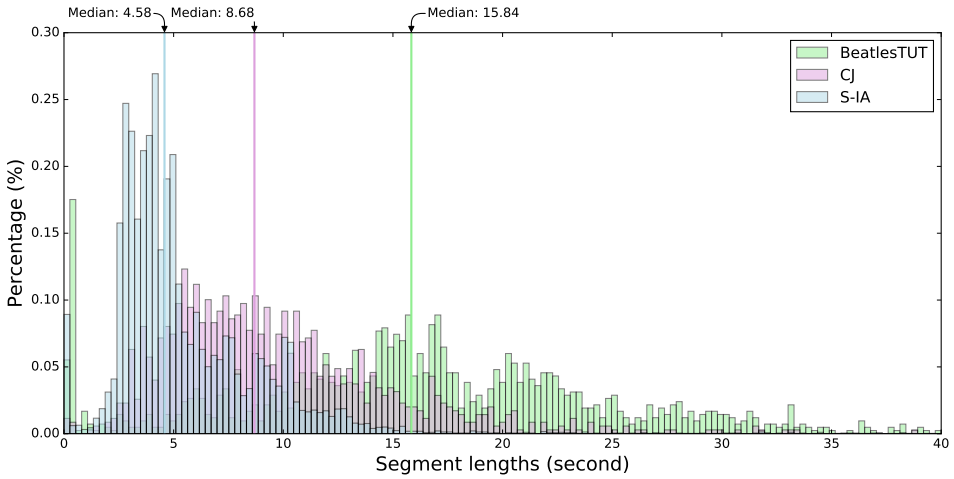


Figure 3.5: Distributions of segment lengths for MSS datasets investigated in this thesis.

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 3-C: Statistics of datasets (standard deviations in 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).

short. One intuitive concern is, would including segments that are too short lead to dubious evaluation results? For example, when a section is shorter than 3 seconds, any boundary detected within this section would be applicable to both borders when the evaluation window is set to 3s as the common practice in related work (see Section 2.8.1). Although the evaluation algorithm will assign a detected boundary only to its closest annotated boundary to avoid duplicated counting, it remains however unexplained which border the segmentation algorithm has originally meant to detect.

In this thesis, we will use these three databases to evaluate investigated audio features and segmentation algorithms in an MSS scenario.

3.5 Summary

This chapter presented the evaluation datasets for the investigated audio onset detection (AOD) and music structural segmentation (MSS) tasks. An overview of the related MIR work for Jingju and its musical background were surveyed to start this chapter. Besides introducing the existing datasets, this chapter was focused on presenting two annotated datasets of Jingju music. The first one consists monophonic recordings of Jingju percussion ensembles. This dataset is designed for the AOD and percussion instrument recognition studies. The second contains 30 excerpts of Jingju songs from commercial CDs with annotations made by expert listeners to evaluate the MSS experiments. This chapter also analysed the inter-annotator agreements from the annotation process. Finally, we discussed different characteristics in the statistics of MSS datasets with different music types.

Chapter 4

Audio Onset Detection based on Fusion

4.1 Introduction

As introduced in Chapter 1, music can be represented and perceived at different levels of abstraction. In this chapter, we will investigate the audio onset detection (AOD) task relying on *information fusion* and *Semantic Web* techniques.

The goals of this chapter can be summarised as follows: *i)* to compare different methods for onset detection and fusion strategies in the context of different music types; *ii)* to investigate the principles for effective fusion; *iii)* to study the effects of signal processing methods as well as their parameters in the scenario of onset detection and *v)* to discover potential interactions between the investigated signal processing methods.

This chapter is organised as below. In Section 4.2, we will investigate new onset detection algorithms based on different fusion strategies. We introduce the experimental platform and present the parameter optimisation experiment in Section 4.3. Section 4.4 will present the results with a detailed analysis of the effects of the investigated fusion

techniques and signal processing methods. In Section 4.5 we draw conclusions for the work presented in this chapter.

4.2 Fusion of Onset Detectors

Onset detectors may be combined at various levels using different fusion policies. Three fusion strategies are investigated in this thesis including the fusion of onset features, which essentially combines the underlying assumptions involved in the original ones hence produces new algorithms; the fusion of different detection functions to create a combined onset features and finally, the combination of onsets detected by different methods leading to decision fusion using heuristics.

In this work, six relatively simple onset detectors were selected that use only spectral information including: *High frequency content* (HFC), *Spectral difference* (SD) *Complex domain* (CD), *Broadband energy rise* (BER), *Phase deviation* (PD) and SuperFlux (SF) [Bel+05; BW13a]. See Section 2.4.3 for a review of these methods. We choose these methods because they are commonly employed in related MIR tasks such as beat and tempo tracking using onset detection as the fundamental step [DP04; GMK10; Tia+15]. The inclusion of SF also represents the more recent improvements. The improvement of these methods can thus be expected to yield improvements in subsequent applications as will be introduced in Section 5.4. Meanwhile, they rely on simple spectral information hence can more straightforwardly reflect the effect of fusion than methods relying on more advanced or black-box processes do.

The basic signal processing modules in our onset detection workflow, including the pre- and post-processing and peak picking, follow the Queen Mary Vamp Plugins (QMVP) [Plu06], as illustrated in Section 2.4.3. We have also re-implemented the entire system to enable the fusion as well as to expose more user configurable parameters as will be introduced in Section 4.3. Note that for SF, we only follow the original algorithm [BW13a] for the calculation of the ODF but use the pre- and post-processing and peak picking

methods from QM Onset Detection Vamp Plugins. We will compare our modified version of SF and the original algorithm in Section 4.4.3.

Here we consider the pairwise fusion of baseline methods. The fusions apply at different stages of the workflow involving the calculation of onset detection function (ODF), peak picking (PP) and onset localisation, as summarised in Figure 4.1. Two kinds of spectral descriptions M1 and M2 can be combined into a single onset feature using *early fusion*; two onset detection functions ODF1 and ODF2 can be fused using *linear fusion*; and finally, two sets of onsets represented by their time stamps TS_1 and TS_2 detected by two algorithms can be combined using *decision fusion*. In the remainder of this section, we will introduce all the investigated fusion algorithms.

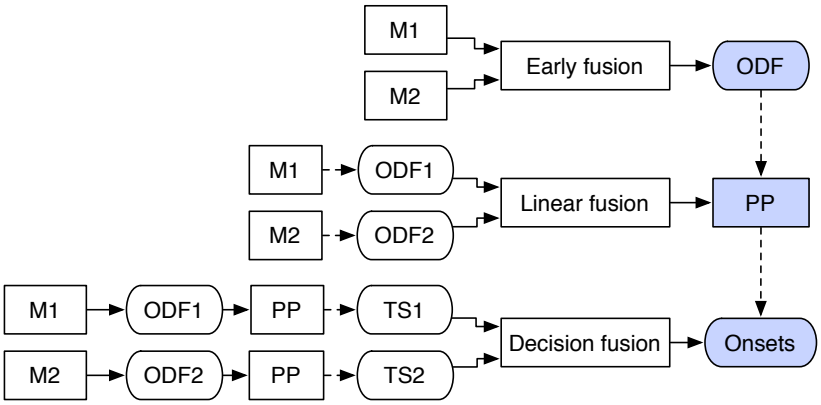


Figure 4.1: Fusion strategies investigated: early fusion, linear fusion and decision fusion.

4.2.1 Early Fusion

In case of early fusion, multiple algorithms are combined to derive fused features. The hypothesis is that by combining different detection methods, different aspects of the onset-related information may be captured at the same time. Assuming that different detection methods are complementary, the fusion will thus improve the results. Practically, this may result in creating new algorithms which rely on the underlying hypotheses of the original methods. Alternatively, we may use one feature, such as the presence of broadband energy, as a condition for modifying the other detection function before the

fused detection function is calculated.

However, not all algorithms can be meaningfully combined using early fusion. For example, CD can be considered as an existing combination of SD and PD, therefore combining CD with either of these methods using this approach is less sensible. Next, we will briefly introduce the early fusion strategies presented in this study. In the remainder of this chapter, fusion algorithms are denoted by combining the acronyms of the constituent methods.

Combinations of BER and Spectral Difference based Methods

We propose to combine BER with three other methods (SD, CD, SF) by calculating the difference between adjacent frames in two different ways. In the first approach, BER can be extended to take multiple frames into account using the spectral difference hence characterise the rate of change of magnitude in each frequency bin:

$$D(n,k) = 20log_{10}\frac{|Y(n,k) - Y(n-1,k)|}{|Y(n-1,k) - Y(n-2,k)|}, \tag{4.1}$$

where $Y(n,k)$ is the magnitude spectrogram of the complex spectrogram $X(n,k)$, $n \in [0:N-1]$ is the frame index ($n \geq 2$ in Equation 4.1), $k \in [0:K-1]$ is the frequency index of the magnitude spectrum. This quantity depends only on the rate of change and it is positive when the local increase in magnitude with respect to time is above linear. We can then define the ODF by counting the number of bins with accelerating rate of magnitude changes:

$$ODF(n) = \sum_{k=0}^{K-1} \begin{cases} 1 & \text{if } D(n,k) > \theta' \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where θ' is a variable threshold parameter.

A similar implementation is applicable in the complex domain, resulting in the early

fusion of CD and BER. The acronym of this fusion method *ends* with BER (e.g. SDBER) in the remainder of the thesis.

The above approach is focused on enhanced percussive onsets detection, however, we can also use BER to suppress the non-onset related local dynamics such as vibratos, in cases when there is no evidence of broadband energy rise. To this end, we mask the detection function $\Theta(n)$, standing for the ODF of either SD, CD, or SF, such that $\Theta(n)$ is used directly when a sharp energy rise is indicated by BER, otherwise it is smoothed by a median filter. This is described in Equation 4.3,

$$ODF(n) = \begin{cases} \Theta(n) & \text{if } BER(n) > \gamma \\ \lambda \cdot \mathsf{median}\left(\Theta(n)\right) & \text{otherwise,} \end{cases} \tag{4.3}$$

where γ is an experimentally defined threshold and λ is a weighting constant empirically set to 0.9. The size of the median window is experimentally set to 3. The acronym of this method *begins* with BER in the rest of the paper (e.g. BERSD).

Combination of High Frequency Content and Spectral Difference Methods

High frequency content (HFC) is one of the earliest methods introduced to detect transient events in audio signals [Mas96]. HFC is closely related to the perceptual brightness (spectral centroid) of a sound, but it simply characterises how dominant the higher frequency components are as opposed to measuring the central tendency. Various studies define the calculation as the sum of frequency-weighted energy [Mas96] or magnitude with linear or exponential weighting [Bel+05; JA03].

Changes corresponding to note onsets can be more prominent in higher frequencies both in case of harmonic and percussive sounds. However, these changes may be masked by energy concentrated in the lower frequency areas. For this reason, we propose an early fusion method which combines high frequency weighting with other detection methods including SD, CD, PD and BER to measure the spectral difference emphasising contents

with higher frequencies. As an example, the ODF for the fusion of SD and HFC is given in Equation 4.4:

$$ODF(n) = \sum_{k=0}^{K-1} \nu(k) H\left(|Y(n,k) - Y(n-1,k)|\right), \tag{4.4}$$

where $n \geq 1$, $\nu(k) = 1 + k$ for $k \in [0, K-1]$ is a linear weight as a function of the frequency index of the STFT bins, $H(x) = \frac{x+|x|}{2}$ is the half-wave rectification function.

Rectified Phase Deviation

For softer onsets which are always characterised by longer transients, the changes in the ODF from frame-by-frame deviation tracking may be less prominent. A possible improvement for detecting such onsets could lie in measuring the difference between frames that are further apart. To this end, we propose the *rectified phase deviation* (RPD) which calculates the phase deviation between complex spectra that are μ frames apart as shown in Equation 4.5. This method is inspired by the calculation of SuperFlux (SF) [BW13b].

$$\begin{aligned} RPD(n) = \frac{1}{K} \sum_{k=1}^{K-1} |&princarg(\phi(n,k) - \phi(n-1,k)) \\ &- princarg(\phi(n-\mu,k) - \phi(n-\mu-1,k))|, \end{aligned} \tag{4.5}$$

where μ is set to 2 following the setting in SF [BW13b] and $n \geq \mu$, and the *princarg* function maps the phase deviation to $[-\pi, \pi]$ range (see Equation 2.16).

4.2.2 Linear Fusion

In *linear combination*, two temporally aligned ODFs, ODF_1 and ODF_2, are used and their weighted linear combination is computed to form a combined detection function,

ODF_L. This process is shown in Equation 4.6,

$$ODF_L(n) = lODF_1(n) + (1-l)ODF_2(n), \tag{4.6}$$

where l is the combination weight ($0 \leq l \leq 1$). We will investigate the optimal value of l in Section 4.3.

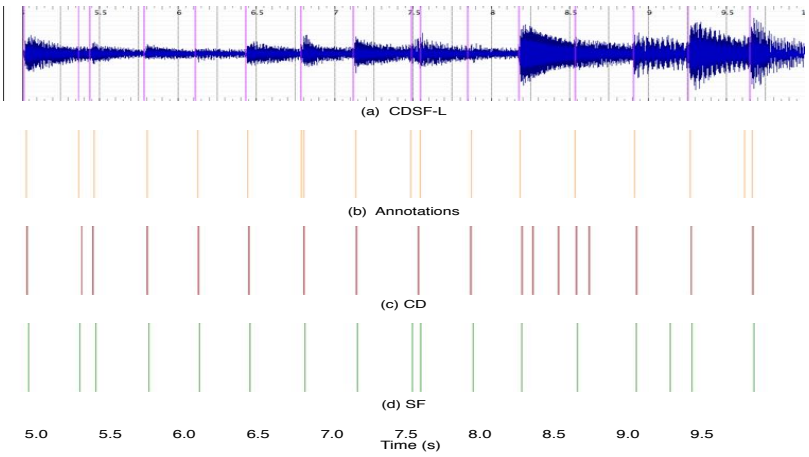


Figure 4.2: Onset detection result demonstrated using Sonic Visualiser. Panes from top to bottom show respectively the waveform and the detection by $CDSF_L$, the ground truth annotations, the detection by CD and the detection by SF.

Figure 4.2 demonstrates the onset detection results using the detector $CDSF_L$ (l=0.3). This is a method based on the linear fusion of Complex domain (CD) and SuperFlux (SF) (see Section 4.2) on a 5-second excerpt of the ballroom dance music "sb_Albums-Chrisanne3-02(12.0-22.0)" from our evaluation dataset *SB* (see Section 3.3). We can notice that many spurious onsets detected by individual methods do not appear in the detection by $CDSF_L$. Additionally, when the two individual methods have indicated slightly different locations for the same onset, the linear fusion has the effect of averaging the two locations hence reducing the deviation in the two inferences.

4.2.3 Decision Fusion

The third fusion policy studied is *decision fusion*, which operates at a later stage and combines prior decisions of two detectors. In case of many onset detection methods we experience a high rate of false positives when the sensitivity of the peak picking is high. To counter this effect, we introduce a constraint that the detection must be reinforced by two different detectors. To be precise, only when a peak is detected by multiple detectors with locations presented in a short tolerance window, it will be accepted as an onset.

Let $\mathcal{TS}_1$ and $\mathcal{TS}_2$ be the lists of onset locations produced by two detectors, i and j be indices of onsets in the candidate lists and τ the tolerance time window. The final onset locations are obtained by taking the average onset times indicated by these two detectors, as described by Algorithm 2.

Algorithm 2 Onset decision fusion.

1: **procedure** DECISIONFUSION($\mathcal{TS}_1, \mathcal{TS}_2, \tau$)
2: $\quad\mathcal{TS} \leftarrow$ *empty list*
3: $\quad$**for all** $i \in \{0, \ldots, len(\mathcal{TS}_1) - 1\}$ and $j \in \{0, \ldots, len(\mathcal{TS}_2) - 1\}$ **do**
4: $\quad\quad$**if** $abs(\mathcal{TS}_1[i] - \mathcal{TS}_2[j]) < \tau$ **then**
5: $\quad\quad\quad$**insert sorted:** $\mathcal{TS} \leftarrow mean(\mathcal{TS}_1[i], \mathcal{TS}_2[j])$
6: $\quad$**return** $\mathcal{TS}$

4.2.4 Fusion Detectors

The three fusion policies investigated are not applied to all combinations of baseline methods. This is mainly due to considerations of how various baseline methods may complement each other as well as the computational cost. Additionally, in some cases, early fusion is not possible or not meaningful to carry out, since it involves modifying the algorithms before the detection function is calculated. The second and third fusion policy are applicable in all cases, but instead of combining all pairs of baseline methods using the two fusion policies, we apply them to selected methods. We exclude some pairs of methods where one technique is considered an existing refinement of another, for instance, SD and PD with CD or SD with SF. The investigated onset detectors based

on fusion are listed in Table 4-A.

In this work, we only consider pairwise fusion of detection methods. The primary purpose of this chapter is to investigate the effect of different fusion strategies in the context of audio onset detection. Including only pairwise fusion makes the experiment more computationally feasible and the comparison of fusion strategies more straightforward.

With the 6 baseline methods, the above introduced fusions will give us 10 early fusion, 13 linear fusion and 13 decision fusion methods. The set of fusions are given in Table 4-A. The names of the fusion algorithms are derived from the abbreviations of the constituent methods and the symbols "E","L" and "D" represent *early* fusion, *linear* fusion and *decision* fusion of pairs of baselines respectively. The implementation of the investigated algorithms as well as the onset detection system are available online[1]. These 42 detectors will be evaluated in Section 4.4.

method	BER	CD	HFC	PD	SF	SD
BER	N/A	E,L,D	E,L,D	L,D	E,L,D	E,L,D
CD	E	N/A	E,L,D	L,D	L,D	L,D
HFC			N/A	E,L,D	L,D	E,L,D
PD				N/A	E	L,D
SF					N/A	
SD	E					N/A

Table 4-A: Onset detection methods with fusion investigated in this thesis.

4.3 Parameter Search

4.3.1 Parameter Specifications

To find the configurations yielding the best onset detection rates and fusion performances in a general case, an exhaustive grid search is carried out for all major parameters involved in the system, including: *adaptive whitening* (WT), *low-pass filtering* (LP) and *cutoff frequency* (f_c), *median filtering* (MF) and its *constant offset* (δ) and the *median*

[1] `https://code.soundsoftware.ac.uk/hg/unused_vampy`

window length (ψ) in the adaptive thresholding, *polynomial fitting* (PF), *detection sensitivity* ($sens$), and *backtracking threshold* (bt), as well as the *linear combination coefficient* (l) used in linear fusion and *tolerance window length* (τ) used in decision fusion. The majority of these eleven parameters are shared by all investigated detectors except for l or τ, which are used only when the detector involves linear or decision fusion.

Among the investigated parameters, LP, MF and PF are treated as binary parameters incorporated in the parameter space taking the values $0, 1$. When a certain signal processing block is disabled, indicated by the binary parameter taking the value 0, the other parameters involved in that process will become ineffective hence are omitted from the evaluation (for example LP and f_c). When PF is disabled, a simpler peak picking method is used which is controlled by the same sensitivity parameter $sens$ (see Equation 2.22).

For each non-binary parameter p, we define a sensible range around experimentally set default values and linearly subdivide this range to obtain a closed set of parameter settings S_p, with $S_p = \{\mathsf{min_p}, \ldots, \mathsf{default_p}, ..., \mathsf{max_p}\}$. The number of subdivisions were chosen to provide a balance between the granularity of the evaluation and the size of the resulting parameter space. The Cartesian product of these sets yields an N_g-dimensional parameter grid, where $N_g = 11$ as the number of parameters tested. The number of elements in each dimension depends on the granularity of the investigated parameter. Parameters investigated in this chapter are listed in Table 4-B.

Intuitively, parameter tuning for an algorithm or system can be done in a one-by-one manner. When one parameter is under assessment, all the others involved in the same system are fixed at experimentally defined values. After the optimal setting for this parameter is found, it is fixed to this setting and this process is repeated for the remaining parameters. However, the effects of two parameters in a system may depend on each other. That is to say, *interactions* may exist among parameters of interest. Consequently, it would be less informative to investigate these parameters individually than in an overall and exhaustive manner with all possible combinations of the investigated parameters and

Parameter	Default	Min	Max	Discrete Points
Adaptive whitening (WT)	1	0	1	2
Low-pass filtering (LP)	1	0	1	2
Cutoff frequency (f_c)	0.34	0.2	0.5	9
Median filtering (MF)	1	0	1	2
Constant offset (δ)	0.0	-1.0	1.0	9
Median window size (ψ)	7.0	5.0	9.0	5
Polynomial fitting (PF)	1	0	1	2
Detection sensitivity (sens)	70.0	10.0	100	10
Backtracking threshold (bt)	0.9	0.4	2.4	9
Linear combination coefficient (l)	0.5	0.0	1.0	11
Tolerance window length (τ)	0.03	0.01	0.05	9

Table 4-B: Parameters investigated in the onset detection system. Min, max, default values and numbers of discrete points are listed.

their individual granularities provided. Additionally, manual parameter tuning can be a tedious process for our experiments involving 42 onset detectors.

To this end, we propose to carry out an automatic parameter configuration and optimise all involved parameters, where all combinations of all parameters with all of their individual possible configurations will be tested. This would enable us to investigate the effects of the signal processing techniques and their parameters on the system performance in a statistical manner, as well as to discover the potential relations between different parameters and consequently enable the selection of parameters suitable for certain instrument or music types.

The 11-dimensional parameter space would generate a huge number of configurations (57,736,800) with the 11 investigated parameters given their individual granularities (see Table 4-B). This means that each of the 42 onset detection algorithms, with the computation complexity of O(n), will be tested around 58 million times under different configurations. The evaluation dataset consists of three individual databases including two public available ones (JPB and SB) and a newly composed one with Jingju percussion instruments (JP), having a total duration of 790 seconds (see Section 3.3). An initial estimation for the execution time of the was over 40 weeks on a 2-Ghz four-core computer.

This prohibiting computation cost is mainly introduced by the scale of the experiment.

The sheer size of this experiment implies that building a robust and manageable experiment framework is necessary for the completion of this work. Therefore, we separate the experiment into two parts involving different stages from the onset detection framework. The first part is focussing on parameters related to standard onset detection processes and fusion, including adaptive whitening WT, linear combination coefficient l, backtracking bt and detection sensitivity $sens$. The second is focussing on the post-processing and peak picking process, including low-pass filtering LP, cutoff frequency f_c, median filter based adaptive thresholding MF, constant offset for the adaptive thresholding δ, median filter window size ψ, polynomial fitting PF and detection sensitivity $sens$. The sensitivity threshold parameter $sens$ is included in both sets of assessment because, as introduced in Section 2.4.3, this parameter is related directly to the acquisition of onsets hence is essential for the evaluation.

The two separate parts yield a five- and a seven-dimensional parameter space respectively. The number of configurations is derived from the multiplication of the number of constituent parameters and their granularities in a test. The first part has 180 configurations for each baseline method and early fusion detector, 1,980 for each linear fusion and 1,620 for each decision fusion detector. The difference in numbers is resulted from the different fusion operation involved. The second is a seven-dimensional parameter space with 324,000 configurations for each onset detector uniformly. The total number of configurations would be the sum of those the two parts yield, which is less than one thousandth of that the original eleven-dimensional space would have yielded, namely, the multiplication of those the two parts yield individually.

During the first part of the test, the peak picking parameters are kept at default values following the settings of the Queen Mary Vamp Plugins [Plu06]. The optimal settings derived would then be used to configure corresponding parameters in the second test. To mitigate potential problems related to overfitting, we choose the optimal settings obtained for the combined datasets instead of individual ones.

4.3.2 Experiment Platform Using Vamp Plugin Ontology

Despite of the existence of numerous software packages for feature extraction, there are issues related to the lack of standardisation of the communication format between the input and output data, feature extraction components and parameter configuration metadata. This might lead to problems hindering the data exchange within different processes in a single experiment. To address these problems, we use the Vamp audio analysis framework (see Section 2.6) to design the onset detection and parameter configuration experiment. Specifically, we use the *Vamp Plugin Ontology* to conceptualise the algorithm specifications and the parameter configurations.

An experiment platform is built to support parallel execution of the large number of onset detectors on multicore servers, as illustrated in Figure 4.3. Rectangular boxes show the four main components in our framework: the generation of the configuration specifications as well as the transform files, the scheduling of the onset detection subprocess on multicore servers and the evaluation of the result for knowledge generation.

Onset detection algorithms are implemented as Vamp plugins. Given a list of onset detectors (see Table 4-A), the algorithms are grouped by parameters they share. The grouping decision is made by the *specification generator* which send queries to the onset detection plugin descriptors for their constituent parameters. The necessity of doing this lies in the different numbers and types of parameters involved in these algorithms. For instance, the tolerance threshold (τ) only applies to onset detectors involving decision fusion. Therefore, separate specification files are generated for the execution of different detectors.

The *transform generator*, fed with the parameter specifications, then produces a pool of *transform files* which are essentially the metadata the Vamp host needs to launch the feature extraction instances. These transform descriptions are encoded based on the *Vamp Plugin Ontology* as introduced in Section 2.6.

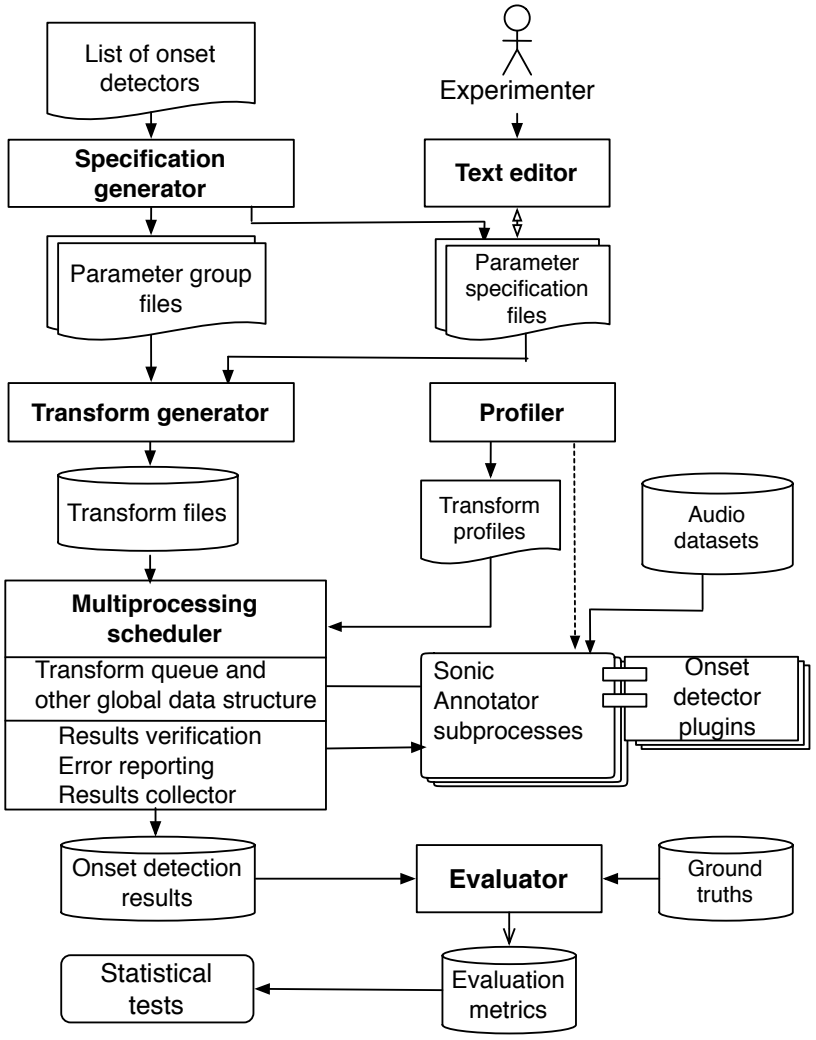


Figure 4.3: Onset detection, parameter optimisation and evaluation workflow using the *Vamp* environment.

Due to the sheer size of the transform files which must cover all configurations to be tested, the feature extraction tasks will be allocated to multiple subprocesses. The transform files are processed by the *multiprocessing scheduler*, which provides dynamic CPU allocation for the involved tasks. The scheduler is responsible for maintaining a global transform queue, execution synchronisation, resource allocation as well as progress and error reporting across several multicore servers, each has multiple Sonic Annotator instances of feature extraction under execution. These instances are launched and monitored by co-operative instances of the scheduler with each instance running on one server.

The inter-process communication between components is implemented using standard Unix pipes [Ste99b].

The outcome of each feature extraction instance is the onset detection result pointed to the specific onset detector and parameter specification used. Detection results are then collected and deposited into a single structured file system location, and finally, assessed giving the ground truths (GT) on investigated datasets by the *evaluator*. We carry out the onset detection experiment on three datasets, *JPB*, *SB* and *JP*, which consist of Western instruments, ballroom dance music and Chinese percussion instruments (see Section 3.3 for introductions). The evaluation will produce the standard *precision*, *recall* and *F-measure* results. A correct match implies that the target and the detected onset are within a 50-ms window (see Section 2.8.1).

All test audio samples are resampled at 44.1 KHz. We use a Hann window for the FFT with the window size and step size set respectively to 46 ms and 23 ms. Here the Hann window is chosen because it offers a good trade-off between the window width and the sidelobe attenuation, providing a good frequency resolution and reduced spectral leakage [Har78]. In the remainder of this thesis, the Hann window is used for Fourier transform unless noted otherwise.

The computation cost of all individual onset detection algorithms is $O(n)$. Two 32-core machines with 3.47 GHz CPU and 96 GB RAM, one eight-core machine with 2.6 GHz CPU and 128 GB RAM and one 12-core machine with 2 GHz CPU and 128 GB RAM were used for the implementation of this experiment. The overall execution took approximately 12 weeks. The use of proposed experiment platform (see Figure 4.3) had greatly reduced the computation time of having only one parameter set without the partition (see Section 4.3.1). In the next section, we will present and analyse the detection results.

4.4 Results and Analysis

4.4.1 Performance of Fusion Detectors

We first investigate the overall performance of our onset detection system including both the fusion and the baseline detectors. Table 4-C shows the detection results of top ten performing detectors for each of the four onset categories, pitched percussive (PP), pitched non-percussive (PNP), non-pitched percussive (NPP) and complex mixture (CM), as well as the overall database with mixed onset types. We also report results of the baseline methods in case they do not appear in the top-ten list. A full set of results is given in Appendix C. Results are ranked using the average F-measures over audio samples in individual datasets, given the best parameter settings for each onset type. A detailed investigation of the employed signal processing components and the optimisation of involved parameters will be presented in Section 4.4.3 and Section 4.4.4. Here we discuss the results concerning the onset detection algorithms and fusion strategies.

While the rankings vary over onset types, some detectors work reliably well overall such as $CDSF_L$ (i.e., the linear fusion of Complex domain and SuperFlux). Compared to our previous study [Tia+14a], the results reported here are improved due to the fact that more parameters were involved in the optimisation. When evaluated on the overall datasets, nine out of the top ten detectors are fusion methods suggesting fusion as an effective technique for designing onset detection algorithms. It can also be noticed that linear fusion appears in the top-ten list more frequently than the other two alternatives.

pitched percussive (PP)				pitched non-percussive (PNP)				non-pitched percussive (NPP)				complex mixture (CM)				Overall			
Method	P	R	F	Method	P	R	F	Method	P	R	F	Method	P	R	F	Method	P	R	F
$CDSF_L$	0.981	0.954	0.968	$CDSF_L$	0.749	0.808	0.777	$CDSF_L$	0.963	0.925	0.946	$BERSF_L$	0.904	0.818	0.859	$CDSF_L$	0.895	0.841	0.867
SF	0.992	0.942	0.966	$BERSF_L$	0.777	0.744	0.760	SF	0.955	0.936	0.943	SF	0.920	0.799	0.855	$BERSF_L$	0.905	0.830	0.866
$CDSF_D$	0.979	0.929	0.953	SF	0.787	0.732	0.758	$BERSF_E$	0.948	0.924	0.936	$CDSF_L$	0.893	0.815	0.852	SF	0.886	0.846	0.860
$BERSF_E$	0.985	0.916	0.950	$CDBER_L$	0.724	0.796	0.758	$BERSF_D$	0.951	0.917	0.934	$BERSF_D$	0.933	0.778	0.849	$BERSF_D$	0.898	0.814	0.854
$HFCCD_L$	0.985	0.910	0.946	$BERSF_E$	0.729	0.789	0.758	$BERSF_L$	0.936	0.906	0.921	$BERSF_E$	0.879	0.809	0.843	$BERSF_E$	0.898	0.811	0.852
$BERSF_L$	0.967	0.923	0.944	$BERSD_L$	0.718	0.764	0.740	$CDSF_D$	0.979	0.926	0.917	$CDBER_L$	0.862	0.813	0.837	$CDBER_L$	0.870	0.831	0.850
$BERSF_D$	0.982	0.903	0.941	$CDSF_D$	0.728	0.751	0.739	$BERSD_E$	0.945	0.874	0.908	$BERSD_L$	0.907	0.773	0.835	$CDSF_D$	0.875	0.819	0.846
$CDPD_L$	0.963	0.918	0.940	$BERCD_E$	0.718	0.741	0.730	$BERSD_L$	0.909	0.905	0.907	$PDBER_L$	0.900	0.778	0.834	$BERSD_L$	0.880	0.765	0.840
CD	0.965	0.915	0.939	$BERSF_D$	0.726	0.728	0.727	$CDBER_L$	0.918	0.893	0.906	$SDBER_E$	0.878	0.789	0.832	$HFCCD_L$	0.871	0.802	0.835
$CDSD_L$	0.961	0.910	0.935	CD	0.717	0.728	0.723	$BERCD_E$	0.831	0.877	0.903	BER	0.845	0.804	0.824	$CDSD_L$	0.862	0.792	0.826
SD	0.960	0.851	0.909	BER	0.770	0.633	0.695	CD	0.949	0.852	0.898	BER	0.900	0.778	0.834	CD	0.887	0.771	0.825
HFC	0.980	0.811	0.907	SD	0.673	0.597	0.658	BER	0.933	0.864	0.897	CD	0.833	0.770	0.801	BER	0.861	0.783	0.820
BER	0.937	0.684	0.885	HFC	0.604	0.706	0.651	SD	0.954	0.836	0.893	HFC	0.848	0.746	0.794	SD	0.865	0.753	0.805
PD	0.935	0.428	0.787	PD	0.655	0.498	0.636	HFC	0.966	0.824	0.889	SD	0.903	0.671	0.791	HFC	0.838	0.768	0.801
N/A				N/A				PD	0.589	0.566	0.577	PD	0.694	0.456	0.672	PD	0.655	0.726	0.689

Table 4-C: *Precision*, *recall* and *F-measure* for the ten best detectors and all rest baseline detectors under optimised configuration. Results are evaluated on the four individual onset types as well as the overall compound dataset. The symbols "E","L" and "D" represent *early* fusion, *linear* fusion and *decision* fusion of pairs of baselines respectively. We also report results of the baseline methods in the case when they de not appear in the top ten list.

Here we discuss the effect of fusion in the onset detection context regardless the strategies it employs. We compare the detection F-measures of each pair of baselines and their best performing fusion detectors for the mixed datasets with all onset types (see Appendix C for a full list of evaluation results). Among all 13 groups of the two baselines and their fusion detectors (see Table 4-A), seven have at least one fusion detector outperforming both their two baselines under their individual best configurations, evaluated on the overall datasets. This shows that fusion can bring improvements to baselines in a general onset detection scenario.

To verify this result, we measure the significance level of the differences among the detection F-measures yielded by all parameter configurations for each pair of baselines and their fusions. The statistical test used is the Kruskal-Wallis H test [McD09]. We use this test because it is applicable for over two samples and does not require normal distribution or equal sample sizes for tested groups, and that the subjects can be considered independent from each as we are comparing different onset detection algorithms (see Section 2.8.2). High significance ($p < 0.001$) has been found for all tested groups for the scenarios of all onset types. To find out which one of the detectors within individual tested groups has introduced the actual difference, pairwise post-hoc comparison of the F-measures of detectors in each group is carried out using the Tukey honestly significant difference (HSD) test [AW10]. We found that, however, the HSD does not consistently present in the pairwise comparisons. When one constituent baseline has already obtained satisfactory results, the improvements provided by fusion can be less significant.

Although most fusion detectors perform better than their constituent baselines, an exceptional case can be observed for *SF*, which yields consistently satisfactory detection across all onset categories and renders better detection than many of its fusions. It however has to be noted that *SF* implemented in our system is not identical to the reference algorithm in [BW13a]. The calculation of the ODF follows [BW13a] (Step ii) in Figure 2.7) while the pre- and post-processing and the peak picking processes uses our framework as described in Section 2.4.3. We report significantly higher detection

F-measures of this detector in our system than the reference implementation [BW13a] for all investigated datasets and onset types ($p < 0.0001$, Wilcoxon signed rank test). We will investigate the effects of the signal processing techniques and of the parameter optimisation in Section 4.4.3 and Section 4.4.4.

4.4.2 Analysis of Fusion Policies

Here we want to find out which investigated fusion policy is the most effective in the onset detection scenario. For the case of early fusion which involves the design of new features, the consideration has to be made whether the two baseline strategies can be meaningfully fused. For linear and decision fusion, the fusion can be applied to any pair or group of baselines. However, the resulting performance is closely related to the setting of the *linear combination coefficient* (l) and the *tolerance window length* (τ).

	pitched percussive (PP)	pitched non-percussive (PNP)	non-pitched percussive (NPP)	complex mixture (CM)	Overall
Early fusion	0.881 0.043	0.679 0.040	0.868 0.103	0.815 0.049	0.790 0.042
Linear fusion	0.921 0.026	0.708 0.041	0.904 0.014	0.816 0.024	0.827 0.023
Decision fusion	0.877 0.058	0.643 0.053	0.855 0.059	0.774 0.047	0.780 0.051

Table 4-D: Average detection F-measures and their standard deviations across audio samples of each detector with specific fusion policy. Results reported are obtained under the best configurations of individual onset detectors.

To compare the performances of the two constituent baselines and their fusions, including early, linear and decision fusion, we note that the difference among their optimal detections provided by the individual best configurations can be only marginal, as can be seen from Table 4-C. However, they vary in terms of the extent of dependency on parameter configurations as suggested by the sizes of their *interquartile ranges* (IQRs). While linear and decision fusion generally maintain the dependency on configurations of the baseline methods, early fusion makes the detector more dependent on system settings.

In order to examine the general performance of the three fusion policies, detection

results for each pair of baselines and all their available fusion detectors are compared, reported in Table 4-D. Results are derived from the average F-measure of all audio samples for individual onset types using the best performing configurations of these detectors.

We can note that linear fusion yields the best performance overall for onset categories with consistently low standard deviation, signifying its stable performance. Both linear fusion and decision fusion leave the original onset detection function of the two baseline methods unchanged. When the evaluation is carried out on all audio samples combining the three datasets, all the linear fusions outperform the decision fusions of the corresponding pairs of baselines in terms of the average F-measure. Among all pairwise fusions, however, few early fusion detectors obtain substantially superior detection than its two baselines when evaluated on individual onset categories (see Appendix C). We therefore conclude that the *linear fusion* is the most effective among the three fusion policies, given an appropriately selected linear combination coefficient.

4.4.3 Performance of Parameter Sets

In order to investigate the overall effects of the parameter optimisation related to the signal processing techniques as summarised in Table 4-B, we compare detection results obtained using the default parameter settings in QM Vamp Plugins [Plu06] and the SF algorithm by Böck and Widmer [BW13a] (denoted *REF*) to results obtained using the optimised configurations in this thesis (denoted *TIAN*) for the six baseline onset detectors, BER, CD, HFC, PD, SD and SF. The technical details of these onset detection algorithms have been introduced in Section 2.4.3.

The average detection F-measures for the tested audio samples of these baseline methods under different parameter configurations are summarised in Table 4-E. A four-fold cross validation was carried out for all detectors, using 75% of the audio samples in each fold as training data. The configuration yielding the optimal detection results for

	pitched percussive (PP)			pitched non-percussive (PNP)			non-pitched percussive (NPP)			complex mixture (CM)		
method	REF	TIAN	TIAN-CV	REF	TIAN	TIAN-CV	REF	TIAN	TIAN-CV	REF	TIAN	TIAN-CV
BER	0.725	0.978	0.885	0.578	0.856	0.695	0.848	0.897	1.000	0.815	0.834	0.978
CD	0.835	0.934	1.0	0.510	0.723	0.546	0.858	0.898	0.928	0.801	0.801	0.900
HFC	0.811	0.959	0.907	0.518	0.691	0.651	0.817	0.939	0.889	0.635	0.801	0.794
PD	0.787	0.787	0.896	0.183	0.577	0.501	0.726	0.636	0.577	0.369	0.672	0.648
SD	0.797	0.909	0.903	0.504	0.658	0.602	0.854	0.893	0.893	0.692	0.791	0.756
SF	0.623	0.966	0.894	0.468	0.713	0.603	0.403	0.947	0.887	0.625	0.848	0.772

Table 4-E: Detection F-measures of baseline detectors in the original systems (REF) and our system (TIAN) after parameter optimisation as well as the cross validation results (TIAN-CV).

the training data was then used to test the rest 25% of the audio samples. The results are shown in Table 4-E denoted *TIAN-CV*. Detection results after the cross-validation outperform the reference configuration *REF* consistently. We observe flawless detections in some cases, for example, when using detector CD on the PP onset category. However, this may partly be due to the small sample size of the test set. It can be noted that using the entire database to find the best setting (TIAN) provides only marginal improvement compared to the mean results of the folds (TIAN-CV). This hence justifies that parameter optimisation can improve the performance of an onset detection algorithm on diverse music types in a general case.

It can be observed that the parameter tuning yields better performance with very notable differences in general. The most substantial improvement is achieved in case of SF and PD. However, the comparison for SF reflects rather the improvement brought by using different signal processing methods to process the ODF and to extract the onsets. This is because in our system, only the calculation of the ODF follows Böck and Widmer [BW13a].

We observe that the original *SF* algorithm (REF) obtains very low detection F-measure on the NPP onset type consisting the Chinese NPP onsets from Dataset *JP* and the NPP onsets from the two Western datasets *SB* and *JPB* (see Section 3.3). This contradicts with the results found previously in Table 4-A. We divided the NPP onset

category into the Chinese and Western one and found while SF (REF) yield reliable detection on the latter with an F-measure averaged on all constituent audio samples of 0.679, it almost failed on the Chinese subset with consistently large numbers of *false positives* hence a low detection *precision* rate. This method is designed especially for pitched non-percussive music to combat the interference of vibratos. As can be noticed from Table C.2, SF and its fusion detectors also require a relatively low sensitivity to yield effective detection in our system, which already employs a more aggressive thresholding strategy using the polynomial fitting (see Section 2.4.3) compared to Böck and Widmer [BW13a].

For PD, as it turns, some of the post-processing steps applied by default in QM Vamp Plugins that all other methods benefit from are not useful. In particular, the optimal choice for this detector is not to use polynomial fitting ($PF = 0$) when evaluated on both individual onset types and datasets (see Section C.2 and Section C.4). The polynomial fitting based peak picking is designed to detect onsets with sharp peaks to compensate for lower amplitudes. However, in the case of PD the peak picking algorithm working in a more aggressive way with a high threshold to combat noise turns out more preferable for an effective detection. Despite being noisier overall, PD produces peaks that are spikier and exhibits less magnitude variation in the ODF. This can be explained by the fact that, as introduced in Section 2.4.3, this method does not rely on energy changes to form the onset detection function (ODF). It is hence less subject to the energy intensity variation in the audio signal as do energy-based methods. The ODF of this method however is noisy overall due to the potential phase distortion from low-quality audio recordings (such as dataset SB) and phase variations of the low-energy noises in the music.

As a conclusion, besides the development of the core onset detection function, the signal processing techniques used for noise reduction, pre- and post-processing and peak picking can also be of significant importance for a successful detection. Fine tuning the parameter configurations can greatly influence the performance of the onset detection

system. Different configurations are needed for different music genres and onset types. The extent of the benefit, however, varies across detectors. Additionally, the signal processing components involved pose diverse effects on different detectors. Next, we will investigate the individual effects of the major signal processing methods involved in our system in the context of different onset types and detection methods.

4.4.4 Analysis of Signal Processing Methods

We first discuss the parameters involved in the first part of the test (see 4.3.1). Adaptive whitening WT has to be *deactivated* for the majority of detectors to obtain their optimal performances on all datasets. For most cases, applying the adaptive whitening leads to higher false negatives hence degraded F-measures overall. This indicates that the method does not improve onset detection performance in general, although it is used by default in the QM Vamp Plugins [Plu06].

The optimal setting for the linear combination weight l varies for each pair of baseline detectors. The parameter is then set accordingly in the second part of the test (see Section 4.3.1). The best setting of the tolerance window τ is consistently 0.05 s for all pairs of late fusion on all datasets, suggesting that the temporal precision of the different detectors varies notably, which requires a fairly wide decision horizon for a successful combination. The parameters in the second set relate to the detection of changepoints from one-dimensional features in general and will be used for music structural segmentation as introduced shortly in Chapter 5.

As introduced in Section 2.4.3, backtracking is introduced to combat the effect of long attacks, in which case the perceived onset locations may *proceed* the peak locations in the ODF. The backtracking threshold bt is set to 0.9 in the QM Vamp Plugins [Plu06]. For most detectors, applying backtracking has yielded improved detection. This hence justifies the hypothesis that the peak locations in the ODF tend to lag behind when a note onset can be perceived. The best bt turns out to be unanimously at a high value

PP	F-measure	LP	f_c	MF	ψ	δ	PF	sens
mean	0.895	0.976	0.39	0.976	7.651	0.064	0.619	0.895
std	0.052	0.152	0.074	0.152	1.379	0.575	0.486	20.933
mode count	- -	1 41	0.350 11	1 41	9 25	0.500 9	1 26	80 9
PNP	F-measure	LP	f_c	MF	ψ	δ	PF	sens
mean	0.680	1	0.273	0.907	7.385	0.106	0.833	51.146
std	0.051	0.00	0.060	0.294	1.168	0.430	0.373	28.880
mode count	- -	1 42	0.238 17	1 38	6 10	0.000 14	1 35	40 6
NPP	F-measure	LP	f_c	MF	ψ	δ	PF	sens
mean	0.867	0.952	0.433	0.976	7.500	0.237	0.643	74.50
std	0.089	0.213	0.092	0.152	1.396	0.393	0.479	24.365
mode count	- -	1 40	0.500 16	1 41	9 13	0.250 14	1 27	80 15
CM	F-measure	LP	f_c	MF	ψ	δ	PF	sens
mean	0.795	1	0.392	0.952	7.225	0.217	0.833	56.93
std	0.049	0	0.086	0.213	1.351	0.335	0.373	32.250
mode count	- -	1 42	0.500 9	1 40	6 13	0.250 13	1 35	20 7

Table 4-F: Statistical summary of configurations of peak picking parameters for all investigated detectors: *mean*, *std*, *mode* and *count* of the occurrence of the mode out of 42. *PP*, *PNP*, *NPP* and *CM* stand for the onset types: pitched percussive, pitched non-percussive, non-pitched percussive and complex mixture.

for all datasets. For a large number of detectors, the optimal *bt* has reached the highest boundary 2.4 set in the experiment. This suggests that in many cases, although back-tracking is proven beneficial, the tracking breaks very soon after the tracking iteration starts (see Algorithm 1), meaning that the final onset locations can be very close to the actual peaks in the ODF. An interesting future direction for this finding could be, to investigate human perception of the presence of note onsets for specific music types.

Table 4-F shows the statistics of optimal settings for the peak picking parameters (see Section 4.3.2) for each detectors on the four onset types. The table reports the mean, standard deviation (std), mode and how many times the mode has appeared out of the 42 detectors. These figures were extracted from the configuration pool when the onset detectors yield the highest detection F-measures in the evaluation. A complete set of configuration results is given in Appendix C. Optimal parameter settings vary across onset types, suggesting that the configuration of parameters for an onset detection system

is target specific. In the remainder of this section, we discuss these signal processing methods and parameters individually.

Low-pass filtering (LP) appears to be beneficial for the overwhelming majority of the cases. With high frequencies in the input signal attenuated, the detection function exhibits much less non-onset related variations. The only exceptions were encountered by detector $HFCCD_E$ (the early fusion of HFC and CD) for the PP onset category and $CDSF_L$ and $BERSF_L$ (linear fusion of CD and SF, and of BER and SF) for the NPP onset category. For percussion onsets, peaks in the ODF tend to correspond to note onsets well. Therefore, it is less crucial that a smoothing operation should be applied. In this case, applying the low-pass filter may result in higher rates of *false negative*s, i.e., more missed onsets, and sometimes a degraded detection F-measure.

The constituent parameter of the low-pass filter is the cutoff frequency (f_c). A moderate value around 0.4 times the Nyquist frequency (22.05 KHz in this experiment) yields the best performance for most detectors. When summarised by onset type, varying f_c leads to more varied detection results for percussive instruments. For the detection of percussive onsets, too strong a smoothing processing can lead to adverse effects with increasing false negatives. For an effective detection, a more aggressive smoothing operation is required for non-percussive onsets or onsets of mixed types of instruments to suppress false positives. However, percussive onsets are detected more successfully using a higher f_c. Overall, for a realistic corpus consisting diverse onset types or music genres, low-pass filtering is proved to be a highly beneficial technique for an effective detection. Whereas for music where note onsets are accompanied by short transients with prominent energy burst, less smoothing is required to avoid a counter effect.

The second procedure applied in the peak picking processing is median filtering based adaptive thresholding (MF). Applying MF is beneficial for the majority of investigated detectors with the evaluation carried out on all onset categories. This is especially notable when evaluating the detection on a composite dataset without differentiating onset types (see Table C.4) or for the complex mixture (CM) onset type with mixed note categories

(see Table C.2). Under such cases, larger diversity is presented in the music making it less reliable to detect individual onsets with a globally defined threshold. To conclude, the median filtering based adaptive thresholding strategy is also generally beneficial in the onset detection system.

The rest a few exceptions mainly include the method *PD* and its fusion detectors, as can be noticed from the Table C.2. The fundamental concept of the median filter is to run through the signal entry by entry such that each entry is replaced by the median of neighbouring entries within a window, leading to a "flattened" signal. Despite the noise removal property, the median filtering would also lead the peaks to appear less spiky. For example, PD may fail to present onset-related peaks in the ODF easily discernable from those related to noise due to similar amplitudes and peak spreads. Applying the median removal hence may have limited effectiveness in reducing noise while preserving the actual onsets.

The two parameters involved in the median filtering process, as shown in Equation 2.20, are the median window size ψ and the constant offset δ. Most detectors benefit from a moderate ψ ranging from 7 to 9 in our work, i.e., around 0.2 seconds, when evaluated on the combined dataset with different onset categories and music types. When investigating the four onset types individually, CM and PNP benefit from slightly lower ψ than PP and NPP. This is because peaks in the raw ODF corresponding to the note onsets are generally of lower magnitudes for the two non-percussive onset types than for the two percussive ones. The constant offset provided by δ does not have a noticeable effect. Although there is a large diversity of the best setting of this parameter for individual detectors and onset types or datasets, differences in detection F-measures under different δ settings are mostly marginal.

The influence of polynomial fitting (*PF*) on the results appears to be more varied compared to low-pass filtering and adaptive thresholding. Its advantage is more prominent in case of non-percussive onsets and complex mixtures. This is indicated by the fact that it is applied for more detectors to yield their individual optimal performances

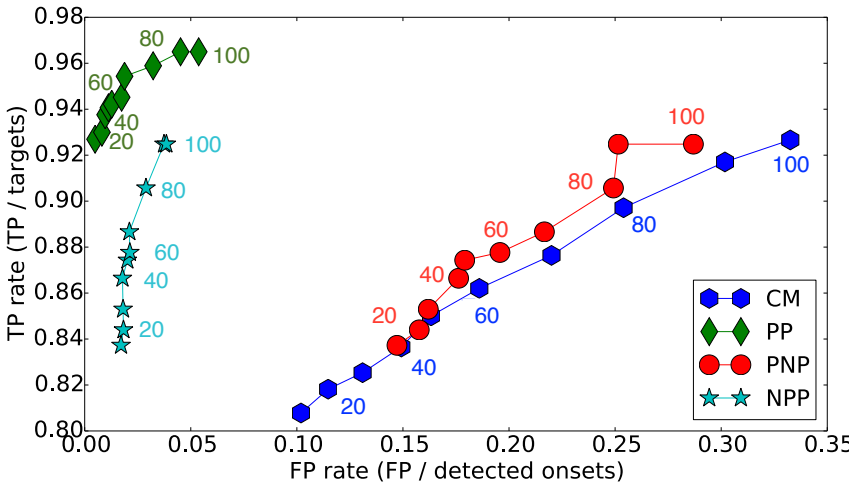


Figure 4.4: Detection *true positive* (TP) rate and *false positive* (FP) rate of $CDSF_L$ under different detection sensitivity (sens) settings (labelled on each curve) for the four onset types (annotated in the side box). *PP*, *PNP*, *NPP* and *CM* stand for the onset types: pitched percussive, pitched non-percussive, non-pitched percussive and complex mixture.

when the evaluation is carried out on the onset type PNP and CM than on PP and NPP, as shown in Table 4-F. More detectors achieve their individual best detection rates with PF applied when evaluated on each dataset (see Table C.4) than on each onset type (see Table C.2). This suggests that the more complex the corpus is, the more advantageous PF is for an effective peak picking.

To conclude, polynomial fitting based peak picking is more effective for the detection of non-percussive and complex onset mixtures than the percussive categories. As for methods that do not rely on the actual spectral energy to calculate the detection function, they are less subject to the dynamics or amplitude variations in the music signal. Consequently, the polynomial fitting based peak picking mechanism turns out less advantageous. Despite these exceptions, polynomial fitting introduces improvements to the majority of the detectors. This is supported also by the fact that the majority of the top-performing detectors have it applied when achieving their best results for each onset category or dataset as illustrated in Table C.2 and Table C.4.

The parameter involved in the peak picking process is *detection sensitivity* (sens). To demonstrate the detection result associated with different *sens* settings, Figure 4.4 shows the detection rates for $CDSF_L$, i.e., the linear fusion of the detection method Complex domain and SuperFlux (see Section 2.4.3), under different choices for *sens* with all the other parameters fixed at their optimal values. The figure illustrates the true positive rate (i.e., correct detections relative to the number of target onsets) and false positive rate (i.e., false detections relative to the number of detected onsets). Better performance is indicated by higher TP and lower FP rate. Significantly different results can be observed for percussive onsets compared to the other categories. It is notable that the optimal performance is obtained using a higher sensitivity setting for percussive onsets, while for the non-percussive onset categories lower sensitivity is needed to avoid accumulating false positives although this is at the expense of losing some actual onsets.

In a more general case in our experiment, when PF is applied, the *sens* parameter controls the assessment of the polynomial coefficients to make the final onset decisions. When PF is excluded such that peak picking is realised using a simpler mechanism, *sens* is responsible for generating a uniform threshold for the ODF after an adaptive median removal as described in Equation 2.22. Higher *sens* (i.e., lower threshold for peak picking) is generally needed when using the latter peak picking approach than the former for each detector or onset type.

Albeit the average detection sensitivity *sens* is close to the centre of the evaluated value range ([0, 100]) except in case of NPP which constantly requires a high detection sensitivity, the standard deviations measured across different detectors are relatively high. Also, near extreme values commonly occur as the optimal setting for this parameter. To investigate how varying the setting of this parameter varies the detection performance, we compare the detection F-measures under different *sens* settings with the remaining parameters unfixed taking all available configurations on all the four onset types. For all the detectors investigated, highly significant difference is shown ($p < 0.001$, Repeated measures ANOVA test) for all onset types. We can conclude that this parame-

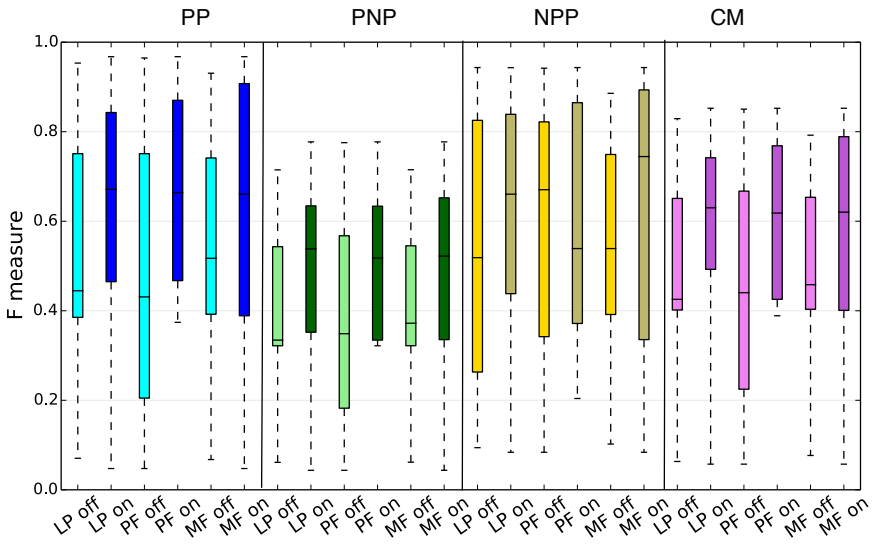


Figure 4.5: Performance of detector $CDSF_L$ under different filter settings. Results are evaluated for each onset type and across all configurations. All other parameters are kept at optimised values.

ter has significant effect on the detection whether it is used in a polynomial fitting based or adaptive thresholding based peak picking (see Section 2.4.3). However, pairwise significance among all granularities tested using a post-hoc Tukey HSD test [AW10] is not always present. This is partly due to the fine granularity of the configuration settings.

Besides investigating the components individually, we also want to understand their effects when used jointly in the peak picking framework. The overall effects of the low-pass filtering (LP), adaptive thresholding (MF) and polynomial fitting (PF) is illustrated in Figure 4.5 for a single detector $CDSF_L$. For the interest of demonstration, we compare the effects of these blocks for one individual detector. This is sensible as most signal processing components exhibit consistent effects on the investigated detectors especially the best performing ones, as shown in Table 4-F. The figure demonstrates the detection F-measures obtained when fixing the investigated parameter fixed to the setting 1 (on) or 0 (off) while leaving all the rest of the parameters to vary. The upper bound of the whiskers, the upper bound of the boxes, the horizontal bar in the boxes, the lower bound of the boxes and the lower bound of the whiskers represent correspondingly the highest, the third quartile, the median, the first quartile and the lowest F-measure rate under

diverse configurations.

Applying all the three techniques leads to better detection results, indicated by the boxes shifting upward in general as well as higher medians. The advantage of each lies also in reducing the dependency on parameter settings even if applying these methods involves using more parameters (for example, only when MF is on, the ψ parameter is used). This is supported by the fact that the majority of other configurations benefit from applying each of these three techniques, presenting smaller interquartile ranges (IQRs). It has to be noted that using MF causes greater variability and larger IQR in general, since enabling the technique implies involving two additional parameters δ and ψ. Another intuitive interpretation of this figure is that the use of one component in this process may influence the behaviour of another.

4.4.5 Analysis of Parameter Interactions

Our results so far suggest that the simultaneous influence of two or more parameters on the results may not simply be additive, in other words, *interactions* exist between the signal processing methods as well as their respective adjustable parameters. Under such situation, the effect of one parameter is dependent on the variation of another.

Figure 4.6 demonstrates an example where the variables cutoff frequency (f_c) and detection sensitivity (*sens*) both interact, with respect to the detection F-measure, with the binary factor adaptive thresholding based on median filtering (MF), whose values correspond to turning the adaptive thresholding technique on or off. When MF is applied, higher f_c and higher *sens* lead to improved results. The trend reverses however when MF is deactivated. This means that there is a *complementary* effect between *MF* and *sens* and between *MF* and f_c. This type of interaction is considered *qualitative*. Here we refer to "qualitative" as the case of interaction where both the magnitude and direction of the effect of each variable interact with the magnitude of the other [UC08].

This indicates that the investigated signal processing techniques mutually interact.

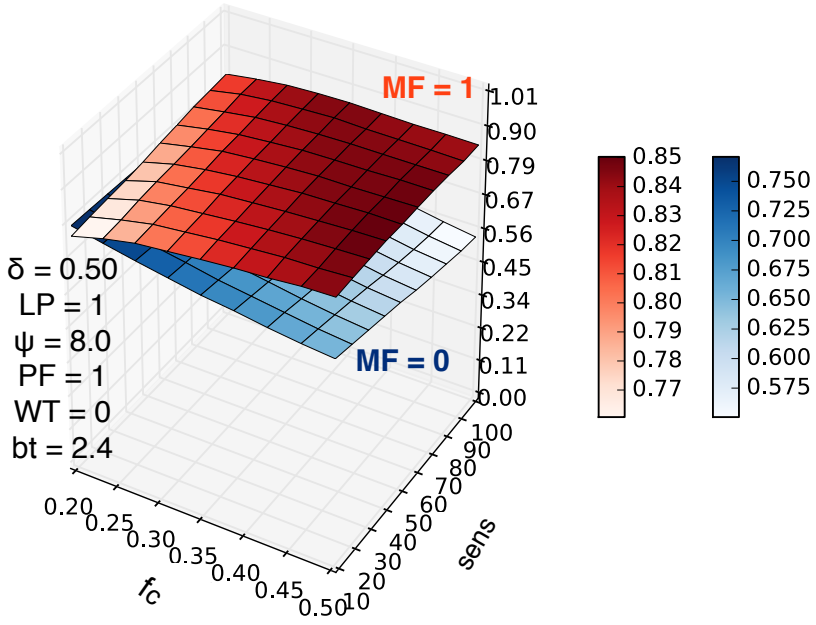
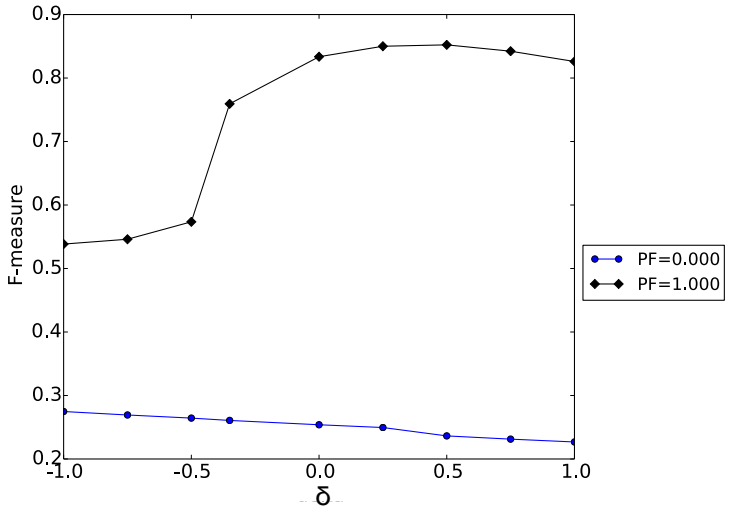


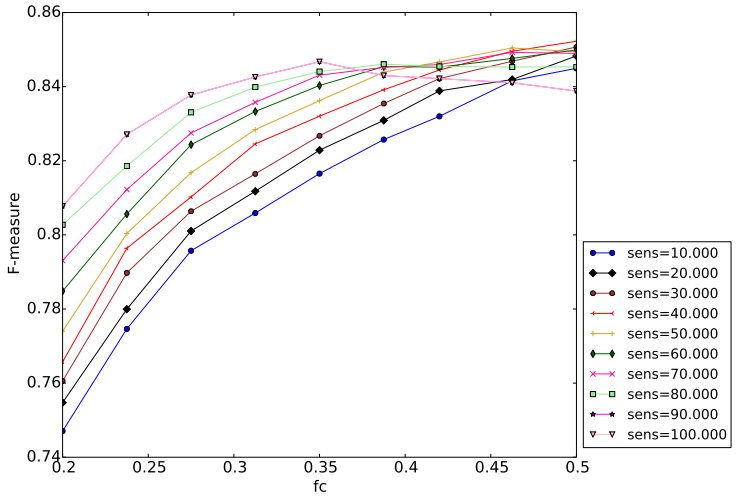
Figure 4.6: Detection F-measure of the linear fusion of Complex domain and SuperFlux $CDSF_L$ under different settings for *detection sensitivity* (sens) and *cutoff frequency* (f_c) for all onset types with *median filtering* (MF) on/off.

While in Figure 4.6 we observe a *complementary* effect, Figure 4.7a and Figure 4.7b demonstrate respectively the *cancellation* interaction between δ and PF and the case between f_c and *sens* where no interaction exists for the same detector [OG91]. When PF is applied, higher δ is preferable for the detection. Since δ introduces a constant offset into the ODF, it influences the comparison of coefficient c in the quadratic function 2.21 with the sensitivity parameter. This trend is reversed when PF is excluded, i.e., the median filter based adaptive thresholding is used, to maintain larger differences between the maximum and minimum value in the ODF after median removal hence higher thresholds (see Equation 2.22). However, the F-measures are lower in general than when PF is applied due to substantially larger number of false positives. As for *sens* and f_c, on the contrary, the change of the former does not alter the direction of the latter. This hence reinforces our motivation of carrying out the parameter optimisation en masse.

Table 4-G shows the pairwise interactions between parameters for the example detector $CDSF_L$ on the combined dataset. We observe the same pattern of interactions for



(a) Cancellation effect between *polynomial fitting* (PF) and *constant offset* (δ) in the median filtering process.



(b) No interaction is found between *detection sensitivity* (sens) and *cutoff frequency* (f_c) in the median filtering process.

Figure 4.7: Interactions between investigated parameters illustrated by the detection F-measure for the combined dataset using the detector $CDSF_L$. All other parameters are kept at optimised values.

all detectors hence illustrate only one as an example. Interactions of their adjustable parameters detailed in Section 4.3 are tested, excluding the combinations of binary factors and the parameters directly controlling the same processing block, indicated by N/A in the table. For instance, since f_c controls the cutoff frequency of the low-pass

filter, when LP is off, f_c becomes deactivated hence should be excluded from the evaluation. Interactions were examined by comparing the average F-measures obtained on the combined dataset. All parameters except for the two under investigation are fixed at their optimal settings. This test can therefore reflect the influence of two independent categorical variables on one continuous dependent variable (detection F-measure in our case) [Sel11]. Statistical significance has been confirmed whenever interactions are present using the 2-way ANOVA test. Similar patterns can be observed for all investigated detectors. Even if two parameters may vary in individual effects on detectors, their relative pairwise interactions remain consistent for all detectors (2-way ANOVA test, $p < 0.05$).

Highly significant interactions are observed between several pairs of parameters providing insight into how the peak picking components work together. We first discuss the interactions of three main components of our peak picking algorithm including low-pass filtering (LP), adaptive thresholding (MF) and polynomial fitting (PF). As also shown in Figure 4.5, the use of each can reduce the dependency of another on system configurations, indicated by higher mean F-measures. When used in combination, LP complements both MF and PF, but using the latter two together does not have a multiplicative effect, as shown in Table 4-G. Increasing the sensitivity parameter (*sens*) improves the results more notably when adaptive thresholding is applied (MF=1). This is mainly because the adaptive thresholding facilitates the use of higher sensitivity without introducing too many false positives. Since δ introduces a constant offset into the ODF after the median removal (see Equation 2.20), its interaction with the polynomial fitting (PF) and median filtering (MF) can reasonably be expected. The effect of using higher f_c for a smoother ODF is proved beneficial when used together with adaptive thresholding. This trend reverses with PF, because peaks in the smoothed ODF become less sharp with higher f_c which affects the selection of coefficient a in the quadratic function (Equation 2.21).

Lessons learned so far includes that interactions do exist between signal processing

	sens	δ	f_c	LP	PF	MF	ψ
sens	N/A	$\times$	$\times$	$\times$	N/A	$+*$	$\times$
δ		N/A	$\times$	$+***$	$-***$	N/A	$\times$
f_c			N/A	N/A	$-***$	$+***$	$\times$
LP				N/A	$+***$	$+***$	$\times$
PF					N/A	$\times$	$\times$
MF						N/A	N/A
ψ							N/A

Table 4-G: F-measures obtained for detector $CDSF_L$ under varying parameter settings evaluated on the combined dataset. When two parameters are assessed, all the other parameters are fixed at the optimal settings. $\times$, $+$ and $-$ represent *no* interaction, *complementary* interaction and *cancellation* interaction. The ratings $*$, $**$, $***$ denote the presence of significance at the level of 0.05, 0.01, 0.001 in the interaction using the 2-way ANOVA test.

parameters in our system. This again justifies our motivation of performing the parameter configuration in an *en masse* manner. In future, a *Multivariate analysis of variance* (MANOVA) can be carried out to uncover the relations among all investigated parameters.

4.4.6 Onset Detection and Music Genres

Considering the fact that the three investigated datasets cover different music types and audio qualities, it is also sensible to assess the performance of our algorithms on a per-dataset basis. Our method $CDSF_L$ obtains F-measures of 0.949, 0.822 and 0.949 on Dataset JPB, SB and JP respectively when using the best configurations for each dataset. However, when using the universally optimised configuration for the mixed datasets, the F-measure results are respectively 0.934, 0.819 and 0.931. Significant difference has been confirmed for Dataset JPB and JP under these configurations by the Wilcoxon signed-rank test ($p < 0.01$). As a conclusion, contextual knowledge about the music type can be used to assist an effective onset detection.

Most onset detection studies report better detection for NPP compared to other onset types. In our evaluation, however, detection results yielded for PP consistently outperform those for other onset types. Here we report results for NPP including the

Chinese instruments. Further analysis into the Chinese NPP onsets (see Section 3.3.2 for details) and the NPP onsets from the two Western datasets shows that detection rates on the former is poorer. The average detection F-measure of all investigated detectors on Chinese NPP onsets is 0.882 and that on the Western ones is 0.890 ($p < 0.01$, Wilcoxon signed-rank test).

As introduced in Section 3.3.2, the cymbal and gong instruments have sustaining notes with temporally varying frequency characteristics. This can also explain the fact that this onset type can be well captured by algorithms that are designed for soft onsets, such as SF and its fusion methods, as illustrated in Table 4-C. However, compared to SF in our system, the algorithm from the original implementation with different post-processing and peak picking methods [BW13a] yielded much worse results as discussed in Section 4.4.3. This again justifies our motivation of carrying out the analysis into the effects of individual signal processing parameters as well as their optimisation in the context of different music types.

4.5 Summary

In this chapter, we investigated the audio onset detection research using fusion and RDF/ontology techniques. The evaluation of the 42 detectors was carried out on both the commonly used Western datasets and the newly presented Chinese percussion dataset with less studied musical characteristics as introduced in the previous chapter.

In the fusion experiment, new onset detection methods were presented based on existing algorithms using three fusion policies, including early fusion applied at a feature level, linear fusion which combines multiple onset detection functions and decision fusion which combines onset decisions from different detectors. We demonstrate significantly improved results for many fusion detectors over the baseline methods. Results show that in general, fusion detectors work better than their constituent baselines. However, the improvement is not necessarily significant if a constituent is already performing well.

Linear fusion turns out to be the most effective out of all fusion policies tested, with $CDSF_L$, the linear fusion of Complex domain and SuperFlux, yielding the best results overall. This onset detector will be used for rhythmic feature extraction presented in the following chapters. The development of new onset detection methods specific to non-Western instruments also constituents our future work.

With respect to specific signal processing methods and parameters, we found that firstly, effective signal processing techniques used for signal enhancement, noise reduction and peak picking greatly assist the onset detection function to achieve an effective detection. Secondly, parameter configuration can be a significant factor to improve the performance of the signal processing methods in an onset detection system. Cross validation results also confirmed that improvements can be expected with diverse instrument types.

Regarding specific components of the peak picking process, the low-pass filter and the adaptive thresholding using a median filter were found to be generally advantageous. The analysis of polynomial fitting yielded more mixed results. Among the investigated parameters, detection sensitivity has the most pronounced effect. We also found there exist significant interactions between different signal processing parameters. Low-pass filtering and adaptive thresholding have a multiplicative effect, i.e., the gain in the onset detection results from using both is higher than the sum of gains provided by each method individually. Another conclusion from this work is that the configuration of an algorithm or a system is genre and instrumentation specific. Different configurations are needed for the investigated algorithms to work effectively on different music genres or onset types.

The overall onset detection, parameter optimisation and the evaluation workflow was implemented in a Vamp Plugin environment based on the Audio Features Ontology and the Vamp Plugin Ontology in an RDF context. With this work, we demonstrated that semantic web tools are highly useful in designing automated MIR experiments and producing reproducible results.

Onset detection is an intermediate stage for many tempo tracking and rhythmic feature extraction processes which will be discussed in the remainder this thesis. In the next chapter, we will investigate novel audio features to retrieve the music structure. The rhythmic features will be based on the onset detection methods presented in this chapter. We will also introduce a music structural segmentation algorithm relying on the signal processing techniques investigated in this chapter.

Chapter 5

Feature Extraction for Music Structural Segmentation

5.1 Introduction

As introduced Section 2.5.2, various types of audio features have been proposed to describe the music structure capturing the timbral, harmonic or rhythmic aspects of the input signal and to derive the music similarity. Among these, MFCCs and chromagram are the most popular ones while the rhythmic features are less frequently used. The selection of audio features for music structural description is mainly based on the heuristics we have for the structure of the Western music. With Jingju newly introduced into the analysis framework, this chapter investigates novel features to summarise the music structure.

Chromagram is originally developed for Western pop music to measure the relative intensity of the 12 pitch classes in an equal-tempered scale (see Section 2.3.2). In this chapter, we will investigate how it may interpret the structural characteristics for Jingju revisiting its *bins per octave* (BPO) settings.

Rhythmic information may identify music structure beyond timbral or harmonic variations. As introduced in Section 3.2.1, Jingju is characterised by vivid metrical patterns. As the second type of features investigated, this chapter presents novel features derived from the tempo spectra of the music. The extraction of relevant features will partly be based on the findings from the onset detection presented in the last chapter.

The application of the Gammatone function for the modelling of human auditory filter response has been studied in previous works (see Section 2.2.2) [GM90; HR88]. Shao and his colleagues applied an auditory-based feature derived from the cepstral analysis of Gammatone filterbank outputs for speech recognition. This feature outperforms the MFCCs and perceptual linear prediction (PLP) features in the evaluation [Sha+09]. Valero and Aliás [VA12] and McKinney and Breebaart [MB03] witnessed improvements of features from auditory perceptual features over standard features when applying to audio and music classification. Although the analysis of music structure can be rather subjective involving human perception and cognition of the music content, little work has investigated feature descriptors based on auditory scales other than Mel in this scenario. In this chapter, we also investigate new psychoacoustics-inspired features derived from the Gammatone filters.

The success of an audio-based music structural segmentation (MSS) system largely depends on the signal processing methods employed. These can be involved in the process of feature extraction, the calculation of the similarity and distance measures, as well as the retrieval of segment boundaries. Harmonic-percussive source separation (HPSS) is a well-studied task concerning separating the input audio signal into harmonic and percussive components [Fit10]. The separation or suppression of the harmonic/percussive source can facilitate the extraction of the percussive/harmonic source for specific MIR tasks. In [ZG13], Zapata and Gómez introduced improvements to beat tracking algorithms by vocal suppression. However, the investigation into its effects for music structural analysis is lacking from the literature. MFCCs have been reported having problems expressing both harmonic and percussive contents when present at the same time in a

music genre classification study [Rum+10]. Furthermore, the heavy use of cymbal and gong instruments in Jingju music has imposed certain amount of masking on the rest of the instrumentation components whose timbral characteristics may hold more delicacy [Tia+14b]. In this chapter, we investigate HPSS as a pre-processing technique for feature extraction in an MSS scenario.

This primary motivation of this chapter is to investigate how different audio features, both existing and new ones, apply to different music genres. This chapter is organised as follows. Section 5.2 introduces a *Harmonic-percussive source separation* algorithm as a feature enhancement technique. The chroma feature is revisited for Jingju in Section 5.3. The design of novel rhythmic features and the extraction of auditory features from the Gammatonegram are presented respectively in Section 5.4 and Section 5.5. We will introduce the segmentation experiment in Section 5.6. Investigated audio features are evaluated and analysed in Section 5.7. Finally, we summarise this chapter in Section 5.8.

5.2 Harmonic-percussive Source Separation

This section investigates harmonic-percussive source separation (HPSS) as a pre-processing step to extract investigated audio features from the spectrogram. Given the complex spectrogram X of the input audio signal, HPSS separates it into its harmonic component X_h and percussive component X_p. We note Y the magnitude spectrogram where $Y = |X|$. The separation can be realised by applying a median filter to Y once in the horizontal direction and once in the vertical direction to derive respectively the harmonic and the percussive spectrogram. This method is presented by Driedger et al. based on FitzGerald [Fit10] with improved separation results [DMD14].

However, this process may be interfered by the existence of vibrato introduced by the singing voice and the bowed string instruments in Jingju, which may yield frequency oscillation over a small time period [YTC15]. A possible solution is to consider widened frequency bins in the neighbourhood when generating the vertical mask. For the sup-

pression of the percussive component, we also propose to widen the masking trajectory across neighbouring time instants in case the transient energies are swaying.

A maximum filter has the capacity of broadening the spectral trajectory to enhance specified components. However, a side effect the maximum filter may introduce is a larger portion of residues mixed in the separated source due to widened trajectories. Instead of targeting a rigorous source separation, this work proposes to use HPSS to generate the spectral basis to extract relevant features carrying the underlying information to describe music structure. We hence argue that preserving certain amount of residues may not be harmful for the subsequent structural analysis. To this end, we propose to use a maximum filter to Y and apply it before the median filter taking its opposite direction. In this way, the harmonic and percussive slices can be individually strengthened before the separation, leading to a highlighting effect of the corresponding sources to combat the possible interference of vibratos or energy sways.

As a modification to Driedger et al. [DMD14], we apply a one-dimensional maximum filter processing to derive the masks before the median filter is applied. The maximum filters are applied vertically and horizontally when the targeting component to separate is respectively the harmonic and percussive source. The median filters are applied subsequently taking the opposite directions of the maximum filters to generate the masks. The whole process is described as follows. First, a maximum filter is applied to strengthen the corresponding slices:

$$Y_h{}^m(n,k) = \mathsf{max}(Y(n, k - m_h : k + m_h)), \tag{5.1a}$$

$$Y_p{}^m(n,k) = \mathsf{max}(Y(n - m_p : n + m_p, k)). \tag{5.1b}$$

A median filter is then applied to both Y and the maximum filtered magnitude

spectrogram ${Y_h}^m(n,k)$ and ${Y_p}^m(n,k)$:

$$\tilde{Y_h}(n,k) = \mathsf{median}(Y(n-l_h : n+l_h, k)), \tag{5.2a}$$

$$\tilde{Y_p}(n,k) = \mathsf{median}(Y(n, k-l_p : k+l_p)). \tag{5.2b}$$

$$\tilde{{Y_h}}^m(n,k) = \mathsf{median}({Y_h}^m(n-l_h : n+l_h, k)), \tag{5.3a}$$

$$\tilde{{Y_p}}^m(n,k) = \mathsf{median}({Y_p}^m(n, k-l_p : k+l_p)). \tag{5.3b}$$

Next, $\tilde{{Y_h}}^m(n,k)$ and $\tilde{{Y_p}}^m(n,k)$ are used to mask X to derive correspondingly the individual source X_h and X_p:

$$X_h(n,k) = X(n,k) \cdot \Big(\tilde{{Y_h}}^m(n,k)/(\tilde{Y_p}(n,k)+\epsilon) > \beta\Big), \tag{5.4a}$$

$$X_p(n,k) = X(n,k) \cdot \Big(\tilde{{Y_p}}^m(n,k)/(\tilde{Y_h}(n,k)+\epsilon) > \beta\Big), \tag{5.4b}$$

where 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 maximum and median filters. ϵ is a small constant to avoid zero division and β is the separation factor to control the ratio of specific component to separate.

Figure 5.1 demonstrates the effects of HPSS showing the harmonic and percussive spectrograms after the source separation. The separation ratio β is set to 0.8 and the maximum filter size is set to 3 samples, i.e., $l_h = 1$, $l_p = 1$. The sample rate f_s, STFT window size and step size are respectively 44.1 KHz, 46 ms and 23 ms. We can notice a more notable horizontal and vertical continuity in the harmonic (Figure 5.1d) and the percussive spectrogram (Figure 5.1e) after the maximum filtering. Additionally, the slow-decaying high frequency components introduced mainly by the percussion instruments are also suppressed in X_h after the maximum filtering, as can be seen when comparing

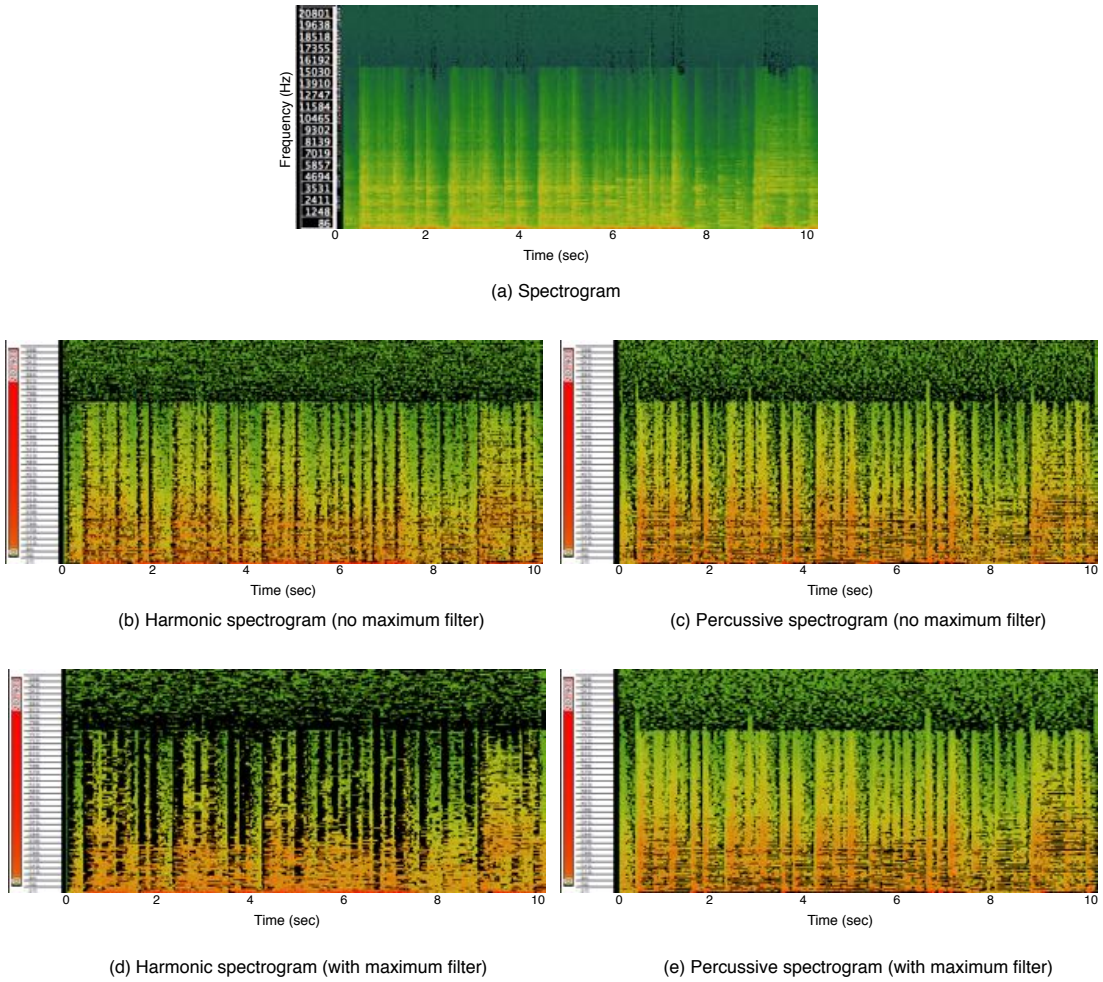


Figure 5.1: Spectrograms derived after Harmonic-percussive source separation (HPSS) of a 10-second excerpt of song "Hideout" from dataset S-IA.

Figure 5.1b to Figure 5.1d. We will discuss the effect of this HPSS technique including the involved parameters in Section 5.7.3.

5.3 Bins per Octave in Chroma for Jingju

As introduced in Section 3.2.1, Jingju uses an anhemitonic pentatonic scale for its main melody structure. The five main notes are built on an equal temperament scale. Two additional notes, which are the 4th and the 7th in its numbered notation system, are also integral in the overall pitched content of Jingju functioning to deliver the musi-

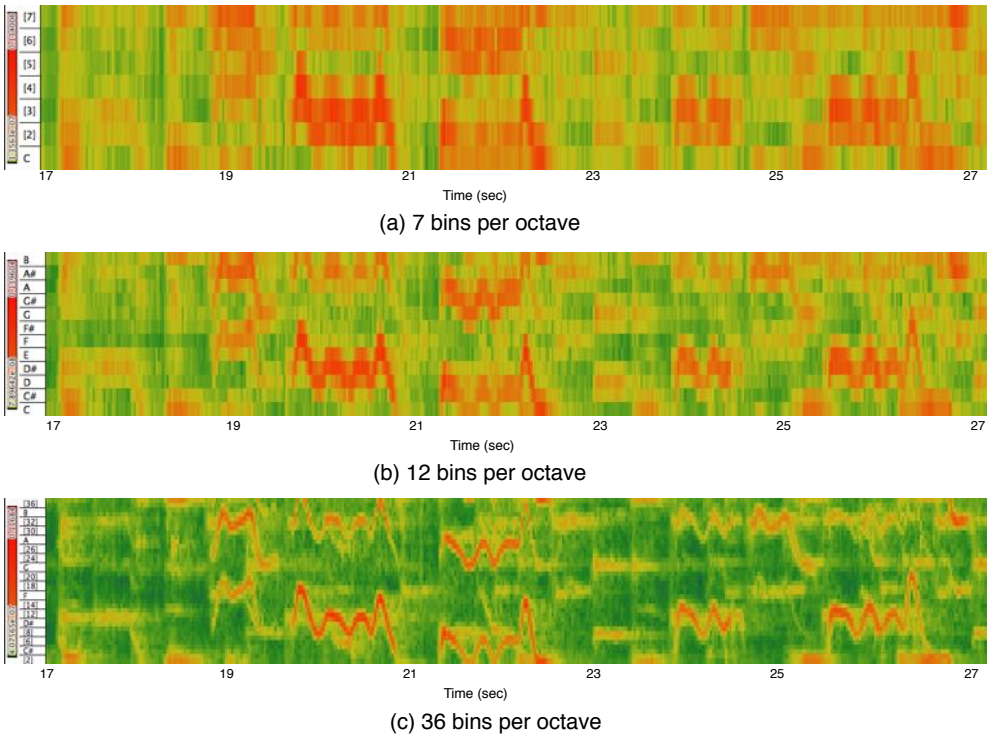


Figure 5.2: Chromagrams with 7, 12 and 36 bins per octave for music "Jin yu nu".

cal expressiveness [Che13; Wic91]. They however use a different tuning system and contribute little to the energy of the entire pitched classes.

Although chroma features are originally designed for chord recognition for Western music, they measure the relative intensity of each pitch class of an equal-tempered scale in a tuning independent way [Fuj99]. Although the 4th and 7th notes do not follow an equal temperament, they are of much lower intensities compared to the other five (see Section 3.2.1). We hereby propose to use chroma features for Jingju music which uses equal temperament except for its 4th and 7th degrees.

Besides the standard $BPO = 12$ setting, some works also use 24 or 36-bin semitone quantisation for fine tuning for Western pop music [HSG06; Lee06] as introduced in Section 2.3.2. Figure 5.2 shows the 7-, 12- and 36-BPO chromagram for a 10-second excerpt of Jingju song "Jin yu nu" from dataset CJ comprising a sung phrase. It can be noticed that when BPO is set to 12, the energies of a major pitch class and its sharp pitch

(for example C and C#) have similar distributions. The one-third semitone resolution is able to capture the pulsating pitch drifts within a sung tone resulted from vibratos, which can span over a few frequency bins, as can be seen in Figure 5.2c. This frequency drifting effect becomes less notable in the 7-BPO chroma, shown in Figure 5.2a, where energies are aggregated into wider bins corresponding only to the major notes indicating the anharmonic nature of Jingju [Liu+09; Che13; Wic91]. Using wider bins may also reduce the interference introduced by fast-decaying high-frequency transients. In this chapter we investigate the 7-BPO chromagram for Jingju and assess its effects in music structural analysis.

5.4 Tempogram Features

5.4.1 Tempogram Revisited

A mid-level rhythmic descriptor, tempogram, is introduced in Section 2.3.3. As shown in Figure 2.4, the tempograms for the Beatles song and the Jingju song reflect different tempo tendencies. The Beatles song shown in Figure 2.4a is a representative example of the Western popular music style, which is commonly characterised by a steady tempo across the whole piece. This property is exactly what defines it to be popular music, as explained by Regev [Reg13]. Frith also notes that the steady tempo and structured beat patterns are necessary to engage the listeners [Fri96]. The tempo of the Jingju music song, on the contrary, is subject to more variations. The tempo at any given time is controlled by the percussion in the music. The structural progressions are often associated with gradual acceleration or deceleration of tempi, as can be seen from Figure 2.4b. Despite less employed than timbral or harmonic features for the music structural analysis, new audio features that are tempo and rhythm descriptive may be discriminative for music structure.

The tempogram is built by using the local periodicity of the onset detection function

(ODF), which identifies the amplitude, phase or other changes in the spectral information of an input audio signal corresponding to the presence of musical notes. As musical notes comprise the most basic hierarchy of a music piece, the tempogram may contain sufficient information of the music structure. Instead of targeting a rigorous tempo or beat tracking, we are interested in the semantic information incorporated in the tempogram with potentials to interpret the music structure. In a recent work, Grosche and Müller extracted the *Predominant local pulse* (PLP) feature from the cyclic tempogram for beat tracking [GM11a] (see Section 2.3.3). However, a single function such as the PLP curve may contain limited information for the structural description. Therefore, we propose to extract new features from the tempogram to reflect the underlying structural patterns of a music signal.

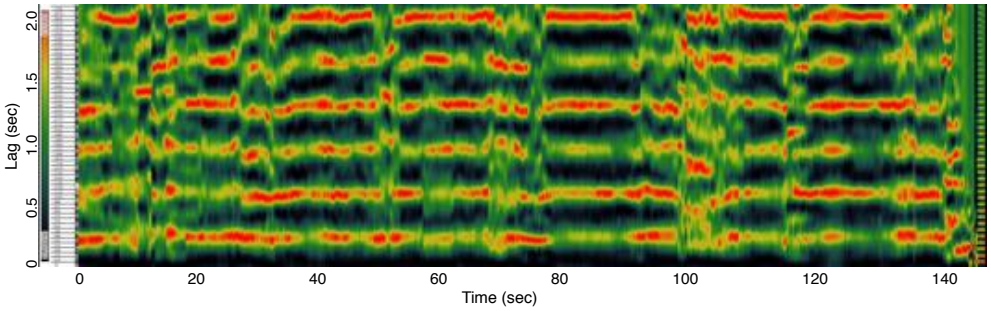
As the calculation of a tempogram relies on the calculation of the ODF, we assume that the accuracy of a tempogram can be improved by using enhanced onset detection techniques. To this end, we adopt a method using the linearly weighted fusion of two algorithms Complex domain (CD) and SuperFlux (SF) [BW13b], which is presented in the last chapter denoted $CDSF_L$. This method shows improvements over all other onset detectors tested in a large-scale evaluation (see Section 4.4). We note the ODF of CD and SF respectively $CD(n)$ and $SF(n)$, the calculation of the ODF of $CDSF_L$ is recapped in Equation 5.5,

$$CDSF_L(n) = l \cdot CD(n) + (1 - l) \cdot SF(n), \tag{5.5}$$

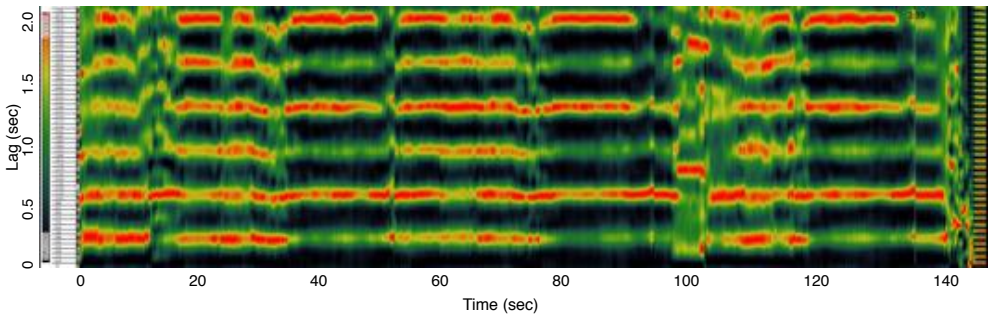
where n denotes the time index and l is the linear combination coefficient set to 0.3 following the parameter optimisation presented in Section 4.4.

As introduced in Section 2.3.3, the two different methods to derive a tempogram are based on the Fourier transform (FT) and the autocorrelation function (ACF) and characterise respectively the *harmonics* and the *subharmonics* of the music content. In this thesis, we are interested in the variation in long-term temporal structure, therefore

we use the ACF-based tempogram, as it emphasises the subharmonics in the tempo spectra corresponding to lower metrical levels. In the remainder of this section, we introduce the extraction of rhythmic features from the ACF-based tempogram.



(a) Tempogram calculated using SD onset detection function

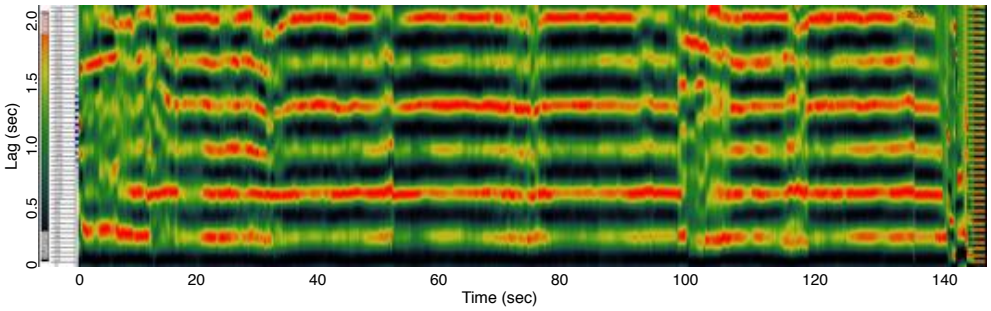


(b) Tempogram calculated using CDSF-L onset detection function

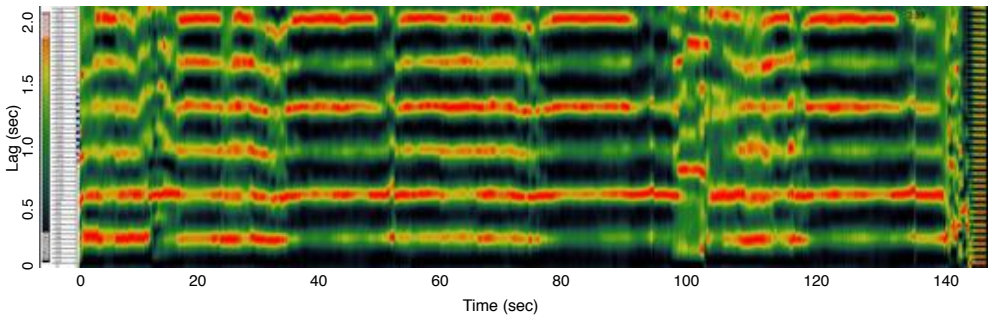
Figure 5.3: Tempogram calculated using the Spectral difference (SD) onset detection function and the linear fusion of Complex domain and SuperFlux ($CDSF_L$) onset detection function.

Figure 5.3 shows the ACF-based tempogram calculated for a piece of Beatles song "Help" from dataset BeatlesTUT (see 3.4.1). Figure 5.3a and Figure 5.3b show respectively the tempogram using ODF derived from the Spectral difference (SD) as the reference approach [GMK10; DP04; Ell07; GM09] and from the proposed $CDSF_L$ method. The ODFs are calculated from the original spectrogram X without HPSS. The sample rate, STFT window size and step size are respectively 44.1 KHz, 46 ms and 23 ms. The tempogram time window used is 6 seconds. We can observe that in the tempogram calculated using $CDSF_L$, lots of local false positives are suppressed. This could have the effect of emphasising the most salient rhythmic component.

Tempograms calculated using the percussive spectrogram X_p calculated after the



(a) Tempogram calculated using SD onset detection function



(b) Tempogram calculated using CDSF-L onset detection function

Figure 5.4: Tempogram calculated using the Spectral difference (SD) onset detection function and the linear fusion of Complex domain and SuperFlux ($CDSF_L$) onset detection function after harmonic-percussive source separation (HPSS).

HPSS (introduced in Section 5.7.3, $\beta = 0.5$, $l_p = 10$, $m_p = 10$) are shown in Figure 5.4. They match closely to the results calculated without HPSS. However, after HPSS, the strength of local pulse appears more steady, especially when accompanied by the accelerando or deccelerando in tempo. The improved tempo salience also indicates the potential advantage of using enhanced onset detection methods in other rhythmic analysis such as beat and tempo tracking.

5.4.2 Dimensionality Reduction based Features

Feature extraction starts from an initial set of measured data and attempts from which to derive more abstract and conceptual representations. The feature extraction process is closely related to dimensionality reduction [Sam06]. An essential nature of the extracted features, compared to the raw data, is a downsized data redundancy. As the first feature

extraction strategy, we apply dimensionality reduction operations to the tempogram.

Principal component analysis (PCA) is a multivariate data analysis technique that aims to minimise the correlation between variables. It provides a linear orthogonal transformation into a new coordinate system such that after the projection the majority of the variance lies in the first few dimensions and the variables become uncorrelated. The values in the remaining dimensions, therefore, tend to hold small variance and can be discarded with minimal loss of information. This technique is commonly used for the analysis of high-dimensional features. To this end, we subject the tempogram to a PCA and keep the first a few dimensions as the first dimensionality reduction based feature investigated in this work, denoted *Tempogram principal coefficients* (TPCs). A 20-coefficient TPCs is used in this thesis [Tia+15].

The *discrete cosine transform* (DCT) can also be used as a dimensionality reduction technique given its property to concentrate high energy components in the lower coefficients. While PCA provides basis functions for the projection ordered by the data, DCT uses a data-independent orthogonal sinusoidal basis. DCT is widely used in the extraction of audio features. For example, it is adopted as the last step in the calculation of the MFCCs which have been proved highly successful in describing the timbral aspect of sound [AP02; DM80; Log00] (see Section 2.3.1).

Inspired by the MFCCs, we introduce a feature called *Tempogram cepstral coefficients* (TCCs). For each tempogram frame we take the logarithm of the energy to emphasise the underlying periodicity in the autocorrelation of the ODF, then apply a DCT to obtain a compressed representation of the rhythmic content of the audio signal. There are different variations of DCTs. In this thesis, we use the most popular one, the *type-II DCT*, which is commonly simply referred to as "the DCT" [NP78]. The algorithm is illustrated in Equation 5.6.

$$TCCs(n) = \sum_{\Lambda=0}^{N-1} log\left(A(n,\Lambda)\right) cos\left(\frac{\pi}{N}(\Lambda + \frac{1}{2})i\right), \tag{5.6}$$

where $i = 0, \ldots, N-1$.

Although the DCT is an orthogonal transform, we apply it to the log-compressed tempogram hence to enable a reduced representation focusing on the overall pulse regularity and helps to suppress noise. A 20-dimensional TCCs is used in this work (C0 included). The number of coefficients is defined experimentally, where it is observed that any value moderately higher than *13* provides similar segmentation results.

5.4.3 Band-wise Processing

The dimensionality reduction techniques such as the DCT and PCA transform reshuffle the tempo spectra such that the majority of the information is expressed in the low dimensions. However, they hardly provide any separation of the underlying classes directly. To this end, we extract additional features with more explicit domain knowledge to characterise the rhythmic patterns encoded in a tempogram.

For most music genres various instruments in a piece play diverse roles in producing the overall rhythmic structure, thus each instrument can be prominent at different metrical level [Par94]. Dawe and his colleagues indicate that if two phenomenal events differ in their perceptual salience, the more prominent event would define the phrase boundary hence the perceived rhythm pattern [DPR93]. Therefore, we propose to describe the rhythmic structure depending on the identification of the perceptually salient events upon which more complex rhythmic patterns may be built.

Perceptually informed audio features are proved very effective in summarising music similarities [Pee04]. The perceptual loudness pattern introduced by Moore and his colleagues [MGB97] has inspired the design of several related audio features. Among these is the *specific loudness*, which approximates Moore's loudness expression by compressing energy components active in specific areas in a relative scale while neglecting the rest [Rod01]. The calculation of this feature is given in Equation 5.7:

$$N'(z) = E(z)^{0.23}, \tag{5.7}$$

where $E(z)$ represents the energy in the zth band. Zwicker also notes the total loudness as the sum of bandwise specific loudness as:

$$N = \sum_{z=1}^{Z} N'(z), \tag{5.8}$$

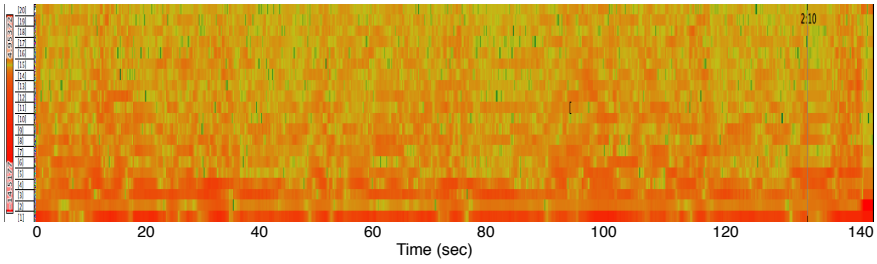
where Z is the total number of the bands [Zwi90].

To this end, we introduce new features incorporating tempo perception cues inspired by Moore's acoustical perception experiments [MGB97]. Previous works indicate that tempo perception occurs in a logarithmically-like space just like pitch and loudness [Sap05]. Studies have also shown that rhythm perception is categorical [Cla87]. To characterise specific tempo strength, we first group the tempogram bins (in lag Λ) into L quasi-logarithmic spaced bands corresponding with the following $L+1$ boundaries in BPM τ_t: $\{440, 240, 170, 130, 110, 90, 80, 65, 55, 40\}$ where $\tau_t = 60/(s_r \cdot \Lambda)$ and $L = 9$. The grouping decision is mainly made from the observations of stable presence of subharmonics in the tempogram [Tia+15].

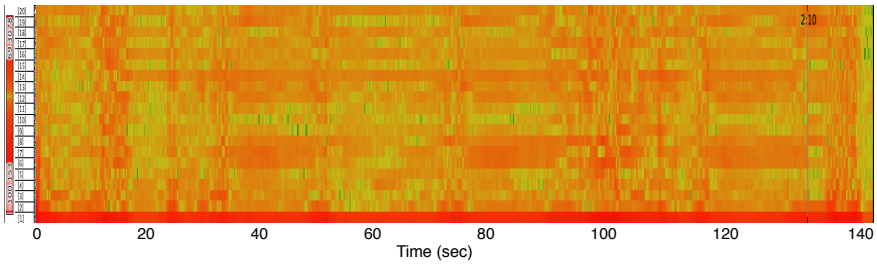
The first feature *Tempo intensity* (TI) is designed to capture the strength of rhythmic components at different metrical levels. In each frame when the tempo t_τ falls into the zth band ($z \in [0, L-1]$), we sum the tempogram magnitude to $T(z)$. Tempo components which drop out of this range will be discarded as they are considered to have less contribution to the overall music structure. Inspired by the specific loudness feature [Rod01], we compress the bandwise intensity values using:

$$TI(z) = T(z)^{\varrho}, \tag{5.9}$$

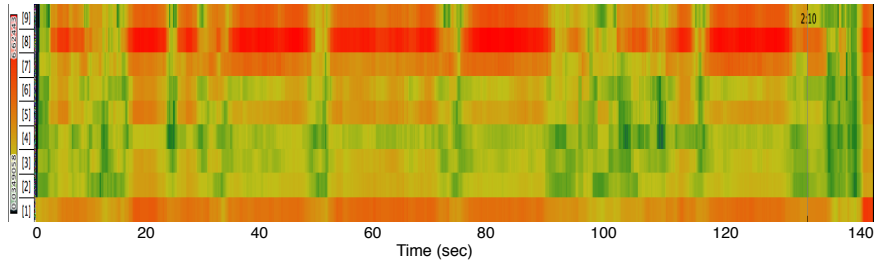
where the exponent ϱ ($\varrho \leq 0.5$) applies a fractional root function to $T(z)$. Manual parameter tuning finds that setting it to 0.4 would yield the optimal segmentation per-



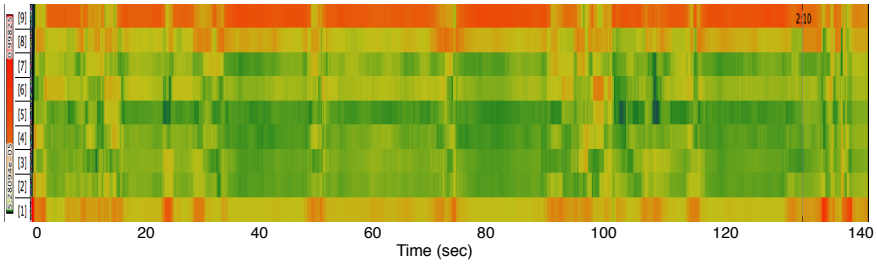
(a) Tempogram principal coefficients (TPCs)



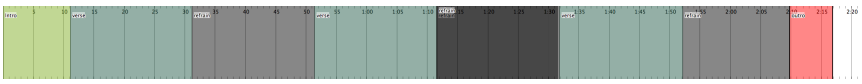
(b) Tempogram cesptral coefficients (TCCs)



(c) Tempo intensity (TI)



(d) Tempo intensity ratio (TIR)



(e) Annnotations

Figure 5.5: Features extracted from the tempogram for audio example "Help" by The Beatles from dataset BeatlesTUT. Panes from top to bottom show respectively the tempogram, Tempogram principal components (TPCs), Tempogram cepstral coefficients (TCCs), Tempo intensity (TI), Tempo intensity ratio (TIR) and the ground truth annotations.

formance [Tia+15].

The second feature, *Tempo intensity ratio* (TIR), describes the perceived relative salience of individual rhythmic components by calculating the intensity ratio of each band as defined above. The formula is given in Equation 5.10:

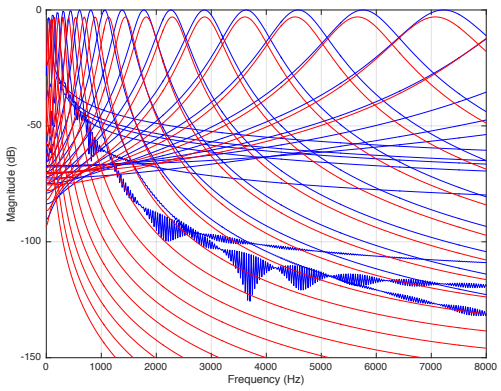
$$TIR(l) = \frac{T(z)}{\sum_{z=0}^{L-1} T(z)}. \tag{5.10}$$

In Figure 5.5, the above introduced features for a popular song "Help" from our dataset BeatlesTUT (see Section 3.4.1) are demonstrated together with the ground truth annotation. Compared to TPCs, TCCs exhibit more visual variations between sections. It can be observed that TI and TIR present patterns correlating to the structural segments without further processing. For these two features, components of higher orders (corresponding to slower tempi) show higher relevance to the sectional characteristics of the music piece. This hence suggests that the slowly-evolving rhythmic components can be more prominent in the perception of segment boundaries. The presented tempogram features will be evaluated in Section 5.6.

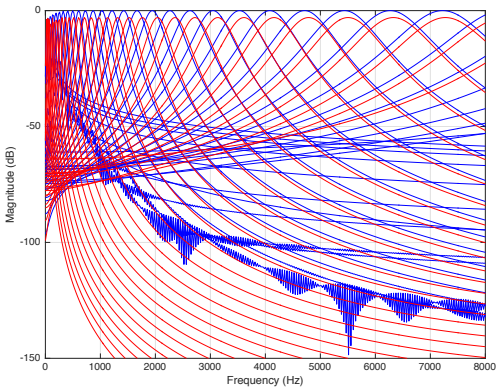
5.5 Gammatone Features

5.5.1 Gammatone Approximation based on Fast Fourier Transform

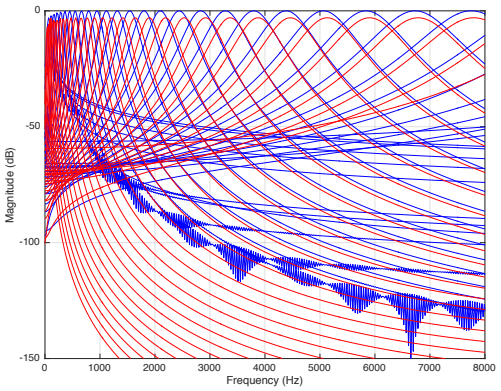
An efficient implementation of the Gammatone filters is provided by Slaney [Sla93] as introduced in Section 2.2.2. However, to process a signal with a bank of M Gammatone filters can still be computationally expensive. Ellis introduced an alternative method using a fast Fourier transform (FFT)-based approximation [Ell09]. In this approach, a conventional spectrogram with fixed bandwidth is first calculated whose frequency bins are then aggregated into coarser resolutions via a weighting function approximating the Gammatone responses.



(a) 20 filters.



(b) 32 filters.



(c) 64 filters.

Figure 5.6: Gammatone filters frequency responses using the accurate method (blue) and the fast method (red).

Figure 5.6 illustrates the frequency responses for each filter using the FFT-based approximation by Ellis [Ell09] (shown by the red curves) and the actual Gammatone filters from Malcolm's Auditory Toolbox [Sla93] (shown by the blue curves) with 20, 32 and 64 filterbanks. The input is an impulse signal with a length of 1000. A 3-dB offset

is introduced to separate the two for visualisation purposes. When the magnitude drops below -100 dB, the approximation becomes less accurate with lots of wiggles mainly resulted from truncating the impulse response at 1000 samples [Ell09]. A better match is obtained with larger M such that the weighting is carried out for narrower bins. As introduced in Section 2.2.2, 32- to 64-channel Gammatone representations are considered appropriate in music and audio processing studies [VA12; HWW14; Qi+13; Sch+07]. In this thesis, we use the 64-channel Gammatone filters for feature extraction. The lower and upper frequency bound are set to 50 Hz and 22.05 KHz (half the sample rate).

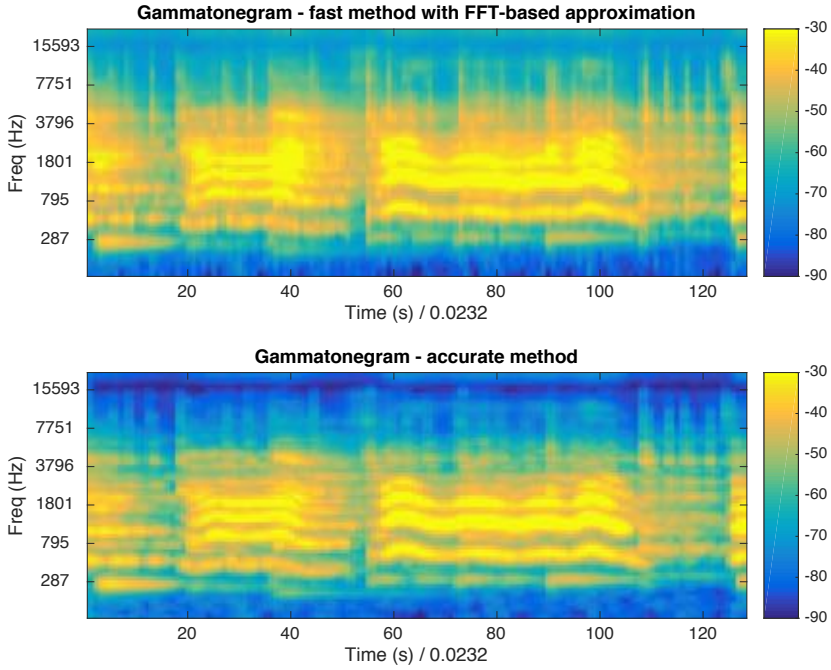


Figure 5.7: Gammatonegrams calculated using the accurate method and the fast method for Jingju song "Jin yu nu" from dataset CJ.

Figure 5.7 demonstrates the 64-channel Gammatonegrams derived using the two methods for a 10-second excerpt of Jingju music "Jin yu nu" from dataset CJ. Although not identical, the fast and the accurate method match well. The main differences noticed include that the fast method accumulates more energy in the highest frequency channels while reduces the time smearing in the lower frequency channels introduced by dispersion of energies [Lyo96], and that it presents smoother frequency responses between channels compared to the accurate one. This is mainly due to the fact that the phase of each

frequency channel is ignored in the summation. The fast method assumes additive coherence in amplitude from different channels, whereas the actual subband energies depend also on how the different frequencies combine [Ell09].

While the accurate Gammatone filter method keeps the original temporal resolution, using the fast method would result in a loss of temporal resolution due to the Fourier transform applied. However, the former will then be subject to temporal summation to derive the Gammatonegram as the time-frequency representation for subsequent feature extraction. Patterson also pointed out in a psychoacoustic experiment on timbre that the phase lag in the low-frequency channels in an auditory system does not affect the perception of a sound [Pat87]. We therefore hypothesise the fast method which relies on the FT and neglects the phase of each frequency channel may lead to trivial influence as compared to the accurate method in the scenario of music structural description. In Section 5.7.4, we will investigate the two methods of Gammatonegram calculation by investigating the features derived, as introduced shortly.

5.5.2 Gammatonegram Feature Extraction

Gammatone Cepstral Coefficients

Note that the dimension of a 64-channel Gammatonegram vector can still be much larger that of a feature commonly used for music structural analysis. Meanwhile, due to the overlap among neighbouring filter channels, the Gammatone filtering outputs are largely correlated with each other. To this end, we introduce a feature called *Gammatone cepstral coefficients* (GCCs) with dimensionality reduction techniques following Shao et al. [Sha+09]. Specifically, a DCT is applied to the Gammatonegram $G(n)$ in order to de-correlate its components:

$$GCCs(n) = \sum_{m=0}^{M-1} G(n) cos\left(\frac{\pi}{M}(m+\frac{1}{2})i\right) \tag{5.11}$$

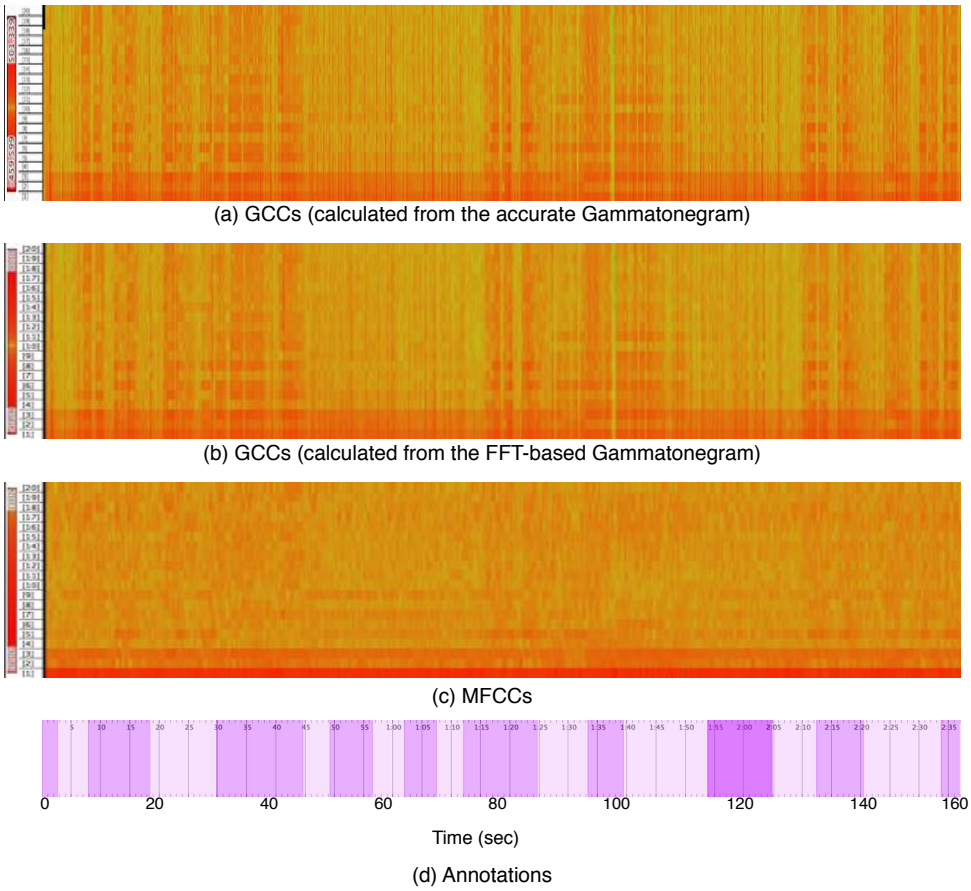


Figure 5.8: Gammatone cepstral coefficients (GCCs) calculated from the accurate Gammatonegram and the FFT-based approximation on a Jingju song "Jin yu nu" from dataset CJ. The segmentation annotation is shown in the bottom pane.

where $i = 0, ..., M - 1$ and M is the number of filters.

However, a log operation is *excluded* as in common cepstral analysis since initial investigation shows degraded segmentation results due to an over-emphasis of the lower frequency components [TS16a]. A 20-dimensional GCCs is used in this thesis.

Figure 5.8 shows the GCCs calculated from the two Gammatonegrams, the accurate one by Slaney (Figure 5.8a) and the fast one by Ellis (Figure 5.8b), MFCCs and the structural annotation for Jingju song "Jin yu nu" from dataset CJ (see Section 3.4.3). We include the MFCCs in the figure as another cepstral feature as a comparison. All features are visualised on a logarithmic scale. The two GCCs match very closely although

the one calculated from the accurate Gammatone method (Figure 5.8a) presents slightly sharper contrast between feature vectors. Both exhibit visible patterns correlating with the annotated segments, which are missing from MFCCs (Figure 5.8c).

Gammatone Contrast

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 at different frequency ranges. 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 [Jia+02].

The calculation of GC is as follows. As the first step, the Gammatone filterbank indices $[0, ..., M-1]$ are grouped into C subbands with linearly equal subdivisions and N_F filterbanks divided into each subband. Note that 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 Jia et al. [Jia+02] in this study yielding a subband frequency division of $[50, 363.198, 1028.195, 2440.148, 5438.074, 11803.409, 22050]$ (Hz). This choice is based on music observations. With such grouping, the vocals will mainly be active in the *second* band, the first and the second band may note the presence of the main pitched instruments and the forth and fifth band can discriminate the presence of drums and cymbals [TS16a].

We note $\mathbf{V_G}$ the Gammatonegram vector of the zth subband $[G_{z,0}(n), G_{z,1}(n), \ldots, G_{z,N_F-1}(n)]^T$ where $z \in [0, C-1]$. $\mathbf{V'_G} = \left[G'_{z,0}, G'_{z,1}, \ldots, G'_{z,N_F-1}\right]^T$ is $\mathbf{V_G}$ sorted in an ascending order such that $G'_{z,0}(n) < G'_{z,1}(n) < \ldots < G'_{z,N_F-1}(n)$. We calculate the difference between the strength of the energy peak and valley for each subband to derive the C-dimensional GC feature:

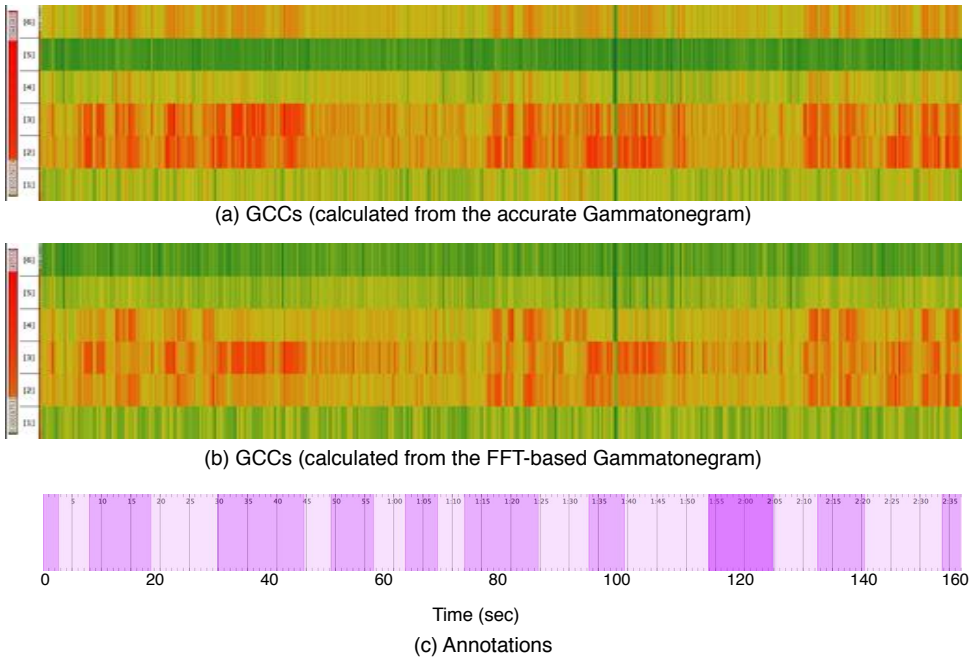


Figure 5.9: Gammatone contrast (GC) calculated from the accurate Gammatonegram and the FFT-based approximation on a Jingju song "Jin yu nu" from dataset CJ. The segmentation annotation is shown in the bottom pane.

$$GC_z(n) = log(G'_{z,N_F-1}(n) - G'_{z,0}(n)), \tag{5.12}$$

Figure 5.9 demonstrates the GC calculated from the accurate Gammatonegram and the FFT-based Gammatonegram on a Jingju song "Jin yu nu" from dataset CJ (see Section 3.4.3). The bottom pane in the figure shows the ground truth annotations. It can be noticed that the majority of the variance is presented in the second to the fourth band, capturing mainly the dynamics in the singing voice and other pitched instruments. GC extracted from both Gammatonegrams match closely to each other in terms of their temporal shapes in the central bands, similarly to the observations for GCCs (Figure 5.8). However, fewer dynamics are presented in the highest band calculated from the fast Gammatone method. This is because as shown in Figure 5.7, the fast Gammatone method accumulates more evenly distributed energies, especially in the high-frequency channels, hence less contrast. Also, a sharper contrast is shown

in the accurate method both frame- and band-wise. We will investigate the presented features GCCs and GC in the next section.

5.6 Segmentation Experiment

5.6.1 Feature Extraction

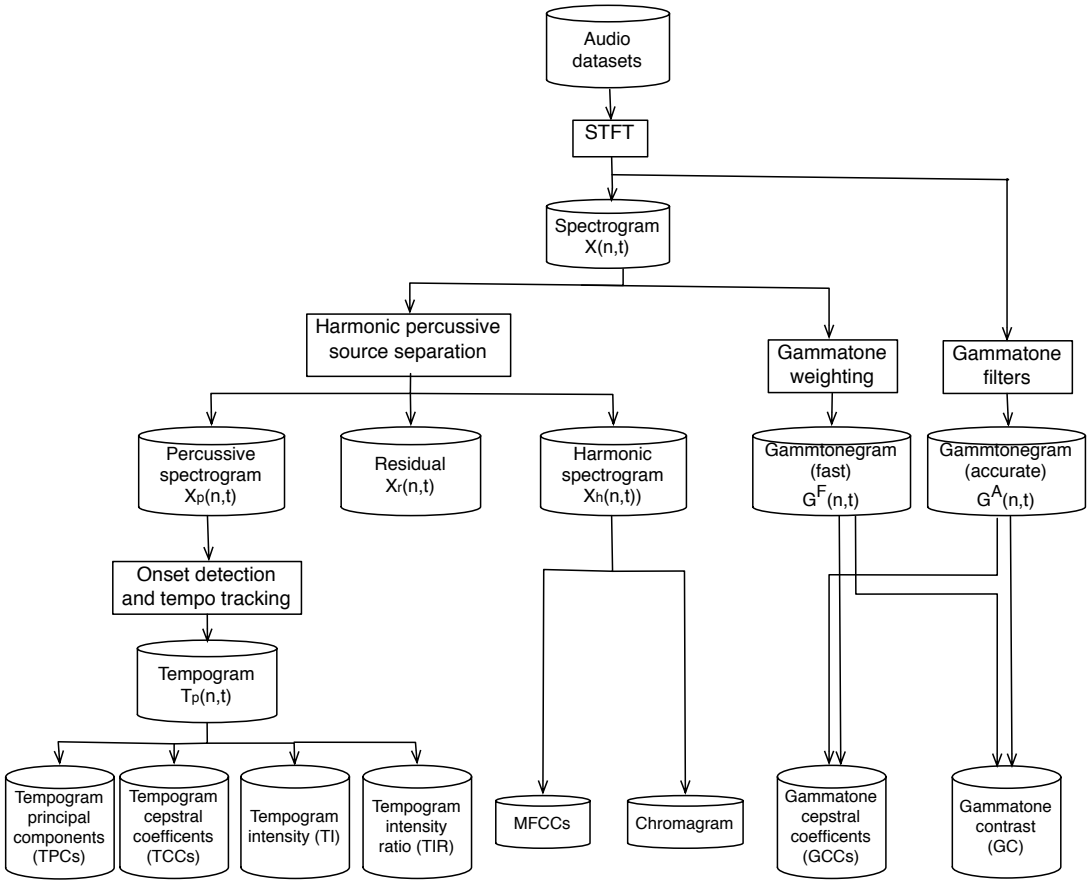


Figure 5.10: Feature extraction workflow.

This section is devoted to evaluate the presented features in an MSS scenario. The feature extraction framework is shown in Figure 5.10. Two existing features, MFCCs and chromagram, are also included to provide a comparison to the new features. Based on the chromagram feature, we will analyse the effect of bins per octave for Jingju. Features investigated are implemented as Vamp plugins (see Section 2.6).

MFCCs and chromagram are extracted on the Queen Mary Vamp Plugins (QMVP) [Plu06]. We modified the implementation to enable HPSS in this process, such that the two features are extracted from the harmonic spectrogram X_h. MFCCs are extracted using 40 Mel filters. We use 13-dimensional MFCCs (including C0) for subsequent processing. Although different variants of the chroma feature exist (see Section 2.3.2), the majority of these variants have hardcoded observations for Western music, for example, use predefined BPO settings [MND09; ME10b]. In this thesis we use the standard chromagram feature based on the constant-Q transform as illustrated in Figure 2.3.

The main parameters involved in the tempogram feature extraction process are the size of the time window W_a used in the autocorrelation. W_a is conventionally set to 6 s [Pee05; GM11b]. In this work, we assess different settings of 3, 5, 6 and 8 s by investigating the segmentation performance of derived features. We will also look for the best usage of the presented tempogram features in the scenario of structural segmentation.

The effect of HPSS will be evaluated using MFCCs, chromagram and the investigated tempogram features. MFCCs and chromagram will be extracted with the harmonic source and the tempogram features will be extracted from the percussive source. Our recent study has shown that when located in a moderate range, the sizes of the median and maximum filters (see Section 5.2) do not introduce significant difference to the segmentation results [TS16b]. In this work, the sizes of the median and maximum filters are experimentally set to 0.23 s, 350 Hz, 0.07 s and 70 Hz, horizontally and vertically. The separation factor β is the most important parameter among the set [DMD14]. We test 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. We also aim to test the effect of the maximum filters as our modification based on Driedger et al. [DMD14]. We extract three variants for a feature investigated: the original feature extracted from the magnitude spectrogram without HPSS; the variant extracted from the separated source after HPSS where the maximum filter is not applied; the variant extracted from the separated source after HPSS with the maximum filter applied.

For the evaluation of investigated Gammatone features, we first analyse the same features extracted from the Gammatonegram derived using the accurate method $G^A(n,t)$ and the that using the FFT-based approximation $G^F(n,t)$ (see Section 5.5.1). For the calculation of the latter, we use the original spectrogram without HPSS to avoid introducing possible interference in the approximation.

The sample rate fs, window size W_N and hop size W_H for feature extraction are respectively 44.1 KHz, 46.4 ms and 23.2 ms. All features are further resampled to obtain a uniform frame rate of 0.2 s. The use of a relatively large window would assist forming the structural description on a musically meaningful scale (see Section 2.5.2). We will situate the evaluation of all the presented features in the scenario of music structural segmentation as introduced shortly.

5.6.2 Music Structural Segmentation

Three evaluation datasets are used in this experiment as introduced in Section 3.4. The first two, BeatlesTUT and S-IA, are publicly available datasets comprising mainly Western music commonly used to evaluate MSS tasks. The third one is presented in this thesis consisting 30 excerpts of Jingju songs with an overall length of 3.6 hour. Due to the relatively small size of the evaluation datasets, we propose a signal processing based segmentation algorithm which does not rely on machine learning.

A novelty-based (for details see Section 2.5.3) segmentation method is used in the experiment. Although different segment strategies exist, the novelty-based approach attempts to detect segment boundaries by locating prominent change points in the feature representations and encodes few heuristics of the music signal itself into the design of the segmentation algorithm. Therefore, assessing the segmentation performances with such methods can provide a straightforward assessment the audio features used.

We first compute the self-similarity matrix (SSM) using the pairwise Euclidean distance of the feature matrix (see Section 2.5.3). A $k_G \times k_G$ Gaussian-tapered "checker-

board" kernel is correlated along the main diagonal of the SSM yielding a novelty curve $NC(i)$ where $i = 1, 2, \cdots, N$ following Foote [Foo00]. In this process, the size of the Gaussian kernel k_G decides the size of the vicinity that the local frame is compared against to derive the associated novelty score. An experimentally defined value of $k_G = 80$ is used in the evaluation. We will discuss the effect of this parameter in Section 6.3.2.

The derived novelty curve is essentially a one-dimensional feature indicating candidate boundaries with peaks. In many novelty-based segmentation algorithms, segment boundaries are typically detected from the NC based on an adaptive thresholding mechanism using median filters, i.e., a boundary will be identified when the novelty score exceeds a local threshold. However, the moving medians may form a "plateau" shape in the presence of a peak in the NC. Therefore, multiple samples around a peak may have magnitudes exceeding the thresholds set by the medians, leading to multiple detections in the vicinity. Additionally, this peak picking method may neglect peaks with small amplitudes overlooking its spikiness especially when peaks with higher amplitudes present in the neighbourhood.

In this work, we propose a new boundary retrieval algorithm using a *polynomial fitting* mechanism inspired by the onset detection algorithm introduced in Section 2.4.3. The main difference between the two is that the backtracking process is excluded here. To be more precise, normalisation is firstly applied to the raw NC. The rescaled NC is then passed through a low-pass filter which is used for noise removal. Subsequently, adaptive thresholds are generated from the smoothed NC using a median filter. Finally, we fit a second-degree polynomial on the smoothed novelty curve centred around each local maximum obtained from the adaptive thresholding using an experimentally defined window of 5 frames. A candidate will be accepted as a segment boundary when both the amplitude and the sharpness of the parabola meet set conditions controlled by a single sensitivity parameter *sens* where $sens \in [0, 100]$. A low *sens* of 20 is used in this experiment. This boundary retrieval algorithm is denoted *Quadratic novelty* (QN) in

this thesis.

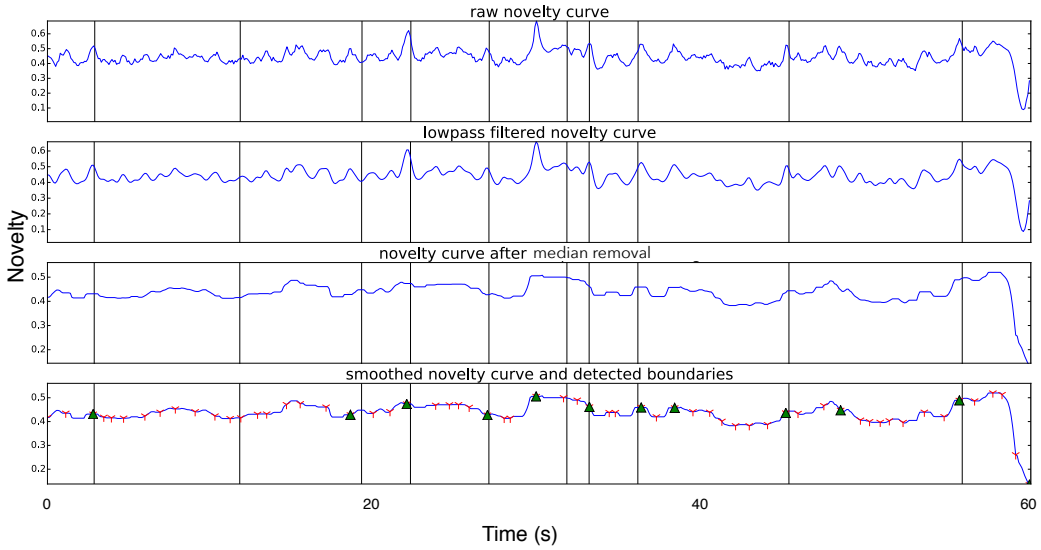


Figure 5.11: Segmentation process on Jingju music excerpt "Hong niang" using the MFCCs feature (extracted after HPSS with the maximum filter applied) by algorithm *QN*. The black vertical lines, green triangles and red crosses represent respectively the annotations, detected boundaries and those would also have been retrieved without the polynomial fitting (using adaptive thresholding).

Figure 5.11 shows the segmentation process of a 60-second excerpt Jingju song "Hong niang" (meaning "The red maid") using feature $hMFCCs_m$ (MFCCs extracted after HPSS with the maximum filter applied) with $\beta = 0.5$, $l_p = 10$, $m_p = 10$. We can see that the novelty scores associated with the annotated segment boundaries in the raw novelty curve can be relatively subtle. The novelties are brought more prominent after the adaptive median removal. Compared to a standard adaptive thesholding based boundary retrieval, polynomial fitting appears effective in deselecting many peaks in the smoothed novelty curve and reducing false positives.

5.7 Results and Analysis

This section is devoted to evaluate the investigated audio features in a music structural segmentation experiment in the contexts of different music genres. Features are evaluated

using the segmentation boundary retrieval precision (P), recall (R) and F-measure (F) measured at a 3-second tolerance (see Section 2.8.1). We first assess the presented features individually in an attempt to find out their optimal use cases and parameter settings then analyse the effect of HPSS on the investigated features.

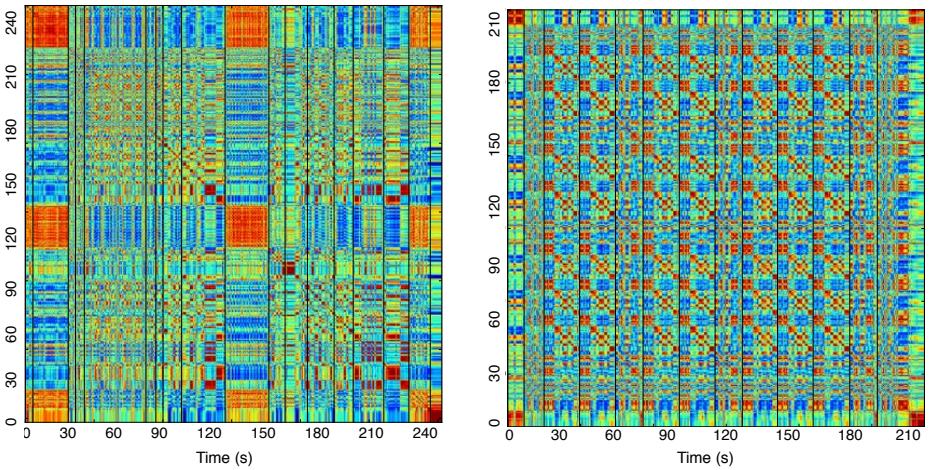
5.7.1 Chroma Features for Jingju

BPO = 12			BPO = 7		
P	R	F	P	R	F
0.387	0.674	0.454	0.416	0.761†	0.503*

Table 5-A: Segmentation precision (P), recall (R) and F-measure (F) using the chromagram feature with 7 and 12 bins-per-octave on dataset CJ. *, † and ‡ denote the presence of significant improvement over the standard versions at the level of 0.05, 0.01 and 0.001 using the Wilcoxon signed-rank test.

Table 5-A shows the segmentation precision (P), recall (R) and F-measure (F) using the chromagram feature with 12 and 7 bins per octave (BPO) on dataset CJ. Significant higher *recall* and the *F-measure* are obtained when $BPO = 7$. As introduced in Section 5.3, a 7-BPO chromagram sets the chromatic energy distribution in an anhemitonic scale which is used by Jingju (see Section 3.2.1). We observe notably higher novelty scores at boundary locations in the novelty curve calculated from the 7-BPO chromagram than that from the 12-BPO chromagram, leading to fewer false natives in the boundary retrieval. Meanwhile, using fewer bins hence coarser resolutions can combat the interference from the short-term frequency oscillation as well as the fast-decaying high-frequency transients, as shown in Figure 5.2. This leads to a reduction of noise in the feature representation and therefore less false positives in the segmentation. In the remainder of this thesis, we use the 7-BPO chromagram for Jingju, i.e., for dataset CJ, unless noted otherwise.

However, it is commonly noticed that chromagram does not form stripes in the subdiagonals in the SSM, contradicting the observations for Western pop music, with one example shown in Figure 5.12. Features used are the 7- and 12-BPO chromagram for the



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

Figure 5.12: SSMs calculated using 7-BPO chromagram for a Jingju song "Tian nu san hua" and 12-BPO chromagram for a Beatles song "Dig a pony". Black vertical lines indicate segment boundaries.

Chinese and Western song respectively with HPSS applied. Although different patterns emerge from the two SSMs, both correlate with individual sectional characteristics. This indicates the potential of chroma features as structural descriptors for genres with an absence of chord repetitions. It is also suggested that different segmentation algorithms may be needed to interpret the structural patterns encoded in the chroma feature for different music styles.

5.7.2 Segmentation with Tempogram Features

	$W_a = 3$ s			$W_a = 5$ s			$W_a = 6$ s			$W_a = 8$ s		
	P	R	F	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.395	0.585	0.446	0.419	0.670	0.464	0.425	0.619	0.501‡	0.478	0.588	0.474
CJ	0.414	0.650	0.456	0.438	0.632	0.460	0.471	0.614	0.484	0.433	0.549	0.451
S-IA	0.404	0.721	0.501	0.468	0.679	0.513	0.495	0.618	0.520	0.501	0.513	0.497

Table 5-B: Segmentation precision (P), recall (R) and F-measure (F) with tempogram features under different time window settings. *, † and ‡ denote the presence of significant difference in F-measure comparing the best performing W_a setting to the second best for each dataset at the level of 0.05, 0.01 and 0.001 using the Wilcoxon signed-rank test.

To obtain the optimal performance for the investigated tempogram features, we first

analyse the effect of the size of the time window W_a, the main parameter involved in the tempogram calculation process (see Section 5.4.1). Segmentation results under different W_a settings are reported in Table 5-B. $W_a = 6$ s achieves the optimal segmentation results for all investigated datasets, which is in agreement with previous studies on Western pop music [Pee05; GM11b]. The difference between the best and the second best performing W_a setting is the most notable for dataset *BeatlesTUT* while the least for *CJ*. However, we can observe that when the time window is set to 5 s or 6 s, lower precision is obtained compared to 3 s or 8 s, with the longest window yielding the highest precision. One explanation is that spurious peaks are suppressed due to longer window, though this is accompanied by a drop of recall due to coarser temporal resolution in the feature representation.

From Table 5-B we can observe a significantly higher recall than precision rate, i.e., an *over-segmentation*, on all investigated datasets when W_a is set to 3, 4 or 5 seconds. This phenomenon is consistent under different sensitivity settings for the boundary detection. This suggests that the tempogram features are relatively robust in detecting true segment boundaries. However, rhythmic variations do not necessarily indicate emergence of new sections. The high false positive rate can also be due to the noise-incurring nature of novelty-based segmentation methods. This is also reflected in Figure 5.11, where we observe the case when there is no annotated boundary whereas it is characterised by a notable peak in the novelty curve. We will investigate different segmentation algorithms in the next chapter.

The calculation of the tempogram is based on the local periodicity of the onset detection function, which identifies the amplitude, phase or other changes in the spectrogram of the input audio signal. Therefore, spectral information is not disregarded; rather, it is presented in an abstracted manner with rhythmic cues emphasised. The fact that the features are tested on a collection of pop music instead of hand selected pieces with well defined rhythmic patterns indicates their general applicability to music content description.

	P	R	F	d_{AD}	d_{DA}
TCCs	0.407	0.612	0.498	0.76	2.03
TPCs	0.402	0.606	0.483	0.85	2.11
TI	0.431	0.527	0.476	0.81	2.15
TIR	0.446	0.551	0.502	0.88	2.08

Table 5-C: Segmentation precision (P), recall (R) and F-measure (F) on S-IA dataset using TCCs, TPCs, TI and TIR measured at 3 seconds (time window size $W_a = 6$ s).

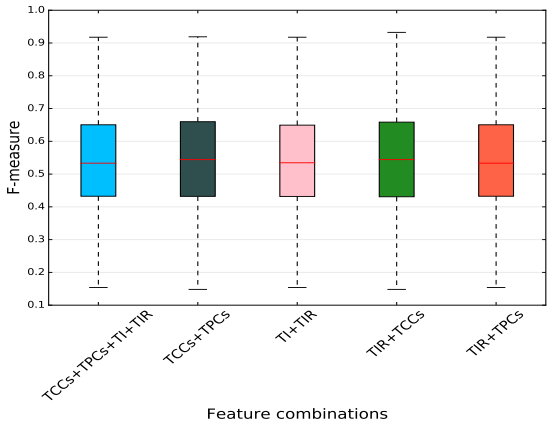


Figure 5.13: Segmentation F-measures of the top five tempogram feature combinations. Boxes contain segmentation F-measures for all samples in the dataset S-IA measured at 3 s. The min, first and third quartile and max value of the data are represented by the bottom bar of the whiskers, bottom and upper borders of the boxes and upper bar of the whiskers respectively. Medians are shown by the red line.

To compare the performance of each feature, we repeat the experiments by using each feature individually under the same experimental conditions on the S-IA dataset. The time window is set to 6 s. Results are given by Table 5-C. All features have similar segmentation rates with an average F-measure of 0.480 when the boundary recovery is measured at 3 s. As pointed out by Smith and Chew [SC13a], retrieving segment boundaries to within 3 s and to within 0.5 s can be two distinct tasks. It is therefore sensible to identify the suitable features for these two tasks individually. When measured at a finer scale (0.5 s), a notable drop of performance of the investigated tempogram features can be observed. This difference is due to the fact that the tempogram calculation employs large window sizes leading to degraded temporal resolutions. We however also identify from this observation that tempogram features may qualify to be used in other MIR task

operated at a track level, for example, music genre classification. No significant difference (Kruskal-Wallis test) can be found among the performances of the four features. A Spearman rank-order correlation test also confirms correlation between the features.

To find out the best usage of the four features, we use all possible combinations of them, i.e., the *power set* of the set of the four features (excluding the empty set), for segmentation under the same experiment conditions for dataset S-IA. Vectors of features to be combined are concatenated. The concatenated feature is subject to PCA where only the lowest 6 dimensions are used. The best five performing feature combinations in terms of segmentation F-measures are shown in Figure 5.13. The concatenation of TCCs and TIR surpasses all the combinations obtaining the highest segmentation F-measure. However, a statistical significance is lacking from different feature combinations.

TCCs and TIR characterise different aspects of the tempogram. While TCCs offer a compact representation of the energy distribution of the tempogram with the cepstral analysis applied, TIR measures the level of tempo intensity from a more musically meaningful perspective. The combination of these two features hence fuses the two aspects, yielding an effective feature descriptor. In the remainder of this thesis, we use the vector-wise concatenation of *TCCs* and *TIR* as the tempogram feature set and denote it *RT*.

5.7.3 Effects of Harmonic Percussive Source Separation

We use MFCCs, chromagram (BPO=7 for CJ and 12 for S-IA and BeatlesTUT) and RT to evaluate the effect of the investigated HPSS technique for music segmentation. Results are given in Table 5-D comparing segmentation precision (P), recall (R) and F-measure (F) measured at 3 seconds. Results are yielded by the original features (Table 5-Da) and the features extracted after the harmonic-percussive source separation (HPSS) with/without the maximum filtering (Table 5-Db/Table 5-Dc). The significance level of the differences in segmentation F-measures between a feature with and without HPSS obtained for each audio sample in a dataset is measured using Wilcoxon signed-rank test.

	Chromagram			*MFCCs*			*RT*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BeatlesTUT	0.367	0.715	0.462	0.346	0.691	0.451	0.383	0.684	0.441
CJ	0.388	0.767	0.448	0.406	0.763	0.490	0.415	0.716	0.467
S-IA	0.417	0.788	0.472	0.478	0.661	0.501	0.465	0.684	0.498

(a) Segmentation results using features extracted without HPSS.

	Chromagram			*MFCCs*			*RT*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BeatlesTUT	0.403†	0.723	0.483	0.365	0.732	0.468	0.395	0.689	0.464
CJ	0.412	0.758	0.462	0.417	0.785	0.491	0.425	0.723	0.480
S-IA	0.454	0.801	0.496	0.499	0.649	0.506	0.469	0.624	0.503

(b) Segmentation results using features extracted with HPSS (maximum filter excluded).

	Chromagram			*MFCCs*			*RT*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BeatlesTUT	0.405†	0.746	**0.504***	0.395	0.745	0.476	0.407	0.683	0.471*
CJ	0.418	0.775	0.491*	0.422	0.791	**0.513**	0.432	0.743	0.495
S-IA	0.521	0.753	**0.517***	0.507	0.624	0.516	0.471	0.581	0.501

(c) Segmentation results using features extracted with HPSS (maximum filter applied).

Table 5-D: Segmentation precision (P), recall (R) and F-measure (F) measured at 3 seconds using selected features on *BeatlesTUT*, *CJ* and *S-IA* dataset. Highest F-measure for each feature is shown in bold. *, † and ‡ denote the presence of significant improvement over the standard versions at the level of 0.05, 0.01 and 0.001 using the Wilcoxon signed-rank test.

The most notable improvements are observed for *chromagram*, with $p < 0.001$ in cases both with and without the maximum filtering in the HPSS, second to which are MFCCs, with $p < 0.05$ in and only in the maximum filtered case. 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 a novelty-based segmentation algorithm may have limited efficacy in discerning the false positives in low-level timbre similarities introduced by MFCCs, and can overlook the longer term repetition structures presented by chroma features, incurring limited precision and recall rate in the first place.

The effect of HPSS is altered by different configurations of its parameters. The most influential parameter in our case is the separation factor β. Driedger et al. report that when $\beta = 1$, the residual is roughly equally distributed in both X_h and X_p while

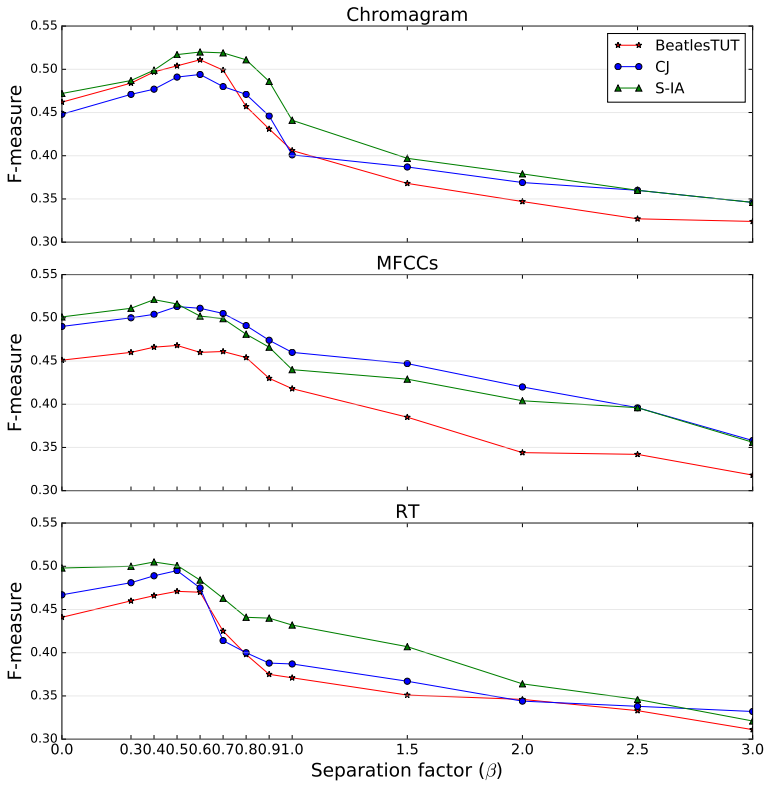


Figure 5.14: Average segmentation F-measures using features extracted with different β settings in the harmonic-percussive source separation (HPSS).

when $\beta = 3$ only clearly horizontal and vertical structures are preserved in the spectrogram [DMD14]. Applying the maximum filter in the separation also has the tendency of leaving residual components in the resulted harmonic and percussion source. As can be seen from Figure 5.14, a β ranging from 0.4 to 0.5 is optimal for all investigated features and datasets (results in Table 5-D are obtained with $\beta = 0.5$). RT requires lower β for an optimal segmentation performance compared to the other two, with the F-measure starts to drop sharply from $\beta = 0.6$. Unlike the chroma descriptions that rely on only the harmonic source, rhythmic information can be expressed also by the pitched components. Removing the harmonic source from the tempo tracking may lead to a loss of subsidiary rhythmic information in the derived features. When β exceeds 0.8, all features yield worse segmentation results than when HPSS is not applied. This shows that in

the case of music structural analysis, it is not desirable to have the opposite source and the residuals tightly removed, given each source may contain complementary structural information.

5.7.4 Segmentation with Gammatone Features

	GCCs (fast)			GCCs (accurate)			GC (fast)			GC (accurate)		
	P	R	F	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.433	0.617	0.502	0.425	0.632	0.503	0.412	0.546	0.434	0.425	0.561	0.451
CJ	0.445	0.785	0.543	0.438	0.726	0.520	0.449	0.634	0.489	0.425	0.624	0.463
S-IA	0.437	0.769	0.511	0.441	0.776	0.515	0.402	0.668	0.445	0.389	0.662	0.441

Table 5-E: Segmentation precision (P), recall (R) and F-measure (F) measured at 3 seconds with Gammatone features extracted from the accurate Gammatone calculation and the FFT-based fast approximation.

As introduced in Section 5.5, the FFT-based Gammatone approximation provides a close match to the accurate Gammatone filtering process. In this section, we evaluate the presented Gammatone features in a segmentation scenario.

Segmentation results using features extracted from both the fast and the accurate Gammatonegrams are shown in Table 5-E. GCCs derived from the two Gammatonegrams give close results on all datasets. As introduced in Section 5.5.1, the fast method assumes coherence of addition of amplitude from neighbouring channels and neglects the variations in phrase, hence yields smoother transitions between bins than the actual Gammatone filters do. The DCT operation, however, re-projects the original spectra for an energy compaction and has the tendency to erase such discrepancies. This would lead to similar representations of GCCs from both methods as also noted in Figure 5.8.

For the Gammatone contrast (GC), higher F-measures are obtained when using the fast Gammatonegram on CJ and S-IA mainly because of improved precisions. As shown in Figure 5.9, using the fast method can reduce the time smearing effect in the lower frequency channels. This can improve the salience of the corresponding boundaries in the feature representations especially when with active presence of vocals. However, better segmentation is obtained when using the accurate Gammatonegram for BeatlesTUT.

One explanation is that the energy additions between channels in the fast Gammatonegram can have a prohibitive effect in discriminating the instruments which could set the structural sections apart.

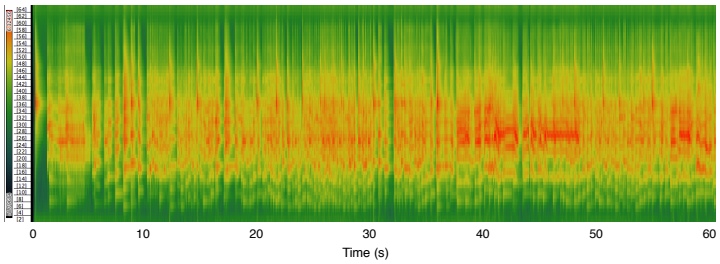
In general, however, the two sets of features yield very similar segmentation results with no statistical significance presented between each pair when evaluated on all the three datasets. Although using the Fourier transform would lead to a loss of temporal resolution, the difference becomes less remarkable after the actual Gammatone filter outputs have been aggregated into fixed temporal windows to derive the time-frequency representation. As a conclusion, using the fast method for feature extraction is proved at least *not harmful* compared to the accurate method in our case, especially with suppressed time smearing effects. In the remainder of this thesis, we will use features derived from the fast method for a reduced computation cost.

	Configuration	BeatlesTUT			CJ			S-IA		
		P	R	F	P	R	F	P	R	F
GCCs	Global	0.388	0.762	0.488	0.458	0.781	0.531	0.453	0.669	0.511
	Individual	0.393	0.766	0.493	0.461	0.791	0.549	0.460	0.787	0.522
GCCs+GC	Global	0.397	0.776	0.495	0.467	0.796	**0.558***	0.425	0.697	0.522
	Individual	0.407	0.787	0.505	0.470	0.797	0.584	0.451	0.693	0.535
MFCCs	Global	0.395	0.745	0.476	0.422	0.791	0.513	0.507	0.624	**0.516**
	Individual	0.401	0.761	0.489	0.430	0.797	0.521	0.512	0.640	0.524
Chromagram	Global	0.405	0.746	**0.504**	0.418	0.775	0.491	0.521	0.753	0.517
	Individual	0.408	0.766	0.506	0.425	0.774	0.509	0.524	0.764	0.522

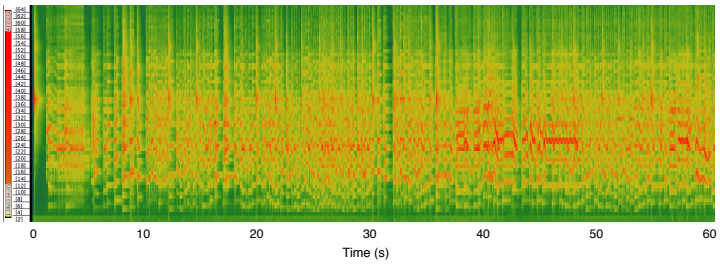
Table 5-F: Segmentation precision (P), recall (R) and F-measure (F) measured at 3 seconds with the Gammatone features, MFCCs and the chromagram. *, † and ‡ denote the presence of significant improvement from the best performing feature over the second best performing feature on individual datasets at the level of 0.05, 0.01 and 0.001 using the Wilcoxon signed-rank test.

Here we compare the Gammatone features extracted from the FFT-based Gammatonegram with chromagram and MFCCs as structural descriptors. The latter features are extracted after harmonic-percussive source separation with maximum filtering (see Section 5.7.3). Individual features are evaluated on each dataset, as shown in Table 5-F. Results are 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 specific music types. To avoid

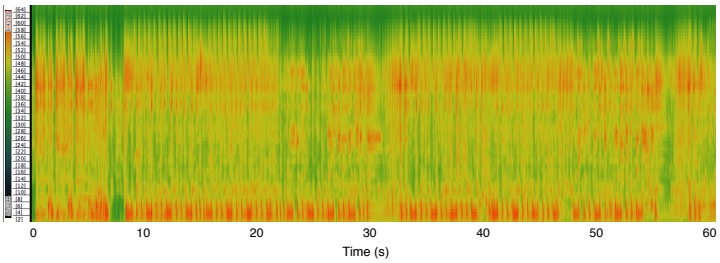
the potential overfitting, we use the global configuration for the Gammatone feature set unless noted otherwise in the remainder of this thesis.



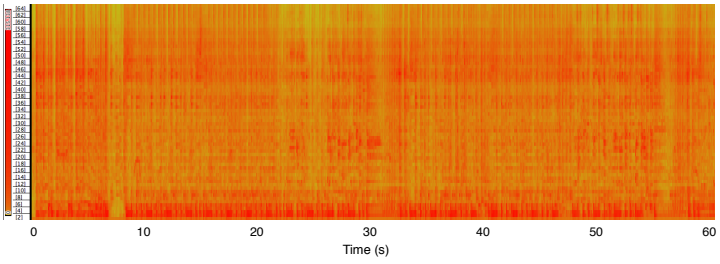
(a) Gammatone (FFT-based) spectrogram for song for song “Ba wang bie ji”.



(b) Mel spectrogram for song “Ba wang bie ji”.



(c) Gammatone (FFT-based) spectrogram for song “Hideout”.



(d) Mel spectrogram for song “Hideout”.

Figure 5.15: Gammatonegram and Mel spectrogram for Jingju song “Ba wang bie ji” from CJ and a rock song “Hideout” (0 - 60 s) from S-IA visualised on a log scale.

Here we compare GCCs to MFCCs, both are based on cepstral analysis of the spectra and related to the music timbre, on individual datasets. The segmentation method used is Quadratic Novelty (QN). For BeatlesTUT and S-IA, GCCs obtain higher recall than

MFCCs. However, there is a significant drop of precision on S-IA ($p < 0.05.$, Wilcoxon signed-rank test). We find that more boundaries have been retrieved, a large portion of which, however, are false positives.

Figure 5.15 shows the Gammatonegram and the Mel-scaled spectrogram for a Jingju song "Ba wang bie ji" from CJ and a rock song "Hideout" from S-IA visualised on a log scale. 64 Mel filters are used with the upper and lower bound set to 50 Hz and half the sample rate, same as the settings for the Gammatone filters. For both songs, only the first 60 seconds are shown in the plot for visualisation purposes. From Figure 5.15c and Figure 5.15d we can notice a rather remarkable difference between the Gammatonegram and the Mel-spectrogram in terms of the level of contrast among vectors and frequency bins. This would lead to larger variations in the derived features hence sharper local peaks in the corresponding novelty curves. For music types with relatively subtle sectional acoustical variations such as the "through-composed" music, this may assist the derived feature representations to become more discriminative against structural segments. However, for music types with distinctive sectional timbral variations, it may be less advantageous for novelty-based segmentation methods to discriminate the false positives from local peaks in the novelty curve. In future work, we propose to use different segmentation methods relying long-term homogeneity or repetition structure principles to better employ this feature.

On CJ, GCCs retrieve less spurious boundaries than MFCCs do, leading to higher precision hence an improved F-measure overall. As introduced in Section 2.2.2, ERB filters yield smoother frequency responses in the low-frequency areas than Mel. This is advantageous for many vocal-driven music works, such as opera, where the singing voice is an important discriminator of the music structure with its salient presence in the overall instrumentation. As can be seen from the Gammatonegram for the Jingju song shown in Figure 5.15a, the dynamics introduced by the singing voice mainly present in the lower-middle frequency bands of the spectra, which can be better captured by the ERB scale than Mel. Meanwhile, by emphasising the lower sound levels, the ERB

warping can be more robust against the high-frequency transients which could interfere with the analysis. Figure 5.16 shows the self-similarity matrices (SSMs) derived from the MFCCs and GCCs on a Jingju song "Ba wang bie ji" from CJ (only the first 60 seconds are shown). It can also be noticed from the SSMs that GCCs yield more distinguished sectional variations than MFCCs.

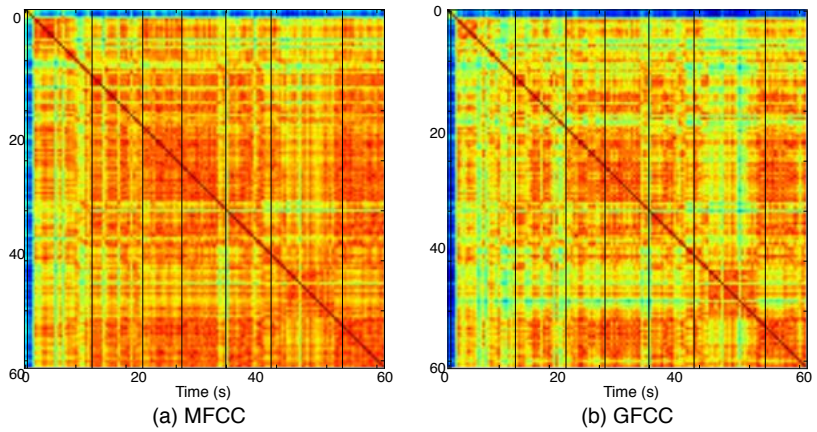


Figure 5.16: SSMs computed using MFCCs and GCCs on the first 60 seconds excerpt of "Ba wang bie ji" from CJ. Vertical lines indicate segment boundaries.

Although GC gives somehow mediocre segmentation rates when used alone (see Table 5-E), combining it with GCCs has introduced improvements over both baselines for most cases on the investigated datasets, as shown in Table 5-F. However, a statistical significance is only present for dataset CJ. The main effect of using this feature by concatenating it to the feature matrix of GCCs is a more pronounced within-SSM variance. This has led to the retrieval of more boundaries as indicated by a higher recall rate in a general case yet sometimes a degrading precision.

In this experiment, GCCs surpass MFCCs notably on the Jingju dataset comprising vocal-driven music. The fact that the Gammatone features also obtain comparable segmentation performance to MFCCs and chromagram on BeatlesTUT and S-IA indicate them as effective alternatives to existing audio features for music structural analysis. In the remainder of this thesis, we use the vector-wise concatenation of GCCs and GC as the Gammatone feature for the analysis of relevant scenarios. We denote this feature set

GF.

5.7.5 Audio Features and Music Genres

Investigated features perform differently on Western and Jingju music. Chromagram and MFCCs work reliably for Western music as shown in Table 5-D and Table 5-F, confirming previous conclusions [PMK10]. For Jingju, timbre features capture its structural characteristics better than the chroma feature, with the auditory-inspired Gammatone features outperforming MFCCs. Structural patterns of different genres may be conveyed differently by individual features, as can be seen in Figure 5.12, suggesting that segmentation algorithms should incorporate different principles to interpret such patterns.

It can be noticed from Table 5-F that chromagram presents lower F-measures than all the timbral features tested, despite that chroma features are well recognised to model the structure of Western pop music [PMK10]. This may partly be due to that QN as a novelty-based segmentation algorithm is less effective to capture the global harmonic structures that chroma features are depicting. It is also observed that algorithms are more dependent on parameter configurations when evaluated on CJ than on S-IA and BeatlesTUT, reflected by the more substantial degradation in segmentation F-measures observed when changing the parameter configuration tuned for the individual dataset to the global setting (see Table 5-F). This implies the need for designing new segmentation methods to bridge the gaps between genres. It also implies that musical knowledge, such as the genre and the hierarchy of music structure to analyse, can assist a segmentation system to obtain better performance. In the next chapter, we will analyse different segmentation methods for specific audio features and music types.

5.7.6 Future Applications of Presented Features

Whilst features presented in thesis chapter are evaluated in an MSS scenario, we also intend to identify their potentials in other applications. Firstly, we propose to apply

the tempogram features to describe the music content on the piece-level for applications such as music classification and genre recognition. In this way, the degraded temporal resolution would become less a limitation for the performance of the features. As shown in Figure 2.4, very distinct temporally evolving patterns emerge in the tempograms for the Jingju and the Beatles song. A recent study [CSC10] shows that the rhythmic aspects of songs, especially the beat intensities and instrument distributions, turn out to be very effective when used for genre recognition. It also shows that such features might be an even better tool to estimate music preferences of humans than the current automatic genre classifiers. We hence are also curious in applying these features for music recommendation.

The sound sources in Jingju, especially the singing voice and the major melody instrument *jinghu* (a bowed-string fiddle), often overlap. This may lead to many pitch tracking algorithms to fail. The Gammatone features have surpassed MFCCs in the structural segmentation for Jingju due to their descriptive capacity for vocals. Applications of this feature set hence lie in the music summarisation as a replacement of the common timbre or spectral features. We also propose to design timbre-invariant chroma features using templates calculated from Gammatone features.

5.8 Summary

This chapter presented novel audio features for music structural segmentation (MSS) in the scenario of both Jingju and Western music. To begin with, although chroma features are originally developed to describe the harmonic properties for Western music with 12 pitch classes, we analysed the bins per octave (BPO) setting in chroma for Jingju and justified the use of the 7-BPO chromagram.

Despite identified as one of the most important indicators of the structure changes, rhythmic features are not as commonly used as harmonic or timbral features. As the second contribution, a set of rhythmic features was presented incorporating perceptual

considerations. Unlike standard rhythmic features from literature which employ only the tempo information, features were extracted from the mid-level tempogram representation consisting complementary spectral information with the rhythmic cues emphasised. The fact that the features are tested on a collection of pop music instead of hand-selected pieces with well defined rhythmic patterns indicates their general applicability to music content description. We also identified that rhythmic components with slower tempo have stronger boundary salience as structure indicators.

Although music structural analysis is considered a high-level task involving human perception, auditory cues are barely incorporated in the commonly used audio features. This chapter presented novel timbre features using the Gammatone filters better modelling the human auditory TF resolutions to describe the music structure. The presented Gammatone features work effectively as an alternative to MFCCs and chromagram on all the investigated datasets with Western and Jingju music, and surpass them on the Jingju dataset with salient presence of vocals.

Besides presenting new features, this chapter also introduced a harmonic-percussive source separation (HPSS) algorithm as a pre-processing step for feature extraction. The separation is applied to different features and evaluated on different datasets in an MSS experiment, bringing significant improvements to the segmentation for the investigated music categories. A low separation factor yields the best segmentation for all features investigated, even for chroma feature which is built only on the harmonic content. This indicates that although applying HPSS in the feature extraction is beneficial, all sound sources contribute to the overall music structure hence cannot be excluded from the structural description.

The tempogram features developed are based on the onset detection methods presented in the previous chapter. The segmentation algorithm in this chapter is also adapted from the peak picking algorithm from the onset detection system investigated. We hereby justified that thorough investigations into MIR algorithms and signal processing methods can shed invaluable perspectives for new applications in a music analysis

system.

In this chapter, we used a single segmentation method to evaluate the investigated features. By doing this, we want to assess the strengths of the features themselves and to derive a fair comparison of them. However, using single segmentation method may not guarantee the optimal performance of all investigated features. This is also supported by the observation that the self-similarity matrices of the chroma feature present different patterns for different music types, as discussed in Section 5.7.1. We also noticed from Figure 5.11 that peaks in the novelty curve do not always reliably indicate the presence of segment boundaries. How do different segmentation methods incorporating different structural principles work when different audio features are used or in the scenario of different music types? We aim to answer this question with the project undertaken in the next chapter.

Chapter 6

Methods for Music Structural Segmentation

6.1 Introduction

The previous chapter presented novel audio features which were evaluated in a music structural segmentation (MSS) experiment using a novelty-based method. The evaluation results also indicate that the same audio features may exhibit the structure of different music types in different manners. Therefore, specific algorithms may be required to interpret the encoded patterns from feature descriptors or similarity representations.

As introduced in Section 2.5.3, techniques for MSS mainly fall into three categories relying respectively on the *novelty*, *homogeneity* and *repetition* principles (see Section 2.5.3). To investigate how the three principles interpret the music structure for different genres, this chapter will investigate several popular segmentation algorithms relying on these principles for three datasets consisting both Chinese and Western music. We will also use different segmentation methods to analyse audio features presented in the last chapter.

With the research carried out in this chapter, we aim to analyse how the choice of audio features, segmentation algorithms, music genres and annotation principles interact in an integrated segmentation system. Does one set of annotations provide consistently reliable reference for the evaluation of an MSS system? What is the pathway towards an automatic MSS system across different music styles? This chapter is organised as follows. Section 6.2 introduces the music segmentation system. Section 6.3 is devoted to evaluate the segmentation results and finally, we conclude this chapter in Section 6.4.

6.2 Segmentation Experiments

Different segmentation methods have been proposed as summarised in Table 2.9. Four recently published ones are investigated in this chapter including *FOOTE* [Foo00], *Constrained clustering* (CC) [LS08], *Convex NMF* (CNMF) [NJ13] and *SERRÀ* [Ser+12]. These methods have been introduced in Section 2.5.3. We also include *Quadratic novelty* (QN) as presented in Section 5.6.2. Altogether, these five methods cover the *novelty*, *homogeneity* and *repetition* structural hypotheses (for introductions see Section 2.5.3).

There are several reasons why these algorithms are chosen. First, they are among the state-of-the-art algorithms from recent works relying on signal processing techniques [PMK10]. Although some recent works using convolutional neural networks (CNNs) have achieved better segmentation results [GS15a; USG14], they are beyond the scope of this thesis. This is because these neural network (NN) based methods use the spectrogram as input to the NNs whereas this thesis is focused on the analysis of more specialised audio features to interpret the music structure. Second, these methods make relatively few assumptions on the music genres with the boundary detection approaches designed to be generic. Therefore, they are considered applicable to all investigated music corpora. Finally, the inclusion of different segmentation methods will enable us to investigate the three underlying structural principles. The only method incorporating the *repetition* principle investigated in this chapter is *SERRÀ* where it is combined

with the novelty and the homogeneity principle. Methods based solely on the repetition principle are however excluded from the investigation in this chapter due to the fact that they make structural discoveries by looking for global repetitions which can be lacking in Jingju. We will however discuss the effect of the repetition structural principle in the scenario of Jingju in Section 6.3.3.

The investigated algorithms are evaluated in an MSS context using *Music Structure Analysis Framework* (MSAF) [NB15] comprising a few recently published segmentation algorithms, including FOOTE, CC, CNMF and SERRÀ which are of interest of our research. These methods are introduced in Section 2.5.3. We also include QN into the MSAF framework to obtain a uniform processing environment.

Although the investigated algorithms use specific features in the original implementations, different features characterising different aspect of the music content are assessed in our evaluation including chromagram, MFCCs and the Gammatonegram features GF, as presented in Section 5.7.4. By doing this, we want to illustrate the relations between the performance of different segmentation algorithms and the selection of audio features.

The feature extraction framework used in MSAF is LibROSA, an open source python library for music and audio analysis [McF+15]. This library provides feature extraction functions for chromagram and MFCCs. We also include the Gammatone feature extraction module as introduced in Section 5.5.2 into this library. Note that for the case of CNMF, we require MFCCs to be rescaled and standardised to ensure non-negativity when used as the input feature.

The segmentation workflow is shown in Figure 6.1. A Hann window is used for the FFT with the sample rate fs, window size W_N and hop size W_H being respectively 44.1 KHz, 46.4 ms and 23.2 ms.

Harmonic-percussive source separation (HPSS) is applied as for feature enhancement. The HPSS implementation provided by LibROSA however differs from our method which applies additionally a maximum filter (see Section 5.2), although both use the median

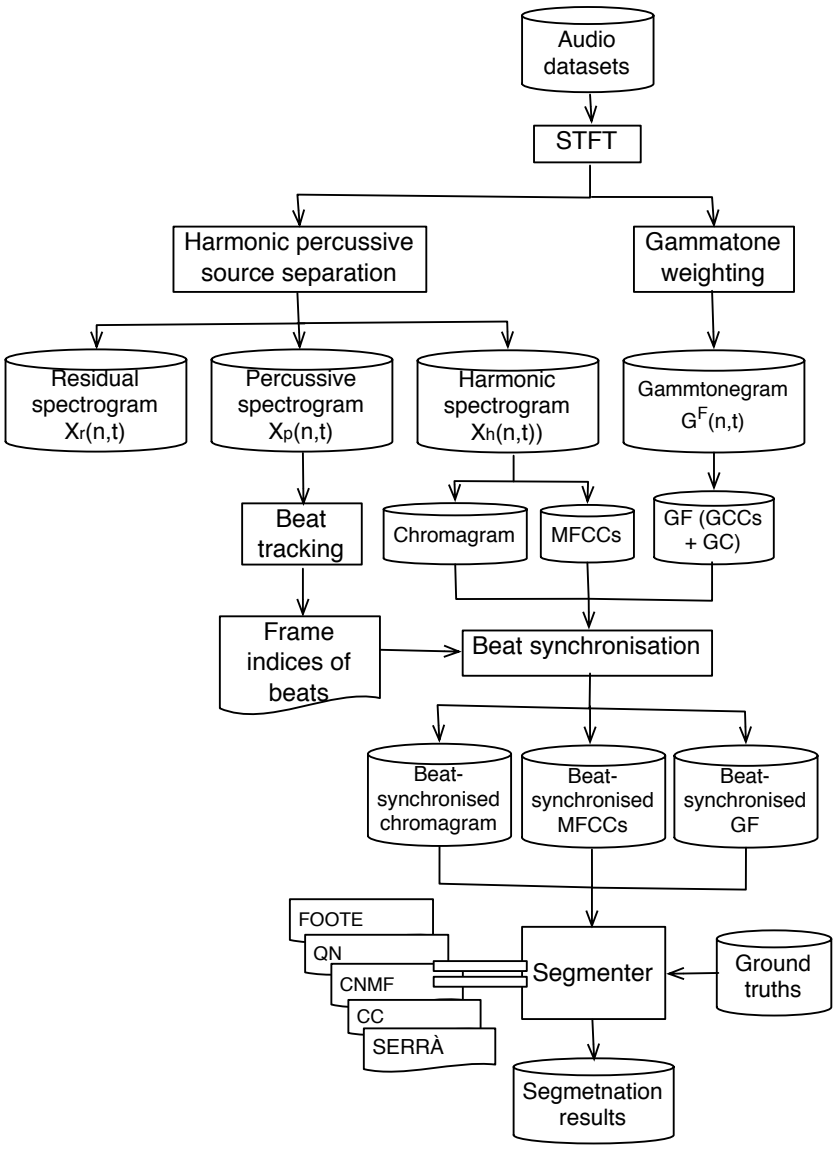


Figure 6.1: Feature extraction and segmentation framework.

filter based approach following FitzGerald [Fit10]. As discussed in Section 5.7.3, the maximum filter has introduced significant improvements in the segmentation compared to features extracted both without HPSS and with HPSS but without the maximum filter following [DMD14]. Therefore, we replace the HPSS implementation in LibROSA with our modified version as introduced in Section 5.2.

MFCCs and chromagram are then extracted from the harmonic spectrogram after the

source separation. The Gammatonegram is calculated using the FFT-based approximation approach (see Section 5.5) where the feature GF is derived. The HPSS is however excluded from this process to obtain a matched approximation. The extracted MFCCs, chromagram and GF are respectively 13, 12 or 7 ($BPO = 12$ for S-IA and BeatlesTUT and $BPO = 7$ for CJ as introduced in Section 5.3) and 19-dimensional features.

The separated percussive source X_p is used for beat tracking following Ellis [Ell07] which employs a dynamic programming approach to search for the optimal beat sequence from the global tempo estimation. Extracted features are then beat-synchronised such that the analysis frames are aggregated onto a beat-level [Ell07]. Features are then used to segment the music on our three evaluation datasets as summarised in Table 3-C using the investigated algorithms. Segmentation results will be presented in the next section.

6.3 Results and Discussion

6.3.1 Segmentation Results

Table 6-A shows the segmentation precision (P), recall (R) and F-measure (F) measured at 3 seconds with investigated methods. While SERRÀ (see Section 2.5.3) performs consistently on the investigated datasets, the other algorithms have varied performances across datasets. QN significantly outperforms all other algorithms on dataset CJ, with $p = 0.023$ (Wilcoxon signed-rank test) between the F-measures by QN and SERRÀ as the second best performing algorithm. SERRÀ surpasses all the other investigated algorithms on BeatlesTUT, with comparable results using the three different features. On S-IA, the novelty- or homogeneity-based methods present superior results than repetition-based ones and the timbre features tend to outperform the chroma feature, with significant differences for most cases. We will analyse the segmentation performance given specific system configurations, feature types and music genres in the remainder of this chapter.

	GF			MFCCs			Chromagram		
	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.661	0.543	0.587	0.658	0.534	0.581	0.612	0.514	0.549
CJ	0.754	0.358	0.475	0.708	0.338	0.449	0.659	0.319	0.421
S-IA	0.562	0.522	0.520	0.556	0.545	0.529	0.516	0.547	0.525

(a) FOOTE.

	GF			MFCCs			Chromagram		
	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.523	0.710	0.587	0.584	0.635	0.580	0.435	0.726	0.530
CJ	0.673	0.521	**0.574**	0.706	0.439	0.521	0.587	0.604	0.574
S-IA	0.438	0.666	0.508	0.478	0.610	0.513	0.394	0.704	0.480

(b) Quadratic novelty (QN).

	GF			MFCCs			Chromagram		
	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.555	0.708	0.612	0.568	0.706	0.618	0.527	0.668	0.579
CJ	0.688	0.436	0.522	0.689	0.441	0.525	0.662	0.453	0.524
S-IA	0.514	0.586	0.533	0.517	0.565	0.526	0.558	0.544	**0.535**

(c) Constrained clustering (CC)

	GF			MFCCs			Chromagram		
	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.563	0.597	0.594	0.540	0.559	0.537	0.491	0.652	0.537
CJ	0.688	0.413	0.492	0.630	0.380	0.463	0.557	0.404	0.450
S-IA	0.528	0.464	0.478	0.502	0.438	0.448	0.472	0.526	0.477

(d) Convex non-negative factorisation (CNMF).

	GF			MFCCs			Chromagram		
	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.603	0.761	0.663	0.621	0.772	0.671	0.612	0.777	**0.679**
CJ	0.664	0.471	0.540	0.676	0.482	0.551	0.674	0.445	0.525
S-IA	0.433	0.635	0.493	0.443	0.635	0.497	0.438	0.630	0.500

(e) SERRÀ

Table 6-A: Segmentation precision (P), recall (R) and F-measure (F) measured at 3 seconds with individual features using investigated methods. The highest F-measure for each dataset is shown in bold.

	FOOTE	QN	CC	CNMF	SERRÀ
BeatlesTUT	0.463	0.588	4.448	1.661	0.754
CJ	1.766	2.438	21.534	22.073	3.632
S-IA	2.136	1.883	15.228	9.270	2.609

Table 6-B: Average computation time (in second) for each sample in a dataset with investigated algorithms. The feature used is chromagram.

These methods are associated with different computation complexities [AB07]. Table 6-B lists the average computation time of each method for each track in a dataset using the chromagram feature. The computation statistics are obtained using a two-core Macintosh machine with 3.2 GHz CPU and 12 GB RAM.

Compared to the novelty-based approaches, the homogeneity- and repetition-based methods tend to be of higher computational cost with their analyses taking a top-down or global approach. While FOOTE, QN and SERRÀ have their computation time correlate quasi-linearly with the length of the track, the computation time of CC and CNMF depends also on the complexity of the feature representations to be encoded or factorised and the pre-defined classification rules. For practical reasons, for an integrated MSS system whose analysis can involve large corpora or playlists, the computation complexity becomes another important factor in the selection of segmentation algorithms. Under these conditions, the relatively low-cost methods may become more advantageous despite of the possible limitations in the segmentation performance.

6.3.2 Algorithms and Parameter Configurations

	FOOTE	QN		CC	CNMF	SERRÀ	
	k_G	k_G	*sens*	C_c	R	m_d	S_n
Tian	66	100	80	8	6	3	0.04
MSAF [NB15]	96	-	-	8	3	3	0.04

Table 6-C: Parameter configurations used to derive the results in Table 6-A.

Results reported in Table 6-A are derived using system configurations differ from the original implementation by Nieto and Bello [NB15], as summarised in Table 6-C. A few additional parameters are left out from the table as they are considered less influential to the segmentation results.

The parameter *Gaussian kernel size* (k_G) is used in both FOOTE and QN. The effect of this parameter for boundary retrieval has been discussed in Foote [Foo00]. We however notice different settings of this parameter for both algorithms resulted from

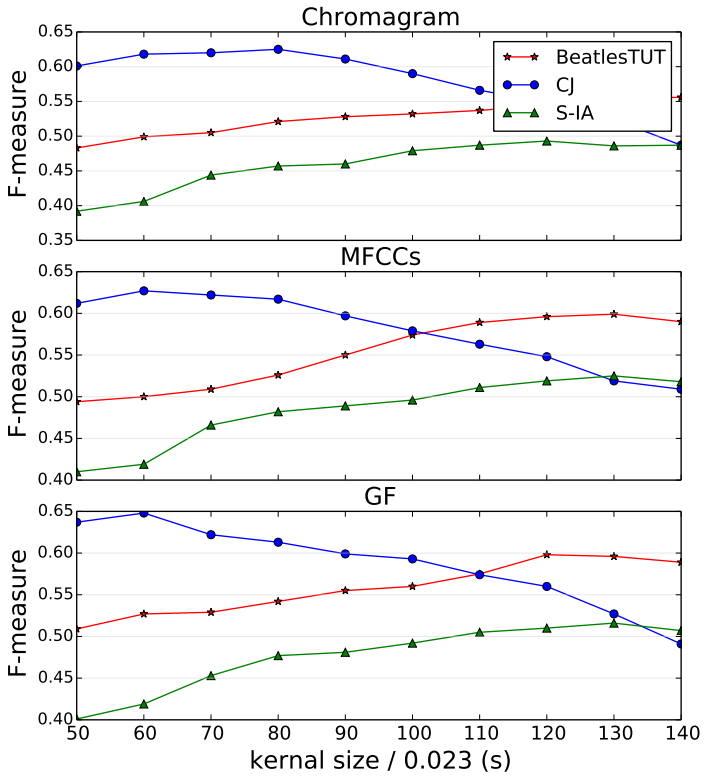


Figure 6.2: Segmentation F-measures measured at 3 seconds using individual features with method QN under different settings of Gaussian kernel size (k_G).

different boundary retrieval methods.

Here we analyse the effect of k_G in the scenario of algorithm QN. Figure 6.2 shows the average segmentation F-measures by algorithm QN under different k_G settings using the three features on each dataset. The rest of the parameters are fixed to the optimal settings. Lower k_G is needed for CJ, with the optimal value ranging from 60 - 80, than S-IA and BeatlesTUT whose optimal k_G is around 120 to 130. For BeatlesTUT which is annotated at a music function level (see Section 3.4.1), a reasonably high k_G is needed to screen out the false positives associated with the local novelties. Although CJ and S-IA are annotated on the same similarity level and that the average segment length of S-IA is shorter than that of CJ (see Table 3-C), smaller kernels are proven beneficial for the latter because of the relatively *lower boundary salience* reflected by the

consistently lower novelty scores in the novelty curve, which also tends to be noisier (see Figure 5.11). This suggests that the boundary salience of acoustical novelty is correlated with the characteristics of specific music genres, as has also been noted for Western music [Smi14].

Another important parameter of QN is the boundary retrieval sensitivity *sens*. This parameter is also involved in the onset detection algorithms as analysed in Chapter 4, where the appropriate *sens* setting differs across different onset types (see Section 4.4.4 for details). In this work, a uniformly high *sens* of around 80 - 90 (out of 100) is required for all datasets. This somehow differs from the findings reported in the last chapter that, when using the frame-synchronised features a low *sens* of around 20-30 is needed to avoid retrieving a large number of false positives.

In FOOTE, boundaries are detected using an adaptive threshold generated by a median filter. The optimal window length of the median filter however varies significantly across the investigated datasets. The boundary retrieval mechanism of QN is to assess not only the amplitude but also the spikiness of a peak in the novelty curve as a boundary candidate. The fact that roughly the same *sens* value yields the optimal segmentation F-measures for all evaluation datasets suggests that this algorithm is less dependent on system configurations. Nevertheless, a novelty-based method alone may not be sufficient enough to prune the false positives introduced by local novelties, as indicated by the lower precision than recall in a common case.

In a standard NMF decomposition, the decomposition rank R is normally chosen such that $(N+P)R < NP$ for the input matrix $\mathbf{V} \in \mathbb{R}^{N \times P}$. Kaiser and Sikora reported a maximum of sectional separability with a rank of 9 for the NMF decomposition where the input matrix $\mathbf{V}$ is the SSM of the audio features for the Beatles dataset [KS10]. In this work, the effect of R is illustrated in Figure 6.3 by the segmentation F-measures. Similar trends emerge from different features where results reported in Figure 6.3 are derived using the Gammatone features GF. The system performance on BeatlesTUT is less dependent on the setting of R compared to the other two datasets. When within

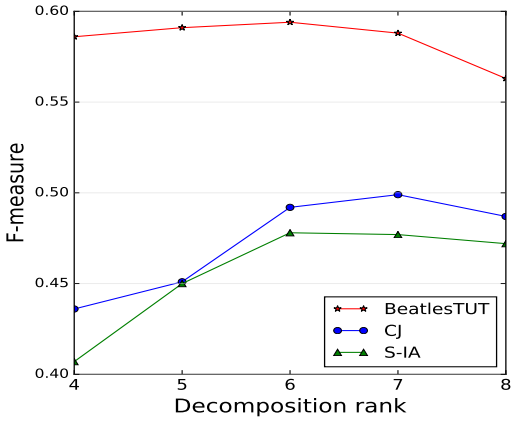


Figure 6.3: Segmentation F-measures measured at 3 seconds with method CNMF under different settings of rank of decomposition using the feature GF.

a moderate range of 4 to 7, no significant difference in boundary retrieval is observed. The main disparity however is that retrieved segment are clustered into different types of sections and labelled differently. For S-IA and CJ, on the contrary, the algorithm shows higher dependency on the configuration of this parameter with the optimal setting being around 6 - 7. Under the optimised condition, the segment types are characterised mainly by the leading instrumentations (vocal, harmonic instruments, etc.).

However, when a large R is used exceeding the optimal range, this algorithm works differently on BeatlesTUT as on CJ and S-IA. For BeatlesTUT, the change of using larger R is mainly a subdivision of the segments retrieved by a smaller R. For CJ and S-IA, however, the most direct outcome of using larger decomposition rank is redefined clusters. That is to say, new segments are retrieved which differ from those formed with a smaller decomposition rank. The new clusters can be rather chaotic with large within-cluster variance. This also indicates that the success of NMF-based segmentation algorithms is dependent on the predefined number of sections, with high homogeneity present in each cluster. Therefore, such segmentation methods can be more effective when the sections are described at a reasonably high hierarchy, describing phenomena such as the musical functions or the lead instruments.

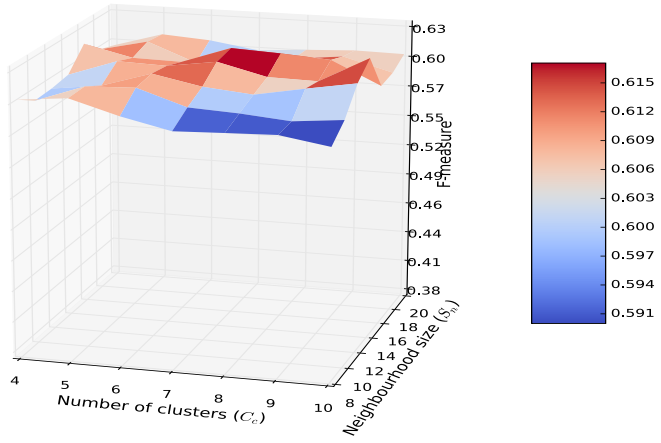
There are three main parameters involved in the segmentation process of CC: the

number of HMMs (C_h), the number of clusters (C_c) and the neighbourhood size (S_n) for the temporal constraints. A relatively large C_h (60 - 90) appears beneficial for all investigated datasets as observed by Levy and Sandler [LS08], due to the fact that segment types can not be comprehensively captured by individual HMM states. We use $C_h = 80$ in this thesis following Levy and Sandler [LS08] and discuss the effect of the other two parameters.

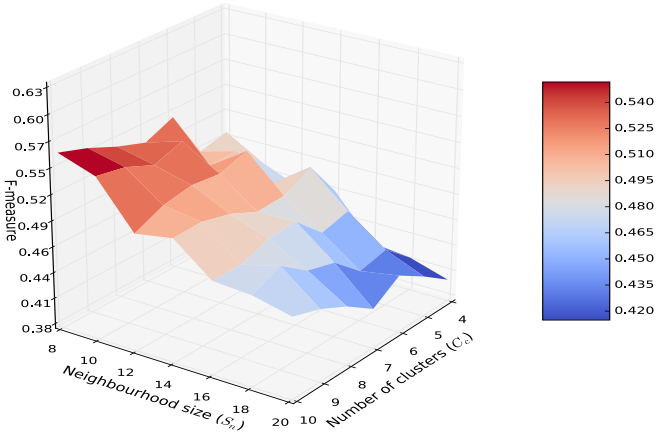
Figure 6.4 demonstrates the segmentation F-measures using CC with different S_n and C_h settings on investigated datasets. The three features, MFCCs, GF (see Section 5.7.4) and chromagram, show similar patterns and here we use GF as an example. In Levy and Sandler [LS08], S_n is set to 16 encoding the musical knowledge observed for Western pop music that a standard phrase length of is 16 beats. This is confirmed by our results that for BeatlesTUT, setting S_n to 16 or 8 which is an integer factor of 16, achieves the highest segmentation F-measures. A moderate C_c ranging from 6 - 8 appears the most beneficial. However, no significant difference is found between pairs of F-measures obtained by different settings of this parameter.

For S-IA, using a moderate C_c and a large S_n yields the optimal segmentation results. For datasets such as S-IA with relatively large within-dataset acoustical variations, constraining the clustering algorithm to operate on a large time scale can effectively reduce the number of false positives. For CJ, larger C_c and lower S_n are beneficial for an effective segmentation with the highest F-measure obtained under the extreme settings ($C_c = 20$, $S_n = 8$). No significant difference in segmentation F-measure is observed when S_n is the range of 1 (i.e., no time constraint is encoded) to 8. To keep increasing it however leads to a steady degradation of F-measure due to the substantially increasing false negatives. This shows that the time constraint does not introduce any solid benefit to the clustering for Jingju where there is no established beat-bar-phrase structure as observed for Western pop music [VF15].

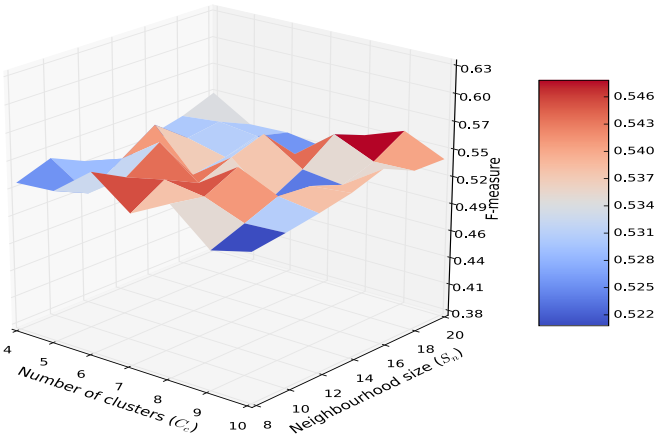
We can also notice from Figure 6.4 that, compared to C_c, S_n has a more profound effect on the segmentation. However, only C_c is made as a user-configurable parameter



(a) Dataset BeatlesTUT.



(b) Dataset CJ.



(c) Dataset S-IA.

Figure 6.4: Surface plot of the segmentation F-measures measured at 3 seconds using feature GF with Constrained Clustering (CC) under different settings of number of clusters (C_c) and neighbourhood size (S_n) on the three datasets.

in the software[1] published by Levy [LS08] while S_n is not.

The parameters employed in SERRÀ are mainly involved in two processes: the calculation of the *structure features* (SF) and the detection of segment boundaries from the feature representations. We found that the setting of parameters involved in the former. i.e., the embedding dimension used for emulating recent past and the number of nearest neighbours used for building the recurrence plot RP, do not vary across datasets significantly. The optimal settings reported in by Serrà et al. are proved appropriate for all investigated datasets [Ser+12], indicating that the *recurrence plot* approach can be relatively invariant to different musical genres.

We observe different optimal parameterisations for the segmentation system for different evaluation datasets. An intuitive application of the knowledge acquired is to build a database holding the metadata for the music signals and parameter configurations such that the user or the segmentation system can query to find the optimal parameterisation for specific algorithms.

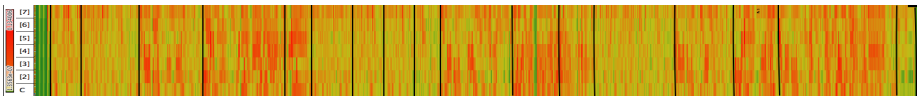
6.3.3 Repetition-based Segmentation Methods and Jingju

Repetition-based methods are developed to capture the chordal repetitions which are observed as important indicators as segment boundaries for Western pop music. Although the methods relying only on the repetition principle are excluded from the evaluation, we are interested in how they perform in the scenario of Jingju. Even for "through-composed" music, it is only the outline of the song which is the "through-composed"; there always exist repetitions with subtle changes of tonality and of theme [SW70].

To this end, we test a few recently published repetition-based methods [MND09; ME14b; WB10] on the CJ dataset. The first method finds repeated chroma sequences in a song, from which the structural segments are identified using a greedy algorithm introduced by Mauch and his colleagues [MND09]. As the second method, McFee and

[1] `http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html`

Ellis propose to detect segment boundaries by applying a spectral decomposition of the Laplacian of the recurrence plot of the chroma features [ME14b]. The third method uses a variant of the sparse convolutive NMF decomposition where the sparsity constraints are introduced to automatically identify the number of segment types and section lengths, presented by Weiss and Bello [WB10]. While these methods generally produce reasonably high precision rates, the average recall rate yielded by each under the optimised configuration is below 0.3 due to a large number of false negatives.



(a) Chroma feature and the annotated segments.

(b) Segmentation with method Mauch.

(c) Segmentation with method McFee.

(d) Segmentation with method Weiss.

Figure 6.5: Segmentation using chromagram with three repetition-based methods Mauch [MND09], McFee [ME14b] and Weiss [WB10] on Jingju song "Jin yu nu". The top pane shows the annotations (shown by black vertical lines) and the 7-BPO chromagram feature.

Figure 6.5 shows the segmentation results on a Jingju song "Jin yu nu" using the chromagram feature (BPO=7) with the three segmentation algorithms. Although harmonic patterns do present correlating with the annotated segments, such patterns hardly reflect any chordal repetition. The boundaries do get detected, however, pinpoint large melodic passages or arias that superimpose the similarity-level structural segments relatively accurately (see Section 3.4.2 for a survey of the different levels that structural segments are annotated on as well as their underlying principles). Therefore, we propose to use the repetition-based methods to detect the lead-instrument level music structure based on timbre variations. Besides reflecting the lack of chord structures in Jingju, this also demonstrates the inherent ambiguity of common segmentation algorithms in

identifying short-term patterns with regard to long-term musical structure. Several ways to address this limitation include to set temporal constraints to the algorithm to avoid it from converging only to large-scale units, or to combine multiple features in the segmentation to increase the variance in the feature representations.

6.3.4 Audio Features and Segmentation Methods

So far we have identified that segmentation algorithms perform differently and require different parameterisations for datasets consisting different music genres. Another question hereby arises: how do we effectively select audio features that can be interpreted by certain algorithms in the scenario of specific music genres?

Previous work has found that the best segmentation methods when using the chroma feature are repetition-based ones [PMK10]. However, this observation is made mainly for Western pop music. As shown in Table 6-A, the novelty-based methods turn out to be the most effective for the Jingju dataset regardless of the feature used. The average segmentation F-measure obtained using QN significantly surpasses that obtained using the second best performing method, both using the chromagram feature. With the absence of chord structure in Jingju, the chroma feature functions mainly to capture its low-level homogeneity in the vicinity, as illustrated in Figure 5.12.

The chroma feature and MFCCs work reliably for the investigated Western music datasets, confirming conclusions of previous studies [BMK06; PMK10]. Among the new features presented in this thesis, Gammatone features are designed to describe the music timbre as a replacement to MFCCs. The features were evaluated using one single segmentation method QN in Section 5.7.4 and showed improvements against MFCCs on the Jingju dataset CJ. This is also confirmed by different segmentation algorithms, as shown in Table 6-A.

SERRÀ obtains the highest segmentation F-measure for BeatlesTUT when the chroma feature is used, surpassing any other method using any feature significantly, confirmed

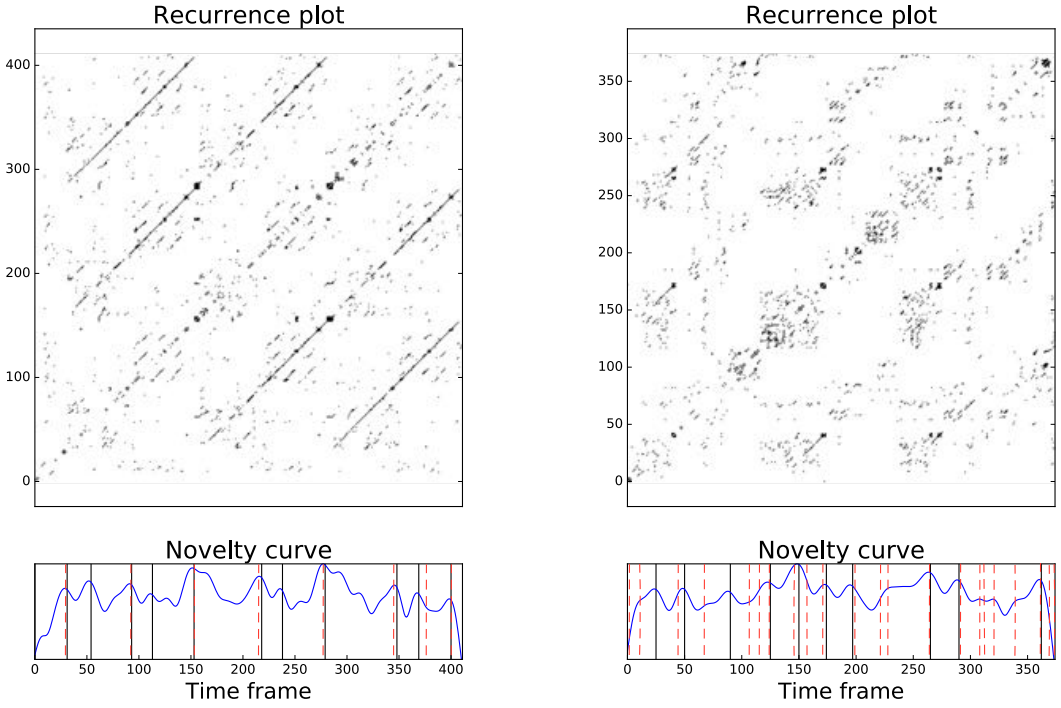


Figure 6.6: Segmentation process using method SERRÀ on Beatles song "Help" (left) and Jingju song "Jin yu nu" (right). The upper pane shows the recurrence plot. The bottom pane shows the novelty curve with the black solid line and red dashed line indicating estimated segment boundaries and annotations.

by the Wilcoxon signed-rank test. However, SERRÀ yields lower F-measure than any other algorithms tested on CJ, with significant difference observed when compared to the second worst performing algorithm for all the three features.

The first step of calculating the Structure features (SF) is to embed time delays in the feature vectors to account for the recent past. This however tends to introduce much chaotic behaviour into the feature representation for Jingju. Figure 6.6 shows the recurrence plot (RP), the derived novelty curve (NC) and the detected segment boundaries on a Beatles and a Jingju song using gammatone features GF (see Section 5.7.4). Here GF is used to replace chromagram because the latter appears to form even more chaos and sparsity in the RP for Jingju music. In contrast to the Beatles song, very little stripe structure is shown in the RP for the Jingju song. Peaks in the NC correspond

mainly to the few blocks in the RP. Although this algorithm is designed to be generic for boundary retrieval tasks [Ser+12], its advantage is more pronounced for music with discernible repetitions [TS16b]. When the music is less repetitive, it may produce low recall rate regardless of the feature used.

The best segmentation rates for S-IA are obtained by CC with the two novelty-based methods performing reasonably well. When annotations are made on a music similarity level, using repetition-based methods tend to lead to an under-segmentation with a large number of false negatives. This suggests that an effective evaluation of segmentation methods and features also relies on the reference annotations as the source knowledge fed into an MSS system.

6.3.5 Music Knowledge for Segmentation

For the structural analysis of Western pop music, the contextual knowledge is encoded in the process of feature extraction and SSM enhancement. A straightforward example is the synchronisation of the feature window into a beat-level under the condition of a reliable beat tracking. In Levy and Sandler [LS08], the neighbourhood size for clustering is set to 16 to approximate the standard musical phrase-length of 16 beats. This approach is originally introduced for Western pop music. However, in Jingju, the beat positions are more flexible and tend to be weakly correlated with the positions of segment boundaries, especially on a music similarity of novelty level (see Section 3.4.2). Meanwhile, accented beats may occasionally be lacking as introduced in Section 3.2.1 which can cause beat tracking algorithms to fail, hence leading to dubious window division for the feature extraction. The experiments carried out in this chapter and the previous chapter used features aggregated to respectively a beat level and a fixed window size (0.2 second). To understand the effect of the beat synchronisation operation on different music styles, we compare the pairwise segmentation F-measures using the same feature each time with the two aggregation approaches on the investigated datasets using the QN algorithm (Table 6-Ab and Table 5-F). While the beat synchronisation operation improves the

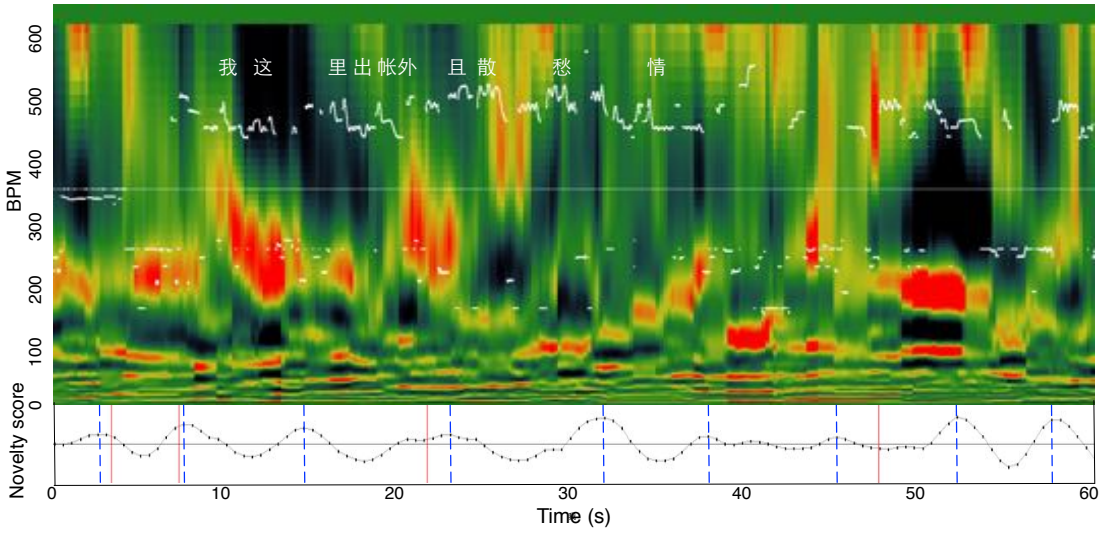


Figure 6.7: Melodic contours and tempogram for Jingju song "Ba wang bie ji". Structural annotations and boundaries detected with the MFCCs feature using method QN are shown respectively in the red solid lines and blue dotted lines in the lower pane.

segmentation significantly for BeatlesTUT using each investigated feature ($p < 0.01$, Wilcoxon signed-rank test), it shows no improvement when evaluated on CJ and S-IA. The variety of music styles in S-IA also invalidates the beat grouping rule occasionally. One solution to this could be to de-activate the beat synchronisation in the absence of accented or salient beat sequences when the confidence of the beat tracking or the frame energy drops below certain thresholds.

Figure 6.7 shows the tempogram calculated using the method described in Section 5.4.1 and the predominant melody [SG12] for a 60-second excerpt of Jingju music introduced in Figure 3.4. 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 calculated using the timbre feature not accompanied by prominent rhythmic or melodic changes is less likely to coincide a structural boundary. This indicates that fusing multiple musical aspects including *timbre*, *rhythm* and *melody* may lead to effective structural descriptions for Jingju, as also found by musicologists [Deu12]. Features of linguistic contours of the lyrics and the singing voice can

also be important factors to shape the structure of singing-driven music [ZRS14].

Instead of targeting a lyric-based segmentation for Jing music exclusively, this thesis is focused on understanding the music structure of different genres from an acoustic perspective using the generally applicable audio features. In future wok, we propose to use the low-level timbre features to derive intermediate structural descriptions and meanwhile, to rely on rhythmic and melodic modelling for verified segmentation in the scenario of Jingju. To summarise, the design of audio features and segmentation methods, the considerations of the underlying annotation principles and the context information related to the music genre should be accounted for together in an effective segmentation system.

6.4 Summary

This chapter investigated various segmentation algorithms covering the novelty, homogeneity and repetition structural hypotheses. By doing so, this chapter was devoted to present a critical evaluation of the audio features presented in the previous chapter and the state-of-the-art segmentation algorithms, as well as to discover the underlying relations between music types, audio features and segmentation algorithms in a music segmentation system. We analysed the effects of the signal processing parameters involved in the investigated algorithms with regard to individual music genres. The purposes of doing this include to identify the properties of the underlying music types as well as to derive general guidelines to devise a music structural analysis system.

The performances of specific segmentation algorithms are highly dependent on the features and the musical properties of the evaluation dataset. Although repetition-based segmentation methods are widely applied for Western pop music obtaining state-of-the-art results, they are proved less effective for Jingju music with its lack of chordal repetitions at a segment-level. We found that these methods are able to detect boundaries corresponding to large melodic passages. We therefore propose to use them to analyse

music structure at a higher level, such as the lead instrument level, to characterise the large-scale timbre or melodic variations. Homogeneity-based methods using unsupervised classification approaches work consistently well for all investigated datasets. Novelty-based methods effectively retrieve segment boundaries which repetition-based methods tend to overlook but are however sensitive to noise. This is supported by the fact that the algorithm Quadratic Novelty works well for the investigated Jingju dataset and S-IA but incurs a large number of false positives for BeatlesTUT whose annotation is charactering repetitions and music functions.

Using beat-synchronised features is very effective for the retrieval of segment boundaries for Western pop music whose musical phrases generally comprise fixed number of beats. Its advantage is however not notable for different music styles. Standard beat tracking algorithms may fail for Jingju due to its occasional lack of accented beats. For the structural analysis of Jingju, we propose to use novelty- or homogeneity-based segmentation methods to derive the basic structural description, and rely on higher-level melodic or rhythmic modelling for further pruning.

We found that different parameterisations for the segmentation algorithms are needed for different evaluation datasets or when different features are used. An intuitive application of the outcomes of this chapter is to build a knowledge-based segmentation system where the selection of the features, algorithms and their parameter configurations can be carried out automatically based on the source knowledge of the music signals.

Chapter 7

Conclusion

7.1 Summary

By investigating different audio features and algorithms in the scenario of different genres, this thesis presented an in-depth analysis into the extraction of note onsets and the song structure. We emphasised the use of audio features to convey semantic meanings to address the given tasks. The major contributions of this thesis can be summarised in three aspects: i) new research corpora are introduced to combat the Western bias in the existing MIR systems; ii) comprehensive onset features and algorithms are designed by fusing different knowledge sources; iii) novel audio features for structural description are presented and the underlying relations between features, segmentation algorithms and music genres are investigated; and iv) a critical evaluation is carried out for the signal processing methods and parameters used in our systems, where we identified their effects in the scenario of different music types.

7.1.1 Music Information Retrieval for Jingju

The majority of the existing datasets used to evaluate MIR tools consist of mainly Western pop music. As introduced in Section 3.2, Jingju is a music genre introduced to the MIR research corpora very recently. The Jingju system consists of very characteristic instrumentations, rhythmic and melodic patterns that collectively distinct it from the Western practice.

In Chapter 2, we presented two Jingju dataset. The first is designed for the audio onset detection (AOD) task and the percussion instrument recognition task consisting ensembles of Jingju percussion instruments. The second is a music structural segmentation (MSS) dataset with audio samples collected from commercial CDs and annotated by professional listeners. The inclusion of these two datasets largely facilitated the research carried out in this thesis combating the cultural bias. However, the relatively small dataset sizes may still confront us with limitations to apply them to evaluate different algorithms and applications in future work.

In the annotation of the MSS dataset, an analysis of the inter-annotator agreement (IAA) was carried out. We however realised that the discussion by the two annotators can produce different boundary decisions to those indicated by both individually. 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 slowly-evolving changes of acoustical properties, for example, the sustaining decaying of cymbal instruments and the fade-out effect of singing. The effect of such temporal proximity of an annotated boundary indicated by different annotators however can be cancelled out in the evaluation of boundary retrieval results given a sufficient acceptance window (3 seconds in this thesis).

7.1.2 Audio Onset Detection with Fusion

Chapter 4 investigated the AOD task relying different fusion strategies with new onset detection features and methods. We demonstrated that fusion of existing onset detection algorithms can introduce significantly improvement over the constituent baselines. Three fusion strategies including early fusion, linear fusion and decision fusion, were tested, among which we found the *linear fusion* is the most effective, which operates at a mid-level of the knowledge representation. The *linear fusion* of the Complex domain (CD) and SuperFlux (SF) onset detection methods, denoted $CDSF_L$ in this thesis, yielded the best detection results overall in a large-scale evaluation. The improvement of this method over its two baseline methods is however not always significant. This method was then used for the extraction of the tempogram and related rhythmic features in this thesis.

The onset detection algorithms were implemented as Vamp plugins and the parameter optimisation was carried out relying on the Vamp Plugin Ontology. By this work, we also highlighted the advantage of semantic web tools in the scenario of audio analysis. We also investigated the signal processing methods used in the detection process and their involved parameters. We found that parameter optimisation reliably improves the performance of existing MIR systems and that parameter configuration can be a critical factor for the success of an onset detection system.

7.1.3 Audio Features for Music Structural Segmentation

Chapter 5 presented novel audio features for music structural description. This chapter also introduced a harmonic-percussive source separation (HPSS) algorithm to facilitate feature extraction for the subsequent structural analysis.

The chroma features are originally developed for Western music with 12 pitch classes. In Section 5.3 we justified the use of 7-BPO chroma feature for Jingju. We also found

that with the absence of chord structure in Jingju, the chroma feature mainly captures its melodic homogeneity in the vicinity.

Rhythmic features are less commonly used for the music structural description than the timbre and chroma features. Section 5.4 presented novel high-level and mid-level rhythmic features extracted from the tempogram. Unlike standard rhythmic features that capture exclusively the rhythm or tempo information, the presented features incorporate also perceptual cues based on the musical observation that rhythmic components of different tempo salience can help setting structural sections apart. This feature set was proved to be an effective alternative to the commonly used MFCCs and chroma features in the evaluation.

HPSS introduced significant improvements to the segmentation for all investigated music categories and feature types. The maximum filter has introduced additional benefits with widened spectral trajectories combating the possible interference from vibratos or energy sways. In this process, the separation factor β is the most influential parameter. Higher β is needed for chromagram and MFCCs than the rhythmic feature RT to obtain their individual best performance. Because for the latter, complementary rhythmic information can be provided also by the harmonic instruments.

Features extracted from the Gammatonegram were applied effectively for MSS. In Section 5.5, we compared features extracted from the actual Gammatone filters and those extracted from the fast Gammatone method by weighting a standard STFT. No notable difference exists between the two sets of features in terms of segmentation rates. Using the fast method for feature extraction is proved at least not harmful compared to the accurate method in the segmentation context. The Gammatone featureset has proved itself an effective alternative to the commonly used MFCCs and chroma feature for music structural analysis, with its advantages more notable on vocal-driven music.

In Chapter 6 we evaluated several segmentation algorithms based on different structural principles: novelty, homogeneity and repetition. The chroma feature forms mainly

block structures in the self-similarity matrices for Jingju instead of the stripes as commonly observed for Western pop music. While previous work found that the most effective segmentation methods when using the chroma feature are repetition-based ones, the novelty and homogeneity-based methods turn out more effective for the Jingju dataset regardless of the feature used. The repetition-based methods typically detect boundaries corresponding to large sung passages. We hence propose to use them to retrieve the music structure at a higher-level such as the lead-instrumentation level. The investigated Gammatone features perform consistently when different segmentation methods are used.

7.1.4 A Critical Evaluation of Signal Processing Methods

The success of an MIR system largely depends on the success of the signal processing methods involved. This thesis carried out critical analyses for the signal processing methods used in the investigated AOD and MSS system.

In the evaluation of the onset detection algorithms using the two Western datasets and the Jingju percussion dataset JP, we found that the selection of onset detection algorithms and the configuration of the system parameters are music type specific. Configurations derived for Western music can be less effective for Jingju music. Despite categorised as a percussion dataset, the Jingju percussion dataset consists of cymbal and gong instruments with slow-evolving time-frequency characteristics. Therefore, when evaluated on this dataset, algorithms designed originally for soft onsets tend to be more effective than those designed for drum sounds. The low-pass filter and the median removal designed for noise reduction are proved generally beneficial for all music types and onset categories. The peak picking method relying on polynomial fitting is more effective for complex note types (percussive and non-percussive, pitched and non-pitched). This is because such onsets may be characterised by peaks with different shapes in the onset detection function hence can be captured by a peak picking method which assesses not only the amplitudes but also the shapes of the peaks. We also found that significant interactions

exist between different parameters.

A novelty-based structural segmentation algorithm was presented in Chapter 5 based on the peak picking mechanism used for onset detection investigated in Chapter 4. We hence highlighted that investigations into signal processing methods for music content analysis can be of general importance in the scenario of different MIR applications. We evaluated different MSS algorithms and their involved signal processing parameters. The analysis of the parameter configurations for different datasets also reflects the underlying genre-invariance of the algorithms. While the performance of some algorithms, such as the NMF-based ones, show high dependence on system configurations, the unsupervised methods and the novelty-based methods such as SERRÀ and Quadratic Novelty are proved more genre-invariant (see Section 6.3.2).

7.2 Future Perspectives

7.2.1 New Features for Jingju Content Description

By combining multiple feature descriptors or knowledge sources, features presented in thesis were characterising the property of *genre-invariance*. Investigated features are applied to detect the note onsets and the song structure of music signals of different genres. An intuitive future direction is, to develop new audio features to better characterise the musical properties for Jingju.

In Section 6.3.5, we discussed that the predominant melodies and the rhythmic patterns may effectively model the sectional transitions for Jingju songs. However, due to the heterophonic nature of this music style and the fact that the vocals and the background instruments always have overlapping frequency ranges, standard pitch tracking and melody tracking methods tend to fail. The occasional lack of metred beats in Jingju also sets pitfalls for common beat and tempo tracking algorithms, as discussed in Section 6.3.5. In future work, we propose to design melodic features for Jingju as well as to

further investigate into the presented tempogram features for a better content description for Jingju. We also intend to combine these aspects with timbre features to model the music structure.

7.2.2 Hierarchical Structure Analysis for Different Music Styles

This thesis presented an audio-based analysis of music structure for different genres featuring the note-level and the song-level. These two hierarchies were chosen because they are shared by the investigated genres. Therefore, they can provide a common ground for the evaluation of the investigated features and algorithms hence kick-start a cross-cultural analysis of the proposed topics. However, the *beat-bar-phrase* structure for Western pop music [Mad+04] and the *melodic couplet* structure for Jingju (see Section 3.2.1) are also indispensable for a comprehensive structural description for the individual music types. In future work, we intend to investigate the melodic couplet structure for Jingju and extend the level of abstractions of music structure in the analysis.

7.2.3 A Knowledge-based Music Structure Analysis System

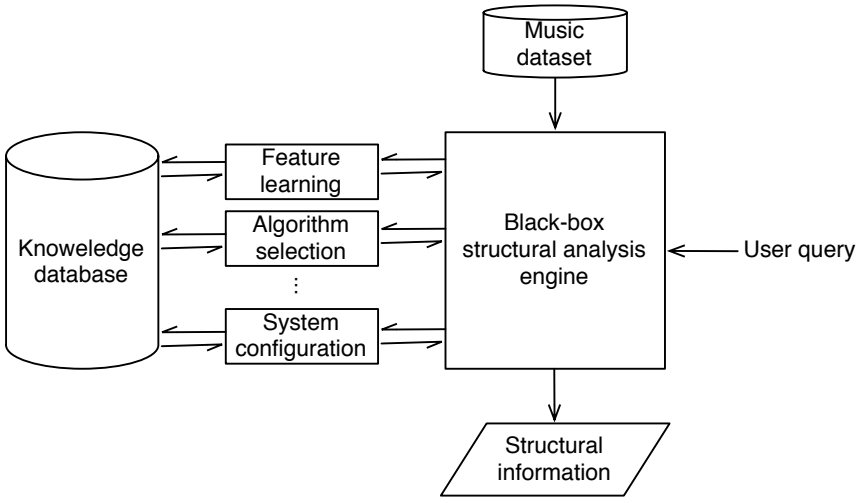


Figure 7.1: Framework of a knowledge-based system for music structural analysis.

This thesis has created a clear pathway towards an automatic analysis of music structure. One future direction is a *knowledge-based* system for music structural analysis. Figure 7.1 illustrates a proposed system which makes structural discoveries given user queries. There are two inputs for this system: the *music database* and the *user queries*. Users will be guided by the front-end interface of the system to indicate what music they want to analyse, for example which song or which playlist, and which hierarchy of the structural information they want to retrieve.

The queries will then be sent to the main analysis system to be reasoned. The results of such reasoning, is the semantic meanings or keywords of the queries which will then be passed to the *black-box analysis engine*. This semantic information will provide essential knowledge regarding the input source and the type of the output. Using these semantic keywords obtained, the analysis engine can query the *knowledge database* for relevant information to configure the analysis. The knowledge database holds contextual metadata related to the music signals, such as the type of the music, audio quality, algorithm specifications and system configurations. Therefore, this knowledge database will provide information to supervise the operation of the system, including the appropriate features and algorithms with optimised settings to select, for a given task.

Subsequently, the information obtained will be fed back to the main analysis engine to make informed structural discoveries. A handy way of realising the communications between different blocks within the system is to *ontologise* the underlying audio features, algorithms and parameter configurations, as did in our onset detection experiment with the exhausted parameter optimisation using the Vamp Plugin Ontology and Audio Features Ontology (see Section 4.3.2).

Compared to the music dataset which is exposed to the users, the knowledge source in the proposed system will be integrated rather passively, used by the main analysis stream only when a certain task has been invoked. In this thesis, the role of such knowledge source was played by the author, who explicitly extracted specific features and used specific algorithms incorporating domain knowledge. By doing this, we aim

to automate the analysis workflow therefore to provide a user-friendly and interactive package to present the research carried out in this thesis.

Appendix A

Onset Detection Databases Used in this Thesis

Three databases are used for the evaluation of onset detection algorithms presented in Chapter 4. These databases are denoted JPB, SB and JP, introduced in Section 3.3. Details of the individual tracks in these three datasets are given in the tables below. The acronyms *PP*, *NPP*, *PNP* and *CM* stand respectively for *pitched percussive*, *non-pitched percussive*, *pitched non-percussive* and *complex mixture*. It has to be noted that there is a 7-onset discrepancy (1058 instead of 1065) between the annotation used in this thesis and that in [Bel+05]. This difference is reported by the author of the dataset due to revisions of the annotations. These 7 onsets belong to the pitched-percussive (PP) class (482 instead of 489).

Track name	format	Length (s)	Onset type	Number of Onsets
arab60s	wav	60	PP	386
dido	wav	12	CM	71
fiona	wav	9	CM	45
Jaillet15	wav	5	PP	4
Jaillet17	wav	3	PP	9
Jaillet21	wav	8	PP	75
Jaillet27	wav	6	PP	10
Jaillet29	wav	3	PP	7
Jaillet34	wav	6	PP	18
Jaillet64	wav	2	NPP	31
Jaillet65	wav	3	CM	15
Jaillet66	wav	5	NPP	55
Jaillet67	wav	4	PP	18
Jaillet70	wav	4	CM	28
Jaillet73	wav	3	PP	10
Jaillet74	wav	1	PP	6
Jaillet75	wav	2	NPP	12
jaxx	wav	7	CM	51
metheny	wav	7	CM	44
PianoDebussy	wav	5	PP	27
violin	wav	13	PNP	97
wilco	wav	15	CM	89
tabla	wav	7	NPP	43

Table 1-A: Dataset JPB

Track name	Format	Length (s)	Onset type	Number of Onsets
JP001	wav	30	NPP	143
JP002	wav	30	NPP	71
JP003	wav	30	NPP	69
JP004	wav	30	NPP	78
JP005	wav	30	NPP	81
JP006	wav	30	NPP	81
JP007	wav	30	NPP	67
JP008	wav	30	NPP	63
JP009	wav	30	NPP	36
JP0010	wav	30	NPP	43

Table 1-B: Dataset JP

Track name	format	Length (s)	Onset type	Number of Onsets
$sb\_Albums - Ballroom\_Magic - 04(6.0 - 16.0)$	flac	30	PNP	21
$sb\_Albums - Cafe\_Paradiso - 05(4.3 - 14.3)$	flac	30	CM	54
$sb\_Albums - Chrisanne2 - 01(16.0 - 26.0)$	flac	30	CM	28
$sb\_Media - 103416(12.0 - 22.0)$	flac	30	NPP	59
$sb\_Media - 106003(0.2 - 10.2)$	flac	30	CM	73
$sb\_Albums - Chrisanne3 - 07(3.0 - 13.0)$	flac	30	PP	70
$sb\_Albums - Chrisanne3 - 02(12.0 - 22.0)$	flac	30	PP	36
$sb\_Media - 103302(15.0 - 25.0)$	flac	30	PP	46
$sb\_Media - 105215(12.0 - 22.0)$	flac	30	NPP	59
$sb\_Media - 105407(6.0 - 16.0)$	flac	30	CM	60
$sb\_Media - 105907(0.0 - 10.0)$	flac	30	CM	58
$sb\_Media - 105801(11.0 - 21.0)$	flac	30	CM	43
$sb\_Media - 103307(4.0 - 14.0)$	flac	30	CM	79
$sb\_Media - 104210(2.0 - 12.0)$	flac	30	PNP	44
$sb\_Media - 106103(4.0 - 14.0)$	flac	30	CM	79
$sb\_Albums - Fire - 13(15.0 - 25.0)$	flac	30	CM	28
$sb\_Media - 105810(5.0 - 15.0)$	flac	30	CM	64
$sb\_Media - 104111(5.0 - 15.0)$	flac	30	CM	64
$sb\_Albums - Chrisanne1 - 08(9.0 - 19.0)$	flac	30	CM	62
$sb\_Albums - I\_Like\_It2 - 01(13.1 - 23.1)$	flac	30	CM	77
$sb\_Albums - Latin\_Jam - 13(6.0 - 16.0)$	flac	30	NPP	56
$sb\_Albums - Cafe\_Paradiso - 07(13.1 - 23.1)$	flac	30	CM	65
$sb\_Albums - Step\_By\_Step - 09(2.0 - 12.0)$	flac	30	PNP	49
$sb\_Media - 106117(7.0 - 17.0)$	flac	30	CM	65
$sb\_Albums - Ballroom\_Classics4 - 01(3.0 - 13.0)$	flac	30	PNP	28
$sb\_Albums - Chrisanne1 - 01(3.0 - 13.0)$	flac	30	PNP	43
$sb\_Media - 100608(3.0 - 13.0)$	flac	30	CM	50
$sb\_Albums - Ballroom\_Magic - 09(4.0 - 14.0)$	flac	30	PNP	48
$sb\_Albums - Latin\_Jam3 - 02(3.0 - 13.0)$	flac	30	CM	69
$sb\_Albums - Ballroom\_Classics4 - 11(7.0 - 17.0)$	flac	30	CM	40

Table 1-C: Dataset SB

Appendix B

Annotators' Guide

B.1 Jingju Music – What You Will be Annotating

We first provide a brief introduction to the music culture. Jingju, also called Peking Opera or Beijing Opera, is a major branch of Chinese traditional music combining singing, dance and theatre art. Despite its rich musical heritage and the size of audience, little work has been done to analyse its music content from an MIR perspective.

The Jingju music system is characterised by three major elements: *melodic phrases* ("qiang"), *metrical patterns* ("banshi"), and *modes ("diaoshi") and modal systems ("shengqiang xitong")*. They are hierarchically related and collectively shape the music structure. 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 accordingly arranged to elaborate specific content of each passage of lyrics. The song lyrics are organised in a *couplet* structure which lays the basis of the music structural framework. A couplet is comprised of two *melodic phrases*, which are sung phrases with tendencies toward certain melodic patterns and are considered the smallest meaningful musical units. A passage of melodic phrases expressing specific music ideas or motifs can be grouped into a *melodic section* ("qiangjie") which can play a rather integrate role in the

overall musical form. The *metrical pattern* is the most expressive characteristic element of Jingju. The transitions of alternating metric patterns in a Jingju song may indicate boundaries between sectional units. There are fixed types of metrical patterns, each is 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 or are free of them. While the first have specific tempi, the latter have no rhythmic regulation. Jingju songs also have instrumental connectives ("Guo men", meaning "through the door") 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 melodic couplets, hence are integral to Jingju structure. The identification of the music structure will be centred around the identification of these events in a temporally ordered pattern.

B.2 How to Annalyse the Music

The term musical form (or musical architecture) refers to the overall structure or plan of a piece of music and it describes the layout of a composition as divided into sections. Your task will be to divide the music you hear into segments by giving it boundaries.

Analysing the structure of a piece of music is no easy task. There might not be a "correct" answer to the questions such as what structure does this piece of music hold or is there a new structural segment arising. It might also be difficult to ping down an exact timestamp as a structural change point. Maybe you familiar with the "chorus-verse" or "ABBA" structure in Western pop music. You may have already had a basic understanding of music background of Jingju. However, you are not required to be able to identify all the musical elements or functions in a song, although your familiarity of it will be recorded in a post-annotation questionnaire. In Section B.1, you have read some introductions of how its structure is formed. By giving you this description, we expect you to have a basic understanding of the fact that the music structure in Chinese Jingju

music is by no means the same with the Western pop music. Don't panic if you have no idea what those melodic phrases or instrumental connectives sound like. What you are invited to do, basically, is to partition each piece into several segments fulfilling different music roles on a *music similarity* level as perceived. You will be paying attention to the prominent changes in music elements such as rhythm, melody, harmony or timbre or sudden similarity breakdowns of these musical phenomena. You will be locating them by assigning boundaries at such locations.

You are asked firstly to listen to each piece from start to end. This will give you a general idea of how the music is like, how "steady" or "fluctuant" it can be throughout the piece. Then you can start annotating the music, following instructions given soon in Section B.3. Feel free to pause and navigate through the music and even to speed up and down the playing anytime. When you finish annotating the whole piece and have created some annotations this time, please listen to the music at least one more times. When you finish the work, save and export your annotation files. We will give some instructions on how you can check your annotations by listening to the songs repeatedly.

You are asked to be CONSISTENT with the same idea of what defines a structural segments and how prominent the changes should be to make the changing point a segment boundary. You will need to listen to each piece multiple times such that each time you can verify your previous annotation decisions and introduce changes whenever applicable. We expect this to help you fix inconsistent decisions you may have made each time. Each time you make some changes to the annotations you have made in the last time, i.e., to delete a boundary or to add a new one, or to change the position of a marked boundary to somewhere over 0.5s away, you are asked to listen to the whole piece again. In other words, the last time you listen to it, before save your annotations and proceed to the next piece, you don't feel dubious of the any boundary you indicated and nothing has changed from the last round of listening and annotating. However, if it reaches the 5th time you listen to the whole song for annotation check purpose and you haven't made substantial changes, please stop it there and save your work without listening to it over

and over again. But if do feel that you are still uncertain with many boundary decisions you have made up until now, let the instructors know. You will need to inform us how many times you have listened to each piece for the annotation. In the next section, we will provide you some instructions regarding the software you will use for the task and involved operations.

B.3 Annotation Procedure

The software used for annotation is Sonic Visualiser: `http://sonicvisualiser.org`. You have already received some instructions on how to use the software from us. You can also open the "Help" tab in the GUI. After finding yourself familiar with the software, you can follow the following steps to start your work.

1. Load the song into Sonic Visualiser, listen to it from start to end. You can press the "space" key or click the play/pause button on the GUI to play or to pause.

2. Go back to the beginning and play the music again to start your annotation. Try to anticipate where boundaries are so that you can press the key exactly when they occur; if you know you missed it by a small amount, pause the song and adjust the boundary.

3. Check your work and verify the boundaries you have indicated for a few times following the instructions given in Section B.2. Here the basic idea is that you should be consistence with the annotation decisions throughout the piece. You can add new boundaries or erase or change the positions of previously noted boundaries by click on the "eraser" button then click on the boundary you want to remove.

4. When you finish your work for a song, select "Export annotation layer" from the "File" tab in the menu and save your work in ".csv" format. You can proceed to the next song and repeat these steps for the annotation.

OK that's almost everything you need to know. Now you can grab your headphone and get started!

Appendix C

Full list of Onset Detection Results and Parameter Configurations

C.1 Best Performing Configurations Grouped by Onset Type (Part I)

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
BERSF-L	0.8495	10.0000	2.4000	off	0.3000	n/a
CDSF-L	0.8454	30.0000	2.4000	off	0.0000	n/a
BERSF-D	0.8448	40.0000	2.1500	off	n/a	0.0500
BERSF-E	0.8406	20.0000	2.4000	off	n/a	n/a
CDSF-D	0.8168	60.0000	0.9000	off	n/a	0.0500
SF	0.7913	40.0000	2.4000	on	n/a	n/a
CDBER-L	0.7877	10.0000	2.4000	on	0.5000	n/a
BERSD-L	0.7873	10.0000	2.4000	on	0.5000	n/a
SDBER-E	0.7752	10.0000	2.4000	off	n/a	n/a
PDBER-L	0.7734	10.0000	2.4000	off	0.9000	n/a
HFCBER-E	0.7721	10.0000	2.4000	on	n/a	n/a
BER	0.7721	10.0000	2.4000	on	n/a	n/a
BERSD-E	0.7707	40.0000	2.4000	on	n/a	n/a
HFCCD-D	0.7701	70.0000	1.4000	on	n/a	0.0300
BERCD-E	0.7683	40.0000	2.4000	off	n/a	n/a
CDBER-E	0.7679	10.0000	2.4000	off	n/a	n/a
HFCSD-L	0.7658	50.0000	1.1500	on	0.7000	n/a
HFCBER-L	0.7657	30.0000	1.6500	off	0.6000	n/a
CDSD-L	0.7649	30.0000	1.4000	on	0.3000	n/a
SDPD-L	0.7647	50.0000	1.9000	on	0.9000	n/a
SD	0.7640	40.0000	1.1500	on	n/a	n/a
HFCCD-L	0.7613	50.0000	1.1500	on	0.9000	n/a
HFC	0.7610	50.0000	1.4000	on	n/a	n/a
CDSD-D	0.7610	70.0000	0.9000	on	n/a	0.0450
CDPD-L	0.7609	30.0000	2.4000	on	0.9000	n/a
HFCSD-E	0.7605	40.0000	1.4000	on	n/a	n/a
BERSD-D	0.7605	40.0000	1.4000	on	n/a	0.0500
HFCSD-D	0.7595	60.0000	0.9000	on	n/a	0.0500
CDBER-D	0.7587	60.0000	1.6500	on	n/a	0.0500
CD	0.7566	50.0000	2.4000	on	n/a	n/a
HFCBER-D	0.7526	50.0000	1.9000	on	n/a	0.0500
HFCCD-E	0.7513	50.0000	2.4000	on	n/a	n/a
HFCPD-L	0.7438	50.0000	0.9000	on	0.9000	n/a
SDPD-D	0.6940	80.0000	2.1500	on	n/a	0.0500
CDPD-D	0.6907	80.0000	0.9000	on	n/a	0.0500
HFCPD-D	0.6805	80.0000	2.1500	on	n/a	0.0500
PD	0.6212	80.0000	1.6500	on	n/a	n/a
HFCPD-E	0.6212	80.0000	1.6500	on	n/a	n/a
RPD	0.6080	80.0000	1.9000	on	n/a	n/a
PDBER-D	0.5239	90.0000	2.4000	off	n/a	0.0500

Table 3-A: Segmentation F-measures of investigated detectors for onset type complex mixture (CM)

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
CDSF-L	0.9414	30.0000	2.4000	off	0.2000	n/a
SF	0.9412	20.0000	2.4000	off	n/a	n/a
BERSF-L	0.9412	20.0000	2.4000	off	0.0000	n/a
CDSF-D	0.9377	50.0000	2.1500	off	n/a	0.0500
BERSF-E	0.9308	30.0000	2.4000	off	n/a	n/a
HFCCD-L	0.9307	50.0000	2.4000	off	0.3000	n/a
CDBER-L	0.9279	10.0000	2.4000	off	0.2000	n/a
BERSF-D	0.9244	50.0000	2.4000	off	n/a	0.0500
CDSD-L	0.9243	30.0000	2.4000	off	0.8000	n/a
CD	0.9209	10.0000	2.4000	off	n/a	n/a
CDPD-L	0.9201	10.0000	2.4000	off	1.0000	n/a
HFCSD-L	0.9152	40.0000	1.1500	off	0.5000	n/a
CDSD-D	0.9128	50.0000	1.1500	off	n/a	0.0500
BERSD-L	0.9117	20.0000	2.4000	off	0.2000	n/a
CDBER-D	0.9100	50.0000	1.4000	off	n/a	0.0500
HFCCD-D	0.9084	70.0000	1.1500	off	n/a	0.0500
HFCPD-L	0.9073	40.0000	2.4000	on	0.9000	n/a
SDPD-L	0.9049	10.0000	2.4000	off	0.9000	n/a
SD	0.9018	10.0000	1.4000	off	n/a	n/a
HFCSD-E	0.8996	40.0000	2.4000	off	n/a	n/a
HFCCD-E	0.8987	40.0000	1.1500	off	n/a	n/a
HFC	0.8980	40.0000	2.4000	off	n/a	n/a
HFCBER-L	0.8973	50.0000	2.4000	off	0.9000	n/a
BERSD-D	0.8929	60.0000	1.1500	off	n/a	0.0500
HFCSD-D	0.8885	60.0000	2.4000	off	n/a	0.0500
HFCBER-D	0.8729	40.0000	0.9000	off	n/a	0.0500
CDBER-E	0.8666	10.0000	2.4000	off	n/a	n/a
HFCBER-E	0.8617	10.0000	2.4000	off	n/a	n/a
BER	0.8617	10.0000	2.4000	off	n/a	n/a
BERCD-E	0.8607	50.0000	1.9000	off	n/a	n/a
PDBER-L	0.8576	10.0000	2.4000	off	1.0000	n/a
BERSD-E	0.8538	50.0000	2.4000	off	n/a	n/a
SDBER-E	0.8486	10.0000	2.4000	off	n/a	n/a
SDPD-D	0.8449	80.0000	1.4000	on	n/a	0.0500
CDPD-D	0.8370	80.0000	2.1500	off	n/a	0.0500
HFCPD-D	0.8309	80.0000	1.6500	on	n/a	0.0500
RPD	0.7709	70.0000	2.4000	on	n/a	n/a
PD	0.7431	70.0000	2.4000	on	n/a	n/a
HFCPD-E	0.7431	70.0000	2.4000	on	n/a	n/a
PDBER-D	0.5682	90.0000	2.4000	off	n/a	0.0500

Table 3-B: Segmentation F-measures of investigated detectors for onset type pitched percussive (PP)

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
CDSF-L	0.7604	10.0000	2.4000	off	0.2000	n/a
CDBER-L	0.7418	10.0000	1.6500	off	0.3000	n/a
CDSF-D	0.7359	10.0000	2.4000	off	n/a	0.0500
BERSF-L	0.7319	10.0000	1.6500	off	0.1000	n/a
SF	0.7296	10.0000	1.6500	off	n/a	n/a
BERSF-D	0.7190	20.0000	2.4000	off	n/a	0.0450
HFCCD-L	0.7180	40.0000	1.4000	off	0.6000	n/a
BERSF-E	0.7173	10.0000	1.6500	off	n/a	n/a
CD	0.7154	10.0000	2.4000	off	n/a	n/a
BERSD-L	0.7084	10.0000	1.9000	off	0.6000	n/a
CDBER-D	0.7077	50.0000	1.1500	off	n/a	0.0500
CDSD-L	0.7071	10.0000	1.4000	off	0.8000	n/a
CDPD-L	0.7036	10.0000	2.4000	off	1.0000	n/a
HFCSD-L	0.6844	60.0000	1.6500	off	0.4000	n/a
CDSD-D	0.6775	60.0000	2.4000	off	n/a	0.0500
PDBER-L	0.6717	10.0000	2.4000	off	1.0000	n/a
HFCBER-E	0.6717	10.0000	2.4000	off	n/a	n/a
BER	0.6717	10.0000	2.4000	off	n/a	n/a
HFCBER-L	0.6567	30.0000	2.4000	on	0.5000	n/a
SDPD-L	0.6543	20.0000	1.4000	off	0.9000	n/a
HFCPD-L	0.6477	70.0000	0.9000	on	0.9000	n/a
SD	0.6474	40.0000	1.9000	off	n/a	n/a
HFC	0.6432	60.0000	0.6500	off	n/a	n/a
HFCCD-D	0.6421	60.0000	1.6500	off	n/a	0.0500
HFCCD-E	0.6418	70.0000	2.4000	off	n/a	n/a
CDPD-D	0.6402	80.0000	2.4000	off	n/a	0.0500
BERSD-D	0.6396	60.0000	2.1500	off	n/a	0.0500
HFCSD-E	0.6348	70.0000	1.1500	off	n/a	n/a
SDBER-E	0.6282	30.0000	2.1500	off	n/a	n/a
SDPD-D	0.6248	80.0000	2.4000	off	n/a	0.0500
HFCSD-D	0.6246	80.0000	1.1500	off	n/a	0.0500
BERCD-E	0.6176	40.0000	2.4000	off	n/a	n/a
CDBER-E	0.6151	10.0000	2.1500	on	n/a	n/a
BERSD-E	0.6126	40.0000	2.4000	off	n/a	n/a
HFCBER-D	0.6052	80.0000	1.9000	on	n/a	0.0500
PD	0.5975	80.0000	2.4000	on	n/a	n/a
HFCPD-E	0.5975	80.0000	2.4000	on	n/a	n/a
HFCPD-D	0.5973	90.0000	2.4000	on	n/a	0.0500
RPD	0.5730	70.0000	2.4000	on	n/a	n/a
PDBER-D	0.4915	90.0000	2.1500	off	n/a	0.0500

Table 3-C: Segmentation F-measures of investigated detectors for onset type pitched non-percussive (PNP)

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
CDSF-L	0.9264	50.0000	2.4000	off	0.2000	n/a
SF	0.9245	50.0000	2.4000	off	n/a	n/a
BERSF-L	0.9245	50.0000	2.4000	off	0.0000	n/a
BERSF-D	0.9194	40.0000	2.4000	off	n/a	0.0500
BERSF-E	0.9174	50.0000	2.4000	off	n/a	n/a
CDSF-D	0.9023	70.0000	2.4000	off	n/a	0.0500
CDBER-L	0.8967	30.0000	2.4000	off	0.2000	n/a
BERSD-L	0.8937	30.0000	2.1500	off	0.2000	n/a
HFCBER-L	0.8906	20.0000	1.9000	off	0.8000	n/a
BERCD-E	0.8885	20.0000	2.4000	on	n/a	n/a
BERSD-E	0.8884	40.0000	2.4000	off	n/a	n/a
CDPD-L	0.8864	50.0000	2.4000	off	0.9000	n/a
CDBER-D	0.8863	60.0000	1.1500	off	n/a	0.0500
SDPD-L	0.8848	20.0000	1.9000	off	0.8000	n/a
CDSD-L	0.8848	40.0000	2.4000	off	0.9000	n/a
HFCCD-L	0.8839	40.0000	2.4000	off	0.1000	n/a
BERSD-D	0.8827	60.0000	1.1500	off	n/a	0.0500
CD	0.8821	40.0000	2.4000	off	n/a	n/a
HFCSD-L	0.8809	50.0000	2.1500	off	0.7000	n/a
CDSD-D	0.8799	40.0000	0.9000	off	n/a	0.0500
CDBER-E	0.8794	10.0000	2.4000	off	n/a	n/a
HFCCD-E	0.8793	30.0000	2.1500	off	n/a	n/a
HFCPD-L	0.8789	40.0000	2.4000	on	0.9000	n/a
HFCBER-D	0.8764	60.0000	1.1500	off	n/a	0.0500
HFCSD-E	0.8758	40.0000	1.6500	off	n/a	n/a
SD	0.8752	40.0000	1.9000	off	n/a	n/a
HFCCD-D	0.8739	50.0000	2.4000	off	n/a	0.0300
HFC	0.8735	40.0000	1.4000	off	n/a	n/a
HFCSD-D	0.8677	50.0000	1.4000	off	n/a	0.0500
PDBER-L	0.8598	10.0000	2.4000	on	0.9000	n/a
HFCBER-E	0.8560	10.0000	2.4000	off	n/a	n/a
BER	0.8560	10.0000	2.4000	off	n/a	n/a
SDBER-E	0.8557	10.0000	2.4000	off	n/a	n/a
SDPD-D	0.7696	90.0000	1.1500	on	n/a	0.0500
CDPD-D	0.7665	90.0000	1.6500	on	n/a	0.0500
HFCPD-D	0.7410	80.0000	2.4000	on	n/a	0.0500
RPD	0.5675	60.0000	0.9000	on	n/a	n/a
PD	0.5228	70.0000	1.1500	on	n/a	n/a
HFCPD-E	0.5228	70.0000	1.1500	on	n/a	n/a
PDBER-D	0.4042	90.0000	0.6500	on	n/a	0.0500

Table 3-D: Segmentation F-measures of investigated detectors for onset type non-pitched percussive (NPP)

C.2 Best Performing Configurations Grouped by Onset Type (Part II)

Method	F-measure	sens	δ	f_c	LP	PF	MF	ψ
BERSF-L	0.8589	40.0000	0.5000	0.5000	on	on	on	8.0000
SF	0.8553	40.0000	0.5000	0.5000	on	on	on	9.0000
CDSF-L	0.8524	20.0000	0.2500	0.5000	on	on	on	7.0000
BERSF-D	0.8487	60.0000	0.2500	0.3500	on	on	on	8.0000
BERSF-E	0.8425	20.0000	0.0000	0.3875	on	on	on	9.0000
CDBER-L	0.8366	70.0000	0.5000	0.4250	on	on	on	7.0000
BERSD-L	0.8346	60.0000	0.5000	0.3875	on	on	on	6.0000
PDBER-L	0.8344	10.0000	0.2500	0.3500	on	on	on	9.0000
HFCBER-E	0.8344	10.0000	0.2500	0.3500	on	on	on	9.0000
BER	0.8344	10.0000	0.2500	0.3500	on	on	on	9.0000
SDBER-E	0.8316	30.0000	0.5000	0.5000	on	on	on	7.0000
CDSF-D	0.8239	50.0000	0.0000	0.4625	on	on	on	5.0000
CDBER-E	0.8203	20.0000	0.0000	0.3125	on	on	on	9.0000
BERSD-E	0.8151	40.0000	0.0000	0.3500	on	on	on	5.0000
BERCD-E	0.8149	40.0000	0.2500	0.3875	on	on	on	6.0000
HFCBER-L	0.8136	80.0000	0.5000	0.4250	on	on	on	6.0000
HFCCD-L	0.8074	60.0000	0.5000	0.5000	on	on	on	7.0000
CDSD-D	0.8051	10.0000	-E.2500	0.3875	on	on	on	8.0000
CDSD-L	0.8050	10.0000	0.0000	0.4250	on	on	on	7.0000
HFCSD-L	0.8049	70.0000	0.5000	0.5000	on	on	on	6.0000
BERSD-D	0.8035	90.0000	0.2500	0.4250	on	on	on	6.0000
CDBER-D	0.8018	20.0000	0.0000	0.4625	on	on	on	6.0000
CD	0.8007	10.0000	0.0000	0.5000	on	on	on	6.0000
HFCPD-L	0.7997	80.0000	0.2500	0.4625	on	on	on	6.0000
CDPD-L	0.7995	50.0000	0.2500	0.5000	on	on	on	6.0000
HFCSD-D	0.7950	90.0000	0.2500	0.3500	on	on	on	9.0000
HFCSD-E	0.7938	80.0000	0.2500	0.4625	on	on	on	7.0000
HFC	0.7937	100.0000	0.5000	0.5000	on	on	on	6.0000
HFCCD-D	0.7921	70.0000	0.0000	0.3125	on	on	on	7.0000
SDPD-L	0.7916	20.0000	0.0000	0.3500	on	on	on	8.0000
SD	0.7913	70.0000	0.5000	0.3875	on	on	on	9.0000
HFCCD-E	0.7886	90.0000	0.2500	0.2750	on	on	on	8.0000
HFCBER-D	0.7885	20.0000	-E.2500	0.3875	on	on	on	5.0000
CDPD-D	0.7417	90.0000	-E.2500	0.2750	on	off	on	6.0000
SDPD-D	0.7408	90.0000	-E.2500	0.3125	on	off	on	6.0000
HFCPD-D	0.7265	90.0000	-E.7500	0.3125	on	off	on	8.0000
RPD	0.6836	90.0000	0.0000	0.2375	on	off	on	9.0000
PDBER-D	0.6737	80.0000	-E.2500	0.4250	on	off	on	6.0000
PD	0.6723	80.0000	1.0000	0.2000	on	off	off	9.0000
HFCPD-E	0.6723	80.0000	1.0000	0.2000	on	off	off	9.0000

Table 3-E: Segmentation F-measures of investigated detectors for onset type complex mixture (CM)

Method	**F-measure**	**sens**	δ	f_c	**LP**	**PF**	**MF**	ψ
CDSF-L	0.9676	70.0000	0.5000	0.5000	on	on	on	8.0000
SF	0.9664	60.0000	0.5000	0.5000	on	on	on	8.0000
CDSF-D	0.9531	100.0000	0.5000	0.4625	on	on	on	5.0000
BERSF-E	0.9495	40.0000	0.2500	0.5000	on	on	on	7.0000
HFCCD-L	0.9462	90.0000	0.7500	0.5000	on	on	on	6.0000
BERSF-L	0.9439	30.0000	0.5000	0.5000	on	on	on	7.0000
BERSF-D	0.9405	90.0000	-L.0000	0.3875	on	off	on	8.0000
CDPD-L	0.9400	50.0000	0.5000	0.5000	on	on	on	5.0000
CD	0.9391	50.0000	0.5000	0.5000	on	on	on	5.0000
CDSD-L	0.9351	100.0000	1.0000	0.5000	on	on	on	6.0000
CDBER-L	0.9335	80.0000	-L.0000	0.4250	on	off	on	8.0000
CDBER-D	0.9277	90.0000	-E.2500	0.4250	on	off	on	9.0000
HFCSD-L	0.9206	70.0000	0.2500	0.3500	on	on	on	9.0000
CDSD-D	0.9193	90.0000	0.2500	0.3500	on	on	on	9.0000
HFCCD-D	0.9167	60.0000	0.0000	0.4250	on	on	on	8.0000
HFCPD-L	0.9145	90.0000	0.5000	0.4250	on	on	on	7.0000
SDPD-L	0.9129	10.0000	0.2500	0.4625	on	on	on	6.0000
HFCSD-E	0.9121	70.0000	0.5000	0.4250	on	on	on	9.0000
BERSD-L	0.9106	30.0000	0.5000	0.4250	on	on	on	9.0000
HFCCD-E	0.9103	80.0000	1.0000	0.5000	off	on	on	5.0000
SD	0.9089	60.0000	0.2500	0.3500	on	on	on	9.0000
HFC	0.9067	70.0000	0.2500	0.4250	on	on	on	8.0000
HFCBER-L	0.8931	80.0000	0.2500	0.3500	on	on	on	8.0000
BERSD-D	0.8880	60.0000	0.0000	0.3875	on	on	on	9.0000
BERCD-E	0.8880	100.0000	0.7500	0.3500	on	on	on	7.0000
HFCBER-E	0.8849	70.0000	-L.0000	0.3500	on	off	on	9.0000
BER	0.8849	70.0000	-L.0000	0.3500	on	off	on	9.0000
PDBER-L	0.8831	80.0000	-E.2500	0.3125	on	off	on	7.0000
CDBER-E	0.8792	70.0000	1.0000	0.2375	on	off	off	9.0000
HFCBER-D	0.8785	80.0000	0.2500	0.3875	on	on	on	8.0000
HFCSD-D	0.8755	90.0000	-E.5000	0.3875	on	off	on	9.0000
BERSD-E	0.8598	50.0000	0.0000	0.2750	on	on	on	5.0000
SDBER-E	0.8580	20.0000	0.0000	0.2750	on	on	on	9.0000
SDPD-D	0.8475	50.0000	-E.2500	0.2750	on	on	on	9.0000
HFCPD-D	0.8449	90.0000	-E.2500	0.3875	on	off	on	7.0000
CDPD-D	0.8411	90.0000	-E.2500	0.2750	on	off	on	5.0000
RPD	0.8024	80.0000	-E.2500	0.2750	on	off	on	9.0000
PD	0.7867	60.0000	-L.0000	0.3500	on	off	on	8.0000
HFCPD-E	0.7867	60.0000	-L.0000	0.3500	on	off	on	8.0000
PDBER-D	0.7246	80.0000	-E.2500	0.3500	on	off	on	8.0000

Table 3-F: Segmentation F-measures of investigated detectors for onset type pitched percussive (PP)

Method	**F-measure**	**sens**	δ	f_c	**LP**	**PF**	**MF**	ψ
CDSF-L	0.7773	90.0000	0.2500	0.2000	on	on	on	8.0000
BERSF-L	0.7602	10.0000	0.0000	0.2375	on	on	on	6.0000
SF	0.7583	100.0000	0.5000	0.2000	on	on	on	7.0000
CDBER-L	0.7580	40.0000	0.0000	0.2375	on	on	on	9.0000
BERSF-E	0.7577	30.0000	0.0000	0.2375	on	on	on	5.0000
BERSD-L	0.7399	10.0000	-E.2500	0.2375	on	on	on	9.0000
CDSF-D	0.7390	20.0000	-E.2500	0.2750	on	on	on	6.0000
BERCD-E	0.7296	100.0000	0.7500	0.3125	on	on	on	9.0000
BERSF-D	0.7273	40.0000	0.0000	0.2750	on	on	on	8.0000
CD	0.7227	80.0000	0.5000	0.3125	on	on	on	9.0000
HFCCD-L	0.7225	60.0000	0.0000	0.2375	on	on	on	6.0000
CDPD-L	0.7220	10.0000	0.2500	0.3125	on	on	on	9.0000
CDBER-D	0.7128	40.0000	0.0000	0.4250	on	on	on	8.0000
CDSD-L	0.7099	40.0000	0.0000	0.2375	on	on	on	7.0000
PDBER-L	0.6947	50.0000	0.2500	0.2375	on	on	on	6.0000
HFCBER-E	0.6947	50.0000	0.2500	0.2375	on	on	on	6.0000
BER	0.6947	50.0000	0.2500	0.2375	on	on	on	6.0000
HFCCD-E	0.6816	80.0000	0.0000	0.2375	on	on	on	8.0000
HFCSD-L	0.6797	20.0000	-E.2500	0.3500	on	on	on	7.0000
CDSD-D	0.6778	90.0000	0.2500	0.2750	on	on	on	8.0000
BERSD-E	0.6760	100.0000	0.5000	0.3125	on	on	on	6.0000
HFCPD-L	0.6657	70.0000	0.0000	0.2750	on	on	on	7.0000
SDPD-L	0.6646	80.0000	-L.0000	0.2000	on	off	on	7.0000
HFCBER-L	0.6587	90.0000	0.2500	0.3125	on	on	on	7.0000
SD	0.6584	40.0000	0.2500	0.5000	on	on	on	6.0000
HFC	0.6510	80.0000	0.0000	0.2375	on	on	on	7.0000
HFCSD-E	0.6487	10.0000	-E.5000	0.2750	on	on	on	7.0000
HFCCD-D	0.6477	40.0000	-E.2500	0.2375	on	on	on	7.0000
SDBER-E	0.6471	30.0000	0.0000	0.2750	on	on	on	6.0000
CDBER-E	0.6399	10.0000	-E.2500	0.2375	on	on	on	8.0000
PD	0.6360	50.0000	1.0000	0.2375	on	off	off	9.0000
HFCPD-E	0.6360	50.0000	1.0000	0.2375	on	off	off	9.0000
RPD	0.6304	70.0000	1.0000	0.2000	on	off	off	9.0000
CDPD-D	0.6218	90.0000	0.0000	0.3125	on	on	on	7.0000
HFCBER-D	0.6205	80.0000	0.0000	0.2375	on	on	on	6.0000
HFCSD-D	0.6184	90.0000	0.0000	0.3500	on	on	on	7.0000
BERSD-D	0.6143	60.0000	1.0000	0.2375	on	on	off	9.0000
SDPD-D	0.6049	20.0000	-E.5000	0.2750	on	on	on	8.0000
HFCPD-D	0.5981	50.0000	-E.2500	0.2750	on	on	on	8.0000
PDBER-D	0.5581	80.0000	-E.5000	0.3500	on	off	on	8.0000

Table 3-G: Segmentation F-measures of investigated detectors for onset type pitched non-percussive (PNP)

Method	**F-measure**	**sens**	δ	f_c	**LP**	**PF**	**MF**	ψ
SF	0.9456	80.0000	0.2500	0.5000	on	on	on	6.0000
CDSF-L	0.9433	90.0000	0.7500	0.5000	off	on	on	5.0000
BERSF-E	0.9358	80.0000	0.2500	0.5000	on	on	on	7.0000
BERSF-D	0.9337	80.0000	0.2500	0.5000	on	on	on	6.0000
BERSF-L	0.9207	50.0000	1.0000	0.5000	off	on	on	5.0000
CDSF-D	0.9167	90.0000	0.2500	0.5000	on	on	on	7.0000
BERSD-E	0.9085	20.0000	1.0000	0.5000	on	on	off	9.0000
BERSD-L	0.9066	60.0000	1.0000	0.5000	on	on	on	7.0000
CDBER-L	0.9056	60.0000	1.0000	0.5000	on	on	on	6.0000
BERCD-E	0.9029	20.0000	-E.2500	0.4250	on	on	on	5.0000
CDBER-E	0.9025	60.0000	1.0000	0.5000	on	on	on	8.0000
SDBER-E	0.9020	80.0000	0.2500	0.5000	on	off	on	5.0000
CDSD-D	0.9009	80.0000	0.2500	0.4625	on	on	on	9.0000
CDBER-D	0.9004	10.0000	-E.2500	0.4625	on	on	on	8.0000
SDPD-L	0.8999	100.0000	0.5000	0.4250	on	on	on	9.0000
HFCBER-L	0.8984	90.0000	0.0000	0.4625	on	off	on	9.0000
CDPD-L	0.8983	80.0000	0.2500	0.4250	on	on	on	8.0000
HFCPD-L	0.8981	90.0000	0.5000	0.5000	on	on	on	6.0000
CDSD-L	0.8981	80.0000	0.2500	0.4250	on	on	on	8.0000
HFCSD-L	0.8979	80.0000	0.2500	0.4625	on	on	on	9.0000
CD	0.8977	80.0000	0.2500	0.4250	on	on	on	8.0000
HFCCD-L	0.8972	70.0000	0.2500	0.5000	on	on	on	9.0000
PDBER-L	0.8969	70.0000	-E.2500	0.4625	on	off	on	7.0000
HFCBER-E	0.8969	70.0000	-E.2500	0.4625	on	off	on	7.0000
BER	0.8969	70.0000	-E.2500	0.4625	on	off	on	7.0000
HFCCD-E	0.8946	100.0000	0.5000	0.5000	on	off	on	8.0000
SD	0.8931	80.0000	0.2500	0.3875	on	on	on	9.0000
HFC	0.8892	90.0000	0.5000	0.5000	on	on	on	8.0000
HFCSD-E	0.8890	70.0000	0.2500	0.4625	on	on	on	9.0000
HFCCD-D	0.8882	80.0000	0.2500	0.4625	on	on	on	8.0000
BERSD-D	0.8861	50.0000	0.0000	0.5000	on	on	on	9.0000
HFCSD-D	0.8824	90.0000	0.2500	0.5000	on	on	on	9.0000
HFCBER-D	0.8818	100.0000	0.5000	0.4625	on	on	on	8.0000
SDPD-D	0.8026	90.0000	-E.2500	0.3125	on	off	on	6.0000
CDPD-D	0.7995	90.0000	-E.2500	0.2750	on	off	on	5.0000
HFCPD-D	0.7993	90.0000	-E.2500	0.3125	on	off	on	8.0000
PDBER-D	0.7333	70.0000	-E.2500	0.4625	on	off	on	6.0000
RPD	0.6372	80.0000	0.0000	0.2000	on	off	on	9.0000
PD	0.5773	80.0000	0.0000	0.2000	on	off	on	9.0000
HFCPD-E	0.5773	80.0000	0.0000	0.2000	on	off	on	9.0000

Table 3-H: Segmentation F-measures of investigated detectors for onset type non-pitched percussive (NPP)

C.3 Best Performing Configurations Grouped by Dataset (Part I)

Method	F-measure	sens	bt	WT	l	τ
CDSF-L	0.8580	10.0000	2.1500	off	0.2000	n/a
BERSF-L	0.8559	10.0000	2.4000	off	0.3000	n/a
BERSF-D	0.8528	40.0000	2.1500	off	n/a	0.0500
BERSF-E	0.8451	30.0000	2.4000	off	n/a	n/a
CDSF-D	0.8392	50.0000	2.4000	off	n/a	0.0500
SF	0.8274	20.0000	2.4000	off	n/a	n/a
CDBER-L	0.8145	10.0000	2.4000	off	0.5000	n/a
BERCD-L	0.8144	10.0000	2.4000	off	0.5000	n/a
SDBER-L	0.8073	10.0000	2.4000	off	0.6000	n/a
BERSD-L	0.8073	10.0000	2.4000	off	0.6000	n/a
HFCCD-L	0.8032	20.0000	1.1500	off	0.5000	n/a
CDBER-D	0.7967	50.0000	1.1500	off	n/a	0.0500
CD	0.7966	10.0000	2.4000	off	n/a	n/a
CDSD-L	0.7957	20.0000	2.4000	off	0.5000	n/a
BERCD-D	0.7949	50.0000	1.1500	off	n/a	0.0500
CDPD-L	0.7929	10.0000	2.4000	off	1.0000	n/a
CDSD-D	0.7912	50.0000	1.1500	off	n/a	0.0500
HFCSD-L	0.7895	40.0000	1.4000	off	0.4000	n/a
HFCBER-L	0.7892	40.0000	1.9000	off	0.7000	n/a
HFCBER-E	0.7883	10.0000	2.4000	off	n/a	n/a
BER	0.7883	10.0000	2.4000	off	n/a	n/a
PDBER-L	0.7870	10.0000	2.4000	off	1.0000	n/a
CDBER-E	0.7819	10.0000	2.4000	off	n/a	n/a
SDPD-L	0.7818	20.0000	2.4000	off	0.9000	n/a
SDBER-E	0.7801	10.0000	2.4000	off	n/a	n/a
HFCCD-D	0.7798	60.0000	1.1500	off	n/a	0.0500
HFCPD-L	0.7798	50.0000	0.9000	on	0.9000	n/a
SD	0.7795	20.0000	2.4000	off	n/a	n/a
SDBER-D	0.7784	70.0000	2.1500	on	n/a	0.0500
BERSD-D	0.7784	70.0000	2.1500	on	n/a	0.0500
HFCSD-E	0.7768	50.0000	1.1500	off	n/a	n/a
HFCCD-E	0.7751	40.0000	1.4000	off	n/a	n/a
BERCD-E	0.7744	40.0000	2.4000	off	n/a	n/a
BERSD-E	0.7739	40.0000	2.4000	off	n/a	n/a
HFC	0.7712	50.0000	0.9000	off	n/a	n/a
HFCBER-D	0.7696	60.0000	1.9000	off	n/a	0.0500
HFCSD-D	0.7649	70.0000	1.9000	on	n/a	0.0500
lpcPD-L	0.7594	60.0000	0.9000	on	0.8000	n/a
lpc	0.7496	50.0000	0.9000	on	n/a	n/a
SDPD-D	0.7262	80.0000	1.6500	on	n/a	0.0500
CDPD-D	0.7205	80.0000	0.9000	on	n/a	0.0500
HFCPD-D	0.7158	80.0000	1.6500	on	n/a	0.0500
lpcPD-D	0.6976	80.0000	2.4000	on	n/a	0.0500
PD	0.6537	80.0000	2.1500	on	n/a	n/a
HFCPD-E	0.6537	80.0000	2.1500	on	n/a	n/a
RPD	0.6476	80.0000	2.4000	on	n/a	n/a
PDBER-D	0.5303	90.0000	2.4000	off	n/a	0.0500

Table 3-I: Segmentation F-measures of investigated detectors for the overall datasets

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
CDSF-L	0.8194	10.0000	2.1500	off	0.2000	n/a
SF	0.8126	20.0000	2.4000	off	n/a	n/a
BERSF-L	0.8126	20.0000	2.4000	off	0.0000	n/a
BERSF-D	0.8089	20.0000	2.1500	off	n/a	0.0500
BERSF-E	0.8025	10.0000	2.4000	off	n/a	n/a
CDSF-D	0.7892	50.0000	2.4000	off	n/a	0.0500
CDBER-L	0.7877	10.0000	2.4000	off	0.5000	n/a
BERCD-L	0.7877	10.0000	2.4000	off	0.5000	n/a
SDBER-L	0.7843	10.0000	2.4000	off	0.6000	n/a
BERSD-L	0.7843	10.0000	2.4000	off	0.6000	n/a
HFCCD-L	0.7802	30.0000	1.1500	off	0.6000	n/a
CDSD-D	0.7728	50.0000	2.4000	off	n/a	0.0200
HFCBER-L	0.7720	30.0000	1.9000	off	0.6000	n/a
CDSD-L	0.7715	10.0000	2.4000	off	0.4000	n/a
CD	0.7692	10.0000	2.4000	off	n/a	n/a
PDBER-L	0.7665	10.0000	2.4000	off	0.9000	n/a
HFCSD-L	0.7660	20.0000	1.6500	off	0.4000	n/a
CDPD-L	0.7629	10.0000	2.4000	off	1.0000	n/a
HFCBER-E	0.7626	10.0000	2.4000	off	n/a	n/a
BER	0.7626	10.0000	2.4000	off	n/a	n/a
CDBER-D	0.7605	50.0000	1.1500	off	n/a	0.0500
BERCD-D	0.7605	50.0000	1.1500	off	n/a	0.0500
SD	0.7604	20.0000	2.1500	off	n/a	n/a
BERSD-E	0.7595	40.0000	2.4000	off	n/a	n/a
BERCD-E	0.7587	40.0000	2.4000	on	n/a	n/a
SDPD-L	0.7585	20.0000	2.1500	off	1.0000	n/a
SDBER-E	0.7552	10.0000	2.4000	off	n/a	n/a
HFCPD-L	0.7506	50.0000	1.1500	on	0.9000	n/a
HFCCD-D	0.7505	60.0000	1.1500	off	n/a	0.0500
SDBER-D	0.7477	40.0000	2.1500	off	n/a	0.0500
BERSD-D	0.7477	40.0000	2.1500	off	n/a	0.0500
HFCSD-E	0.7443	50.0000	1.4000	off	n/a	n/a
CDBER-E	0.7439	10.0000	2.4000	off	n/a	n/a
HFCSD-D	0.7435	70.0000	1.9000	on	n/a	0.0450
HFC	0.7411	50.0000	1.1500	off	n/a	n/a
HFCBER-D	0.7344	40.0000	1.1500	off	n/a	0.0500
HFCCD-E	0.7324	40.0000	1.1500	off	n/a	n/a
SDPD-D	0.6840	80.0000	1.9000	on	n/a	0.0500
CDPD-D	0.6809	90.0000	0.9000	on	n/a	0.0500
HFCPD-D	0.6799	80.0000	2.1500	on	n/a	0.0500
RPD	0.6145	80.0000	2.4000	on	n/a	n/a
PD	0.6144	80.0000	1.6500	on	n/a	n/a
HFCPD-E	0.6144	80.0000	1.6500	on	n/a	n/a
PDBER-D	0.5224	100.0000	1.9000	off	n/a	0.0450

Table 3-J: Segmentation F-measures of investigated detectors for dataset SB

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
CDSF-L	0.9286	50.0000	2.4000	off	0.2000	n/a
BERSF-L	0.9283	40.0000	2.4000	off	0.1000	n/a
BERSF-D	0.9230	60.0000	2.4000	off	n/a	0.0500
BERSF-E	0.9175	40.0000	2.4000	off	n/a	n/a
CDSF-D	0.9165	70.0000	2.4000	off	n/a	0.0500
CDBER-L	0.8560	20.0000	2.4000	off	0.3000	n/a
BERCD-L	0.8547	20.0000	2.4000	off	0.3000	n/a
CDBER-D	0.8498	50.0000	1.9000	off	n/a	0.0500
SF	0.8488	40.0000	2.4000	off	n/a	n/a
BERCD-D	0.8458	50.0000	1.9000	off	n/a	0.0500
SDBER-L	0.8420	20.0000	1.9000	off	0.7000	n/a
BERSD-L	0.8420	20.0000	1.9000	off	0.7000	n/a
HFCCD-L	0.8416	50.0000	2.4000	off	0.3000	n/a
CDPD-L	0.8356	10.0000	2.4000	off	0.9000	n/a
CDSD-L	0.8341	30.0000	2.4000	off	0.7000	n/a
CDBER-E	0.8327	10.0000	2.4000	off	n/a	n/a
HFCCD-E	0.8326	40.0000	1.4000	off	n/a	n/a
CD	0.8320	10.0000	2.4000	off	n/a	n/a
HFCSD-L	0.8262	40.0000	1.6500	off	0.6000	n/a
SDPD-L	0.8261	60.0000	1.9000	on	0.9000	n/a
HFCBER-E	0.8254	20.0000	2.4000	off	n/a	n/a
BER	0.8254	20.0000	2.4000	off	n/a	n/a
CDSD-D	0.8226	70.0000	2.4000	on	n/a	0.0450
HFCBER-L	0.8226	40.0000	0.9000	off	0.7000	n/a
PDBER-L	0.8226	20.0000	2.4000	off	1.0000	n/a
SDBER-D	0.8217	70.0000	2.4000	on	n/a	0.0500
BERSD-D	0.8217	70.0000	2.4000	on	n/a	0.0500
HFCSD-E	0.8216	40.0000	1.1500	off	n/a	n/a
SD	0.8210	70.0000	1.6500	on	n/a	n/a
HFCCD-D	0.8200	70.0000	0.9000	off	n/a	0.0500
HFCPD-L	0.8197	40.0000	0.9000	on	0.9000	n/a
HFCBER-D	0.8183	60.0000	1.9000	off	n/a	0.0500
HFC	0.8159	60.0000	1.6500	off	n/a	n/a
SDBER-E	0.8132	10.0000	2.4000	off	n/a	n/a
HFCSD-D	0.8069	70.0000	0.9000	off	n/a	0.0500
BERCD-E	0.8067	50.0000	2.4000	off	n/a	n/a
BERSD-E	0.8016	50.0000	2.4000	on	n/a	n/a
CDPD-D	0.7875	80.0000	1.6500	on	n/a	0.0500
SDPD-D	0.7827	80.0000	1.6500	on	n/a	0.0500
HFCPD-D	0.7667	80.0000	1.4000	on	n/a	0.0500
PD	0.7114	80.0000	2.4000	on	n/a	n/a
HFCPD-E	0.7114	80.0000	2.4000	on	n/a	n/a
RPD	0.6958	70.0000	1.9000	on	n/a	n/a
PDBER-D	0.5507	90.0000	2.4000	off	n/a	0.0500

Table 3-K: Segmentation F-measures of investigated detectors for dataset JPB

Method	**F-measure**	**sens**	**bt**	**WT**	**l**	τ
CDSF-L	0.9368	40.0000	2.4000	off	0.3000	n/a
SF	0.9350	40.0000	2.4000	off	n/a	n/a
BERSF-L	0.9350	40.0000	2.4000	off	0.0000	n/a
BERSF-E	0.9307	20.0000	2.4000	off	n/a	n/a
BERSF-D	0.9300	40.0000	2.4000	off	n/a	0.0500
CDSF-D	0.9146	70.0000	1.4000	off	n/a	0.0500
CDBER-L	0.9081	30.0000	0.9000	off	0.1000	n/a
BERCD-L	0.9081	30.0000	0.9000	off	0.1000	n/a
SDBER-L	0.9074	30.0000	0.6500	on	0.2000	n/a
BERSD-L	0.9074	30.0000	0.6500	on	0.2000	n/a
SDPD-L	0.9071	20.0000	1.1500	off	0.8000	n/a
HFCBER-L	0.9037	10.0000	2.4000	on	0.8000	n/a
BERCD-E	0.9017	20.0000	2.4000	off	n/a	n/a
CDPD-L	0.9010	30.0000	2.4000	off	0.9000	n/a
BERSD-E	0.9007	30.0000	2.4000	off	n/a	n/a
CDBER-D	0.8998	60.0000	1.1500	off	n/a	0.0500
BERCD-D	0.8998	60.0000	1.1500	off	n/a	0.0500
CDSD-L	0.8985	40.0000	0.4000	off	0.9000	n/a
SDBER-D	0.8978	60.0000	0.9000	off	n/a	0.0500
BERSD-D	0.8978	60.0000	0.9000	off	n/a	0.0500
HFCCD-L	0.8963	30.0000	0.9000	off	0.0000	n/a
CD	0.8963	30.0000	0.9000	off	n/a	n/a
SD	0.8956	20.0000	0.6500	off	n/a	n/a
HFCSD-L	0.8956	20.0000	0.6500	off	0.0000	n/a
CDSD-D	0.8956	40.0000	0.9000	off	n/a	0.0500
HFCPD-L	0.8932	10.0000	2.4000	on	0.8000	n/a
HFCBER-D	0.8905	30.0000	0.9000	on	n/a	0.0500
HFCCD-D	0.8896	50.0000	1.4000	off	n/a	0.0200
HFCCD-E	0.8896	30.0000	2.1500	off	n/a	n/a
HFCSD-D	0.8893	20.0000	2.4000	on	n/a	0.0500
HFCSD-E	0.8882	10.0000	2.4000	on	n/a	n/a
HFC	0.8878	30.0000	2.4000	on	n/a	n/a
CDBER-E	0.8856	10.0000	2.4000	off	n/a	n/a
PDBER-L	0.8763	10.0000	2.4000	on	0.9000	n/a
HFCBER-E	0.8731	10.0000	2.4000	on	n/a	n/a
BER	0.8731	10.0000	2.4000	on	n/a	n/a
SDBER-E	0.8691	10.0000	2.4000	off	n/a	n/a
SDPD-D	0.7634	90.0000	0.9000	on	n/a	0.0500
CDPD-D	0.7610	80.0000	2.4000	on	n/a	0.0500
HFCPD-D	0.7332	80.0000	1.1500	on	n/a	0.0500
RPD	0.5609	60.0000	0.9000	on	n/a	n/a
PD	0.5042	70.0000	0.4000	on	n/a	n/a
HFCPD-E	0.5042	70.0000	0.4000	on	n/a	n/a
PDBER-D	0.3799	80.0000	1.9000	off	n/a	0.0500

Table 3-L: Segmentation F-measures of investigated detectors for dataset CP

C.4 Best Performing Configurations Grouped by Dataset (Part II)

Method	F-measure	sens	δ	f_c	LP	PF	MF	ψ
CDSF-L	0.8669	40.0000	0.5000	0.5000	on	on	on	7.0000
SF	0.8668	10.0000	0.2500	0.5000	on	on	on	8.0000
BERSF-L	0.8656	20.0000	0.5000	0.5000	on	on	on	7.0000
BERSF-D	0.8541	50.0000	0.2500	0.4250	on	on	on	8.0000
BERSF-E	0.8522	40.0000	0.5000	0.5000	on	on	on	8.0000
CDBER-L	0.8501	70.0000	0.5000	0.3875	on	on	on	8.0000
CDSF-D	0.8457	60.0000	0.2500	0.5000	on	on	on	6.0000
BERSD-L	0.8398	60.0000	0.5000	0.3875	on	on	on	8.0000
HFCCD-L	0.8346	80.0000	0.5000	0.4625	on	on	on	7.0000
CDSD-L	0.8256	90.0000	0.7500	0.4625	on	on	on	7.0000
CD	0.8247	70.0000	0.5000	0.3875	on	on	on	8.0000
CDBER-D	0.8220	40.0000	0.0000	0.4250	on	on	on	8.0000
CDPD-L	0.8219	70.0000	0.5000	0.3875	on	on	on	8.0000
CDSD-D	0.8208	70.0000	0.2500	0.4250	on	on	on	8.0000
HFCBER-E	0.8202	20.0000	0.2500	0.3500	on	on	on	9.0000
BER	0.8202	20.0000	0.2500	0.3500	on	on	on	9.0000
HFCSD-L	0.8201	70.0000	0.5000	0.5000	on	on	on	6.0000
PDBER-L	0.8182	20.0000	0.2500	0.3500	on	on	on	9.0000
HFCBER-L	0.8123	30.0000	0.0000	0.3500	on	on	on	7.0000
CDBER-E	0.8116	10.0000	0.0000	0.3125	on	on	on	8.0000
SDBER-E	0.8100	30.0000	0.5000	0.5000	on	on	on	7.0000
HFCPD-L	0.8090	100.0000	0.5000	0.4625	on	on	on	6.0000
HFCCD-D	0.8068	50.0000	0.0000	0.3875	on	on	on	7.0000
SDPD-L	0.8066	80.0000	0.5000	0.3875	on	on	on	7.0000
HFCSD-E	0.8059	90.0000	0.2500	0.3125	on	on	on	8.0000
SD	0.8054	80.0000	0.5000	0.3875	on	on	on	8.0000
HFCCD-E	0.8046	70.0000	0.2500	0.4625	on	on	on	7.0000
HFC	0.8014	90.0000	0.2500	0.3500	on	on	on	9.0000
BERCD-E	0.7993	40.0000	1.0000	0.2750	on	on	off	9.0000
BERSD-D	0.7981	30.0000	-E.2500	0.3500	on	on	on	8.0000
BERSD-E	0.7979	40.0000	-E.2500	0.3500	on	on	on	6.0000
HFCBER-D	0.7865	20.0000	-E.2500	0.3500	on	on	on	6.0000
HFCSD-D	0.7829	70.0000	0.0000	0.3500	on	on	on	8.0000
lpcPD-L	0.7817	20.0000	-E.2500	0.3500	on	on	on	7.0000
lpc	0.7723	80.0000	0.2500	0.3875	on	on	on	9.0000
SDPD-D	0.7450	50.0000	-E.2500	0.3125	on	on	on	7.0000
lpcPD-D	0.7405	40.0000	1.0000	0.2000	on	on	off	9.0000
CDPD-D	0.7402	50.0000	-E.2500	0.3125	on	on	on	7.0000
HFCPD-D	0.7342	60.0000	1.0000	0.2375	on	on	off	9.0000
RPD	0.7036	90.0000	0.0000	0.2375	on	off	on	9.0000
PD	0.6887	90.0000	0.0000	0.2000	on	off	on	6.0000
HFCPD-E	0.6887	90.0000	0.0000	0.2000	on	off	on	6.0000
PDBER-D	0.6650	80.0000	-E.2500	0.3500	on	off	on	8.0000

Table 3-M: Segmentation F-measures of investigated detectors for the overall datasets

Method	**F-measure**	**sens**	δ	f_c	**LP**	**PF**	**MF**	ψ
CDSF-L	0.8217	40.0000	0.5000	0.4625	on	on	on	7.0000
BERSF-L	0.8207	40.0000	0.5000	0.3875	on	on	on	8.0000
SF	0.8193	20.0000	0.5000	0.5000	on	on	on	8.0000
BERSF-D	0.8145	40.0000	0.0000	0.2750	on	on	on	8.0000
BERSF-E	0.8117	20.0000	0.5000	0.5000	on	on	on	9.0000
CDBER-L	0.7997	40.0000	0.2500	0.3125	on	on	on	9.0000
CDSF-D	0.7970	20.0000	0.0000	0.4625	on	on	on	6.0000
BERSD-L	0.7970	30.0000	0.2500	0.2750	on	on	on	8.0000
PDBER-L	0.7827	10.0000	0.2500	0.3125	on	on	on	8.0000
HFCBER-E	0.7827	10.0000	0.2500	0.3125	on	on	on	8.0000
BER	0.7827	10.0000	0.2500	0.3125	on	on	on	8.0000
HFCCD-L	0.7812	70.0000	0.5000	0.3875	on	on	on	7.0000
SDBER-E	0.7752	30.0000	0.2500	0.3125	on	on	on	8.0000
CDSD-D	0.7730	10.0000	-E.2500	0.3875	on	on	on	7.0000
CD	0.7725	80.0000	0.2500	0.2375	on	on	on	9.0000
CDSD-L	0.7724	10.0000	0.0000	0.4250	on	on	on	7.0000
HFCSD-L	0.7696	60.0000	0.2500	0.3875	on	on	on	6.0000
HFCBER-L	0.7689	30.0000	0.0000	0.3500	on	on	on	7.0000
CDPD-L	0.7689	50.0000	0.0000	0.2375	on	on	on	7.0000
CDBER-D	0.7656	60.0000	0.2500	0.4625	on	on	on	6.0000
SD	0.7617	20.0000	0.0000	0.3500	on	on	on	7.0000
SDPD-L	0.7596	20.0000	0.0000	0.3125	on	on	on	7.0000
BERCD-E	0.7574	40.0000	0.2500	0.3875	on	on	on	6.0000
BERSD-E	0.7572	40.0000	0.0000	0.3500	on	on	on	5.0000
HFCPD-L	0.7560	60.0000	0.0000	0.2750	on	on	on	8.0000
CDBER-E	0.7533	10.0000	0.0000	0.3125	on	on	on	8.0000
HFCCD-D	0.7532	100.0000	0.2500	0.3125	on	off	on	7.0000
HFCSD-E	0.7512	20.0000	-E.2500	0.2750	on	on	on	9.0000
BERSD-D	0.7479	50.0000	0.0000	0.4250	on	on	on	9.0000
HFC	0.7470	20.0000	-E.2500	0.3125	on	on	on	9.0000
HFCSD-D	0.7464	100.0000	0.2500	0.3125	on	off	on	9.0000
HFCCD-E	0.7421	60.0000	0.0000	0.2375	on	on	on	9.0000
HFCBER-D	0.7386	70.0000	0.0000	0.2750	on	on	on	9.0000
lpcPD-L	0.7343	20.0000	-E.2500	0.3125	on	on	on	9.0000
lpc	0.7295	60.0000	0.0000	0.2375	on	on	on	9.0000
lpcPD-D	0.7008	40.0000	1.0000	0.2000	on	on	off	9.0000
SDPD-D	0.6985	90.0000	-E.2500	0.3125	on	off	on	7.0000
CDPD-D	0.6978	90.0000	-E.2500	0.3125	on	off	on	7.0000
HFCPD-D	0.6828	90.0000	-E.2500	0.3125	on	off	on	9.0000
RPD	0.6549	90.0000	0.0000	0.2000	on	off	on	9.0000
PD	0.6325	90.0000	0.0000	0.2000	on	off	on	6.0000
HFCPD-E	0.6325	90.0000	0.0000	0.2000	on	off	on	6.0000
PDBER-D	0.6325	80.0000	-E.5000	0.3500	on	off	on	8.0000

Table 3-N: Segmentation F-measures of investigated detectors for dataset SB

Method	**F-measure**	**sens**	δ	f_c	**LP**	**PF**	**MF**	ψ
CDSF-L	0.9485	100.0000	0.5000	0.5000	on	on	on	6.0000
SF	0.9448	60.0000	0.2500	0.5000	on	on	on	7.0000
BERSF-L	0.9407	10.0000	0.0000	0.5000	on	on	on	7.0000
BERSF-E	0.9311	100.0000	0.5000	0.5000	on	on	on	6.0000
BERSF-D	0.9286	40.0000	0.0000	0.4250	on	on	on	8.0000
CDSF-D	0.9275	50.0000	0.0000	0.5000	on	on	on	5.0000
HFCCD-L	0.9269	80.0000	0.5000	0.5000	on	on	on	6.0000
CDBER-L	0.9259	60.0000	0.5000	0.5000	on	on	on	7.0000
HFCCD-E	0.9181	90.0000	0.5000	0.4625	on	on	on	5.0000
CDPD-L	0.9115	60.0000	0.2500	0.5000	on	on	on	6.0000
BERSD-L	0.9104	20.0000	0.2500	0.4625	on	on	on	7.0000
CD	0.9094	70.0000	0.7500	0.5000	on	on	on	7.0000
CDBER-D	0.9072	10.0000	-E.2500	0.4625	on	on	on	8.0000
CDSD-L	0.9055	80.0000	0.2500	0.4250	on	on	on	7.0000
HFCPD-L	0.9001	100.0000	0.5000	0.5000	on	on	on	6.0000
HFCCD-D	0.8991	70.0000	0.0000	0.4250	on	on	on	7.0000
HFCSD-L	0.8981	80.0000	0.5000	0.4625	on	on	on	6.0000
CDBER-E	0.8967	60.0000	0.5000	0.4625	on	on	on	7.0000
HFCSD-E	0.8957	80.0000	0.5000	0.5000	on	on	on	7.0000
CDSD-D	0.8899	80.0000	0.2500	0.3875	on	on	on	9.0000
HFC	0.8891	50.0000	0.0000	0.4250	on	on	on	7.0000
HFCBER-L	0.8843	80.0000	0.2500	0.3500	on	on	on	8.0000
SDPD-L	0.8835	80.0000	0.5000	0.4250	on	on	on	7.0000
HFCBER-E	0.8826	40.0000	0.2500	0.3500	on	on	on	9.0000
BER	0.8826	40.0000	0.2500	0.3500	on	on	on	9.0000
PDBER-L	0.8806	10.0000	0.2500	0.4250	on	on	on	7.0000
SD	0.8769	80.0000	0.2500	0.3875	on	on	on	8.0000
SDBER-E	0.8714	30.0000	0.2500	0.3875	on	on	on	8.0000
BERSD-D	0.8712	90.0000	0.2500	0.4625	on	on	on	7.0000
BERCD-E	0.8705	50.0000	1.0000	0.2375	on	on	off	9.0000
HFCBER-D	0.8680	60.0000	0.0000	0.3875	on	on	on	7.0000
BERSD-E	0.8657	50.0000	1.0000	0.2750	on	on	off	9.0000
HFCSD-D	0.8536	90.0000	-E.5000	0.4250	on	off	on	5.0000
CDPD-D	0.8328	10.0000	-E.5000	0.3500	on	on	on	8.0000
SDPD-D	0.8297	10.0000	1.0000	0.2375	on	on	off	9.0000
HFCPD-D	0.8164	40.0000	-E.2500	0.4250	on	on	on	7.0000
RPD	0.7949	80.0000	-E.2500	0.2750	on	off	on	9.0000
PD	0.7801	90.0000	0.0000	0.2375	on	off	on	7.0000
HFCPD-E	0.7801	90.0000	0.0000	0.2375	on	off	on	7.0000
PDBER-D	0.7175	80.0000	-E.2500	0.4250	on	off	on	9.0000

Table 3-O: Segmentation F-measures of investigated detectors for dataset JPB

Method	**F-measure**	**sens**	δ	f_c	**LP**	**PF**	**MF**	ψ
SF	0.9486	80.0000	0.2500	0.5000	on	on	on	6.0000
CDSF-L	0.9455	100.0000	0.5000	0.5000	on	on	on	6.0000
BERSF-E	0.9440	20.0000	0.0000	0.5000	on	on	on	9.0000
BERSF-D	0.9396	10.0000	-E.2500	0.5000	on	on	on	5.0000
BERSF-L	0.9232	90.0000	-L.0000	0.3875	on	off	on	5.0000
CDSF-D	0.9224	70.0000	0.0000	0.3875	on	on	on	8.0000
BERSD-E	0.9134	10.0000	-E.2500	0.4625	on	on	on	7.0000
BERSD-L	0.9120	40.0000	1.0000	0.4250	on	on	on	6.0000
CDBER-D	0.9115	20.0000	-E.2500	0.4625	on	on	on	9.0000
CDBER-L	0.9103	40.0000	0.7500	0.3500	on	on	on	9.0000
SDPD-L	0.9103	60.0000	0.2500	0.3875	on	on	on	9.0000
BERCD-E	0.9102	10.0000	-E.2500	0.4625	on	on	on	7.0000
PDBER-L	0.9087	70.0000	-E.7500	0.4250	on	off	on	5.0000
HFCBER-E	0.9087	70.0000	-E.7500	0.4250	on	off	on	5.0000
BER	0.9087	70.0000	-E.7500	0.4250	on	off	on	5.0000
CDBER-E	0.9084	20.0000	1.0000	0.3500	on	on	on	9.0000
HFCBER-L	0.9075	80.0000	-L.0000	0.3125	on	off	on	9.0000
CDPD-L	0.9068	40.0000	0.0000	0.3875	on	on	on	9.0000
CD	0.9068	40.0000	0.0000	0.3875	on	on	on	9.0000
SDBER-E	0.9064	80.0000	0.2500	0.5000	on	off	on	5.0000
CDSD-D	0.9050	80.0000	0.2500	0.3500	on	on	on	9.0000
HFCSD-L	0.9041	40.0000	0.0000	0.3500	on	on	on	9.0000
CDSD-L	0.9039	40.0000	0.0000	0.2750	on	on	on	8.0000
SD	0.9013	100.0000	0.5000	0.3500	on	on	on	9.0000
HFCPD-L	0.8991	10.0000	0.0000	0.5000	on	on	on	8.0000
HFCCD-L	0.8974	30.0000	0.0000	0.4250	on	on	on	9.0000
HFCCD-D	0.8956	40.0000	0.0000	0.4250	on	on	on	9.0000
HFCCD-E	0.8953	90.0000	-E.2500	0.4625	on	off	on	9.0000
HFC	0.8944	90.0000	0.0000	0.4625	on	off	on	8.0000
BERSD-D	0.8932	60.0000	0.2500	0.4250	on	on	on	9.0000
HFCSD-E	0.8930	40.0000	0.0000	0.3875	on	on	on	9.0000
HFCBER-D	0.8927	100.0000	0.5000	0.3875	on	on	on	9.0000
HFCSD-D	0.8918	100.0000	0.5000	0.3125	on	on	on	9.0000
CDPD-D	0.8261	80.0000	-L.0000	0.2375	on	off	on	6.0000
SDPD-D	0.8239	80.0000	-L.0000	0.2375	on	off	on	7.0000
HFCPD-D	0.8203	80.0000	-E.7500	0.2750	on	off	on	6.0000
PDBER-D	0.7639	80.0000	-E.5000	0.2750	on	off	on	6.0000
RPD	0.6305	80.0000	0.0000	0.2000	on	off	on	9.0000
PD	0.5776	60.0000	-E.2500	0.2000	on	off	on	9.0000
HFCPD-E	0.5776	60.0000	-E.2500	0.2000	on	off	on	9.0000

Table 3-P: Segmentation F-measures of investigated detectors for dataset CP

Bibliography

[AB07] Sanjeev Arora and Boaz Barak. *Computational complexity: A modern approach*. Cambridge University Press, 2007.

[ANR74] Nasir Ahmed, T Natarajan, and Kamisetty R Rao. "Discrete cosine transform". In: *Computers, IEEE Transactions on* 100.1 (1974), pp. 90–93.

[AP02] Jean-Julien Aucouturier and Francois Pachet. "Music similarity measures: What's the use?" In: *Proceedings of the 3rd International Society of Music Information Retrieval Conference*. Paris, France, 2002.

[APS05] J-J Aucouturier, François Pachet, and Mark Sandler. ""The way it Sounds": timbre models for analysis and retrieval of music signals". In: *Multimedia, IEEE Transactions on* 7.6 (2005), pp. 1028–1035.

[AS01] Jean-Julien Aucouturier and Mark Sandler. "Segmentation of musical signals using hidden Markov models". In: *Proceedinds of Audio Engineering Society Convention 110*. Amsterdam, Netherlands, 2001, pp. 5379–5387.

[AW10] Hervé Abdi and Lynne J Williams. "Tukey's honestly significant difference (HSD) test". In: *Encyclopedia of Research Design* (2010), pp. 1–5.

[Bar+05] Dan Barry, Derry Fitzgerald, Eugene Coyle, and Bob Lawlor. "Drum source separation using percussive feature detection and spectral modulation". In: *Proceedings of IEE Irish Signals and Systems Conference*. Dublin, Ireland, 2005.

[BC02] Donald Byrd and Tim Crawford. "Problems of music information retrieval in the real world". In: *Information processing & management* 38.2 (2002), pp. 249–272.

[BC11] Tim Berners-Lee and Dan Connolly. *Notation3 (N3): A readable RDF syntax*. Online. 2011. URL: `http://www.w3.org/TeamSubmission/n3/` (visited on 2011).

[BCL10] Luke Barrington, Antoni B Chan, and Gert Lanckriet. "Modeling music as a dynamic texture". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 18.3 (2010), pp. 602–612.

[BDC15] Donald Byrd, J.Stephen Downie, and Tim Crawford, eds. *Current research in Music Information Retrieval: Searching Audio, Midi and Notation*. Kluwer Academic Publishers, 2015.

[Bel+04] Juan P. Bello, Chris Duxbury, Mike Davies, and Mark Sandler. "On the use of phase and energy for musical onset detection in the complex domain". In: *IEEE Signal Processing Letters* 11.6 (2004).

[Bel+05] Juan Pablo Bello, Laurent Daudet, Samer Abdallan, Chris Duxbury, and Mike Davies. "A tutorial on onset detection in music signals". In: *Audio, Speech and Language Processing, IEEE Transactions on* 13.5 (2005).

[Bel03] Juan P. Bello. "Towards the automated analysis of simple polyphonic music: A knowledge-based approach". PhD thesis. Queen Mary University of London, 2003.

[Ben08] Jacob Benesty. *Springer handbook of speech processing*. Springer Science & Business Media, 2008.

[BHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. "The Semantic Web". In: *Scientific American* 284.5 (2001), pp. 28–37.

[BKS12] Sebastian Böck, Florian Krebs, and Markus Schedl. "Evaluating the online capabilities of onset detection methods". In: *Proceedings of the 13th International Society for Music Information Retrieval Conference*. Porto, Portugal, 2012, pp. 49–54.

[BLT14] Dawn A. A. Black, Ma Li, and Mi Tian. "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*. Seoul, South Korea, 2014, pp. 250–255.

[BMK06] Michael J Bruderer, Martin F McKinney, and Armin Kohlrausch. "Structural boundary perception in popular music". In: *Proceedings of the 7th International Society of Music Information Retrieval Conference*. Victoria, Canada, 2006, pp. 198–201.

[Böc+12] Sebastian Böck, Andreas Arzt, Florian Krebs, and Markus Schedl. "Online realtime onset detection with recurrent neural networks". In: *Proceedings of International Conference on Digital Audio Effects*. York, UK, 2012.

[BR99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. 1st ed. Addison Wesley, 1999.

[Bro91] Judith C. Brown. "Calculation of a constant Q spectral transform". In: *The Journal of the Acoustical Society of America* 89.1 (1991), pp. 425–434.

[Bru08] Michael J Bruderer. "Perception and modeling of segment boundaries in popular music". PhD thesis. J.F. Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven, 2008.

[BW05] Mark A Bartsch and Gregory H Wakefield. "Audio thumbnailing of popular music using chroma-based representations". In: *Multimedia, IEEE Transactions on* 7.1 (2005), pp. 96–104.

[BW13a] Sebastian Böck and Gerhard Widmer. "Local group delay based vibrato and tremolo suppression for onset detection". In: *Proceedings of the 14th International Society for Music Information Retrieval Conference*. Curtiba, Brazil, 2013, pp. 361–366.

[BW13b] Sebastian Böck and Gerhard Widmer. "Maximum filter vibrato suppression for onset detection". In: *16th International Conference on Digital Audio Effects*. Maynooth, Ireland, 2013.

[Can+06] Chris Cannam, Christian Landone, Mark B Sandler, and Juan Pablo Bello. "The Sonic Visualiser: A Visualisation Platform for Semantic Descriptors from Musical Signals". In: *Proceedings of the 7th International Society for Music Information Retrieval Conference*. Victoria, Canada, 2006, pp. 324–327.

[Can09] Chris Cannam. *The Vamp Audio Analysis Plugin API: A Programmer's Guide*. Availble online: `http://vamp-plugins.org/guide.pdf`. 2009.

[Cas+08] Michael A. Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. "Content-Based Music Information Retrieval: Current Directions and Future Challenges". In: vol. 96. IEEE Proceedings, Apr. 2008.

[CF81] Paul R. Cohen and Edward A. Feigenbaum, eds. *The handbook of artificial intelligence*. Vol. 3. William Kaufman, 1981.

[Cha05] Wei Chai. "Automated analysis of musical structure". PhD thesis. Massachusetts Institute of Technology Media Laboratory, 2005.

[Che13] Kainan Chen. "Characterization of Pitch Intonation of Beijing Opera". MA thesis. Universitat Pompeu Fabra, 2013.

[chu92] Shanghai wenyi chubanshe. *Collection of jingju scores ("Jingju qupu jicheng")*. Shanghai wenyi Press, 1992. URL: `http://www.lib.cam.ac.uk/mulu/fb93106.html`.

[Cla87] Eric F. Clarke. "Categorical rhythm perception: An ecological perspective". In: ed. by A. Gabrielsson. 55. Stockholm: Publications issued by the Royal Swedish Academy of Music, 1987.

[CMG10] China Music Group (CMG). *Peking Opera Box set, Limited Edition*. Audio CD. 2010.

[Col05a] Nick Collins. "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions". In: *Proceedings of Audio Engineering Society Convention 118*. Barcelona, Spain, 2005, pp. 6363–6374.

[Col05b] Nick Collins. "Using a pitch detector for onset detection". In: *Proceedings of 6th International Society for Music Information Retrieval Conference*. London, UK, 2005, pp. 100–106.

[Coo63] Meyer Cooper. *The Rhythmic Structure of Music*. University of Chicago Press, 1963.

[Cro98] Malcolm J. Crocker. *Handbook of Acoustics*. Spinger, 1998.

[CSC10] Debora C. Correa, Jose H. Saito, and Luciano da F. Costa. "Musical genres: beating to the rhythms of different drums". In: *New Journal of Physics* 12.5 (2010).

[Deg+09] Norberto Degara-Quintela, Antonio Pena, Soledad Torres-Guijarro, and Laboratorio Oficial de Metroloxıa. "A comparison of score-level fusion rules for onset detection in music signals". In: *Proceedings of the 10th International Society for Music Information Retrieval Conference*. Kobe, Japan, 2009, pp. 117–122.

[Deg+10] Norberto Degara, Antonio Pena, Matthew EP Davies, and Mark D Plumbley. "Note onset detection using rhythmic structure". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, USA, 2010, pp. 5526–5529.

[Deu12] Diana Deutsch. *The psychology of music*. Academic Press, 2012.

[Dix06] Simon Dixon. "Onset detection revisited". In: *Proceedings of the 9th International Conference on Digital Audio Effects*. Montreal, Canada, 2006, pp. 133–137.

[DM80] Steven B Davis and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 28.4 (1980), pp. 457–366.

[DMD14] Jonathan Driedger, Meinard Müller, and Sascha Disch. "Extending harmonic-percussive separation of audio signals". In: *Proceedings of the 15th Inter-*

national Society for Music Information Conference. Taipei, Taiwan, 2014, pp. 611–616.

[Dow03a] J. Stephen Downie. "Music information retrieval". In: *Annual Review of Information Science and Technology* 37 (2003), pp. 295–340.

[Dow03b] J. Stephen Downie. "Toward the scientific evaluation of music information retrieval systems". In: *Proceedings of the 4th International Society for Music Information Retrieval Conference*. Maryland, USA, 2003.

[DP04] Matthew E.P. Davies and Mark D. Plumbley. "Causal tempo tracking of audio". In: *Proceedings of the 5th International Society for Music Information Retrieval Conference*. Barcelona, Spain, 2004, pp. 164–169.

[DP07] Matthew EP Davies and Mark D Plumbley. "Context-dependent beat tracking of musical audio". In: *Audio, Speech, and Language Processing, IEEE Transactions on* (2007).

[DPR93] Lloyd A. Dawe, John R. Platf, and Ronald J. Racine. "Harmonic accents in inference of metrical structure and perception of rhythm patterns". In: *Perception & psychophysics* 54.6 (1993), pp. 794–807.

[Dux+03] Chris Duxbury, Juan Pablo Bello, Mike Davies, and Mark Sandler. "Complex domain onset detection for musical signals". In: *Proceedings of International Conference on Digital Audio Effects*. Madrid, Spain, 2003, pp. 6–9.

[EB04] Jana Eggink and Guy J. Brown. "Extracting melody lines from complex audio". In: *Proceedings of the 5th International Society for Music Information Retrieval Conference*. Barcelona, Spain, 2004, pp. 124–131.

[EKR87] Jean-Pierre Eckmann, S. Oliffson Kamphorst, and David Ruelle. "Recurrence plots of dynamical systems". In: *Europhysics Letters* 4.9 (1987), pp. 973–977.

[Ell07] Daniel P.W. Ellis. "Beat tracking by dynamic programming". In: *Journal of New Music Research* 36.1 (2007), pp. 51–60.

[Ell09] Daniel P. W. Ellis. *Gammatone-like spectrograms*. 2009. URL: `http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/`.

[EP07] Daniel P.W. Ellis and Graham E. Poliner. "Identifying cover songs with chroma features and dynamic programming beat tracking". In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. Honolulu, HI, 2007, pp. 1429–1432.

[Ero07] Antti Eronen. "Chorus detection with combined use of MFCC and chroma features and image processing filters". In: *Proceedings of International Conference on Digital Audio Effects*. Bordeaux, France, 2007, pp. 229–236.

[Eyb+10] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. "Universal onset detection with bidirectional long short-rerm memory neural networks". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands, 2010, pp. 589–594.

[FC03] Jonathan T Foote and Matthew L Cooper. "Media segmentation using self-similarity decomposition". In: *Electronic Imaging*. 2003.

[Fit10] Derry FitzGerald. "Harmonic/percussive separation using median filtering". In: *Proceedings of the 13th International Conference on Digital Audio Effects*. Graz, Austria, 2010, pp. 15–18.

[Fle06] Arthur Flexer. "Statistical Evaluation of music information retrieval experiments". In: *Journal of New Music Research* 35.2 (2006), pp. 113–120.

[Fle40] Harvey Fletcher. "Auditory patterns". In: *Reviews of modern physics* 12.1 (1940), p. 47.

[Foo00] Jonathan Foote. "Automatic audio segmentation using a measure of audio novelty". In: *Proceedings of IEEE International Conference on Multimedia and Expo*. New York, NY, 2000, pp. 452–455.

[Fra87] Jr. Franklin E. White. *Data Fusion Lexicon, Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego*. Tech. rep. 1987.

[Fri96] Simon Frith. “Music an Identity”. In: *Questions of cultural identity* 108.27 (1996).

[FU01] Jonathan Foote and Shingo Uchihashi. “The beat spectrum: A new approach To rhythm analysis”. In: *Proceedings of IEEE International Conference on Multimedia and Expo*. Tokyo, Japan, 2001, pp. 881–884.

[Fuj99] Takuya Fujishima. “Realtime chord recognition of musical sound: A system using common lisp music”. In: *Proceedings of the International Computer Music Conference*. Beijing, China, 1999, pp. 464–467.

[Gab46] Dennis Gabor. “Theory of communication. Part 1: The analysis of information”. In: *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering* 93.26 (1946), pp. 429–441.

[GD05] Fabien Gouyon and Simon Dixon. “A review of automatic rhythm description systems”. In: *Computer Music Journal* 29.1 (2005), pp. 34–54.

[GM09] Peter Grosche and Meinard Muller. “Computing predominant local periodicity information in music recordings”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, 2009, pp. 33–36.

[GM11a] Peter Grosche and Meinard Müller. “Extracting predominant local pulse information from music recordings”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 19.6 (2011), pp. 1688–1701.

[GM11b] Peter Grosche and Meinard Müller. “Tempogram Toolbox: MATLAB tempo and pulse analysis of music recordings”. In: *Late breaking session, 12th International Society for Music Information Retrieval Conference*. Miami, ML, USA, 2011.

[GM90] Brian R. Glasberg and Brian C. J. Moore. “Derivation of auditory filter shapes from notched-noise data”. In: *Hearing research* 47.1 (1990), pp. 103–138.

[GMK10] Peter Grosche, Meinard Müller, and Frank Kurth. “Cyclic tempogram—A mid-level tempo representation for musicsignals”. In: *Proceedings of IEEE*

International Conference on Acoustics, Speech and Signal Processing. Dallas, TX, USA, 2010, pp. 5522–5525.

[Góm06] Emilia Gómez. "Tonal description of polyphonic audio for music content processing". In: *INFORMS Journal on Computing* 18.3 (2006).

[Got+02] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. "RWC Music Database: Popular, Classical and Jazz Music Databases." In: *Proceedings of the 3rd International Society for Music Information Retrieval Conference*. Paris, France, 2002, pp. 287–288.

[Got03] Masataka Goto. "A chorus-section detecting method for musical audio signals". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Hong Kong (cancelled), 2003, pp. 437–440.

[Got06] Masataka Goto. "AIST annotation for the RWC Music Database". In: *Proceedings of the 7th International Society for Music Information Retrieval*. Victoria, Canada, 2006, pp. 359–360.

[Gro+13] Harald Grohganz, Michael Clausen, Nanzhu Jiang, and Meinard Müller. "Converting path structures into block structures using eigenvalue decompositions of self-similarity matrices". In: *Proceedings of the 14th International Society for Music Information Retrieval Conference*. Curtiba, Brazil, 2013, pp. 209–214.

[Gro12] Peter Matthias Grosche. "Signal processing methods for beat tracking, music segmentation, and audio retrieval". PhD thesis. Max-Planck-Institut für Informatik, 2012.

[GS15a] Thomas Grill and Jan Schlüter. "Music boundary detection using neural networks on combined features and two-level annotations". In: *Proceedings of the 16th International Society for Music Information Conference*. Malaga, Spain, 2015.

[GS15b] Thomas Grill and Jan Schlüter. "Music boundary detection using neural networks on spectrograms and self-similarity lag matrices". In: *Proceedings*

of the 16th International Society for Music Information Retrieval Conference. Malaga, Spain, 2015.

[Har78] Fredric J. Harris. "On the use of windows for harmonic analysis with the discrete Fourier transform". In: *Proceedings of the IEEE* 66.1 (1978), pp. 51–83.

[Hel63] Hermann Von Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover, 1863.

[Hol+10] André Holzapfel, Yannis Stylianou, Ali Cenk Gedik, and Baris Bozkurt. "Three dimensions of pitched instrument onset detection". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 18.6 (2010), pp. 1517–1527.

[Hon02] Henkjan Honing. "Structure and interpretation of rhythm and timing". In: *Tijdschrift voor Muziektheorie* 7.3 (2002), pp. 227–232.

[HR88] John Holdsworth and Peter Rice. *Spiral Vos Final Report, Part A: The Auditory Filter bank (Annex C)*. Tech. rep. Cambridge University, 1988.

[HS08] André Holzapfel and Yannis Stylianou. "Beat tracking using group delay based onset detection". In: *Proceedings of the 9th International Society for Music Information Retrieval Conference*. Pennylvania, USA, 2008, pp. 653–658.

[HSG06] Christopher Harte, Mark Sandler, and Martin Gasser. "Detecting harmonic change in musical audio". In: *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. Santa Barara, CA, 2006, pp. 21–26.

[HSL13] Hoon Heo, Dooyong Sung, and Kyogu Lee. "Note onset detection based on harmonic cepstrum regularity". In: *Proceedings of IEEE International Conference on Multimedia and Expo*. San Jose, CA, 2013, pp. 1–6.

[HWW14] Kun Han, Yuxuan Wang, and DeLiang Wang. "Learning spectral mapping for speech dereverberation". In: *Audio, Speech and Language Processing, IEEE Transactions on* 23.6 (2014), pp. 982–992.

[JA03] Kristoffer Jensen and Tue Haste Andersen. "Beat Estimation On the Beat". In: *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* Hong Kong (cancelled), 2003, pp. 87–90.

[JA04] Kristoffer Jensen and Tue Haste Andersen. "Real-Time Beat EstimationUsing Feature Extraction". In: *Proceedings of International Symposium on Computer Music Modeling and Retrieval.* Montpellier, France: Springer Berlin Heidelberg, 2004, pp. 13–22.

[Jen+06] Jesper Hojvang Jensen, Mads Graesboll Christensen, Manohar N Murthi, and Soren Holdt Jensen. "Evaluation of MFCC estimation techniques for music similarity". In: *Signal Processing Conference, 2006 14th European.* 2006.

[Jia+02] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. "Music type classification by spectral contrast feature". In: *Proceedings of IEEE International Conference on Multimedia and Expo.* Naples, Italy, 2002, pp. 113–116.

[Jia+11] Nanzhu Jiang, Peter Grosche, Verena Konz, and Meinard Müller. "Analyzing chroma feature types for automated chord recognition". In: *Proceedings of the Audio Engineering Society International Conference on Semantic Audio.* Illmenau, Germany, 2011.

[Jia15] Nanzhu Jiang. "Repetition-based Structure Analysis of Music Recordings". PhD thesis. Friedrich-Alexander-Universität Erlangen-Nürnberg, 2015.

[Jia95] Jing Jiang. *Zhongguo Qiqu Yinyue.* Renmin Yinyue Chubanshe, 1995.

[Jid81] Lium Jidian. *Jingju Yinyue Gailun.* Renmin Yinyue Chubanshe, 1981.

[KD06] Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription.* Springer, 2006.

[KK83] Gottfried Köthe and Gottfried Köthe. *Topological vector spaces.* Springer, 1983.

[Kla99] Anssi Klapuri. "Sound onset detection by applying psychoacoustic knowledge". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. Phoenix, AZ, 1999, pp. 3089–3092.

[KP04] Emir Kapanci and Avi Pfeffer. "A hierarchical approach to onset detection". In: *Proceedings of the International Computer Music Conference*. Miami, ML, USA, 2004, pp. 438–441.

[KS10] Florian Kaiser and Thomas Sikora. "Music structure discovery in popular music using non-negative matrix factorization". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands, 2010, pp. 429–434.

[LC00] Beth Logan and Stephen Chu. "Music summarization using key phrases". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Istanbul, Turkey, 2000, pp. 749–752.

[LD04] Pierre Leveau and Laurent Daudet. "Methodology and tools for the evaluation of automatic onset detection algorithms in music". In: *Proceedings of the 5th International Society for Music Information Retrieval Conference*. Barcelona, Spain, 2004.

[LE06] Alexandre Lacoste and Douglas Eck. "A supervised classification algorithm for note onset detection". In: *EURASIP Journal on Advances in Signal Processing* 2007.1 (2006).

[Lee06] Kyogu Lee. "Automatic chord recognition from audio using enhanced pitch class profile". In: *Proceedings of the International Computer Music Conference*. New Orleans, USA, 2006, pp. 26–33.

[Liu+09] Yuxiang Liu, Qiaoliang Xiang, Ye Wang, and Lianhong Cai. "Cultural style based music classification of audio signals". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan, 2009, pp. 57–60.

[Liu89] Guojie Liu. *Xipi Erhuang Yinyue Gailun*. Shanghai Yinyue Chubanshe, 1989.

[LK06] W-C Lee and C-CJ Kuo. "Musical onset detection based on adaptive linear prediction". In: *Proceedings of IEEE International Conference on Multimedia and Expo*. Toronto, Ont., 2006, pp. 957–960.

[LNS07] Mark Levy, Katy Noland, and Mark Sandler. "A comparison of timbral and harmonic music segmentation algorithms". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Honolulu, HI, 2007, pp. 1433–1436.

[Log00] Beth Logan. "Mel frequency cepstral coefficients for music modeling". In: *Proceedings of the 1st International Conference on Music Information Retrieval*. Plymouth, Massachusetts, USA, 2000.

[LS06] Mark Levy and Mark B. Sandler. "Lightweight measures for timbral similarity of musical audio". In: *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. Santa Barara, CA, 2006, pp. 27–36.

[LS08] Mark Levy and Mark B. Sandler. "Structural segmentation of musical audio by constrained clustering". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 16.2 (2008), pp. 318–326.

[LS09] Ora Lassila and Ralph R. Swick. *Resource Description Framework (RDF) Model and Syntax Specification*. Availble online: `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`. 2009.

[LS99] Yuan-Yuan Lee and Sinyan Shen. *Chinese Musical Instruments (Chinese Music Monograph Series*. Chinese Music Society of North America Press, 1999.

[LSC06] Mark Levy, Mark B. Sandler, and M. A. Casey. "Extraction of high-level musical structure from audio data and its application to thumbnail generation". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Toulouse, France, 2006, pp. 13–16.

[Luk08] Hanna M Lukashevich. "Towards Quantitative Measures of Evaluating Song Segmentation". In: *Proceedings of the 9th International Society for Music Information Retrieval Conference*. Pennylvania, USA, 2008, pp. 375–380.

[LWW10] Hung-Yi Lo, Ju-Chiang Wang, and Hsin-Min Wang. "Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval". In: *Proceedings of IEEE International Conference on Multimedia and Expo*. Suntec City, 2010, pp. 304–309.

[Lyo96] Richard F. Lyon. "The all-pole Gammatone filter and Auditory models". In: *Proceedings of Forum Acusticum*. 1996.

[Mac03] David J. C. MacKay. "Information Theory, Inference, and Learning Algorithms Information theory, inference, and learning algorithms". In: 4th ed. Cambridge University Press, 2003. Chap. 4.

[Mad+04] Namunu C. Maddage, Changsheng Xu, Mohan S Kankanhalli, and Xi Shao. "Content-based music structure analysis with applications to music semantics understanding". In: *Proceedings of the 12th annual ACM international conference on Multimedia*. New York, NY, 2004, pp. 112–119.

[Mar+14] Eric Marchi, Giacomo Ferroni, Florian Eyben, Leonardo Gabrielli, Stefano Squartini, and Björn Schuller. "Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy, 2014, pp. 2164–2168.

[Mas96] Paul Masri. "Computer modelling of sound for transformation and synthesis of musical signals". PhD thesis. University of Bristol, 1996.

[Mau+09] Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Christopher Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler. "OMRAS2 metadata project 2009". In: *Late breaking session, 10th International Society for Music Information Retrieval Conference*. Kobe, Japan, 2009.

[Mau10] Matthias Mauch. "Automatic chord transcription from audio using computational models of musical context". PhD thesis. Queen Mary University of London, 2010.

[MB03] Martin F McKinney and Jeroen Breebaart. "Features for audio and music classification". In: *Proceedings of the 4th International Soviety for Music*

Information Retrieval Conference. Washington, D.C., USA, 2003, pp. 151–158.

[MB96] Paul Masri and Andrew Bateman. "Improved modelling of attack transients in music analysis-resynthesis". In: *International Computer Music Conference*. Singapore, 1996, pp. 100–103.

[MBH12] Hassan MLA Ezzaidi, Mohammed Bahoura, and Glenn Eric Hall. "Towards a characterization of musical timbre based on chroma contours". In: *Proceedings of International Conference on Advanced Machine Learning Technologies and Applications*. Cairo, Egypt, 2012, pp. 162–171.

[McD09] John H McDonald. *Handbook of biological statistics*. Baltimore, MD: Sparky House Publishing, 2009.

[McF+15] Brian McFee, Matt McVicar, Colin Raffel, Dawen Liang, Oriol Nieto, Eric Battenberg, Josh Moore, Dan Ellis, Ryuichi YAMAMOTO, Rachel Bittner, Douglas Repetto, Petr Viktorin, João Felipe Santos, and Adrian Holovaty. "Librosa: Python library for audio and music analysis". In: *Proceedings of the 14th Python in Science Conference*. Austin, Texas, 2015.

[McK+05] Cory McKay, Rebecca Fiebrink, Daniel McEnnis, Beinan Li, and Ichiro Fujinaga. "ACE: A framework for optimizing music classification". In: *Proceedings of the 6th International Society on Music Information Retrieval Conference*. London, UK, 2005, pp. 42–49.

[MD10] Matthias Mauch and Simon Dixon. "Approximate Note Transcription for the Improved Identification of Difficult Chords". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands, 2010, pp. 135–140.

[ME10a] Meinard Müller and Sebastian Ewert. "Towards timbre-invariant audio features for harmony-based music". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 18.3 (2010), pp. 649–662.

[ME10b] Meinard Müller and Sebastian Ewert. "Towards timbre-invariant audio features for harmony-based music". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 18.3 (2010), pp. 649–662.

[ME14a] Brian McFee and Daniel P. W. Ellis. "Analyzing song structure with spectral clustering". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014, pp. 205–410.

[ME14b] Brian McFee and Daniel P. W. Ellis. "Learning to segment songs with ordinal linear discriminant analysis". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy, 2014, pp. 5197–5201.

[MEK09] Meinard Müller, Sebastian Ewert, and Sebastian Kreuzer. "Making chroma features more robust to timbre changes." In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan, 2009, pp. 1877–1880.

[MG83] Brian C. J. Moore and Brian R. Glasberg. "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". In: *The Journal of the Acoustical Society of America* 74.3 (1983), pp. 750–753.

[MGB97] Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer. "A model for the prediction of thresholds, loudness, and partial loudness". In: *Journal of the Audio Engineering Society* 45.4 (1997), pp. 224–240.

[MIR] MIREX. *MIREX2013 Results*. URL: `http://www.music-ir.org/mirex/wiki/2013:MIREX2013_Results` (visited on 2013).

[MIR14] MIREX. *MIREX2014 Results*. 2014. URL: `http://www.music-ir.org/mirex/wiki/2014:MIREX2014_Results` (visited on 2014).

[MIR15a] MIREX. *Audio onset detection*. 2015. URL: `http://www.music-ir.org/mirex/wiki/2015:Audio_Onset_Detection`.

[MIR15b] MIREX. *MIREX2015 Results*. 2015. URL: `http://www.music-ir.org/mirex/wiki/2015:MIREX2015_Results`.

[MIR15c] MIREX. *Music structural segmentation*. 2015. URL: `http://www.music-ir.org/mirex/wiki/2015:Structural_Segmentation`.

[MJG14] Meinard Müller, Nanzhu Jiang, and Harald Grohganz. "SM Toolbox: MATLAB Implementations for Computing and Enhancing Similarity Matrices". In: *Proceedings of Audio Engineering Society Conference on Semantic Audio*. London, UK, 2014.

[MNB15] Brian McFee, Oriol Nieto, and Juan P. Bello. "Hierarchical evaluation of segment boundary detection". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference*. Malaga, Spain, 2015.

[MND09] Matthias Mauch, Katy Noland, and Simon Dixon. "Using musical structure to enhance automatic chord transcription". In: *Proceedings of 10th International Society for Music Information Retrieval Conference*. Kobe, Japan, 2009, pp. 231–236.

[Moo12] Brian C. J. Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[Nar91] Eugene Narmour. "The top-down and bottom-up systems of musical implication: Building on Meyer's theory of emotional syntax". In: *Music Perception* 9.1 (1991), pp. 1–26.

[NB14] Oriol Nieto and Juan Pablo Bello. "Music Segment Similarity Using 2D-Fourier Magnitude Coefficients". In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy, 2014, pp. 664–668.

[NB15] Oriol Nieto and Juan P. Bello. "MSAF: Music Structure Analytis Framework". In: *Late breaking session, 16th International Society for Music Information Retrieval Conference*. Malaga, Spain, 2015.

[Nie+14] Oriol Nieto, Morwaread M Farbood, Tristan Jehan, and Juan Pablo Bello. "Perceptual analysis of the F-measure for evaluating section boundaries in music". In: *15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014, pp. 265–270.

[Nie15] Oriol Nieto. "Discovering Structure in Music: Automatic Approaches and Perceptual Evaluations." PhD thesis. New York University, 2015.

[NJ13] Oriol Nieto and Tristan Jehan. "Convex non-negative matrix factorization for automatic music structure identification". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, 2013, pp. 236–240.

[NP78] Madihally J. Narasimha and Allen M. Peterson. "On the computation of the Discrete cosine transform". In: *Communications, IEEE Transactions on* 26.6 (1978), pp. 834–936.

[NS07] Katy Noland and Mark Sandler. "Signal processing parameters for tonality estimation". In: *Proceedings of Audio Engineering Society Convention 122*. vienna, Austria, 2007.

[NS12] Michael J. Newton and Leslie S. Smith. "A neurally inspired musical instrument classification system based upon the sound onset". In: *The Journal of the Acoustical Society of America* 131.6 (2012), pp. 4785–4798.

[OG91] Jerome D. Oremlad and Merton M. Gill, eds. *Interpretation and interaction: psychoanalysis or psychotherapy?* The Analytic Press, Inc, 1991.

[OGS06] Bee Suan Ong, Emilia Gómez, and Sebastian Streich. "Automatic extraction of musical structure using pitch class distribution features". In: *Proceedings of Workshop on learning the semantics of audio signals*. 2006, pp. 53–65.

[OH05] Bee Suan Ong and Perfecto Herrera. "Semantic segmentation of music audio contents". In: *Proceedings of the International Computer Music Conference*. Barcelona, Spain, 2005, pp. 61–64.

[Ong06] Bee Suan Ong. "Structural analysis and segmentation of music signals". PhD thesis. Universitat Pompeu Fabra, 2006.

[Onl09] Online. *The Vamp Plugin Ontology*. 2009. URL: `http://omras2.org/VampOntology`.

[Onl16] Online. *Peking opera*. 2016. URL: `https://en.wikipedia.org/wiki/Peking_opera`.

[Par94] Richard Parncutt. "A perceptual model of pulse salience and metrical accent in musical rhythms". In: *Music Perception* 11 (1994), pp. 409–464.

[Pat+87] RD Patterson, Ian Nimmo-Smith, John Holdsworth, and Peter Rice. *An efficient auditory filterbank based on the gammatone function.* Tech. rep. IOC Speech Group on Auditory Modelling at RSRE, 1987.

[Pat87] Roy D. Patterson. "A pulse ribbon model of monaural phase perception". In: *The Journal of the Acoustical Society of America* 85.2 (1987), pp. 1560–1586.

[PB14] Geoffroy Peeters and Victor Bisot. "Improving Music Structural Segmentation Using Lag-priors". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference.* Taipei, Taiwan, 2014, pp. 337–342.

[PD09] Geoffroy Peeters and Emmanuel Deruty. "Is music structure annotation multi-dimensional? A proposal for robust local music annotational music annotation". In: *Proceedings of the 10th International Society for Music Information Retrieval Conference.* Kobe, Japan, 2009, pp. 337–342.

[Pee03] Geoffroy Peeters. "Deriving musical structures from signal analysis for music audio summary generation:"Sequence" and "State" approach". In: *Computer Music Modeling and Retrieval.* Singapore, 2003, pp. 143–166.

[Pee04] Geoffroy Peeters. "A large set of audio features for sound description (similarity and classification) in the CUIDADO project". In: *Rapport technique, CUIDADO I.S.T.* (2004).

[Pee05] Geoffroy Peeters. "Time variable tempo detection and beat marking". In: *Proceedings of the International Computer Music Conference.* Barcelona, Spain, 2005.

[Pee07] Geoffroy Peeters. "Sequence Representation of Music Structure Using Higher-Order Similarity Matrix and Maximum-Likelihood Approach". In: *Proceedings of the 8th International Society for Music Information Retrieval Conference.* Vienna, Austria, 2007, pp. 35–40.

[PK06] Jouni Paulus and Anssi Klapuri. "Music structure analysis by finding repeated parts". In: *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. Santa Barara, CA, 2006, pp. 59–68.

[PK08] Jouni Paulus and Anssi Klapuri. "Labelling the structural parts of a music piece with Markov models". In: *Proceedings of Computers in Music Modeling and Retrieval Conference (CMMR)*. 2008.

[PK09] Jouni Paulus and Anssi Klapuri. "Music structure analysis using a probabilistic fitness measure and a greedy search algorithm". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 17.6 (2009), pp. 1159–1170.

[PLR02] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. "Toward Automatic Music Audio Summary Generation from Signal Analysis". In: *Proceedings of the 3rd International Society for Music Information Retrieval Conference*. Paris, France, 2002, pp. 1–7.

[Plu06] Queen Mary Vamp Plugins. *Queen Mary Vamp Plugins*. 2006. URL: `http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html` (visited on 2006).

[Plu07] Vamp Plugins. *The Vamp audio analysis plugin system*. 2007. URL: `http://vamp-plugins.org/`.

[PMK10] Jouni Paulus, Meinard Müller, and Anssi Klapuri. "State of the art report: audio-based music structure analysis". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands, 2010, pp. 625–636.

[Pol00] Allan Pollack. *Alan W. Pollack's 'Notes On' Series*. 2000. URL: `http://www.icce.rug.nl/~soundscapes/DATABASES/AWP/awp-notes_on.shtml`.

[PP07] Bhanu Prasad and S. R. Mahadeva Prasanna. *Speech, audio, image and biomedical signal processing using neural networks*. Springer, 2007.

[PP13] Kaiser Pauwels J. and Geoffroy Peeters. "Combining harmony based and novelty based approaches for structural segmentation". In: *Proceedings of*

the 14th International Society for Music Information Retrieval Conference. Curtiba, Brazil, 2013,

[Qi+13] Jun Qi, Dong Wang, Yi Jiang, and Runsheng Liu. "Auditory features based on gammatone filters for robust speech recognition". In: *Proceedings of IEEE International Symposium on Circuits and Systems*. Beijing, China, 2013, pp. 305–308.

[Rai07] Yves Raimond. *The audio feature ontology*. 2007. URL: `http://motools.sourceforge.net/doc/audio_features.html`.

[Ras79] Rudolf A. Rasch. "Synchronization in performed ensemble music". In: *Acustica* 43.2 (1979), pp. 121–131.

[RBH13] Bruno Rocha, Niels Bogaards, and Aline Honingh. "Segmentation and Timbre Similarity in Electronic Dance Music". In: *Proceedings of the Sound and Music Computing Conference*. Stockholm, Sweden, 2013.

[RC07] Christophe Rhodes and Michael Casey. "Algorithms for determining and labelling approximate hierarchical self-similarity." In: *Proceedings of the 8th International Society for Music Information Retrieval Conference*. Vienna, Austria, 2007, pp. 41–46.

[Reg13] Motti Regev. *Pop-Rock music: Aesthetic cosmopolitanism in late modernity*. John Wiley & Sons, 2013.

[Rep+14] Rafael Caro Repetto, Ajay Srinivasamurthy, Sankalp Gulati, and Xavier Serra. *Jingju music: concepts and computational tools for its analysis*. Tech. rep. Taipei, Taiwan: Tutorial session, International Socity for Music Information Retrieval Conference, 2014.

[Rep+15] Rafael Caro Repetto, Rong Gong, Nadine Kroher, and Xavier Serra. "Comparison of the singing style of two Jingju schools". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference*. Malaga, Spain, 2015.

[Rho+06] Christophe Rhodes, Michael Casey, Samer Abdallah, and Mark Sandler. "A Markov-chain Monte-Carlo approach to musical audio segmentation".

In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Toulouse, France, 2006.

[Rod01] Xavier Rodet. *Project Ecrins: calcul des descripteur de bas-niveaux*. Tech. rep. Ircam, 2001.

[Ros+99] Stéphane Rossignol, Xavier Rodet, Joël Soumagne, J-L Collette, and Philippe Depalle. "Automatic characterisation of musical signals: Feature extraction and temporal segmentation". In: *Journal of new music research* 28.4 (1999), pp. 281–295.

[Ros07] Thomas Rossing. *Handbook of Acoustics*. Springer, 2007.

[RRM12] Carlos Rosão, Ricardo Ribeiro, and David Martins de Matos. "Influence of peak selection methods on onset detection". In: *Proceedings of the 13th International Society for Music Information Retrieval Conference*. Porto, Portugal, 2012, pp. 517–522.

[RS14] Rafael Caro Repetto and Xavier Serra. "Creating a corpus of Jingju (Beijing Opera) music and possibilities for melodic analysis". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014.

[Rum+10] Halfdan Rump, Shigeki Miyabe, Emiru Tsunoo, and Nobutaka Ono. "Autoregressive MFCC models for genre classification improved by harmonic-percussion separation". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands, 2010, pp. 87–92.

[Sam06] Hanan Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[San10] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

[Sap05] Craig Stuart Sapp. *Tempo change JND experiment, Mazurka Project*. 2005. URL: `http://www.mazurka.org.uk/experiments/tempojnd/`.

[SB03] Paris Smaragdis and Judith C Brown. "Non-negative matrix factorization for polyphonic music transcription". In: *Proceedings of IEEE Workshop on*

Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, 2003, pp. 177–180.

[SB13] Jan Schlüter and Sebastian Böck. "Musical onset detection with convolutional neural networks". In: *Proceedings of the 6th International Workshop on Machine Learning and Music*. Prague, Czech Republic, 2013.

[SBV11] Gabriel Sargent, Frédéric Bimbot, and Emmanuel Vincent. "A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, ML, USA, 2011.

[SC13a] Jordan B. L. Smith and Elaine Chew. "A meta-analysis of the MIREX structure segmentation task". In: *Proceedings of the 14th International Society for Music Information Retrieval Conference*. Canada, 2013, pp. 251–256.

[SC13b] Jordan B. L. Smith and Elaine Chew. "Using quadratic programming to estimate feature relevance in structural analyses of music". In: *21st ACM international conference on Multimedia*. ACM. 2013, pp. 113–122.

[Sch+07] R. Schlüter, L. Bezrukov, H. Wagner, and H. Ney. "Gammatone features and feature combination for large vocabulary speech recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Honolulu, HI, 2007, pp. 649–652.

[SDP09] Adam M Stark, Matthew EP Davies, and Mark D Plumbley. "Real-time beat-synchronous analysis of musical audio". In: *Proceedings of the 12th International Conference on Digital Audio Effects*. Como, Italy, 2009.

[Sel11] Howard J Seltman. *Experimental design and analysis*. Online, 2011. URL: `http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf`.

[Ser+12] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Lluis Arcos. "Unsupervised detection of music boundaries by time series structure features". In: *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Toronto, Ont., 2012, pp. 1613–1619.

[Ser11] Xavier Serra. "A multicultural approach in music information research". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, ML, USA, 2011.

[Ser12a] X. Serra. "Opportunities for a Cultural Specific Approach in the Computational Description of Music". In: *2nd CompMusic Workshop*. Istanbul, Turkey, July 2012. ISBN: 978-84-695-4958-2. URL: `http://mtg.upf.edu/system/files/publications/1-Xavier-Serra-2nd-CompMusic-Workshop-2012_0.pdf`.

[Ser12b] Xavier Serra. *Computational Models for the Discovery of the World Music*. http://compmusic.upf.edu/node/2. 2012.

[SF16] Tony C. Smith and Eibe Frank. "Statistical Genomics: Methods and Protocols". In: New York, NY: Springer, 2016. Chap. Introducing Machine Learning Concepts with WEKA, pp. 353–378.

[SG12] Justin Salamon and Emilia Gómez. "Melody extraction from polyphonic music signals using pitch contour characteristics". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20.6 (2012), pp. 1759–1770.

[SGU14] Markus Schedl, Emilia Gómez, and Julián Urbano. "Music information retrieval: recent developments and applications". In: *Foundations and Trends in Information Retrieval* 8.2-3 (2014), pp. 127–261.

[Sha+09] Yang Shao, Zhaozhang Jin, DeLiang Wang, and Soundararajan Srinivasan. "An auditory-based feature for robust speech recognition". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan, 2009, pp. 4625–4628.

[SJK05] Yu Shiu, Hong Jeong, and C. -C. Jay Kuo. "Musical structure analysis using similarity matrix and dynamic programming". In: *Proceedings of International Society for Optics and Photonics*. 2005, p. 601516.

[Sla93] Malcolm Slaney. *An efficient implementation of the Patterson-Holdsworth auditory filter bank*. Tech. rep. Apple Technical Report, 1993.

[Smi+11] Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. "Design and creation of a large-scale database of structural annotations". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, ML, USA, 2011, pp. 555–560.

[Smi10] Jordan B. L. Smith. "A comparison and evaluation of approaches to the automatic formal analysis of musical audio". MA thesis. McGill University, 2010.

[Smi14] Jordan B. L. Smith. "Explaining Listener Differences in the Perception of Musical Structure". PhD thesis. Queen Mary, University of London, 2014.

[SP07] Dan Stowell and Mark Plumbley. "Adaptive whitening for improved real-time audio onset detection". In: *Proceedings of International Computer Music Conference*. Copenhagen, Denmark, 2007.

[Sri+14] Ajay Srinivasamurthy, Rafael Caro Repetto, Harshavardhan Sundar, and Xavier Serra. "Transcription and recognition of syllable based percussion patterns: the case of Beijing opera". In: *Proceedings of the 15th Society for Music Information Retrieval*. Taipei, Taiwan, 2014, pp. 431–436.

[SSC14] Jordan B. L Smith, Isaac Schankler, and Elaine Chew. "Listening as a creative act: Meaningful differences in structural annotations of improvised performances". In: *Music Theory Online* 20.3 (2014).

[Ste99a] Charlie Steinberg. *Steinberg virtual studio technology plug-In specification 2.0, Software development kit*. Tech. rep. Steinberg Media Technologies AG, 1999. URL: `http://jvstwrapper.sourceforge.net/vst20spec.pdf`.

[Ste99b] Richard Stevens. *UNIX network programming: Interprocess communications*. 2nd. Vol. 2. Prentice Hall, 1999.

[Sto99] Jonathan P. J. Stock. "A reassessment of the relationship between text, speech tone, melody, and aria structure in Beijing Opera". In: *Journal of Musicological Research* 18.3 (1999), pp. 183–206.

[Sun+12] Johan Sundberg, Lide Gu, Qiang Huang, and Ping Huang. "Acoustical study of classical Peking Opera singing". In: *Journal of Voice* 26.2 (2012), pp. 137–143.

[SVN37] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. "A scale for the measurement of the psychological magnitude pitch". In: *The Journal of the Acoustical Society of America* 8.3 (1937), pp. 185–190.

[SW70] Percy Alfred Scholes and John Owen Ward. *The Oxford Companion to Music*. Oxford University Press, 1970.

[Swi97] Andrew Swift. *A brief Introduction to MIDI*. 1997. (Visited on 1997).

[SWW08] Malcolm Slaney, Kilian Weinberger, and William White. "Learning a metric for music similarity". In: *Proceedings of the 9th International Society for Music Information Retrieval Conference*. Philadelphia, PA, 2008.

[Tia+13] Mi Tian, György Fazekas, Dawn A. A. Black, and Mark B. Sandler. "Towards the representation of Chinese traditional music: A state of the art review of music metadata standards". In: *Proceedings of International Conference on Dublin Core and Metadata Applications (DC-13)*. Lisbon, Portugo, 2013, pp. 71–81.

[Tia+14a] Mi Tian, György Fazekas, Dawn A. A. Black A. A. Black, and Mark Sandler. "Design and evaluation of onset detectors using different fusion policies". In: *Proceedinds of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014.

[Tia+14b] Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. "A study of instrument-wise onset detection in Beijing opera percussion ensembles". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy, 2014, pp. 2159–2163.

[Tia+15] Mi Tian, György Fazekas, Dawn A. A. Black, and Mark Sandler. "On the use of tempogram to describe audio content and its application to music structural segmentation". In: *Proceedings of IEEE International Confer-*

ence on Acoustics, Speech and Signal Processing. Brisbane, Australia, 2015, pp. 256–260.

[Tra90] Harmut Traunmüller. "Analytical expressions for the tonotopic sensory scale". In: *The Journal of the Acoustical Society of America* 88.1 (1990), pp. 97–100.

[Tra97] Harmut Traunmüller. *Auditory scales of frequency representation*. 1997. URL: `http://www2.ling.su.se/staff/hartmut/bark.htm`.

[TS16a] Mi Tian and Mark B. Sandler. "Music structural segmentation across genres with Gammatone features". In: *Proceedings of the 17th International Conferece on Music Information Retrieval*. New York, NY, 2016.

[TS16b] Mi Tian and Mark B. Sandler. "Towards Music Structural Segmentation Across Genres: Features, Structural Hypotheses and Annotation Principles". In: *Special Issue on Intelligent Music Systems and Applications, Intelligent Systems and Technology, ACM Transactions on.* 8.2 (2016), pp. 23–41.

[TSB05] Hiroko Terasawa, Malcolm Slaney, and Jonathan Berger. "The thirteen colors of timbre". In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, 2005, pp. 625–636.

[Tur+07] Douglas Turnbull, Gert RG Lanckriet, Elias Pampalk, and Masataka Goto. "A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting". In: *Proceedings of the 8th International Society for Music Information Retrieval Conference*. Vienna, Austria, 2007, pp. 51–54.

[TZW08] Chee-Chuan Toh, Bingjun Zhang, and Ye Wang. "Multiple-feature fusion based onset detection for solo singing voice". In: *Proceedings of the 9th International Society for Music Information Retrieval Conference*. Philadelphia, PA, 2008, pp. 515–520.

[UC08] Graham Upton and Ian Cook. *A dictionary of statistics (Oxford quick reference)*. 2nd. Oxford University Press, 2008.

[Urb+12] Julián Urbano, Brian McFee, J. Stephen Downie, and Markus Schedl. "How significant is statistically significant? The case of audio music similarity

and retrieval". In: *Proceedings of the 13th International Society for Music Information Retrieval Conference*. Porto, Portugal, 2012, pp. 181–186.

[USG14] Karen Ullrich, Jan Schlüter, and Thomas Grill. "Boundary Detection in Music Structure Analysis using Convolutional Neural Networks". In: *proceeding of the 15th International Society for Music Information Retrieval*. Taipei, Taiwan, 2014, pp. 417–422.

[VA12] Xavier Valero and Francesc Alías. "Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification". In: *Multimedia, IEEE Transactions on* 14.6 (2012), pp. 1684–1689.

[Vas08] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. Wiley, 2008.

[VF15] Ralf Von Appen and Markus Frei-Hauenschild. "ABBA, refrain, chorus, ridge, prechorus – song forms and their historical development". In: *Samples. Online-Publikationen der Gesellschaft für Popularmusikforschung* (2015).

[Wan+06] Wenwu Wang, Yuhui Luo, Jonathon Chambers, and Saeid Sanei. "Non-negative matrix factorization for note onset detection of audio signals". In: *Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*. Arlingto, VA, 2006, pp. 447–452.

[WB10] Ron J. Weiss and Juan P. Bello. "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands, 2010.

[Wic91] Elizabeth Wichmann. *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press, 1991.

[Wil12] Gareth Williams. *Linear algebra with applications*. 8th ed. Jones & Bartlett Publishers, 2012.

[WL06] Claus Weihs and Uwe Ligges. "Parameter optimization in automatic transcription of music". In: Springer, 2006.

[WM15] Christof Weiss and Meinard Müller. "Tonal complexity features for style classification of classical music". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Brisbane, Australia, 2015, pp. 688–692.

[WM16] Cheng-i Wang and Gaurham J. Mysore. "Structural Segmentation with the Variable Markov Oracle and Boundary Adjustment". In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Shanghai, China, 2016, pp. 291–296.

[XMK06] Changsheng Xu, Namunu C. Maddage, and Mohan S. Kankanhalli. "Automatic structure detection for popular music". In: *Multimedia, IEEE Transactions on* 13.1 (2006), pp. 65–77.

[YTC15] Luwei Yang, Mi Tian, and Elaine Chew. "Vibrato characteristics and frequency histogram envelopes in Beijing opera singing". In: *Proceedings of the 5th International Workshop on Folk Music Analysis*. 2015, pp. 139–140.

[ZB14] Tang Zheng and Dawn A. A. Black. "Melody extraction from polyphonic audio of Western opera: a method based on detection of the singer's formant". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014, pp. 161–166.

[ZCS15] Shuo Zhang, Rafael Caro Repetto, and Xavier Serra. "Predicting pairwise pitch contour relations based on linguistic tone information in Beijing opera singing". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference*. Malaga, Spain, 2015, pp. 107–113.

[ZG11] Jose R. Zapata and Emilia Gómez. "Comparative evaluation and combination of audio tempo estimation approaches". In: *Proceedings of Audio Engineering Society Conference on Semantic Audio*. Illmenau, Germany, 2011.

[ZG13] Jose R. Zapata and Emilia Gómez. "Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals". In:

Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE. Vancouver, BC, 2013, pp. 51–55.

[Zha+13] Zhengchen Zhang, Dong-yan Huang, Renbo Zhao, and Minghui Dong. "Onset detection based on fusion of Simpls and SuperFlux". In: *MIREX Results*. 2013.

[ZMZ08] Ruohua Zhou, Marco Mattavelli, and Giorgio Zoia. "Music onset detection based on resonator time frequency image". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 16.8 (2008), pp. 1685–1695.

[ZRS14] Shuo Zhang, Rafael Caro Repetto, and Xavier Serra. "Study of the similarity between linguistic tones and melodic pitch contours in Beijing opera singing". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014, pp. 343–348.

[Zwi90] Eberhard Zwicker. *Psychoacoustics*. Berlin, Springer-Verlag, 1990.

[ZZ03] Yibin Zhang and Jie Zhou. "A study on content-based music classification". In: *Proceedings of IEEE International Symposium on Signal Processing and Its Applications*. Paris, France, 2003, pp. 113–116.