Computer Science
Department
University of Crete

University of Crete

A thesis submitted for the degree of
Doctor of Philosophy

Similarity methods for computational ethnomusicology

Author:
André Holzapfel

Supervisor:
Yannis Stylianou

March 17, 2010

Abstract

The field of computational ethnomusicology has drawn growing attention by researchers in the music information retrieval community. In general, subjects are considered that are related to the processing of traditional forms of music, often with the goal to support studies in the field of musicology with computational means.

Tools have been proposed that make access to large digital collections of traditional music easier, for example by automatically detecting a specific kind of similarity between pieces or by automatically segmenting data into partitions that are either relevant or irrelevant for further investigation.

In this thesis, the focus lies on music of the Eastern Mediterranean, and specifically on traditional music of Greece and Turkey. At the beginning of the thesis related work, the task was defined which directed the aspects of the necessary research activities.

The task was motivated by the geographical location of the author, the island of Crete in Greece, but in the course of the thesis this task proved to have strong relevance for a much wider musical context: Given a polyphonic recording of a piece of Cretan traditional dance music, find a recording that is similar to it. Theory of musicology provided us with the way to approach this task.

The traditional music encountered in Greece and in wide parts of the Balkan states and Turkey as well, follows the logic of *parataxis*, which means that pieces are constructed by temporally aligning short musical phrases, without the existence of structures present in classical music or popular music. Thus, a system that is designed to cope with the above mentioned task has to be able to estimate the similarity of such phrases. As we deal with polyphonic audio signals of music that has not been written to a score, at least not before the performance, we need to do some simplification.

This is because the exact transcription of the main melody from a polyphonic mixture into a score is still an unsolved problem. And on the other side, the transcription of traditional music even by human experts is an extremely complex and difficult process. For that reason, a system has been designed that considers aspects of rhythm, timbre and melody for approaching the task.

The central aspect that has been considered in this thesis is rhythm. For this, a point of major interest is the estimation at which time instances within an audio signal a musical instrument starts playing a note. This estimation is referred to as onset detection, and has been approached in this thesis using novel group delay and fundamental frequency based approaches, and with a fusion of these characteristics with an spectral amplitude criterion. With these findings in the field of onset detection, improved beat trackers and rhythmic similarity estimation techniques are developed. The proposed beat tracker applies the group delay based onset detection method in the context of a state-of-the-art approach for beat tracking. Results show clear improvements when applying this method for beat tracking on a dataset of traditional music.

The rhythmic similarity estimation is based on scale transformation, which avoids the influence of tempo differences between pieces of music that are to be compared. On datasets containing Greek and Turkish traditional music high accuracies in a classification task are achieved, and the validity of the proposed measure as a similarity measure is supported by the results of listening tests.

Apart from rhythm, also the aspect of instrumental timbre has been addressed. A novel feature set based on Non-negative Matrix Factorization (NMF) is proposed to describe the characteristic spectral bases of a piece of music. These bases are modelled using statistical methods, and it is shown that these models describe the spectral space of musical genres and instrumental classes in a compact and discriminative way.

Finally, melodic aspects have been considered as well by combining state-of-the-art approaches for cover song detection in popular music and fundamental frequency detection from polyphonic signals. This combination is shown to tackle the central task of the thesis work in a satisfying way on a small exemplary dataset. A morphological analysis framework that combines the aspects of rhythm, timbre and melody is proposed, which can be used to detect similarities in traditional music.

For the development of the algorithms presented in this thesis, evaluation data had to be collected. This was a task of major difficulty and much effort has been made by the author to understand well the musical context that is investigated in this thesis. For many datasets, the ground truth was achieved in cooperation with local musicians in time-consuming but very informative interviews. The knowledge obtained in these interviews and the resulting datasets are another important contribution of this thesis.

Περίληψη

Ο τομέας της Ψπολογιστικής Εθνομουσικολογίας έχει προσελκύσει την προσοχή των ερευνητών που δραστηριοποιούνται στην περιοχή της Ανάκτησης Μουσικής Πληροφορίας. Στην Υπολογιστική Εθνομουσικολογία εξετάζονται θέματα που συσχετίζονται με την επεξεργασία παραδοσιακής μουσικής, συχνά με το στόχο για να υποστηρίξουν τις μελέτες στον τομέα της μουσικολογίας με υπολογιστικά μέσα.
Ο στόχος της παρούσας διατριβής είναι να ορίσει την ομοιότητα μεταξύ μουσικών κομματιών. Για παράδειγμα, λαμβάνοντας υπόψη μια πολυφωνική καταγραφή ενός χορευτικού κομματιού της κρητικής παραδοσιακής μουσικής, ψάχνουμε μια καταγραφή που είναι παρόμοια με αυτήν. Η εστίαση βρίσκεται στη μουσική της ανατολικής Μεσογείου, και συγκεκριμένα στην παραδοσιακή μουσική της Ελλάδας και της Τουρκίας. Θεωρούμε ότι η ομοιότητα περιγράφεται από τρεις διαστάσεις: ρυθμό, χροιά και μελωδία.
Η θεωρία της μουσικολογίας παρέχει τον τρόπο να πλησιάσουμε αυτό το στόχο. Η παραδοσιακή μουσική που συναντάται στην Ελλάδα, στην ευρύτερη Βαλκανική χερσόνησο καθώς και στην Τουρκία, ακολουθεί σε πολλές περιπτώσεις τη λογική της παράταξης. Αυτό σημαίνει ότι τα κομμάτια κατασκευάζονται με σύντομες μουσικές φράσεις, χωρίς την ύπαρξη δομών όπως συμβαίνει στην κλασική μουσική ή στο Pop/Rock. Κατά συνέπεια, ένα σύστημα που σχεδιάζεται για να αντιμετωπίσει τον παραπάνω στόχο στην παραδοσιακή μουσική πρέπει να είναι σε θέση να υπολογίσει την ομοιότητα τέτοιων φράσεων. Δεδομένου ότι εξετάζουμε πολυφωνικά ακουστικά σήματα μουσικής που δεν έχουν καταηραφεί σε παρτιτούρα, τουλάχιστον όχι πριν από την ερμηνεία, πρέπει να κάνουμε κάποια απλοποίηση, λαμβάνοντας υπόψη ότι η ακριβής καταγραφή της κύριας μελωδίας από μία πολυφωνική μίξη είναι ακόμα ένα άλυτο πρόβλημα. Από την άλλη πλευρά, η καταγραφή σε παρτιτούρα της παραδοσιακής μουσικής, ακόμη και από εμπειρογνώμονες, είναι μια εξαιρετικά σύνθετη και δύσκολη διαδικασία. Λόγω των παραπάνω δυσκολιών, στην παρούσα διατριβή προτείνεται ένα σύστημα αυτόματης εκτίμησης ρυθμού, χροιάς και μελωδίας, ώστε στη συνέχεια να μπορεί να ορισθεί και να ελεγχθεί η έννοια της ομοιότητας μεταξύ των μουσικών καταγραφών.

Για την εκτίμηση του ρυθμού είναι ιδιαίτερα σημαντική η εκτίμηση των στιγμών στις οποίες ξεκινάει μία νότα (onset detection). Στην εργασία αυτή προτείνεται μια πρωτότυπη τεχνική ανίχνευσης έναρξης χρησιμοποιώντας καθυστέρηση ομάδας και θεμελιώδη συχνότητα, ενώ εξετάζονται θέματα συγχώνευσης αυτών των χαρακτηριστικών με χαρακτηριστικά ενέργειας (φάσμα πλάτους). Με αυτά τα συμπεράσματα στον τομέα της ανίχνευσης έναρξης, βελτιωμένη παρακολούθηση ρυθμού και τεχνικές εκτίμησης ρυθμικής ομοιότητας αναπτύσσονται.

Σχετικά με τη χροιά, προτείνεται ένα νέο σύνολο χαρακτηριστικών βασισμένο στη παραγοντοποίηση με μη αρνητικούς πινάκες (Non negative Matrix Factorization, NMF) για να περιγράψει τις χαρακτηριστικές φασματικές βάσεις ενός δείγματος. Αυτές οι βάσεις μοντελοποιούνται χρησιμοποιώντας στατιστικές μεθόδους, και αποδεικνύεται ότι αυτά τα πρότυπα περιγράφουν το φασματικό διάστημα των μουσικών ειδών και των κατηγοριών των μουσικών οργάνων με έναν συμπαγή και χαρακτηριστικό τρόπο.

Η εκτίμηση της μελωδίας έχει εξεταστεί επίσης με το συνδυασμό προσεγγίσεων που έχουν προταθεί για την ανίχνευση τραγουδιού ερμηνείας (Cover Song Detection) και την ανίχνευση συχνότητας από πολυφωνικά σήματα. Τέλος, ένα μορφολογικό πλαίσιο ανάλυσης που συνδυάζει το ρυθμό, τη χροιά και τη μελωδία, προτείνεται, το οποίο μπορεί να χρησιμοποιηθεί για να ανιχνεύσει τις ομοιότητες στην παραδοσιακή μουσική.

Για το σχεδιασμό, την ανάπτυξη, και το έλεγχο των αλγορίθμων που παρουσιάζονται σε αυτή τη διατριβή, αρκετά μουσικά δεδομένα έπρεπε να συλλεχθούν και να προ-επεξεργαστούν (π.χ. ετικετοποίηση). Οι βάσεις δεδομένων που παράχθηκαν αποτελούν επίσης μια σημαντική συμβολή της εργασίας στη μελέτη της παραδοσιακής μουσικής.

Acknowledgements

I would like to thank my supervisor Yannis Stylianou for his continuous support. He enabled me to take big steps towards relevant work in the field of science. I am delighted that I had the honour to work and discuss with him throughout my years in Crete.
I would also like to thank Athanasios Mouchtaris and George Tzanetakis for supporting me with advice throughout the duration of the thesis.
Also, I am grateful for the support in the improvement of my work that I obtained from the other members of the committee.
Special thanks go to all the people who helped me in getting access to musical culture that was previously unknown to me. In this context, I want to thank Barış Bozkurt for his support, and Bora Uymaz for his inspiring lessons.
Ein Dank auch an meine Familie, die ihren verlorenen Sohn immer mit Stolz unterstuetzt hat.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scope of the Thesis

In the field of ethnomusicology, computer based methods can be used for simplifying musicological studies, or even to make them feasible at all. Useful methods include the recognition of intervals played by an instrument or determining the meter structure of a signal. Using such methods, a search engine can be developed that can detect similarities between different pieces. Such a tool is valuable for research in ethnomusicology, because it enables a faster access to pieces that are interesting for a comparison due to similar structure in their compositions.

In this context, the scope of this thesis is to provide a set of approaches to determine similarities between pieces of music, which are adapted to the context of traditional forms of music. Traditional music makes some demands on the applied methods, which differ from the demands of popular western music. These differences are caused by various aspects, such as the morphology of the pieces, the instruments that are contained and parameters such as the tempo and the tonal space of the melodies. As this thesis focusses on the rhythmic and timbral properties, a general framework for the morphological analysis of traditional music is proposed and the parts of the framework that are related to rhythm and timbre will be worked out in detail. While the tools presented in this thesis are not restricted to a specific kind of music, emphasis will be given to the traditional form of music encountered in the area of the eastern Mediterranean.

In general, morphology of music is defined as the methodical description of the structure of the form of musical works [148]. The word is derived from the German word Formenlehre. According to the musicologist Hugo Riemann (1849-1919) [132], the technical and aesthetic analysis of a musical opus is based on its morphological organization. The elements of this organization are themes, phrases and motives, which themselves are made up of characteristics of sound like tonal height, duration, intensity and timbre. The analysis aims at the discovery of the sentence structure (Periodenbau) and the transformative structure of these elements. This discovery is the core of morphological analysis. For example, the musicologist Hugo Leichtentritt (1874-1951) emphasizes the antithesis between the forms of *fuga* and *sonata*, which follow the schemes $AA'A''\cdot$ and ABA, respectively. By considering the nominal form of an opus, one can locate all the characteristics and particularities of the piece, by examining the causal relations between the form and the particular opus. Another example would be the analysis of the content of a pop song into chorus and verse and their variations, and analyzing

possible deviations from usual composition schemes for pop songs.
Recently, the research presented in Sarris *et al.* [135] shed light on the difficulty of understanding traditional music in the eastern Mediterranean area: to a great extent, it is following a different kind of morphology, the logic of *parataxis*. The term *parataxis* stems from the field of linguistics, where it denotes a way of forming phrases using short sentences, without the use of coordinating or subordinating conjunctions [113]. In music following this logic, the tunes (*skopos*) are built from small melodic phrases which do not follow a specific morphologic structure. This means, that there is no composed elaboration of a theme like for example in a fuga, neither is there a clear periodic structure, according to which a musical theme is repeated, like the repeating element of a chorus in popular western music. In the context of traditional music of the island of Crete, Theodosopoulou [149] observes that the same motifs are often found in different pieces. It is observed that in some traditional dances themes of four bar length predominate, while in other dances themes of two bar length predominate. Theodosopoulou [149] introduces a way of numbering motifs and a methodology of morphological analysis that can be applied to music of the same region, for example to specific forms of dances from the island of Crete. Melodies from other regions like Dodekanes, Kyklades, or from Thrace, Macedonia or Epirus could also be examined with this method. The extensibility of this approach has also been underlined by Amargianakis in [2], and is supported by the relations described by Baud-Bovy in [7]. At this point, it has to be clarified that the goal of this work is not the achievement of research results in ethnomusicology, but the development of computational tools that make such a musicological research feasible for experts. As mentioned in Theodosopoulou [149], it is a major effort to transcribe and analyze a big number of pieces. The goal is to derive at least some conclusions about the content and similarity between pieces in an automatic way. Thus, a concept is presented that is aimed to discover recurring elements in a musical signal. These recurring elements are the melodic phrases that are the characteristic themes of the music following the logic of *parataxis*. The recognition of these phrases and their assignment to a specific dance by a human being appears to be a complex task. In interviews the author conducted with local musicians, repeatedly the recognition of a dance was connected with the recognition of a specific melodic phrase. This process is also described similarly in Tsouchlarakis [156]. Also, in all listening test conducted in the course of this thesis, it was observed that dancing teachers had memorized almost all melodies they have been presented with. With this knowledge they were able to conduct *e.g.* assignments to a class of dance much faster and with higher accuracy than their students. It is apparent that the similarity estimation between the used motifs is important for a search engine for this music. While further details concerning the structure and elements of Cretan music are given in the dataset descriptions in Appendix A.3, we will now proceed in showing the problems in developing such a system.

1.2 Problem of Transcription

Given the above description of the problem, the procedure that imitates the human expert would proceed using the following steps. First, the polyphonic mixture of a signal has to be analyzed and the instrumental sources have to be separated. Then, a decision has to be found about which separated signal contains the main melody. This signal has to be analyzed according to its fundamental frequency, and these estimates have to be formed into a stream of continuous frequency estimates that make sense in terms of their tonal steps and durations. Then, this estimate has to be transcribed into a musical score. From this score, the motifs have to be understood and located, and in the final step compared with each other. However, each step of this imitation procedure, which is commonly referred to as transcription of polyphonic music, is infeasible in this musical context. To clarify this, two publications that examine the problem of transcription will be contrasted, one publication from the field of ethnomusicology and one from the field of music information retrieval (MIR). This comparison is important especially because the problem of transcription is often considered as a solely technical problem, while its complexity as a cognitive process is disregarded.

In this text, transcription will be understood as the process of transferring music as a sound event into the symbolic form of a score [146]. In western music the score usually contains a number of staves, one for each timbre present in the piece. The complexity of this problem, for an algorithmic approach but to a certain extend also for a human being, depends on the complexity of the musical sound event that we want to transcribe. The state of the art in MIR will be outlined by referring to Klapuri [91], leading to the conclusion that current systems deal fairly well with monophonic signals but face difficulties on polyphonic inputs. However, even for the human expert, transcription can be a matter of high complexity. These problems gained importance with the possibility of recording musical performances, because it became possible to do field recordings of improvised music that has never been written in a score. In Stockmann[146], the problems for musicologists in this context have been described in detail, and this publication shall be outlined in order to clarify the existent gap between the state of the art in MIR and the demands of the musical signals that will be considered in this thesis. The problem of transcription in ethnomusicology origins from the fact that a big part of the musical heritage is being passed from one generation to the next in oral form. A scientific investigation of these forms of music makes their transcription necessary. For this, the impact of progress in recording and analysis technology has been very important throughout the last century ([146],p.205): Before the availability of sound recording techniques, a transcription had to be done either immediately when listening to a piece, or afterwards by transcribing from memory. This introduced a high grade of simplification and insecurity into the transcription. With the development of recording techniques, complex styles of music could also be transcripted, and the validity of the result could be evaluated by referring to the sound source ([146],p.207). One of the first musicologists who observed the problem of notating music, which origins from other countries but Europe, was Erich Moritz von Hornbostel, who suggested standard annotations for transcribing exotic melodies into a stave [1] ([146],p.209). It can be observed that many of these notational problems also appear for the traditional music of Crete. This can be proved by examining the transcriptions of Samuel Baud-Bovy in [8], where for example in transcription 53 of the dance tune *Syrtos rethymniotikos* many of these annotations appear indicating deviations in pitch from the notes as written in the score.

The process of transcription is a process of abstraction, which transforms from a sensory to a rational field of cognition. This transition does not completely transform the acoustic content, or better the perceived content, to the symbolic representation in the score. This is due to the limited expressiveness of the notational system, but also to the difficulty of transforming a complex psychophysical process into a musical score ([146],p.210). Compared to the tran-

scription of spoken text, the transcription of music is much more demanding, even though the process is similar. This is because the diversity of the means and methods used for the production of music is much larger than those for the production of speech ([146],p.211). As well, the criteria for differentiating in phonological, morphological and syntactical levels, are much more immanent in speech, and much more sparse in music. Because of that, there is no existing symbolic system applicable to all different kinds of music, like it exists for example for speech by phonetic annotation ([146],p.211).
In Stockmann [146],p.212, the author compares the process of transcription with a communication channel. The source is the musician and the receiver is the transcriber. In order for this channel to work without big loss, it is not enough to establish an acoustic connection between source and receiver, but also to have a common codebook of musical norms and rules, such as scales and rhythmic patterns. The transcription can then be expressed as the transcoding into another code, which is improved when the communication code between source and receiver is well understood by the receiver.
Because of the high subjectivity of the transcription process, two transcriptions of the same piece by two authors are very unlikely to be exactly the same ([146],p.213). This has been examined in List [104] as well. There, problems appeared more often in the annotation of duration than in the annotation of pitch, especially when no clear reference beat was present. Nevertheless, the experiments in resulted in at least equivalent transcriptions in most cases [104]. This variability can be considered as an equivalent to the variability in the interpretations of a piece of classical music. In the context of traditional music, the order of notating to a score and performing a piece is exactly opposite, and thus the variability due to the subject happens in the performance ([146],p.214). Another source of variation in the transcriptions is the intention of the transcriber: when intending to describe the general form of a piece, a small amount of details of the performance needs to be transcribed, while when emphasis is placed on the personal style of a player, each ornament can be of importance. However, the decision on what is an important detail is difficult and demands a familiarity with the examined musical culture. Furthermore, it must not be forgotten, that the piece to be transcribed is always just a snapshot of a socio-cultural development. As such, one must be careful in over-interpreting details, and if possible, a big number of performances have to be considered in order to draw valid conclusions. This is very time demanding for a human subject, and indicates an advantage of applying computational methods in this fields of research. However, deciding about the level of detail that is to be transcribed in an automatic procedure is an open problem.
In order to capture all details considered important, four different approaches are mentioned ([146],p.215):

1. The enhancement of the notational system, by using accents to indicate a deviation from the pitch prescribed by the note. There have been many different proposals to do such annotations, with little effort to form some kind of standard. Nevertheless, indicators like ↓or ↑ over the affected note to indicate a slight decreased/increased pitch are common. Also for rhythmic deviations, there are different kinds of annotations available. Note that for example in the context of classical ottoman music a system proposed by Arel [5] is applied that uses additional accidentals for intervals different from a half note.

2. Technical analysis methods, which *e.g.* enable to replace the arrow symbols by exact values of how many cent deviation a note has.

3. Different notational systems, which make sense especially when the examined music differs strongly from European norms. Besides notational systems adapted to the particular culture, also the verbal description of the music plays an important role here.

4. Improved recording systems, such as multichannel recordings or parallel video capturing.

In the following, Stockmann lists some important conditions and procedures for a meaningful transcription result. Note that a part of these clues holds as well for computational approaches of the transcription of music, as will be shown when examining the paper by Klapuri [91]. At first, the tonal extension of an instrument must be known. Also, it is helpful if the transcriber is familiar with playing the instrument. In performances with multiple voices, the exact number, position and timbres should be known. The procedure of a transcription process is in general as follows:

1. Establishing a general structure, *e.g.* A-B-A-B'...

2. Choosing a characteristic starting point for transcribing, which is typically NOT the beginning.

3. Which is the tonal extension, which is the general morphology and metric structure

4. Pitch annotation:
   (a) determining the central tone
   (b) determine the transposition
   (c) determine the key
   (d) preliminary melody annotation

5. Determination of durations

6. Determination of structure:
   (a) finding verses, pauses, accents;
   (b) setting the bars

7. Analysis of performance characteristics
   (a) Ornaments, dynamics, register changes etc.
   (b) Decision if these elements are assigned to the performer or the structure of the music performed

Given the above list, the central steps and problems of transcription by an expert are clarified and a structure is build up, which would have to be followed by an automatic procedure.
Contrasted with the transcription from a musicological point of view, in [91] Klapuri gives an overview of the transcription task as it is approached in MIR. Here, the task is constrained to the transformation of harmonic sounds from the acoustic domain to a symbolic description like the MIDI note format. Thus, the pitch annotation task is simplified, because no decision about the musical key or transpositions of the score annotation has to be made. In

the paper, Klapuri refers to transcription as the detection of the recipe of a piece of music, as he considers only music that has been composed in written form before performing. Even though the author mentions the application of musicological analysis of improvised music, no further comment about the differing demands of this task is made. It has to be noted, that in general these problematics have not been addressed systematically by computer science yet. Similar to Stockmann, also Klapuri points out the similarity to the task of speech recognition, while denoting that transcription of music has not received comparable interest, yet. The transcription of music by computational means is mentioned to be feasible only when constraining sounds regarding polyphony and instrumental timbres. However, even the transcription of a single singing voice is not perfectly possible, with the difficulty consisting in assigning an estimated pitch to a note value. Klapuri divides the transcription problem into a multiple fundamental frequency (F0) estimation task and a rhythmic parsing task. While physiological representations in the human ear are mentioned, the highly subjective aspect of transcription as mentioned by Stockmann remains unobserved. The systems that represent the state of the art, are mostly pure signal processing systems, musicological knowledge about morphologies has not been included in any way to such systems. In many approaches for the multiple F0 analysis, a sinusoidal model [141] is used for representing the tonal components of a signal which are then grouped into streams based on the principles of *Auditory Scene Analysis* [21]. These principles include parameters like common onsets and modulations of components and frequency proximity. Regarding the rhythmic parsing task, Klapuri points out the importance of regular accents and stresses for the analysis of the meter structure, like it was mentioned by Stockmann as well. As the author presented a state of the art system to tackle the problem of beat detection in the MIREX 2006 beat detection contest, this system will be outlined and used in this thesis in Chapter 5.
Summing up, the author states that the problem of transcribing polyphonic music is far from being solved. He suggests the automatic generation of music in order to train transcription systems. It has to be noted, that he assigns the difficulty of the task in the combination of multiple sources, while he assumes that each of the sources has a complexity smaller than those of speech sounds. This assumption clearly contradicts with the findings from musicology as documented by Stockmann, which assign the higher complexity to the sound elements in music.

Concluding the summaries, it has to be stated that transcription of music is in general not tractable using computational approaches. Apart from the difficulties already observed in the MIR community, the confrontation of two important publications from the fields of MIR and musicology sheds light on the following facts:

- Transcription of traditional music is more complex due to its non-written character
- Part of the complexity of the transcription task is due to the best possible presentation in a staff, not just as a MIDI note
- Transcription is difficult not only because it has high demands in terms of signal processing, but also because it is a highly subjective process
- There is a disagreement concerning the complexity of the elementary sounds encountered in music and speech

Due to these conclusions, in this thesis the task of transcription will be reduced in complexity as outlined in Section 1.3.

1.3 Reducing Complexity

Regarding the types of recording that can be processed using the methods proposed in this thesis, the goal is to be able to process polyphonic mixtures. This means that we want to enable musicologists to use whatever kind of field recording, monophonic recordings of a single instrument as well as polyphonic recordings of a lead instrument that is accompanied with various other instruments. In order to compute characteristics of a piece and to compare them to other pieces, three aspects of the music are proposed to be taken under consideration: rhythm, melody and timbre. These aspects are going to be connected in a way depicted in Figure 1.1. The rhythmic properties of a piece are derived from an Onset Strength Signal (OSS), melodic descriptors will be derived from melody histograms (MELHIST), and timbre description is based on Non-negative Matrix Factorization (NMF), as will be explained in Section 1.4. The aspects contained in Blocks A to D in Figure 1.1 can be grouped together to a system of morphologic analysis. The first aspect, rhythm, has to capture characteristic periodicities present in the signal. It can be assumed that by taking rhythmic properties into account, it is likely to get a more accurate description of the signal than by using melodic properties only. For example such an improvement has been reported when measuring the similarity of symbolic representations of folk melodies [161]. This assumption is also supported by the statements of local musicians in interviews that have been conducted by the author, where beside the melody the importance of the rhythmic intonation is underlined for the recognition of specific pieces. The rhythmic similarity measure has to be capable of describing the rhythmic content of a piece independent of its tempo. This is desirable because frequently in signals encountered in

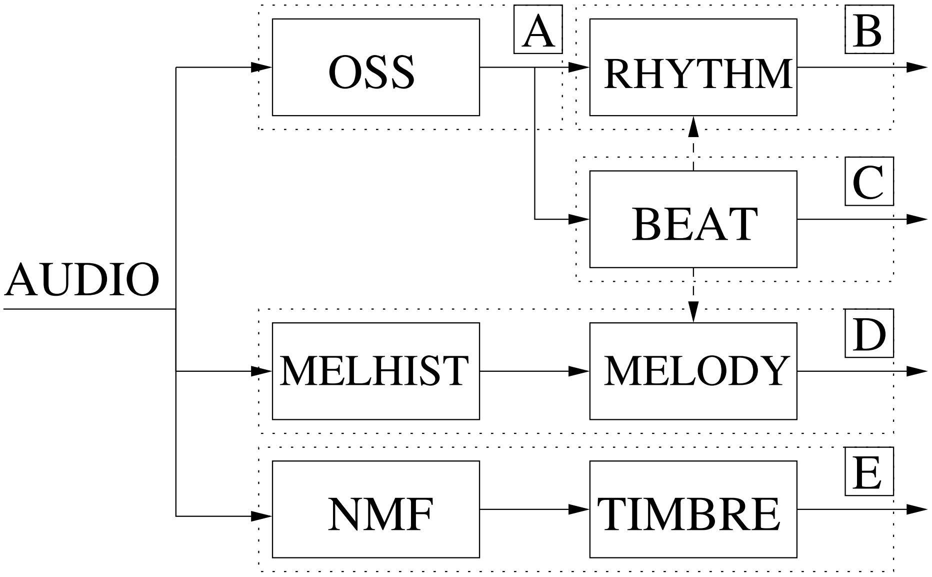


Figure 1.1: Characteristics computed from a music signal

traditional music, little percussive content is present, which makes the accurate estimation of meter properties a difficult task [92]. For that reason, it has been decided to approach this problem in two ways. First, a tempo-robust description of the rhythm of a piece is desired even if no accurate tempo estimation and beat tracking can be guaranteed. And, second, present state of the art systems for beat tracking are to be evaluated on data exemplary for the traditional styles of music encountered in the eastern Mediterranean, and possible ways to improve their performance should be indicated. As indicated in Figure 1.1, the estimated beat positions can then be used to synchronize the rhythmic and melodic description of the piece. The second aspect that has to be considered in a morphological analysis system is melody. It has been shown that a beat synchronous computation of melodic content enables for an improved retrieval of cover songs of western pop music [46]. In Ellis and Poliner [46], chroma features have been used as the descriptors of the melodic content. For that reason, these features are used in this thesis as a baseline. However, it is likely that including methods for tracking the lead melody in a mixture like the one presented in Klapuri [90] can further improve measurements of melodic properties. For that reason, such an approach will be combined with a melody histogram computation as proposed for example in Bozkurt [19], and results will be compared with the reference system based on chroma features.

The last aspect, timbre, helps to understand about the differences concerning the used instruments. This can help in categorizing music into various regions. For example in regions of northern Greece the clarinet is the dominating lead instrument, while this instrument is not used in the traditional music of the island of Crete, where the Cretan *lyra* is the dominating lead instrument. While this aspect is helpful for understanding the instrumental content of a mixture, it is not strictly related to the morphology of a piece.

1.4 Contributions

As mentioned above, the main focus of this thesis is to explore the rhythm aspects that are needed for a computational morphological analysis of traditional music. Apart from that, the timbre and melodic similarity has been be approached as well. The main contributions of this thesis are:

1. Onset Detection

   In order to get a description of rhythmical properties of music, a first processing step is a computation of an onset strength signal (OSS) from the input sample. OSS are supposed to have large values in the vicinity of a note onset. For their computation, different aspects of the signal can be considered, such as magnitude changes, phase characteristics or fundamental frequency changes. OSS can then be used for the deriving descriptors of rhythmical properties of a piece, for the estimation of the time instances of note onsets, or for beat tracking tasks. The contributions of this thesis concerning onset detection are:

   (a) The compilation of a dataset for the evaluation of onset detectors. This dataset contains monophonic recordings of various pitched instruments, and it enables for the first time to investigate the performance of onset detection in dependence of the chosen instrument type. It has been determined which kind of onset detector is preferable for a specific kind of instrument.

(b) As another contribution of this thesis, a novel onset detection method based on group delay is introduced. While previous approaches mostly use a time derivative of phase, the frequency derivative of phase, *i.e.* the group delay, is shown to represent an interesting characteristic for onset detection. The usage of group delay for onset detection is examined from a theoretical point of view as well as regarding its detection performance on musical instrument signals.

(c) The output of a state-of-the-art F0 estimator is used to derive an OSS that has the ability to detect onsets that are related to note changes with a constant excitation, as encountered for example when changing the finger position on a violin while constantly moving the bow.

(d) A fusion of magnitude, group delay and F0 information is proposed, and its performance is shown to improve compared to using each characteristic individually.

During the course of the thesis, the work related to onset detection lead to the following journal and conference contributions:

- Andre Holzapfel and Yannis Stylianou and Ali C. Gedik and Baris Bozkurt, “Three dimensions of pitched instrument onset detection”, accepted for publication in IEEE Transactions on Audio, Speech and Language Processing, 2009.
- Emmanouil Benetos and Andre Holzapfel and Yannis Stylianou, “Pitched instrument onset detection based on auditory spectra”, Proceedings of ISMIR - International Conference on Music Information Retrieval, Kobe, Japan, 2009.

2. Rhythm similarity
Using the OSS, a method based on the scale transform is proposed for the description of the rhythmic properties of a piece. As mentioned in Section 1.1, due to the difficulty of the beat tracking task, the proposed method works without requiring beat tracking or any tempo estimation. However, it is robust to tempo changes within one piece and between different pieces of similar rhythmic content that are to be compared. The accuracy of the proposed rhythm description has been evaluated on three datasets. Two of them contain traditional music and one contains western ballroom dances. On these datasets, the classification accuracy into a certain kind of rhythm class serves as an indicator of the accuracy of a rhythm similarity measurement. Regarding rhythmic similarity, the contributions provided in this thesis are:

(a) A rhythmic similarity measurement based on representations in the scale domain is proposed.

(b) The compilation of two datasets of traditional music for the evaluation of rhythmic similarity. One dataset contains audio samples of traditional Cretan dances, and the other is made up of symbolic representations of traditional and classical Turkish music.

(c) The evaluation of the proposed rhythmic similarity measure using the mentioned datasets as well as its correlation with human perception has been determined in a listening test.

During the course of the thesis, the work related to rhythmic similarity was documented in the following journal and conference contributions:

- Andre Holzapfel and Yannis Stylianou, “Rhythmic similarity of music based on dynamic periodicity warping”, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, USA, 2008.
- Andre Holzapfel and Yannis Stylianou, “A scale transform based method for rhythmic similarity of music”, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taiwan, 2009.
- Andre Holzapfel and Yannis Stylianou, “Rhythmic similarity in traditional Turkish music”, Proceedings of ISMIR - International Conference on Music Information Retrieval, Kobe, Japan, 2009.
- Andre Holzapfel and Yannis Stylianou, “Scale transform in rhythmic similarity of music”, Accepted for publication in IEEE Transactions on Audio, Speech and Language Processing, 2010.

3. Beat Tracking
Using a dataset of beat annotated traditional music samples the accuracy of beat tracking using a state of the art approach is evaluated. By using the novel group delay based OSS it is shown that the accuracy of the beat tracking system can be improved. Again, for this evaluation a dataset of beat annotated Cretan dance samples has been compiled. These achievements have been partly documented in

- Andre Holzapfel and Yannis Stylianou, “Beat tracking using group delay based onset detection”, Proceedings of ISMIR - International Conference on Music Information Retrieval, Philadelphia, USA, 2008.

4. Timbre Similarity
Timbre similarity bears a likeness to the task of music genre classification, as explained for example in Li and Ogihara [97]. This is because at least for genres such as classical music, rock or jazz, the musical content varies a lot in respect to the instruments contained in the signals. For that reason, a standard set of timbre descriptors was compared with a novel feature set in a genre classification task on two datasets, and the classifier found to perform best on this data was evaluated in the classification of traditional music from various regions of Greece and Turkey that vary regarding their instrumental content. The contributions of this thesis regarding timbre similarity are:

(a) Introduction of a new feature set that is derived from spectrogram representations by using Non-negative matrix factorization (NMF).

(b) Compilation of a dataset of traditional music of various regions of Greece and Turkey, that differ by the instruments used in their traditional context.

The work related to timbre similarity lead to the following journal and conference contributions during the course of the thesis work:

- Andre Holzapfel and Yannis Stylianou, “Musical genre classification using Nonnegative Matrix Factorization based features”, IEEE Transactions on Audio, Speech and Language Processing, vol. 16, nr. 2, pp 424-434, 2008.
- Andre Holzapfel and Yannis Stylianou, “Singer Identification in Rembetiko Music”, Proceedings of SMC 2007, Conference on Sound and Music Computing, Lefkada, Greece, 2007.

5. Melodic Similarity / Integration
   A beat synchronous computation of melody histograms is shown to provide an approach that is able to locate morphological similarities between pieces of traditional Cretan music. An integration of the rhythmic similarity measures into this approach is proposed. Even though the results related to melodic similarity and integration have been obtained on a rather small dataset, the proposed method represents an interesting tool for investigating morphologic similarities in music following the logic of *parataxis*.

Chapter 2

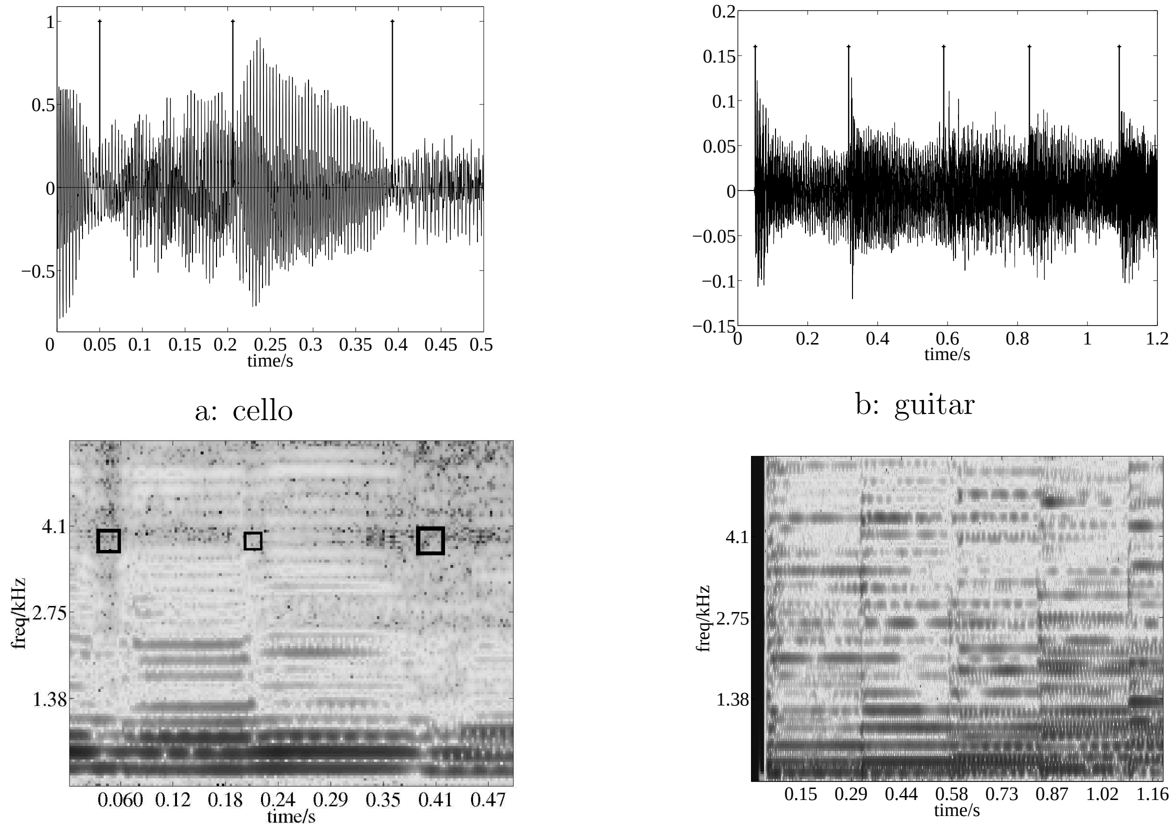
Related Work

2.1 Onset Detection

In this thesis, various ways to compute Onset Strength Signals (OSS) from a signal are compared. As mentioned above, the OSS can be used for various tasks such as beat tracking, rhythm description, or the detection of musical instrument note onsets. As a starting point, the performance of various OSS for onset detection will be evaluated. Onset detection, the detection of the starting instant of an event in a signal, is an extensively studied topic in various domains of signal processing. Musical onset detection, the detection of the starting point of a musical note transient [10], is one of the sub-domains with a large literature. Various methods have been proposed, evaluations are taking place [110], and tutorials are available [10, 41]. The main challenge in a musical onset detection problem is to build a robust algorithm that can detect onsets of various types of signals, *i.e.* the notes of the tune played by the violin player as well as the accompaniment played by the guitar. Considering also the variations in musical styles (classical, pop, jazz to folk musics, etc.) and performance styles (playing with a pick, finger picking, ornamentation styles in folk musics etc.) the variability is so large that it is problematic and time-consuming to collect a representative dataset and to evaluate various methods comparatively. Two examples that display the difficulty of the task are shown in Figure 2.1. In the time signals Figures 2.1.a and 2.1.b, the onset time instances have been marked with impulses. It can be seen that for the guitar signal the envelope of the time signal gives a good clue about the position of the onsets. Also the spectrogram of the guitar signal in Figure 2.1.d is characterized by sudden increases in energy in the vicinity of all onsets. However, the onsets of the cello signal are not so easy to spot. In the envelope of the time signal only the second onset stands out, and the spectrogram shows even clearer that onsets for this bowed string instruments differ in their characteristics from guitar onsets. In Figure 2.1.c it can be seen that onsets in the cello signal are rather characterized by gaps in energy especially for higher frequency bands. The time span of these gaps for the depicted example is symbolized by rectangles of varying size. Thus, it is clear from these two simple examples that an onset detection approach applicable for a wide variety of instruments has to be capable to detect onsets that differ regarding their characteristics.

The algorithmic steps of an onset detection system are:

1. Pre-processing of the audio signal (the raw time-series data)



a: cello

b: guitar

c: cello

d: guitar

Figure 2.1: Time signals of a cello and a guitar (a and b), and their spectrograms (c and d).

2. Computation of an onset strength signal (OSS), which is mainly the time-series of a computed parameter at a sampling frequency lower than that of the audio signal. The term OSS has been used in Ellis [47], while OSS is referred to as novelty function in Foote [48] and as detection functions in Bello [10].

3. Detection of transients in the OSS typically by applying a peak-picking algorithm [10].

The pre-processing is an optional step as in many other signal processing applications. The most common form of pre-processing used for musical onset detection is multi- band decomposition. Most of the studies using multi-band processing, approach the problem in a similar fashion: dividing the signal into several frequency bands, estimating OSS for each band, combining either at the OSS level or the onset decision level to achieve a final decision for the onsets [56, 136, 88, 43, 52]. It has been reported that the robustness of onset detection is improved by using such a methodology [136].

The core of the design of the onset detection system is the OSS estimation part for which a large variety of methods exist. Comprehensive reviews of these methods are available [10, 41, 28].

One type of approach for the OSS computation is the use of temporal feature variations such as the time-domain amplitude envelope of the signal [137], or short-time energy [56]. These relatively old approaches are successful for processing clean recordings of instruments with percussive character; however, they have problems in processing, for example, bowed instrument sounds where musical note change does not always imply a sudden change in the energy or amplitude. Another common approach for OSS computation is the use of spectral features since spectrograms of recordings very often reveal clear visual clues of the onset locations. Due to the difficulties involved in phase processing, amplitude processing is much more common in spectral methods. Spectral flux, the amplitude spectra difference computed for consecutive frames using various distance functions (L-1 norm, L-2 norm, Kullback-Leibler distance, etc.), is used in many studies due to the simplicity of computation and robustness in detection of onsets of pitched-percussive sounds, for example in [43, 62, 28, 41].

Although less common, phase processing is also used for OSS computation. In Bello and Sandler [13] a phase-based OSS is presented for the first time. Their approach is based on the segmentation of a signal into transient and steady-state (TSS) frames by detecting fast instantaneous frequency changes. In Bello *et al.* [11], the previous phase based approach is improved by using a mean absolute phase deviation function or alternatively a difference function on the complex Fourier coefficients from consecutive short-time frames. Phase-based OSS are used in combination with energy-based detection functions in a number of other studies [43, 28, 41]. However, computing reliable phase deviation information (or similarly a complex Fourier coefficients deviation) from consecutive frames is problematic. The main problem is the phase unwrapping operation or window synchronization [20]. It has been shown in various studies ([20, 165, 116]) that a large number of very high jumps in the phase slope (i.e. the negative group delay function) of audio signals exist due to zeros of the z-transform closely located around the unit circle. Steiglitz and Dickinson [145] have shown that the roots of the z-transform of a short-time signal tend to be evenly distributed in angle and tightly clustered near the unit circle as the degree of the polynomial (length of the time domain signal) increases. Hence, zeros of the z-transform for 20-30ms short-time audio signals are clustered around the unit circle resulting in many spikes in the phase slope [20]. This leads to the conclusion that reliable phase processing is very difficult to achieve unless certain synchronization rules (such as pitch synchronization in speech processing) are applied. The alternative to the direct usage of phase information is the processing of either some modified version of the group delay [66] or the average of group delay which can be used for detection of events like Glottal Closure Instants (GCI) [144, 22]. The average group delay has been applied for other types of transient detections as well, for example in the detection of clicks from marine mammals [83]. Because detecting onsets in music is a transient detection problem as well, phase information can be used in a similar way as in click or GCI detection.

In this thesis, a new onset detection method is proposed. This method is based on processing the average of the group delay function which will be referred to as *phase slope function.* The derivative of phase with respect to *time* is referred to as instantaneous frequency, and has been used in Bello [13], among others, for onset detection. In this work, the usage of group delay will be proposed, which is the derivative of phase with respect to *frequency.* It is interesting to note that the observations made on phase plane plots in Lacoste and Eck [94] showed that onsets appear more clearly when computing the derivative of the phase over frequency than over time. However, these observations were not developed further into any onset detection system. Recently, an onset detection method proposed by the author based on group delay was shown to improve the performance of beat tracking in music with little percussive content [71].

Another type of approach which specifically targets improvement of onset detection for non-percussive sounds is the usage of the fundamental frequency or pitch of the signal [29]. It was shown that previously presented approaches based on spectral features perform worse for pitched non-percussive than for pitched percussive sounds [28].

Onset detection can therefore be performed using spectral amplitude, phase, and pitch information. These three features or cues therefore define a three dimensional space[1]. We suggest that a human makes use of all these dimensions for onset detection and the importance of each dimension (weight) depends on the type of musical signal. Thus, the second major contribution of this thesis, beside the usage of group delay for onset detection, will be an appropriate combination of the information contained in these three dimensions. So far, only amplitude and phase information have been combined in various studies (for example [12]), where the phase information considers the instantaneous frequency changes and not the group delay as proposed in this thesis. In Zhou and Reiss [166], depending on the type of signal, either an energy based or a pitch based detector is applied. In Toh *et al.* [152], statistical models are built for different features (Mel-frequency Ceptral Coefficients (MFCC), Linear Prediction Coefficients (LPC) and others), and the decisions derived from the different models are combined to a single decision function. To the best of our knowledge, it has not been tried yet to combine the three dimensions of pitch, spectral amplitude and phase in order to get an improved onset detector. In this thesis, a combination of the decisions derived from the three individual dimensions (decision fusion) is proposed. This simple combination works without training complex statistical models for the feature distributions like in Toh *et al.* [152], and can easily be improved or extended by either changing one of the OSS or by adding a new one. Such a late fusion concept was shown to improve onset detection accuracy for OSS derived from phase and magnitude characteristics of a signal [39].

In order to determine the performance of different OSS, it is necessary to study the three feature dimensions and their fusion on a large enough dataset that is publicly available. The lack of common databases of pitched instruments is an important obstacle for further improvement. Thus, another major contribution of this thesis is the compilation of such a publicly available database, and studying the above mentioned three dimensions on this dataset. Despite the fact that signal characteristics (hence the onset detection performance) vary largely for different types of instruments, very few studies include performance styles or instruments of traditional forms of music in their databases (for example in the context of Irish instruments [85, 51]). In the database compiled in the context of this thesis, traditional Turkish music instruments have been included (*ud*, *tanbur*, *ney* and *kemençe*) to also study variations between western and non-western music. As a result, a dataset containing a diverse set of pitched instruments is available for the evaluation of onset detection systems. The dataset can be provided to interested researchers on request to the author.

2.2 Rhythm Similarity

As a first step what is meant when using the term rhythm has to be clarified. In Cooper and Meyer [32], rhythm is defined as the way one or more unaccented beats are grouped in relation to an accented one. Furthermore, meter is defined as the measurement of the number of pulses between more or less regularly occurring accents. The music encountered in the context of this

[1] we use the term "dimension" loosely without its formal definition

thesis can be assumed to have meter, *i.e.* it has a hierarchical structure of levels at different note values, such as half-note and quarter-note levels [96]. While it is possible to encounter rhythm without the existence of meter this is not the case for the traditional forms of music that are investigated in this thesis. In literature, various views exist regarding the different aspects of temporal organization of music, such as grouping, meter and rhythm. But it can be concluded that the form of meter is one important aspect of the rhythmic characteristic of music. Meter is an hierarchical structure with the beat level, the tempo which the human listener is likely to tap his/her foot to, somewhere in the middle. An example is shown in Figure 2.2. The shown piece has a $\frac{2}{4}$ time signature, and the beat level is positioned at the next level ($\frac{1}{4}$). The smallest inter onset interval in the shown piece is related to $\frac{1}{32}$ notes. This lowest meter level is usually referred to as tatum [17]. Note that the number of levels between measure, beat and tatum level depend on the piece of music, for example a piece in $\frac{4}{4}$ time signature would have one more intermediate level between measure and beat level.

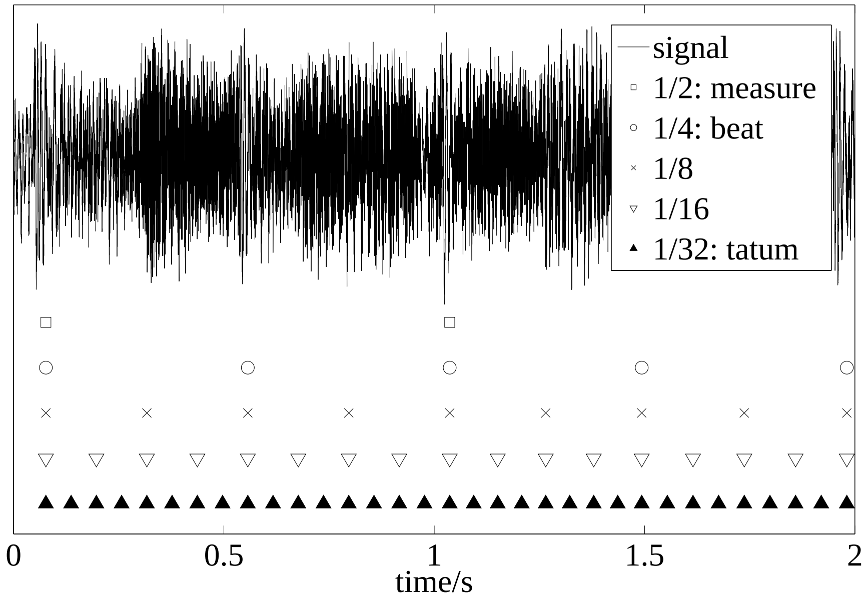


Figure 2.2: Example for the hierarchical structure of meter in music

Another aspect of rhythm is tempo, which is divided in Berry [16] into the frequency of pulsation (*i.e.* the pulse-tempo) and the eventfulness of music. While the first aspect is not of importance for rhythmic similarity within certain boundaries, because a rhythm or theme will be recognizably the same whether played faster or slower [32], the second aspect has to be taken into consideration when measuring rhythmic properties. Another important aspect is the rhythmic grouping, which is organized hierarchically like the meter structure. The basic level of this hierarchy is made up of rhythmic motives [96], which are grouped together on the higher level to form longer rhythmic themes. In the ideal case, an algorithmic approach for the description of the rhythmic properties of music should be able to capture the characteristics of this organization. As referred to in Lerdahl and Jackendoff [96], the phase relations between the metrical organization and the grouping structure is an important aspect for the perception of rhythmic complexity. However, this understanding is a highly difficult task even when a symbolic representation of music is available, as it is a complex cognitive procedure that differs widely and depends highly on the musical education of the listener.

There have been various approaches for the estimation of rhythmic similarity using computational means. In Foote *et al.* [49], a self similarity measure is used to derive beat spectra. These beat spectra exhibit high energy at periodicities that appear in the signal. The beat spectra are compared by using a cosine distance. This measure is shown to work well within a narrow range of tempo variation only. Other approaches do work in presence of different tempi [125, 124], but for this either the tempo or meter characteristics have to be estimated. As indicated in Klapuri [92], this type of estimation is not very reliable for music signals without strong percussive content or with complex rhythmic structure, such as Folk or Jazz. The findings in Holzapfel and Stylianou [71] indicate that these type of estimation is difficult on traditional forms of music. Furthermore, state of the art meter tracking approaches have not been applied yet to music forms with time signatures unusual in Western popular music, such as $\frac{9}{8}$ rhythms frequently encountered in Greek and Turkish music. In [59, 121, 100], some features are presented that do not need any tempo estimation, such as periodicity histograms, inter onset interval histograms or temporal modulation patterns. The common shortcoming of these descriptors is that they cannot be directly compared in presence of tempo differences, and for that reason characteristics of the descriptors such as their flatness or energy have to be used. To improve the robustness to tempo changes in a music signal, in Holzapfel and Stylianou [73] periodicity spectra have been computed from onset strength signals [47] and have been used in a method referred to as Dynamic Periodicity Warping (DPW). There, a matrix of point wise distances between periodicity spectra is computed, and a minimum cost warping path through this matrix is found. This path is compared to an ideal warping path to get a distance measure. In Antonopoulos *et al.* [4], warping with different kind of step criteria than in Holzapfel and Stylianou [73] is applied to periodicity representations derived from self similarity measures; thereafter simply the cost of the warping is taken as distance.

In the case when a discrimination between different rhythm classes that differ regarding their time signature (*i.e.* $\frac{4}{4}$, $\frac{7}{8}$) is desired, the problem can be reduced to a time signature identification task. This will be of significance as one of the datasets used in this thesis for rhythmic similarity has this characteristic. In Toiviainen and Eerola [154], an approach was presented to estimate the time signature of a piece of music based on symbolic descriptions (MIDI). This approach uses autocorrelation coefficients (ACF) derived from the annotated onsets. For audio signals a time signature estimation system was proposed and evaluated on a set of percussive music in Uhle and Herre [158]. The system estimates the tatum [17] of the signal using inter-onset intervals (IOI) and in parallel, ACF are computed from the amplitude envelope of the signal. Beat and bar length are chosen from the peaks of the ACF, taking into account the estimated tatum. In Gouyon and Herrera [60], the determination of musical meter was reduced to a classification into either binary or ternary meter. Beat indexes are extracted in a semi-automatic way and then ACF on a chosen set of features are used to decide on the meter type.

As the next contribution of this thesis, a novel method for the measurement of rhythmic similarity in music is presented. In Western music, tempo changes appear within certain boundaries, as observed on the example of dance music [125]. In traditional dances the tempo of the performance usually varies between different performances but also within the duration of the piece [105, 3]. Thus, in order to compare dance music that accompanies the same dance but is performed in different tempo, a similarity measure robust to these changes is necessary. Apart from traditional dances, other forms of traditional music are also characterized by wide tempo changes. An example is classic Ottoman music, where compositions are categorized by their melodic scheme, the *makam*, and their rhythmic scheme, the *usul*. As these rhythmic categories are not in general connected to a certain form of dance, they can vary widely in tempo. Furthermore, the *usul* can have complex or compound time signatures.

For types of music signals with varying tempo, we recently proposed a rhythmic similarity measure [75] which is based on the scale transform [27]. Scale transform is scale invariant, or equivalent in music, is not sensitive to tempo changes. In Holzapfel and Stylianou [75], it was shown that it can be applied in rhythmic similarity of music without previous tempo or meter estimations, which makes its usage for music with compound and complex time signatures feasible as well [74]. Until now, the scale transform has been applied in various fields of signal processing in order to compare signals that have been changed by a scale factor. For example, the scale transform has been applied to vowel recognition in speech [159]. The usage of the scale transform is motivated by the fact, that between two speakers uttering the same vowel there is a scaling in frequency domain due to the different vocal tract lengths (VTL). Similar observations can be found in Irino and Patterson [79], where the scaling of the impulse response of the vocal tract due to different VTL's is shown to disappear when applying a Mellin transform. Apart from these speech processing applications, the scale transform was applied in order to estimate the speed gaps between mechanical systems, which are assumed to cause the related signals to be different by a scale factor [30]. To the best of our knowledge, scale transform has been applied to music signals only for audio effects [138]. However, two studies have observed improvements when including a scale invariance into their approaches. In Saito *et al.* [134], scale invariance helped to investigate multiple fundamental frequencies with common harmonic structure. In terms of rhythm, the authors of [81] presented a method to compensate for tempo changes between two pieces of music by applying a logarithmic scale, which is closely related to the relation between the scale transform and the Fourier transform as will be denoted in Section 4.1.1.

In this thesis, scale transform is applied for the analysis of music signals, by using autocorrelation sequences as descriptors for the rhythmic content of a piece of dance music. When the same piece of music is performed at a different tempo, its autocorrelation is scaled in time. Thus, the scale transform magnitudes of the autocorrelations remain essentially the same and can be compared in a straightforward way. In this thesis, this method will be detailed and extended so that it can be used for different types of signals. Signals are used that are different regarding their musical content, and audio signals are used as well as MIDI files. In order to allow for a well-founded evaluation of the proposed method, three datasets will be used. The first dataset is a set of ballroom dances that was used in the rhythm classification contest in the ISMIR conference 2004 [80]. The other two datasets have been compiled by the author in the course of the work on this thesis. One of these datasets contains Turkish traditional music which is available in a symbolic description format (MIDI). The other dataset contains audio data of traditional dances encountered in the island of Crete. The influence of critical system parameters will be analyzed in detail and insights into the characteristics of the obtained scale transform descriptors will be given. The scale transform based methods proposed in this thesis will also be compared to the Dynamic Periodicity Warping based rhythmic similarity measure that was presented by the author [73], in order to clarify the advantages of using scale transform to achieve robustness to tempo changes.

2.3 Beat Tracking

The task of estimating the times at which a human would tap his foot to a musical sound is known as beat tracking [47]. This periodicity, which is also referred to as *tactus*, finds itself in the somewhere in the middle of the meter hierarchy of the piece of music, see Figure 2.2 for an example. All state-of-the-art approaches ([47, 92, 40, 34]) for this task first conduct an onset detection. The signal used for onset detection, as explained in Chapter 1.4, is an onset strength signal (OSS) with a lower time resolution than the input signal, which has peaks at the time instances where a musical instrument in the input started playing a note. Usually, this OSS is derived from the amplitude of the signal [47, 92, 40]. Less frequently, phase information is considered, by computing the phase deviation between neighboring analysis frames [34]. As can be seen in the results depicted in Davies and Plumbley [35] (Table II), the state-of-the-art approaches for beat tracking decrease significantly in accuracy, when applied to folk music. These music signals contain weaker percussive content than music of rock or disco styles. This problem is of particular importance when dealing with traditional dances as well, as they are often played using string or wind instruments only [73]. Based on the results obtained in Davies and Plumbley [35], it is necessary to improve beat tracking on music with little percussive content. While a decrease in the case of jazz and classical music can partly be attributed to rhythmic complexity, meter structure of folk music is simpler, and thus the decrease in this forms of music may be attributed solely to the problem of detecting onsets. Thus, in this thesis improved beat tracking results on musical signals with simple rhythmic structure and little or no percussive content is achieved by using the phase slope based OSS as outlined in Section 2.1 and detailed in Section 3.2.1. This OSS has been combined with the state of the art beat tracking approach presented in Klapuri [92]. This way, a beat tracking approach is proposed that is more sensitive to instrument onsets of non-percussive sounds, and the beat tracking performance for polyphonic recordings of traditional music is shown to improve compared to the state-of-the-art approach as described in Klapuri [92].

2.4 Timbre Similarity

The musical instruments encountered in traditional music vary depending on the area where the music is being played. For example, in Cretan traditional music the most popular lead instrument is the Cretan lyra, while on the Greek mainland clarinet is a widely used instrument. It is considered useful to include an automatic way to capture this kind of differences in a system for a morphological analysis, in order to recognize the regional context of the piece of music based on the contained instrument timbres. This task bares similarity with the task of musical genre classification, in the case when musical genres differ regarding their characteristic timbres, as it is the case when for example discriminating between classical music and rock or disco music. The notes contained in the signals are reproduced by organs with characteristic frequency structures, which is referred to as the formant structure of an instrument [129]. These sounds have all been processed individually and/or together in a studio environment, thus changing their spectral characteristics. In Music Information Retrieval (MIR) this is often referred to as the timbre of music. Unlike rhythm, which is a structure that develops in time and is a characteristic of the horizontal structure of a music score, timbre is an instant characteristic which is observable in the vertical structure of the sound. Experimental results lead to the conclusion that, apart from timbre, musical style is a characteristic found in the vertical structure as well. For example in Perrott and Gjerdingen [126], listeners were able to assign a piece of music to a style given an excerpt of duration less than one second. Recently Li and Ogihara [97] received improved results in a genre classification task by using only spectral descriptors and neglecting temporal information. This can be interpreted as a supporting result for Perrott and Gjerdingen [126], since a musical genre is defined as a category of pieces that share a certain style [160]. Therefore, a system to automatically retrieve information about the vertical structure of music will be capable of describing style, genre, and timbre of the composition. In many publications the vertical dimension of music has been described by using a feature set consisting of Mel Frequency Cepstral Coefficients (MFCC). These features have been successfully applied to the task of speech recognition [36]. They have also found wide application in the classification of music into genres or in developing measures for the similarity of musical pieces [119]. In Pachet and Aucouturier [119] it has been shown that systems following the general model of using MFCC based features are upper bounded in their recognition performance.

An aspect that has not been considered in the development of the previously reported representation approaches is the fact that the characteristic timbre of the recordings is usually created by mixing several instruments into a single signal. Thus an approach to derive descriptions of these components from the mixture signal could provide a more versatile feature set for the genre classification task. In Casey [24], a method for the classification of sounds has been presented, where the spectral space of a signal is described using techniques based on Independent Component Analysis (ICA, [31]) applied to the spectrogram of the signal. Considering musical signals, methods based on a Non-negative Matrix Factorization (NMF, [95]) have recently shown success in separating instruments from a mixture [143, 162]. NMF has been used as well for the classification of sounds [14, 86, 26]. The classification approaches based on these techniques follow a deterministic path by first defining a set of spectral bases for the sounds and then projecting new sounds into these spaces.

In Holzapfel [69], NMF is shown to yield a compact representation and, compared to ICA, superior results in a mean squared error sense for some selected sound samples. In this thesis, these results are confirmed systematically using large datasets. Thus, a signal spectrogram is described with the spectral space spanned by the vectors computed by this factorization approach. For a given musical genre, a Gaussian Mixture Model (GMM) is built on all the spectral base vectors that have been computed for the spectrograms of the training data for a particular class. In this way we get a description for the spectral base of the particular genre. The classification is based on the Maximum Likelihood (ML) considering all the spectral base vectors from a test signal. Extended classification tests were conducted on two widely used datasets for music classification (Tzanetakis *et al.* [99] and from the ISMIR 2004 contest[2])

[2] http://ismir2004.ismir.net/ISMIR_Contest.html

comparing the performance of the proposed NMF based features and that of MFCCs. The proposed NMF based features constantly outperformed the MFCCs in terms of classification score. The proposed classification system was also compared to reference systems [99, 120, 15] for the task of music genres classification. The proposed classification system achieved higher classification score compared to these systems, in most of the conducted experiments, although Li and Tzanetakis [99] employs features that model both the vertical and horizontal structure of music. For that reason, the NMF based features are evaluated on a dataset of traditional music of various regions of Greece, resulting in a method to discriminate traditional music based on the musical instruments that are used in the specific regions.

Note that for the classification of instruments contained in a mixture methods exist that are likely to further improve the performance compared to the approach presented in this thesis. For example, in Heittola *et al.* [67] a method is proposed that complements the usage of NMF with a multiple fundamental frequency estimation and a source filter model. This way, the search space for the spectral bases is reduced. A recognition of the instruments is then performed on the separated signals using MFCC. However, in Heittola *et al.* [67] artificially mixed signals are used, while in our case the estimation of the necessary polyphony (*i.e.* the number of synchronous notes) could not be addressed using this method. This is because the accompaniment in the used samples consists of string instruments at some time instances play chords with an unknown number of contained notes. A different kind of source separation based on graph theory was presented in Martins *et al.* [109] that avoids the polyphony determination by using a different kind of source separation based on graph theory.

A problematic issue for the instrument classification in traditional music is the difficulty of obtaining a sufficient amount of data. In order to get a sufficient amount of data, it has been proposed to use data that has been synthesized from MIDI files using sample banks of the instruments under consideration [63]. However, for traditional music neither such sample banks nor MIDI files of the pieces are available. For that reason, realistic experiments will have to use original recordings, which makes the compilation of datasets a time consuming and expensive task. As it was shown in Fuhrmann *et al.* [50], the usage of such real world recordings in combination with a pattern recognition system without any source separation yields sufficiently good results in instrument recognition, given a sufficient amount of data. Thus, it will be evaluated if the NMF system proposed in this thesis shows comparable or improved accuracies to the ones obtained using a standard MFCC representation.

2.5 Melodic Similarity

Recently, similarity in folk song and traditional melodies has drawn increasing attention of the Music Information Retrieval research community. Most of the related publications investigate symbolic transcriptions of melodies. Juhász [82] proposed a system for the recognition of characteristic phrases from melodies of various musical cultures. He shows that by learning characteristic melody contours conclusions about the relation between various musical traditions can be drawn. In Kranenburg *et al.* [161], it was shown that by aligning folk song melodies using criteria derived from pitch, rhythm and segmentation-based scoring functions good retrieval of similar melodies can be achieved. One difficulty in the retrieval of melodic similarity in traditional music is that the same song can differ widely from interpretation to interpretation, as pointed out by Bohak and Marolt [18]. For a set of symbolic representations of folk music samples they propose using various statistics derived from the melodies to get a retrieval that is robust to these changes. However, it is pointed out that the achieved accuracies are not sufficient to result in a fully automatic variant classification system. Statistics derived from symbolic representations of melody are also used by Toiviainen and Eerola [153] in order to provide an analysis tool based on Self Organizing Maps (SOM) [93]. These two dimensional representations enable for a comparison of various songs according to their distance on the map. In Grachten *et al.* [61] an approach is proposed for the symbolic similarity between melodies, which is based on measures derived from the Implication/Realization model [117]. Even though their approach performed well on non-traditional music and the Implication/Realization model has been used to analyze folk music by Thompson and Stainton [151], to the best of the author's knowledge it has not been applied to similarity tasks in folk music. An overview of methods for symbolic melody similarity is given by Typke [157], who proposes the usage of transportation distances.

The concentration of research on symbolic representation can lead to loss of valuable information, as pointed out by Müller *et al.* [115]. There, the authors propose a system that is able to segment folk song recordings into its constituent stanzas and to visualize their contents. For this, chroma features [46] are derived both from the available symbolic representation and from the monophonic audio, and Dynamic Time Warping [130] techniques are used to get alignments. On audio signals, Moelants *et al.* [111] and Bozkurt [19] derive pitch histograms from monophonic recordings, the former using African music and the latter in the context of Turkish music. Both methods are aimed towards the recognition of underlying tonal concepts (*i.e.* scales or *makams*, respectively), and stress the importance of a finer frequency solution than the one provided by the chroma features. Cabrera *et al.* [23] investigate the estimation of melodic similarity on a set of mainly monophonic vocal Flamenco recordings. They use pitch and note duration estimations as proposed by Gomez and Bonada [54], and use them to derive *e.g.* displays of similarity relations based on phylogenetic trees, which were previously proposed by Daz-Banez *et al.* [37] for the analysis of relations in the rhythmic aspect of Flamenco music.

The focus in this thesis will be the comparison of polyphonic audio signals of traditional music. To the best of the author's knowledge, this task has not yet been approached in literature. However, it bears similarity of a well known task for popular music, which is the detection of cover songs. This means that given two songs we would like to determine the likelihood that one is a different interpretation of the other. This task has been approached by Ellis and Poliner [46] using beat-synchronous chroma features, which are compared using two dimensional correlation between the feature matrices. A different signal representation has been chosen for the task of cover song detection in [140], and instead of beat tracking a dynamic programming procedure was chosen to get similarity estimations. However, in Liem and Hanjalic [103] it is shown that the chroma features enable for higher accuracies in cover song detection.

In Chapter 7 of this thesis, the system presented by Ellis and Poliner [46] will serve as a baseline system for the detection of melodic similarity in traditional music. It will be evaluated if usage of a finer frequency resolution and the estimation of the fundamental frequency of the main melody can lead to an improved similarity measure on a dataset of polyphonic recordings of traditional music. Furthermore, the integration of melodic and rhythmic similarity measures into a system for the estimation of morphological similarity will be proposed.

Chapter 3

Onset Detection

3.1 Dataset

In order to evaluate the performance of a musical instrument onset detector an annotated dataset is necessary. Although in recent years many publications have treated the problem of onset detection, experiments are usually performed on small datasets with uneven class distributions [10], or on datasets containing samples with several instruments playing at the same time. No sufficiently large onset annotated dataset of different pitched musical instruments is publicly available. Such a dataset would make fundamental research on the accuracy of onset detection techniques feasible. For this thesis, a dataset has been compiled, which is described in the Appendix A in more detail. Non-pitched percussive instruments, such as drums and percussion, have not been included in this dataset as their onsets can be considered easy to detect, for example by using criteria derived from their energy envelope [10]. Since for onset detection the characteristic of the excitation is a crucial point, the instruments have been grouped into the following classes according to this aspect: pitched-percussive instruments, wind instruments and bowed string instruments. All samples are monophonic. Effort has been made, such that each of the above classes is represented by a similar number of samples and instruments. Furthermore, besides the choice of instruments commonly used in western music, also instruments of Turkish music are included. This enables to compare the influence of the musical style on the accuracy of onset detection systems. As detailed in Appendix A, for annotating new samples a supervised procedure was adapted [33], in which each annotation is cross-checked by a second person. For the annotation the wavesurfer[1] software was used. Spectrogram, waveform and the F0 curves were used simultaneously to locate the onsets that were perceived in the sample.

As depicted in Tables A.2 and A.3, the dataset is divided into a main set (MS) and a development set (DS). MS contains 1829 annotated onsets in 57 files of 11 musical instruments, 21 more samples of the instruments guitar, *ud*, piano and violin are contained in the DS. These files were used for parameter evaluations and development, and therefore this dataset will be referred to as development set (DS) in the following Sections. DS contains 674 onsets, see Table A.3 for details.

Note that the focus lies on evaluating onset detection methods on monophonic pitched musical

[1] http://www.speech.kth.se/wavesurfer/

instrument sounds, in order to determine performance of onset detectors at a basic level using simple signals. Therefore, we recall that all the 78 collected samples of MS and DS contain only one instrument each. The datasets used in Daudet *et al.* [33] and in Bello *et al.* [10] contain samples with several instruments playing together, which is referred to as complex mixture in Bello *et al.* [10]. In order to get a broader perspective, the complex mixture samples from these two publications were combined to a dataset of thirteen complex mixture samples with an overall number of 498 onsets. This data will be referred to as complex mixture dataset (CMS).

3.2 Onset Strength Signals

As detailed above, it is the goal of this chapter to evaluate three characteristics of musical instrument signals for their efficiency in onset detection: phase spectra (in terms of the phase slope function), magnitude spectra, and fundamental frequency contour. For this, the audio waveforms at a sampling frequency of $44.1kHz$ are used to derive onset strength signals (OSS), which are expected to have local maxima at the samples which are related to musical onsets in the waveform. These OSS are computed using the sampling frequency of $f_{ons} = 175Hz$ (5.7ms). This sampling frequency guarantees a temporal solution which is equal to the minimal distance at which two sound events can be perceived separately, which was found to be at most $10ms$ [112].

3.2.1 Phase Slope

A signal $x[n]$ can be described in frequency domain by its Fourier transform $X(\omega) = A(\omega)e^{j\phi(\omega)}$, with ω denoting frequency and $A(\omega)$ being the amplitude spectrum and $\phi(\omega)$ being the phase spectrum. The basic motivation for using the phase spectrum $\phi(\omega)$ of a signal for onset detection arises from properties of the group delay which is defined as

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{3.1}$$

The group delay of a delayed unit sample sequence $x[n] = \delta[n - n_0]$ is $\tau(\omega) = n_0$. This holds because $x[n]$ has the Fourier Transform $X(\omega) = e^{-j\omega n_0}$ with the phase component $\phi(\omega) = -\omega n_0$. Computing its derivative regarding frequency (*i.e.* the group delay) results in $\tau(\omega) = n_0$, $\forall\omega$. This means that computing the average value of the group delay results in a value equal to the temporal distance between the center of the analysis window (at zero) and the position of the impulse (at n_0). This holds in general for the output of a minimum phase system excited by a delayed unit sample sequence as shown in Smits and Yegnanarayana [144]. Two simple examples are depicted in Figure 3.1. The upper two panels (a) and (b) show the delayed unit sample sequence and its group delay, respectively. In the lower two panels in Figure 3.1, the sequence shown in panel (a) is convolved with a minimum phase system, resulting in the signal shown in (c) and the corresponding group delay shown in (d). Note that in (d), the average of the group delay is again equal to the displacement between analysis window and the delay of the unit sample sequence. The peaks that appear in (d) are caused by the poles of the minimum phase system. Computing the average group delay, the influence of these poles

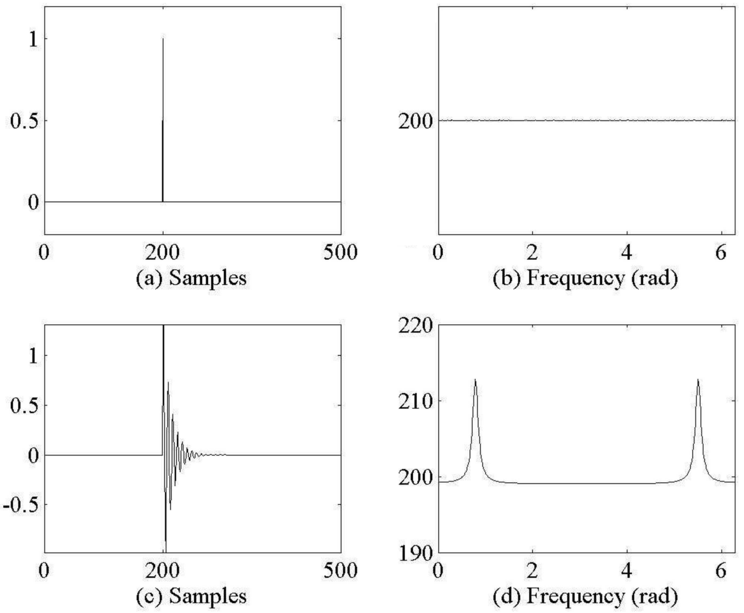


Figure 3.1: (a) A unitary sample sequence delayed by 200 samples. (b) The group delay function of the signal in (a). (c) A minimum phase signal with an oscillation at $\pi/4$. (d) The group delay function of the signal in (c).

dissapears. This basic observation leads to the assumption that the onset of a note played by an instrument can be determined using group delay, because an instrument can be sometimes considered as a minimum phase system excited by an impulse. This impulse can be caused by, for example, a hammer, a bow, the finger of a guitar player. An exception to this model is for example a violin player changing the left hand position while not changing the excitation, *i.e.* the movement of the bow. It is important to note that it has been shown that impulses can be detected with little impact of their actual amplitude by using group delay [83]. Furthermore, it has been shown, that onset (click) detectors based on the group delay are robust to additive noise as well [71, 83]. This means that even onsets that cannot be observed at all in magnitude can be detected using group delay.

In order to get a meaningful descriptor for onset strength from group delay as depicted in (3.1), the negative of its average is determined at each position of the analysis window. This value corresponds to the negative of the slope of the phase spectrum of the examined signal, and will thus be referred to as the phase slope $\tilde{\tau}$. In Figure 3.2, the dashed lines show sequences of phase slopes obtained when shifting analysis windows over the depicted signals. It can be observed that at all points where the center of the analysis window coincides with the position of an impulse, the phase slope has a positive zero crossing. In Figure 3.2 changing the length of the analysis window from signal period (long window) to a length shorter than the signal period (short window) does not affect this property of the phase slope. Also, as it was mentioned before, the efficiency of the phase slope function is not affected by the amplitude of the onset as it is clearly shown in 3.2b. Thus, the fundamental idea is to detect the onsets of musical instruments by determining the positions of the positive zero crossings of the phase slope.

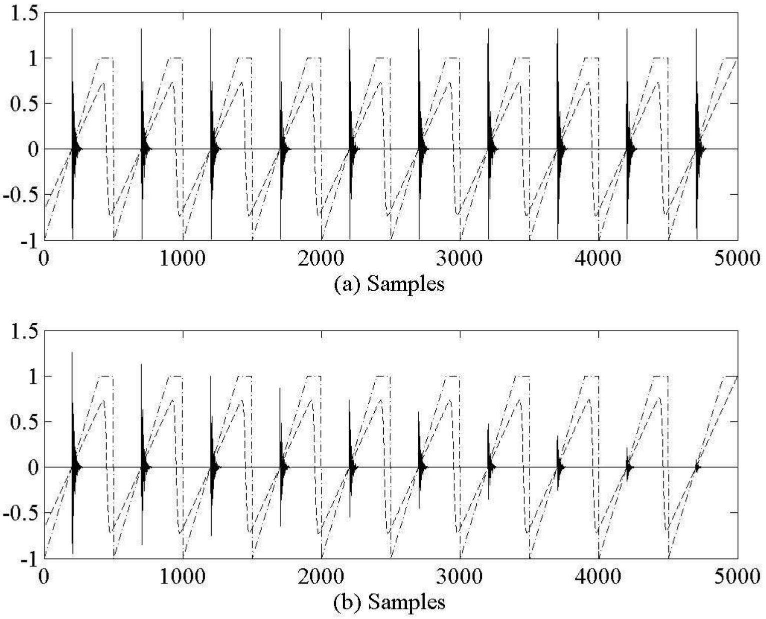


Figure 3.2: (a) A sequence of impulses of constant amplitude and the associated phase slope function using long (dashed line) and short (dash-dotted line) window (b) A sequence of impulses with linearly time varying amplitudes and the associated phase slope function using long (dashed line) and short (dash-dotted line) window.

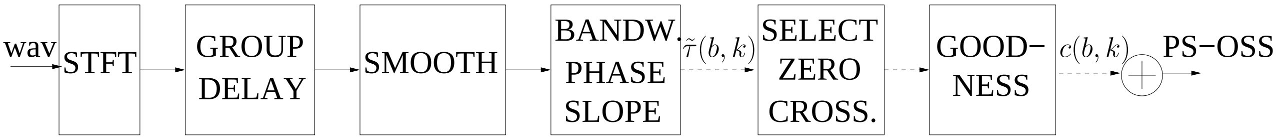


Figure 3.3: Block diagram of the PS_OSS computation

Details on the computation of the slope function can be found in Kandia and Stylianou [83]. As observed by the author in [71], group delay cannot be used in such a straightforward way when dealing with music signals. Specifically, it was found that the group delay had to be computed in several frequency bands separately, while a selection of zero crossings was necessary [71]. In this work, parameters like the number of bands or criteria for the zero crossing selection are determined using only the development dataset DS described in Table A.3.

The block diagram in Figure 3.3 shows the processing steps for the computation of the phase slope onset strength signal (PS_OSS). The first processing block consists of the computation of the Short-time Fourier Transform (STFT) of the signal $x[n]$

$$X(\omega, k) = \sum_{m=0}^{N-1} x[m + kh]w[m]e^{-j\omega m} \tag{3.2}$$

where hop size h is set to $5.6ms$ in order to achieve the sampling frequency $f_{ons} = 175Hz$. The window length N of the applied Hanning window $w[n]$ has been set to $0.1s$. In order to apply the FFT algorithm the signal is zero padded. Note that in the context of the beat tracking task presented in Holzapfel and Stylianou [71], the analysis window length had been set to $0.2s$. It was found that reducing the window size to $0.1s$ leads to detecting more of the annotated onsets, while the number of false positive detections slightly increased. Thus the accuracy of the detection is increased using the smaller window, while for beat tracking a cleaner OSS is preferable that can be obtained by using a larger analysis window. The second processing block contains the computation of group delays. To avoid the problems of unwrapping the phase spectrum of the signal for the computation of group delay, it is computed as in Oppenheim *et al.* [118]:

$$\tau(\omega, k) = \frac{X_R(\omega, k)Y_R(\omega, k) + X_I(\omega, k)Y_I(\omega, k)}{|X(\omega, k)|^2} \tag{3.3}$$

where

$$\begin{aligned} X(\omega, k) &= X_R(\omega, k) + jX_I(\omega, k) \\ Y(\omega, k) &= Y_R(\omega, k) + jY_I(\omega, k) \end{aligned}$$

are, respectively, the STFT of $x[n]$ and $nx[n]$ in analysis frame k, respectively.

In the third processing block, each frequency bin in the group delay vector $\tau(\omega, k)$ is median-filtered in time: $\tau(\omega, k) = \mu_{1/2}(\tau(\omega, k - i))$ for $i = [-4, -3, ..., 4]$. This 9-th order median filtering is necessary due to the presence of many instruments with soft onsets in the dataset. It has been observed that especially for bowed string instruments onsets have a temporal extent of up to 50ms, which is about the length of the median filter at the sampling period of $t_{ons} = 5.6ms$ (observe *e.g.* the cello signal shown in Figure 2.1). Thus, this value represents an upper bound for the precision achievable on this dataset. Next, the group delay vectors are divided into 21 non overlapping frequency bands, as proposed in Klapuri [88]. This transition to bandwise processing is indicated by dashed lines in Figure 3.3. In each band the negative of the median of the group delay values is determined, resulting in $[b = 1...21]$ phase slope values $\tilde{\tau}(b, k)$ for each frame k. The exact number of bands was found to be uncritical, if this is chosen to be bigger than 5. Also, dividing the bands as proposed in Klapuri [88] leads to a linear division for low frequencies. This was found to be crucial, because choosing for example logarithmic frequency bands causes the group delays of the lower bands to contain too few coefficients, and the medians are too noisy in these bands.

As mentioned by the author in [71], in each band the selection of zero crossings is necessary. In Figure 3.4 the phase slope computed in the third band of a guitar signal is shown, along with the manually annotated onsets depicted as impulses. It can be observed that the phase slope has some spurious zero crossings, for example short after sample 100. It was observed that accepting only the positive zero crossings that are surrounded by large oscillations improves the accuracy of the detection. Such oscillations can easily be detected by thresholding, as shown by the dotted lines in Figure 3.4. The positive threshold was determined by the mean of the absolute values of the phase slope for a whole sample; the negative threshold was simply the negative of this value. A positive zero-crossing is selected if the minimum and the maximum amplitude of the phase slope function, before and after the zero-crossing, respectively, pass the corresponding thresholds. Note that this way in the example shown in Figure 3.4, there is no false positive detection. There are two missed onsets (at samples 384 and 609), because the phase slope has no peaks before and after these zero-crossings that cross the dotted threshold

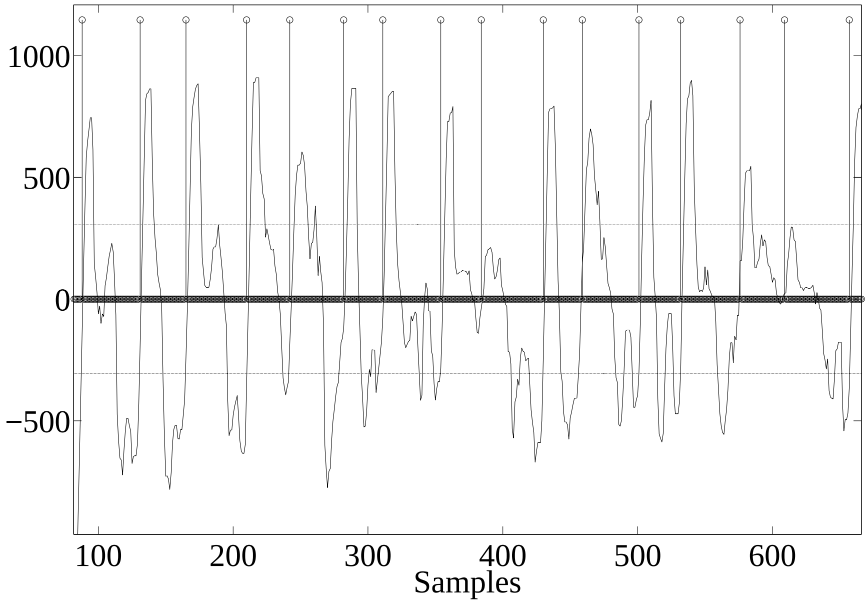


Figure 3.4: Phase slope computed from the third band of a guitar signal

line. Different values of the threshold and the usage of adaptive thresholding have been tried, but the presented method was found to be sufficient.
The next processing block in Figure 3.3 is the goodness computation: for each of the above selected positive zero crossings, a value is assigned that denotes how much confidence can be given that this zero crossing coincides with an onset. In the context of speech excitation a method was proposed that measures the deviation of the computed phase slope from an ideal one (*i.e.*, straight line) [144]. This approach was evaluated but a simpler solution was found: the confidence value is set to the value of the derivative of the phase slope in the vicinity of the zero crossing. High value of this derivative signifies high confidence. The output of this final block for the k-th analysis window is the confidence level vector $c_b(k)$, that contains either the value zero in the b-th band, when no zero crossing was selected, or the computed confidence value for this zero crossing. The final onset strength signal PS_OSS is then computed by summing $c(b,k)$ over the 21 bands: $\text{PS\_OSS}(k) = \sum_{b=1}^{21} c_b(k)$. It has been considered to use different weighting schemes for this summation as proposed for example in Klapuri *et al.* [92]. However, no weighting scheme could be determined that lead to consistently improved onset detection performance.

3.2.2 Spectral Flux

Spectral Flux (SF) is based on the detection of sudden positive energy changes in the signal which indicate attack parts of new notes. Mainly there are two kinds of spectral flux OSS based on L1-norm and L2-norm as presented below:

$$\begin{aligned} \mathsf{SF\_OSS}_{L1}(k) &= \sum_{\omega} H(|X(\omega,k)| - |X(\omega,k-1)|) & (3.4) \\ \mathsf{SF\_OSS}_{L2}(k) &= \sum_{\omega} H(|X(\omega,k)| - |X(\omega,k-1)|)^2 & (3.5) \end{aligned}$$

where $H(x) = \frac{x+|x|}{2}$ is the half-wave rectifier function, and $X(\omega, k)$ is the STFT of the signal with $5.6ms$ hop size and a window length h of $46ms$. For the experiments in this thesis, the L1-norm SF is used, because it was shown that L1-norm outperforms L2-norm [41]. The accuracy of onset detection using SF_OSS and its computational simplicity were presented in [10, 41].

3.2.3 Fundamental Frequency Change

Considering the fact that note onsets are often difficult to observe in the amplitude in the case of pitched non-percussive instruments, it was decided to evaluate an additional onset strength signal. When playing for example a bowed string instrument, it is possible to create a new note onset by changing the position of the finger on the fingerboard while keeping the excitation caused by the bow constant. Because of the constant excitation, these onsets will be difficult to be observed in the phase slope as well. The only clear change is then the fundamental frequency (F0). Thus, it was decided to compute an OSS using the F0 estimations produced by the YIN algorithm [38].
At first, F0 estimations were calculated every $t_{ons} = 5.6$ ms. It is common practice to use the cent unit (obtained by the division of an octave into 1200 logarithmically equal partitions) for musical F0 analysis. Most of the musical pitch perception studies use this logarithmic measure of relative pitch which can be easily computed by:

$$F0_c = H(1200 \log_2(F0_{Hz}/c_{ref})) \tag{3.6}$$

where the reference frequency, the frequency of note lowest-C, is $c_{ref} = 440 \cdot 2^{-69/12} \approx 8.1758Hz$ and $H(x)$ is again the half wave rectifier as introduced in (3.4). The application of the rectifier sets all values smaller than c_{ref} to zero, including the points were the YIN estimator did not compute any pitch ($F0_{Hz} = 0Hz$). The computed sequence of pitch values is checked at the points where no F0 has been computed by YIN (*i.e.* $F0_c = 0$). This is either the case in silence parts, or at unstationary parts of the sound like instrument onsets. For this, a simple silence detector was applied. Whenever missing pitch values coincide with silence, the pitch was set to the pitch of the previous frames. Otherwise, the missing pitch values were set to the next computed pitch. This way, the robustness to silence parts and the accuracy of the onset estimation was improved. An example for this improvement is shown in Figure 3.5: a typical example of the F0 estimation in the proximity of an onset for a cello signal. In this Figure, the F0 change at sample 245 is related to an onset, but the pitch estimator gives a correct F0 just after sample 260. In this example, samples 245 to 260 are non-silent frames and the pitch estimation is corrected, as shown by the dotted line, avoiding one false detection at sample 260. The final onset strength signal F0_OSS is computed from the silence-filtered

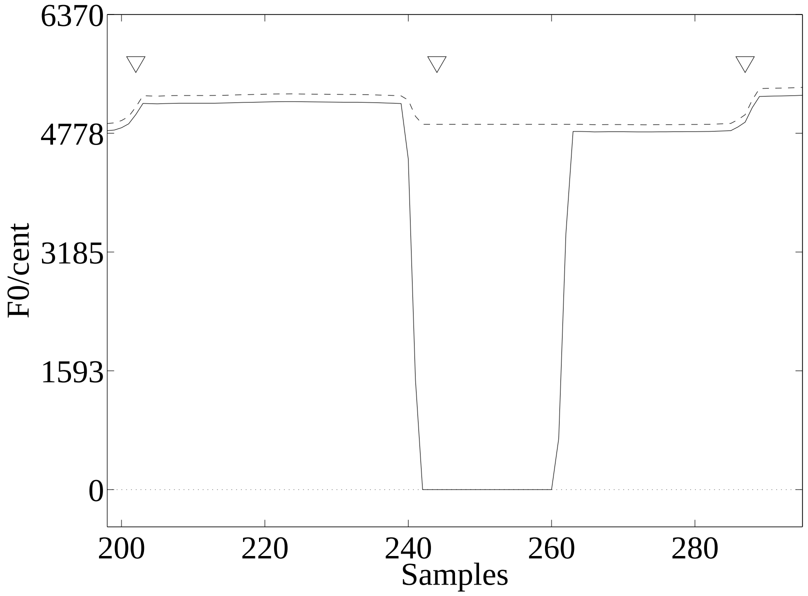


Figure 3.5: F0 estimation for a cello sample before (bold) and after (dashed) silence filter, onsets at positions marked with arrows (samples 202, 245 and 288, respectively)

F0 estimations F_0 as

$$\Delta F_0(t) = \min \begin{cases} \mathrm{mod}_{1200}(|F_0(t) - F_0(t-1)|)/600 \\ \mathrm{mod}_{1200}(-|F_0(t) - F_0(t-1)|)/600 \end{cases} \tag{3.7}$$

This difference curve will have positives peaks at the instants of F0 changes. The magnitude of the peak depends on the change of F0. The modulo operator was applied to prevent from octave errors (note that 1200 cent is equivalent to an octave). The division by 600 normalizes the range of the F0_OSS from zero to one.

3.2.4 Fusion

It can be assumed that using each of the three OSS, it is possible to detect different types of onsets: SF captures onsets that are observable in magnitude change (hard onsets), F0 can detect note changes that happen in presence of a constant excitation, and PS can detect onsets that are characterized by the start of an excitation but that are not detectable in amplitude (soft onsets). As it is desirable not to select the optimal detector manually depending on the signal, a fusion of the three OSS can combine their advantages. In experiments combining the features to a three dimensional space was tried (feature fusion), as well as the linear combination of the features to a single dimensional descriptor. Both approaches were not successful due to the following reasons: The sparseness of the OSS (many zero values) causes problems when trying to train classifiers in the three dimensional space. Apart from that, the onsets are not exactly aligned in the three OSS: the beginning of an excitation is detected by PS while the maximum change in energy will be detected by SF, which happens typically with a temporal

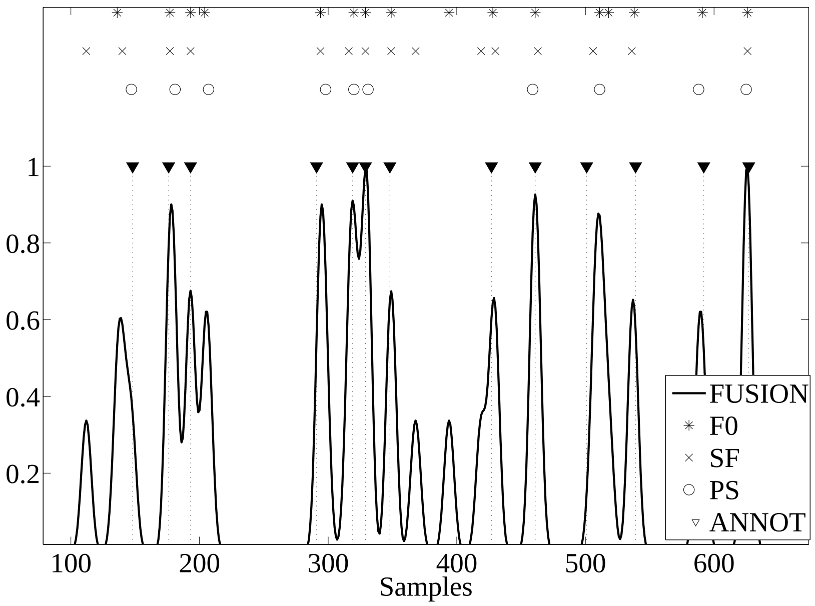


Figure 3.6: Example for the decision fusion using a *ney* sample. Dotted lines with arrow markers show reference onset annotations, above the FUSE_OSS the positions of the onsets determined by F0_OSS, SF_OSS and PS_OSS are marked.

delay. A simpler solution that avoids these traps is the fusion at the decision level: Using each OSS separately onsets are determined. This results in three vectors: for each OSS, one decision vector with sampling frequency f_{ons} is obtained that has value one at the detected onsets and zero value elsewhere. By first summing these three vectors and then smoothing with a $75ms$ Hanning window a fusion onset strength signal (FUSE_OSS) is obtained. Onsets can then be defined by a peak picking on FUSE_OSS. An example FUSE_OSS is shown in Figure 3.6. The dashed impulses show the reference onset annotations. The positions of the onsets determined by F0_OSS, SF_OSS and PS_OSS are also marked. It can be seen that the resulting FUSE_OSS has the largest maxima when all three OSS detect an onset close to this point. When only one OSS detects an onset this leads to a small amplitude in the FUSE_OSS, as for example at samples 370 and 390, where the onsets have only been detected by SF_OSS and F0_OSS, respectively. In the example shown in Figure 3.6 there is an improvement by using the FUSE_OSS for onset detection. The general performance of FUSE_OSS will be provided in Section 3.5.

3.3 Evaluation Methods

In the MIREX onset detection evaluation, the F-measure of the detection is computed as the main criterion. F-measure is defined as

$$F = \frac{2PR}{P+R} \tag{3.8}$$

with Precision, P, and Recall, R, being computed from the number of correctly detected onsets (N_{tp}), the number of false alarms (N_{fp}), and the number of missed onsets (N_{fn}) as

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}}\,,\; and\, R = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{3.9}$$

According to the MIREX specifications, onsets are counted as correct detections when they are within a window of $t_{tol} = \pm 50ms$ around the onset annotation. If there are several onset detections in this tolerance window, only one is counted as true positive, the others are counted as false positives (double detections). If a detection is within the tolerance window of two annotations one true positive and one false negative are counted (merged onsets).

In order to get a more detailed description of the accuracy of an onset detector, the threshold δ applied to the OSS in the peak picking process (see Section 3.4) can be varied in small steps. This way precision P and recall R values can be obtained for different threshold values, and P/R-curves are created by putting R values on the abscissa and P values on the ordinate. This representation has been proposed in the MIREX 2007 onset detection evaluation as well. In P/R-diagrams, the best onset detector, in terms of F-measure, is the one whose P/R-curve is closer to the upper right corner of the diagram. Furthermore, given a fixed threshold δ for the peak picking, the F-measures can be computed for varying time tolerances t_{tol}, to get an impression of how close the true detections are to the annotation in time. This gives a second representation besides the P/R-curves: plotting the F-measures over different tolerances in ms, which will be referred to as F/T-curve.

3.4 Peak Picking

In order to determine the time instants of onsets from the OSS described in Section 3.2, the salient maxima in the OSS need to be detected. As mentioned in [10], this process is of major significance for the accuracy of the result. In Figure 3.7 the basic blocks, as described in Bello *et al.* [10], of a peak picking process are depicted. In order to smoothen the onset strength signals they were filtered using a Hanning window of short length. Normalization refers to the subtraction of the means and the division by the variance of the OSS (z-score). The low pass filter is a simple third order FIR filter with a cutoff frequency at $f_{ons}/5$. The adaptive threshold is computed by applying a moving median filter to the OSS. This threshold is then subtracted from the OSS to cancel dynamic changes. The length of the moving median filter was set to 17 samples ($97.1ms$). The peak picking is a simple selection of local maxima. The performance of the OSS depends on the setting of a parameter δ, that defines the threshold

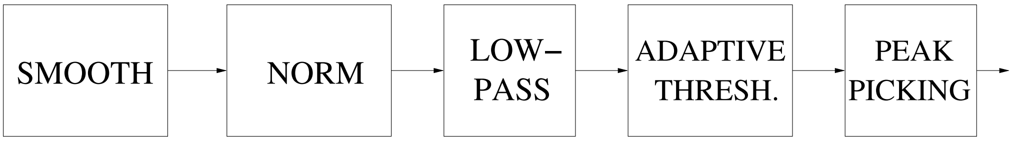


Figure 3.7: Schematic of peak picking

that a local maximum has to excess in order to be selected as an onset. Threshold δ can be varied in small steps in order to create the P/R-curves described in Section 3.1. An optimum threshold for an OSS can be obtained from the corresponding P/R-curve by determining which threshold leads to the best F-measure. Using this optimum threshold, the F/T-curves can be generated by changing the desired tolerance.

For each of the three OSS, the peak picking has been optimized by evaluating its accuracy according to the F-measure described in (3.8) on the development set. The resulting optimum peak picking procedure for the SF_OSS and F0_OSS are the same as described in Bello *et al.* [10], except of the additional computation of a smoothing (first block in Figure 3.7) which is not mentioned in [2] as crucial; however, we found that this smoothing improved our results and therefore it was decided to include it. The length of the applied Hanning window was $51ms$. This degrades the possible resolution in time, but as detailed in Section 3.1 the required resolution is $100ms$ because of the temporal extent of onsets of non-percussive instruments. Furthermore, the locations of the peaks can be preserved by applying zero-phase filtering.

For PS_OSS the application of an adaptive threshold (fourth block in Figure 3.7) computed using a moving median filter was found to degrade the accuracy. This is due to the different characteristic of this OSS: it is not immediately derived from the temporal change of a signal property, but it is a time series of confidence values at some candidate onset instants, and contains more zero values than the other two OSS. Applying a moving median leads to the removal of too many onset candidates. Apart from that, the adaptive thresholds compensates changes in the strength signals due to changes in signal amplitude. These changes obviously do not affect the PS_OSS.

Thus, while for the peak picking in SF_OSS and F0_OSS all blocks of the diagram in Figure 3.7 are active, for PS_OSS the fourth block (adaptive threshold) was left out.

3.5 Results

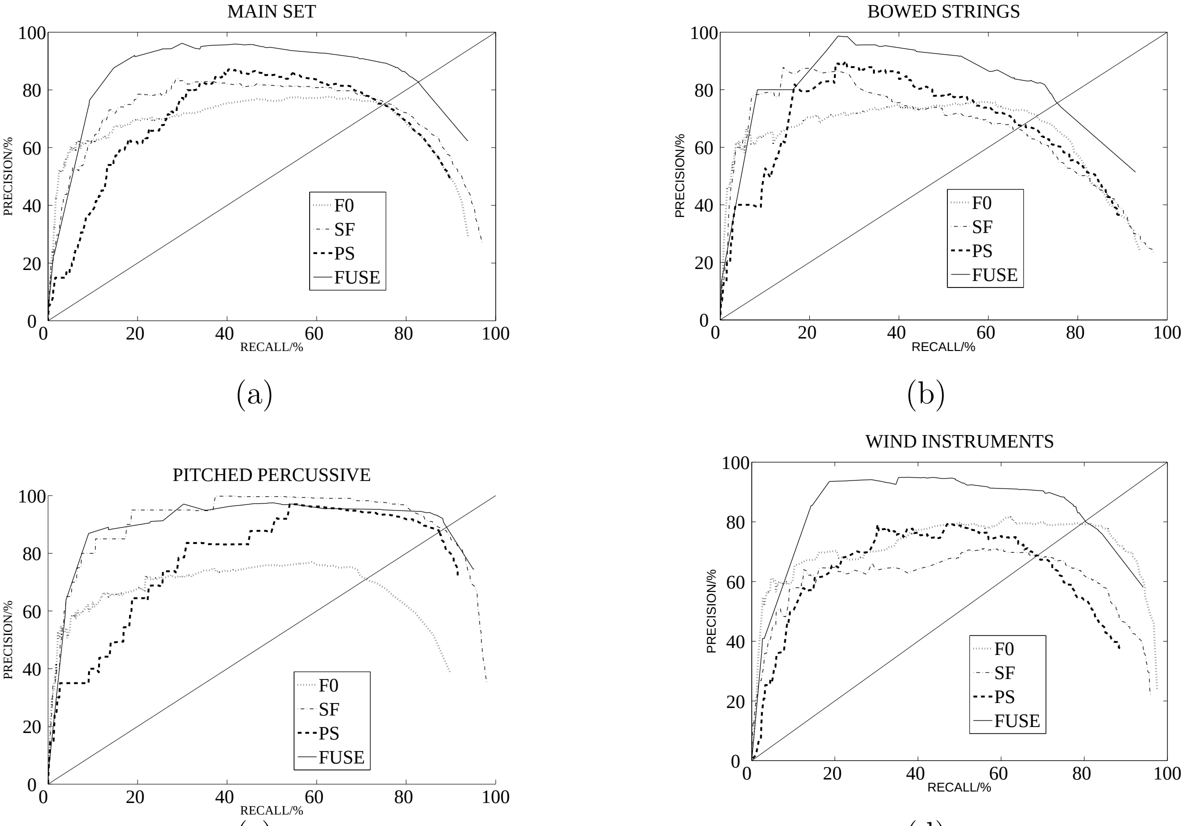
The performance of the OSS on the main dataset is shown in the P/R-curves in Figure 3.8. Figure 3.8.(a) shows the performance on the complete MS as described in Table A.2. Regarding their optimum F-measure, all three single OSS perform almost equally well, which can be seen from the fact that they cross the diagonal in the graph at almost the same point. The PS_OSS achieves higher precision values, while the SF_OSS is able to achieve higher recall rates. This means that when a low false alarm rate is of importance, as for example in beat tracking tasks, PS is superior to SF, which confirms the findings in Holzapfel and Stylianou [71]. Combining the decisions of the three OSS leads to a clear improvement of the performance. This can be seen from the large distance of the corresponding P/R-curve in Figure 3.8.(a), and from the best F-measures on the main set as listed in Table 3.1. Here, the F-measure of the decision fusion (82.1%) improves the best single OSS F-measure (74.1%) by 8.0%. It is important to note that the three thresholds δ used in F0_OSS, SF_OSS and PS_OSS for the decision fusion are the ones that resulted in the best F-measure on the main dataset MS. No significant difference was observed when these threshold values have been derived from either DS or the data from Bello *et al.* [10]. The three threshold values were left unchanged in the experiments conducted on the various subsets of the data (wind instruments, Turkish instruments, *etc.*) and on other datasets, in order not to present over-optimistic results for FUSE_OSS.

Table 3.1: F-measures of the OSS, along with threshold values δ

OSS	F0	SF	PS	FUSE
F-measure	74.1%	73.9%	73.7%	82.1%
δ	0.012	0.051	0.027	7.78

All the plots in Figure 3.8 were produced using a $\pm 50ms$ tolerance window. In Figure 3.9, the F-measures are shown as a function of the tolerance value t_{tol}. This plot was generated using the MS dataset, and the threshold values that produced the optimum F-measures as listed in Table 3.1. It can be seen that for bigger tolerances, PS and F0 are superior to SF, but their performance decreases when demanding higher accuracy in time. For PS_OSS, this is due to the usage of the time median filtering in the phase slope computation, as described in Section 3.2.1. The accuracy of PS_OSS can be improved by using a shorter median filter, which is possible when only hard onsets are considered. The decision fusion results in F-measures that are clearly superior for all desired tolerance values. The decreasing F-measures for F0_OSS for low tolerance values is caused by the uncertainty of the pitch estimation close to the onsets. However, when considering subfigures (b) and (d) in Figure 3.8 the advantages of using a fundamental frequency estimation for onset detection can be observed: for both wind and bowed string instruments, F0_OSS achieves clearly improved F-measures compared to SF_OSS and PS_OSS; It is characterized by a curve that is closest to the upper right corner in both cases. Furthermore, for both wind and bowed string instruments the decision fusion can improve onset detection. The best F-measures of FUSE_OSS are higher than the best F-measures achieved with any single OSS. Moreover, FUSE_OSS improves the maximum precision. Note that for both instrument groups, using FUSE_OSS best precision values of more than 90% are reached. This leads to a very low false alarm rate when missing a number of onsets is accepted, which is typically desired in a beat tracking task as in Holzapfel and Stylianou[71]. For percussive onsets a well-known finding is confirmed: Because these types of onsets can be captured well from the magnitude spectrum, SF_OSS performs very well. Nevertheless, including also the information of the other OSS in the decision fusion leads to further improvement of the F-measure. The marginality of the improvement can be assigned to the bad performance of F0_OSS on this type of instruments. It was observed that the Yin F0 estimator had problems on these types of instruments, which are characterized by estimation errors especially in the vicinity of onsets.

As the dataset presented for onset detection evaluation in this chapter contains some western instruments and some Turkish instruments, experiments could be conducted to judge the influence of the style of performance on the onset detection. For this, a set of instruments was chosen which contains only western performance styles (clarinet, guitar, piano, trumpet). The Turkish performance style is represented by the instruments: *kemençe* (bowed string), *ney* (wind instrument), *ud* and *tanbur* (both plucked string instruments). Note that both groups contain two instrument types with percussive and two types with soft onsets. Thus, the influence of the instrument types in this comparison is small, as the differences affect only the instrument timbre and not the type of onset. Decision fusion produced clearly superior results for both western and Turkish performance style. The resulting P/R-curves for the



(a)

(b)

(c)

(d)

Figure 3.8: Performance of the onset detection, Precision/Recall values in % are plotted on Ordinate/Abscissa, respectively. Curves have been obtained by changing threshold δ in small steps.

decision fusion are shown in Figure 3.10. It can be seen that onset detection on the Turkish instruments is much more difficult. While similar maximum precision values can be achieved, the curve decreases rapidly when lowering the threshold of the detection (*i.e.* moving to higher recall rates). This coincides with an observation made in the onset annotation process: Turkish playing consists of many ornamentations which are difficult to annotate. These less salient onsets appear in the lower amplitudes of the OSS, and lead to the fall-off of the P/R-curve. This form of the curve causes a decrease of the F-measure from 89.8% for the western instruments to 77.8% for the Turkish instruments. This shows that not only the type of instrument but also the style of performance affects the accuracy of an onset detection system. However, in order to specify exactly how much of this decrease is caused by playing style, comparative studies with the same instruments played in both styles must be conducted. This decrease in the performance is likely to be encountered in other improvised forms of music as well, such as the folk tunes investigated in Collins [29].

For the complex mixture files in the CMS set described in Section 3.1 F0_OSS completely failed. This has to be expected since the Yin algorithm has been developed for F0 estimation from single sources. When complex mixtures are considered, an OSS will have to be derived from a multiple F0 estimator like the one described in Klapuri [89]. On CMS, the PS and

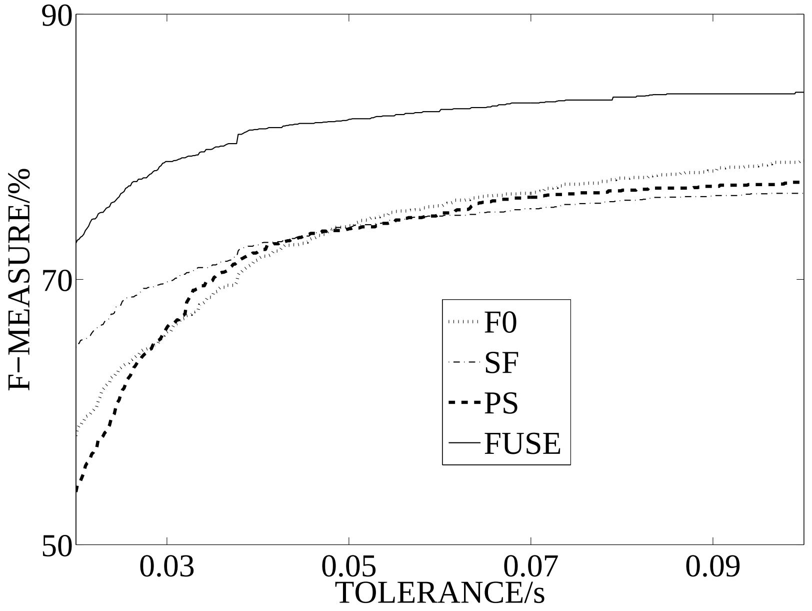


Figure 3.9: F/T-curve for the three OSS and the decision fusion on the MS

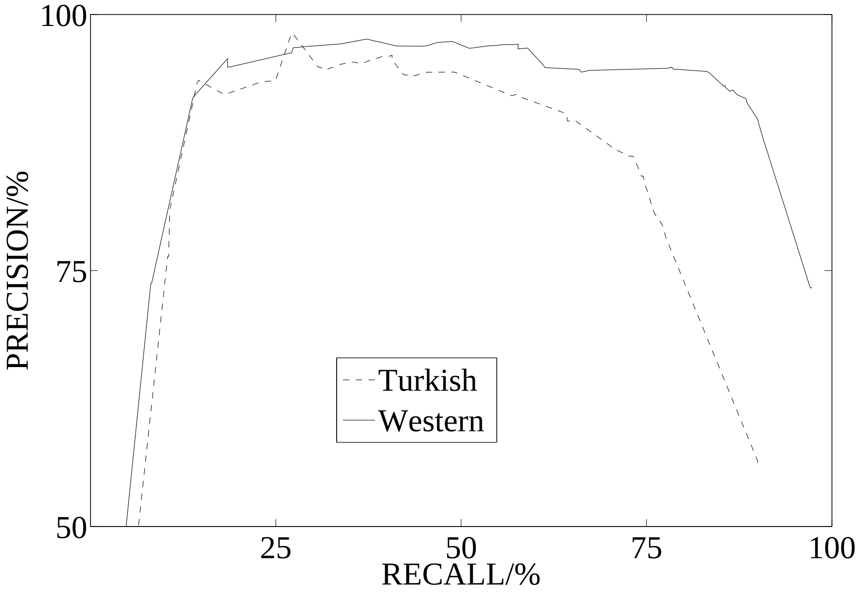


Figure 3.10: Precision/Recall curves for the decision fusion, using western and Turkish performance style separately

the SF onset strength signals were compared. Results show that about the same accuracy is obtained by using either of the two OSS. The obtained best F-measures on the data were 78.3% for the SF and 77.6% for the PS onset strength signals. The computed P/R-curves did not differ significantly. A decision fusion of only those two OSS resulted in a small improvement to an F-measure of 80.1%. Using a decision fusion of all three OSS for the CMS data results in an F-measure of 78.3%, *i.e.* the same F-measure as for SF_OSS alone. This shows that the proposed fusion method is robust even if one OSS completely fails. The influence of applying all subsets of OSS for decision fusion was evaluated on MS and all separate instrument groups, and the results are shown in Table 3.2. CMS data represents the only case in which decision fusion using all three OSS does not improve the detection performance. For all monophonic signals using all OSS leads to the best F-measures.

The dataset presented in Bello *et al.* [10] was used for experiments as well. As this dataset contains only 23 samples, it is difficult to use it for comparison with the results obtained on single instruments in this work. This is because out of the 23 samples only one is a bowed string instrument, while wind instruments are not contained in this dataset. It was decided to determine the average performance using all samples, and to exclude F0_OSS in these experiments. This is because there are seven complex mixture files and six non-pitched instrument files, and thus the usage of F0 on this data would be meaningless. On this data PS_OSS was slightly superior to SF_OSS in the sense of F-measure (90.4% compared to 89.0%). A fusion of these two OSS was not found to further improve results on this data.

Table 3.2: F-measures when using all combinations of OSS for decision fusion

	MS	P.PERC	WIND	BOWED	CMS
ALL	82.1	90.1	80.2	76.3	78.3
SF+PS	76.0	88.4	70.1	66.6	80.3
F0+PS	75.8	83.6	74.8	68.4	70.5
F0+SF	76.5	84.2	74.8	69.0	69.7

3.6 Conclusion

In this chapter a novel phase slope based onset strength signals (PS_OSS) was introduced. PS_OSS is able to reach good performance when considering soft onsets, while high precision values can be reached using this descriptor. The proposed F0_OSS performs very well for soft onsets in the sense of F-measure, but has problems for hard onsets due to inaccuracies of the F0 estimator. Because of that, and in order to use F0_OSS on complex mixtures as well, an appropriate F0 estimator must be used. The decision fusion of the onset detections derived from SF_OSS, F0_OSS and PS_OSS was shown to improve independently from the type of signal. Thus, it constitutes a method to detect onsets from pitched musical instruments without the necessity of choosing any signal dependent parameters.

Considering the dataset, it can be concluded that a diverse dataset of pitched instruments is now available for the evaluation of onset detection systems. Requests can be addressed to the author of this thesis. Comparing the best F-measures of the presented dataset using single OSS (73.3%) with the best F-measure of 90.4% achieved with a single OSS on the dataset presented in Bello *et al.* [10] it can be concluded that the dataset compiled for this work is more difficult and we expect it to be a valuable tool for researchers working in this area.

Chapter 4

Rhythm Similarity

As described in Section 2.2, the task of rhythmic similarity of music is approached in this thesis by using scale transform based descriptors. These descriptors are widely tempo invariant and enable to compare two pieces of music regarding their rhythmic content even when their differences in tempo are large. Furthermore, the proposed approach has no need of estimating meter properties of the piece such as the tempo or the beat, which is an advantage whenever music is encountered for that such properties are difficult to estimate. This holds especially for various forms of traditional music. This is because in many cases the absence of percussive instruments makes beat tracking an error-prone procedure for these signals, and for compound meters, which frequently occur in the music of the eastern Mediterranean, there is currently no reliable procedure available to perform this task. This chapter is organized in the following way: Section 4.1 introduces the proposed method, by giving a general overview in Section 4.1.1. The methods for computing the scale invariant rhythm descriptors for audio signals and for MIDI signals will be presented in Sections 4.1.2 and 4.1.3, respectively. In order to facilitate a better understanding of the proposed scale domain descriptors, in Section 4.1.4 some of their characteristics are detailed. In Section 4.2.1, the music collections will be described that have been used for evaluation. The characteristics of these datasets will be outlined, and their different demands to a rhythmic similarity measure will be described. Section 4.2.2 describes previously proposed measures that will serve as a baseline for comparison, and the evaluation method is detailed in Section 4.2.3. The experimental results are discussed in Section 4.3 and the chapter is concluded in Section 4.4. Note that in this Chapter, to a wide extend the accuracy of rhythmic similarity measurements will be evaluated in classification tasks on the presented datasets. In order to confirm the validity of relating the obtained classification results with the subjective term of rhythmic similarity, also listening tests have been conducted and their results will be shown in Section 4.3.

4.1 Suggested Rhythm Descriptors

In this section, first we provide general background of scale transform. Then, we describe the suggested method of measuring rhythmic similarities in music by distinguishing the cases of music representation by an audio waveform and by the MIDI format. More specifically, the necessary background will be provided in Section 4.1.1, and thereafter in Sections 4.1.2 and 4.1.3 the different demands of the waveform and the MIDI data will be addressed. Section

4.1.4 gives further information about characteristics of the proposed features.

4.1.1 Scale Invariant Rhythm Descriptor

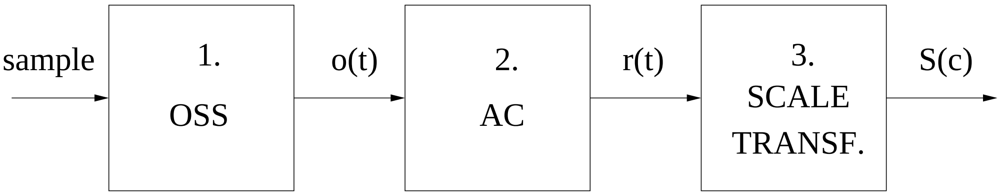


Figure 4.1: Computational steps of scale invariant rhythm descriptors

In Figure 4.1, the three steps in the computation of scale invariant rhythm descriptors are shown. As a pre-processing step towards a scale invariant description of rhythm, onset strength signals (OSS) at a sampling frequency of $f_{ons} = 50Hz$ are computed. This sampling period ensures that only frequencies related to the perception of rhythm are contained in the OSS, and was found to be sufficient compared to the higher sampling frequency of $175Hz$ that has been used for the OSS evaluations in Chapter 3. OSS have salient peaks at the instants where a musical instrument starts playing a note. For example, in Ellis [47] OSS have been computed from audio signals by using a method based on spectral magnitude differences, and in Parncutt [123] a method to compute OSS from a MIDI file was proposed. From the computed OSS, salient periodicities that are characteristic of the rhythm of the sample have to be found. In Holzapfel and Stylianou [73], STFTs of the onset strength signals were computed, referred to as periodicity spectra. If $X(\omega)$ is the Fourier transform of $x(t)$, then it is well known that:

$$\sqrt{a}x(t) \leftrightarrows \frac{1}{a}X(\omega/a) \tag{4.1}$$

In Figure 4.2, a periodicity spectrum of a Cretan dance sample of the class *Siganos* is shown in bold lines, while the periodicity spectrum of its time scaled version is depicted in dotted lines. The scaled version was obtained using the *audacity*[1] software, by applying the included plug-in for changing tempo of an audio file with a scale factor of $a = 1.1$. The scaling in the frequency domain representation can be recognized in Figure 4.2. The immediate computation of a point wise distance between the depicted periodicity spectra is affected by the time scaling caused by the different tempi.

In this work, the use of the scale transform is suggested to overcome the tempo differences between music pieces that are similar in terms of their rhythm. The scale transform is a special case of the Mellin Transform, defined as [27]:

$$X(c) = \frac{1}{2\pi}\int_0^\infty x(t)e^{(-jc-1/2)\ln t}dt \tag{4.2}$$

[1] http://audacity.sourceforge.net/

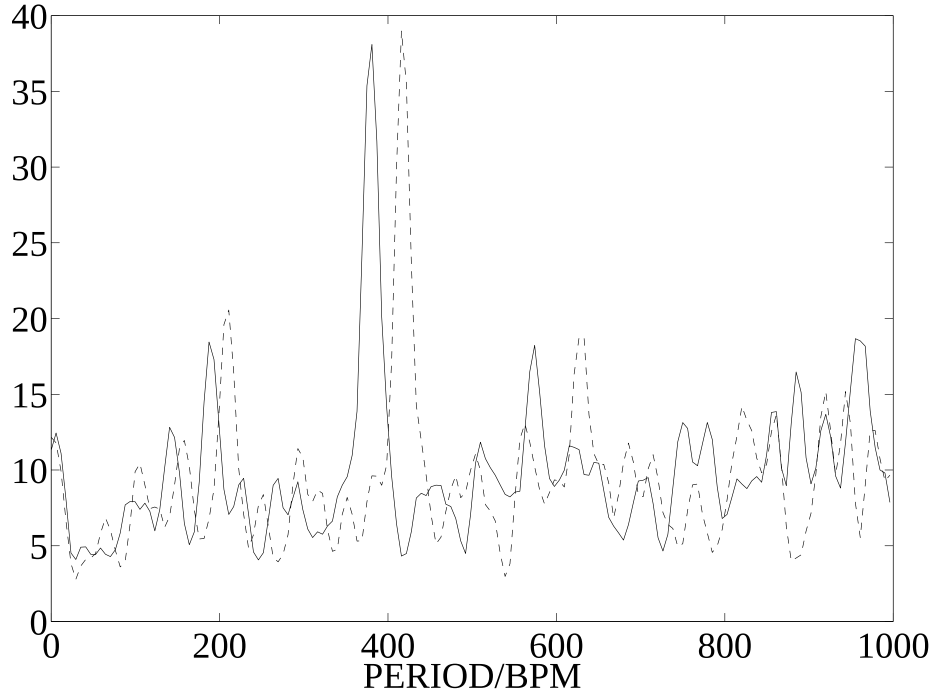


Figure 4.2: Periodicity spectra of original (bold) and time scaled (dashed) Cretan dance sample, Time scale factor: $a = 1.1$

and it can be shown to be scale-invariant, which means that the magnitude distributions of the scale transforms of signals $x(t)$ and $\sqrt{a}x(at)$ are equal [27]. Although the scale transform is scale invariant, it is not shift invariant. This means that x(t) and $x(t-a)$ have different scale transform magnitudes. Instead of using OSS, as usually suggested in this context (i.e., [73] and references there in), and motivated by the approach described in Combet*et al.* [30], we suggest to use the autocorrelation function $r(t)$ of OSS as a descriptor for the rhythm. It is worth noting that the autocorrelation function of a scaled signal is equal to the scaled (by the same scale factor) version of the autocorrelation of the original signal. By using the autocorrelation function of OSS we overcome the shift-variant property of the scale transform. Therefore, the suggested approach is scale (or tempo) and shift invariant. Throughout the chapter, the computed autocorrelations were normalized, so that their value at the zero lag equals to one. In Figure 4.3, the scale magnitudes for the same examples used in Figure 4.2 are depicted. It is evident that their scale magnitudes are essentially the same and they can be compared by a point to point distance measure in a straightforward way, avoiding the dynamic programming procedure proposed by the author of this thesis in [73].

The computation of the scale transform can be performed efficiently by using its relation to the Fourier transform [139]:

$$R(c) = \int_0^\infty r(e^t)e^{1/2t}e^{-jct}dt \tag{4.3}$$

which is the Fourier transform of the exponentially warped signal weighted by an exponential window. Since the autocorrelation computed from OSS is a real signal, this relation to the

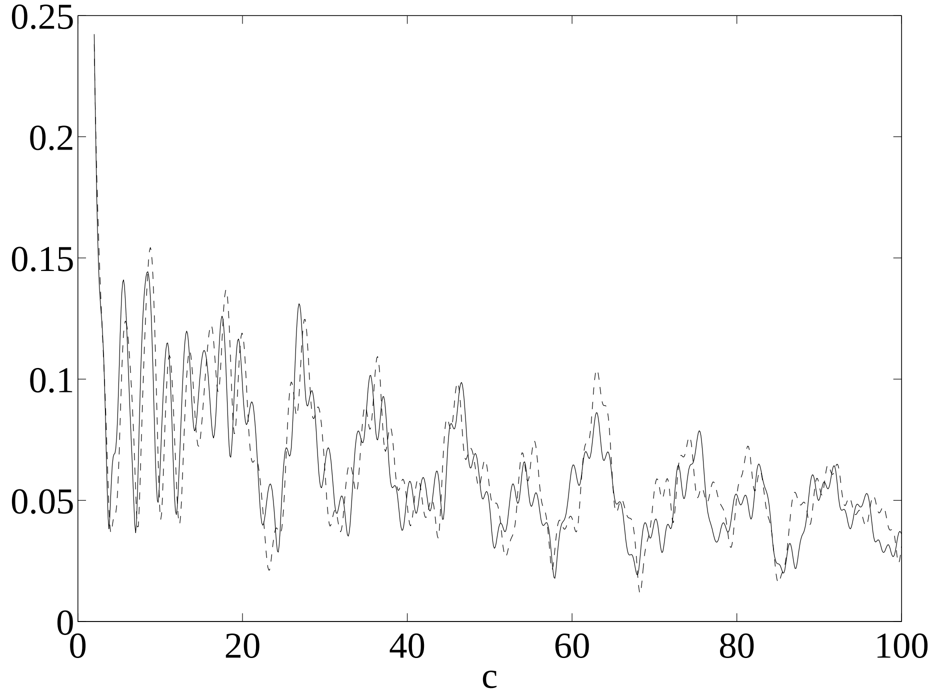


Figure 4.3: Mean scale transform magnitudes of original (bold) and time scaled (dashed) Cretan dance sample, Time scale factor: $a = 1.1$

Fourier transform clarifies that negative scale values need not to be considered since the magnitude spectrum is an even function of frequency. While in Holzapfel and Stylianou [75] the implementation of the scale transform based on (4.3) was used, in this chapter the algorithm for computing the discrete scale transform (DST) as presented by Williams and Zalubas in [163] was applied. DST is derived from (4.2), by approximating the integral in (4.2) as follows:

$$R(c) \approx \frac{\sum_{k=1}^{\infty}[r(kT_s - T_s) - r(kT_s)](kT_s)^{1/2-jc}}{(1/2 - jc)\sqrt{2\pi}} \tag{4.4}$$

where T_s denotes the minimum lag size of $r(t)$, which is equal to the sampling period of the OSS. Compared to the implementation presented in De Sena and Rocchesso [139], the way of computation depicted in (4.4) avoids the interpolation that is necessary to get exponentially spaced samples from signal $r(t)$. The highest scale value C computed in (4.4) will be determined in the experiments shown in Section 4.3. The scale resolution Δc, which defines at which scale values the scale transform in (4.4) is computed, was not found critical. In Combet *et al.* [30], a value of $\Delta c = 1$ was referred to be sufficient for their application. In general, Δc is related to the time domain as:

$$\Delta c = \frac{\pi}{\ln \frac{T_{up}+T_s}{T_s}} \tag{4.5}$$

where T_{up} is the maximum retained lag time of the used autocorrelation [30]. For example, if $T_{up} = 8s$ and $T_s = 0.02s$ then a value of $\Delta c = 0.52$ is obtained, which means that the n-th scale coefficient is computed for $c = n\Delta c$. In this chapter (4.5) will be applied for the computation of Δc. The scale resolution was found to be of minor importance for the accuracy of the system.

4.1.2 Computation from Audio Signals

The datasets used for the evaluation of the rhythmic similarity measure in this chapter contain complex mixture signals of various musical instruments. For that reason, meaningful OSS can be computed using either spectral flux or phase slope based methods, SF_OSS and PS_OSS, respectively. The third OSS that was introduced in Chapter 3 is based on a fundamental frequency estimation derived from monophonic inputs and can not be applied to complex mixtures as shown in the experiments in Chapter 3. For that reason, SF_OSS and PS_OSS as described in Section 3.2 are computed at a sampling frequency of $50Hz$. The difference regarding their performance for a rhythmic similarity task will be evaluated in Section 4.3. As the next step after the computation of the OSS, the sample autocorrelation r_a is computed from the OSS, $o(t)$, as

$$r_a(t,k) = \sum_{n=0}^{T_{win}-t-1} o(n+t+kh_{rth})o(n+kh_{rth}) \tag{4.6}$$

where T_{win} denotes the length of the rectangular analysis window in seconds, k denotes the index of the analysis frame, and h_{rth} the analysis hop size, which was set to $0.5s$. The maximum lag T_{up} of the autocorrelation was set equal to T_{win}. For each analysis frame k the sample autocorrelation is transformed into scale domain by applying the DST as denoted in (4.4), and only the magnitude values for scales $c < C$ are kept. This way, slight tempo changes within the piece are compensated, as they cause a scaling between autocorrelations computed in different analysis windows, which does not affect the scale transform magnitudes. To get a single description vector for a song i, the mean of the scale transform magnitudes is computed, which will be denoted by S_i^C. In Figure 4.3, the mean scale transform magnitudes (STM) computed using the described method are depicted.

4.1.3 Computation from MIDI data

For MIDI data, there are mainly two differences in computing the STM:
First, the onset times and the note durations are exactly known as they can be read from a MIDI file. For that reason, tools from the miditoolbox [44] could be used to derive the sample autocorrelations. As will be described in the following, two types of ACF computation will be considered. The first is described in Section 4.1.3 and uses the tempo information included in the MIDI file to get rid of the scaling in the ACF that is caused by varying tempi. The second way of computing ACF is described in Section 4.1.3 and ignores this information and thus shows the scale changes caused by tempo differences.
The second difference compared to audio is that the windowed computation of the autocorrelation as defined in (4.6) has been found to cause problems. This is related to two facts: OSS derived from MIDI data are much more sparse than OSS derived from waveform data, as the onsets are discrete impulses of varying height. Furthermore, the tempo of pieces in MIDI format remains absolutely constant. No noise is induced by the way humans play musical instruments, which can cause the peaks in OSS to deviate from the position determined

by the meter. Because of that, one sample autocorrelation is obtained using the whole onset strength signal as input. The autocorrelation is then transformed into scale space by using (4.4), resulting in the STM descriptor for a MIDI signal.

Tempo-invariant ACF

In order to describe and compare the content of the samples in D3, an autocorrelation based method as presented in Toiviainen and Eerola [154] has been applied. The onset times are read from the MIDI files and each onset is assigned a weight. In Toiviainen and Eerola [154], different methods to set the weights were evaluated, and in this thesis the three most successfull weighting schemes have been applied: the weight of an onset can either be related to the note duration [123], to characteristics of the melody [150], or all onsets are assigned the same weight. The best weighting scheme will be determined in Section 4.3. In the method presented in Toiviainen and Eerola [154], an onset strength signal (OSS) is generated at a sampling frequency related to the eighth note of the piece. This OSS has an impulse of height according to the assigned weight at the positions related to the onset time. From an OSS $o(n)$ an ACF $r_c(m)$ can be derived

$$r_c(m) = \frac{\sum_n o(n)o(n-m)}{\sum_n o(n)^2} \tag{4.7}$$

Note that the autocorrelations are not affected by tempo differences, when the OSS are computed at a sampling frequency that changes with the tempo (eighth note). Because of this, changing the tempo will result in constant ACF, which will be denoted as r_c. These representations will not be transformed into the scale domain, and they can be compared immediately using a point to point distance measure. In this thesis, for comparing two r_c the cosine distance measure will be applied which was shown to achieve better results in similar cases [49, 72]. Similarity measures obtained this way represent a valuable tool to estimate the influence of scale transform in the computation described in the next Section.

Tempo-variant ACF

As mentioned in Toiviainen and Eerola [154], beat tracking is a necessary step when applying the above described approach to audio. It is necessary to correctly estimate all metric levels in order to determine the eighth note pulse of the piece. When dealing with compound rhythms of different type as they are commonly encountered in the music of Turkey and the whole eastern Mediterranean, no method has been presented yet to perform this task. For that reason, the MIDI data contained in the data set as described in Section 4.2.1 is used to compute OSS using a constant sampling frequency of $f_s = 50Hz$. From the OSS autocorrelations are derived. For two pieces having the same time signature but different tempi, their autocorrelations will differ by an unknown scaling factor, as can be seen in Figure 4.4 for an autocorrelation sequence derived from a MIDI file. This is particularly critical for the type of music encountered in this thesis due to the large tempo deviations (see Section 4.2.1 for details). In order to overcome this scaling problem, typically the beat tracking would be necessary in order to estimate the tempo difference between the pieces. However, in this thesis the usage of the scale transform is proposed to avoid the intractable problem of beat tracking in the presence of complex and compound time signatures. In Figure 4.5, the two scaled autocorrelations from Figure 4.4 have

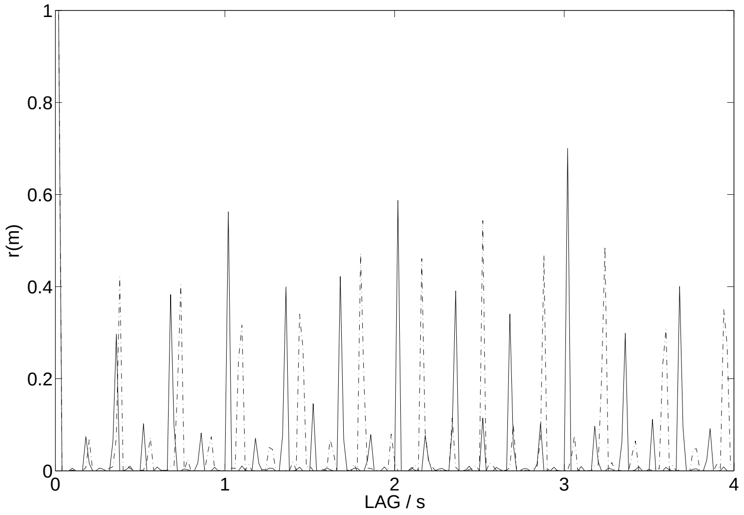


Figure 4.4: Autocorrelations r_u derived from two MIDI samples belonging to the same class of rhythm

been transformed to scale space. Due to the scale invariance property they are aligned and can be directly compared, like for the examples derived from audio as described in Section 4.1.1.

Thus, in this thesis OSS will be computed from the MIDI files using a constant sampling

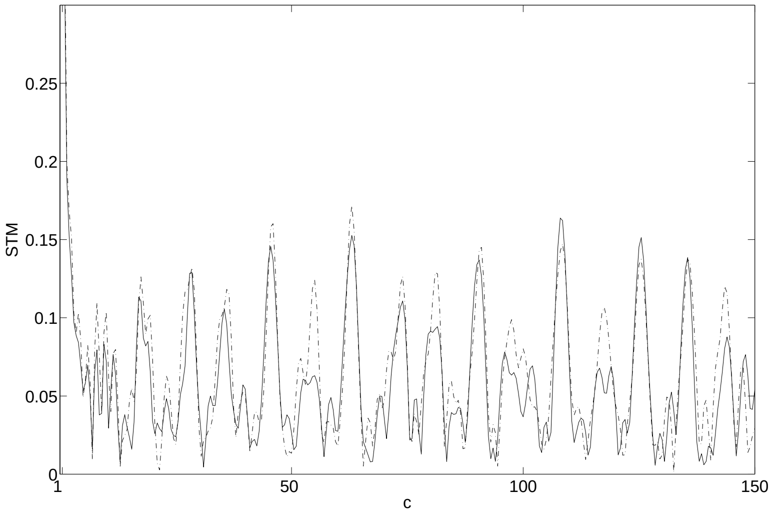


Figure 4.5: Two STM derived from the two *aksak* examples shown in Figure 4.4

frequency of $f_s = 50Hz$. Then, scale transform magnitudes (STM) are computed from the autocorrelations r_u using the discrete scale transform algorithm proposed in Williams and Zalubas [163]. This results in a STM vector that describes the rhythmic content of the signal. The accuracies on the MIDI dataset when using either scaling free autocorrelations r_c or the STM derived from r_u will be compared. The results will indicate if by using a scale transform, the unsolved problem of meter estimation in complex time signatures can be avoided and the similarity between pieces could be determined by using this method.

4.1.4 Some Properties of STM

In order to enable better understanding of the features in the scale domain, some more details about the scale transform will be provided in this Section. Two autocorrelation sequences of OSS computed from audio (a) and MIDI data (b) are depicted in Figure 4.6. Note that both autocorrelations show a periodicity that is related to the tatum, *i.e.*, the smallest metrical unit in the piece [92]. Especially the autocorrelation sequence computed from MIDI data shows a similarity with a pulse train of the tatum period. Considering a pulse train $\sum_{n=1}^{\infty} \delta(t - nt_0)$ with period $t_0 > 0$, the scale transform pair of this pulse train is given by [128]:

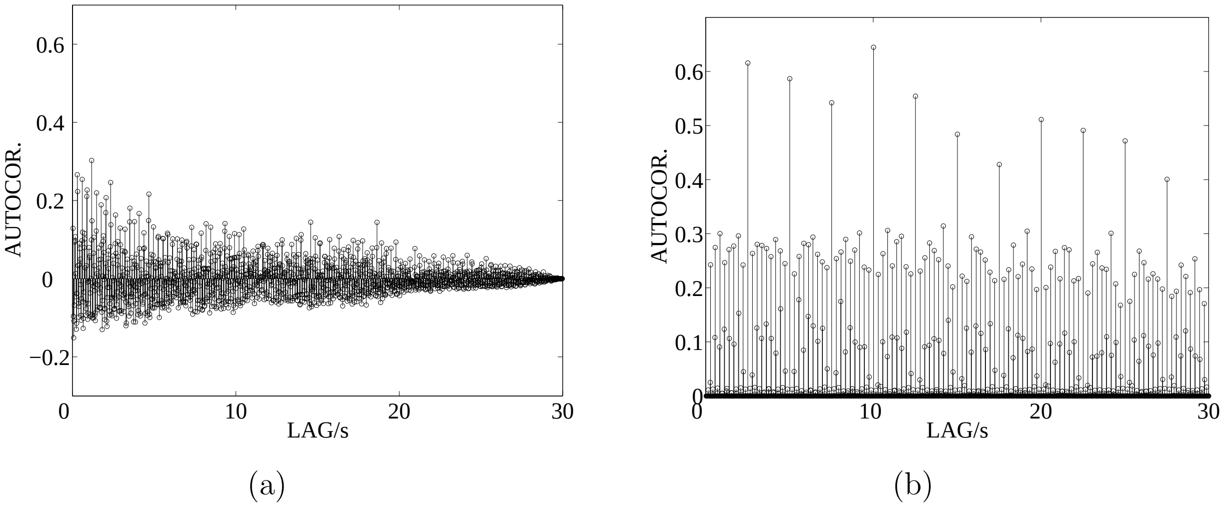
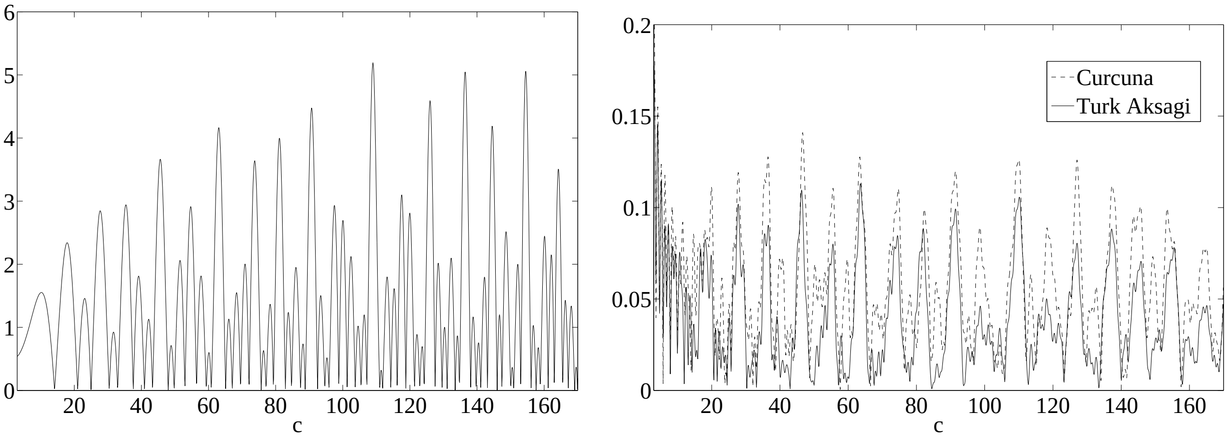


Figure 4.6: Two examples of autocorrelation vectors for waveform (panel (a)) and MIDI data (panel (b))

$$\sum_{n=1}^{\infty} \delta(t - nt_0) \Longleftrightarrow t_0^{-jc-0.5} \zeta(jc + 0.5) \tag{4.8}$$

where $\zeta(s)$ denotes the Riemann Zeta function [131]. In panel (a) of Figure 4.7, the magnitude of the Riemann Zeta function $\zeta(jc + 0.5)$ is depicted. In panel (b) of Figure 4.7, two STM derived from autocorrelations of samples from two traditional Turkish songs represented in MIDI format are shown. It is apparent that these STM have similarities with the envelope of the Riemann Zeta function. Note that for the STM computed on the autocorrelation sequences obtained from audio waveforms (see an example in Figure 4.3) depicted in Figure 4.3, this similarity is not so distinct. This is because, as it was shown in Figure 4.6, the autocorrelation sequences derived from waveform data are less spiky than the corresponding sequences computed from MIDI data. Note that the magnitude of the transform in (4.8) does not depend on period t_0, and thus leads to a similar shape of the STM envelope for pieces with different tempi. In practice, one more problem we have to face is the energy compensation between scaled signals. In theory, because of the energy normalization factor $\sqrt{a}$ the scale transform magnitude remains the same for scaled signals. However, in our case, the autocorrelation functions cannot easily be normalized since they are derived from different signals, with unknown scale relation. This infeasibility of correct normalization in the time domain would lead to a constant factor change in scale magnitude. For that reason a Euclidean distance measure between STM is not applicable. As the appearance of t_0 in the scale transform of a pulse train constitutes a constant factor in magnitude, instead of measuring Euclidean distance we suggest to measure the angle between two STM.

It is worth to clarify the effect of choosing some range of scale coefficients $c < C$ at this point.



(a)

(b)

Figure 4.7: Comparison of the Riemann Zeta function in panel (a) and two STM computed from two autocorrelations of MIDI samples in panel (b)

As mentioned above, autocorrelation sequences derived from musical signals are typically characterized by the period defined by the tatum of the piece. In Figure 4.8, three pulse trains, as a simplified model for such type of autocorrelation sequence, are reconstructed using the complex scale coefficients smaller than $C = \{50, 100, 200\}$. The pulse train has a length of $5s$ and a period length of $100ms$, and it was sampled at a sampling period of $T_s = 20ms$. It can be seen that by using more scale coefficients for the reconstruction, the approximation of samples at large time values gets improved. This is caused by the type of the base function applied in the scale transform as denoted in (4.2): functions $e^{(-jc-1/2)\ln t}$ are chirp functions for which the period is increasing as time increases.. This increment is realized faster for small scale values. Thus, the base functions of c_1 will match the period of the pulse train earlier in time than the base function of c_2, if $c_1 < c_2$. This leads to an interesting interpretation: Fixing the maximum lag T_{up} of the autocorrelation results in a vector of a given length, and increasing the number C in the STM descriptors equals to giving more weight to higher lag values within this vector.

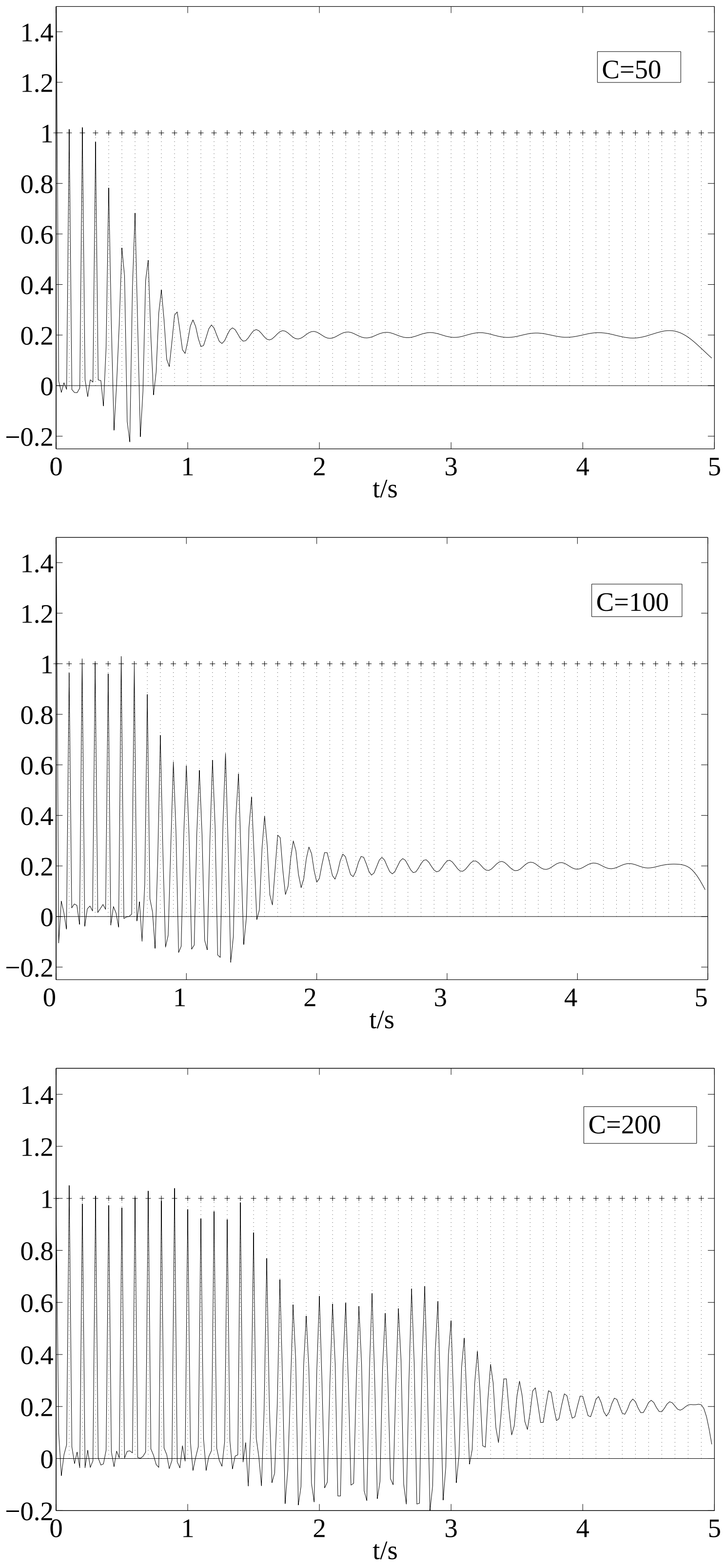


Figure 4.8: Reconstruction of an impulse train by filtering in scale domain

4.2 Experimental Setup

4.2.1 Evaluation Data

For the evaluation of the proposed rhythmic similarity measure, three different datasets are used: The first dataset, which will be referred to as D1, is a set of ballroom dances that was used in the rhythm classification contest in the ISMIR conference 2004 [80]. It has been used for the evaluation of dance music classification for example in [125, 59]. In Peeters [125], it was found that a classification accuracy of 78% can be achieved given the true tempo of the pieces as the only input to the classifier. Because there is a small overlap in the tempo distribution of the classes, this dataset can be considered as simple and it was chosen in order to prove the general validity of the approach presented in this chapter. The second dataset, D2, is a dataset of traditional dances encountered in the island of Crete in Greece, and the third dataset, D3, consists of samples of traditional Turkish music. The latter two datasets were compiled by the author of this thesis. The distribution of tempi per dataset is provided in Table 4.1.

Dataset D2 was used previously by the author of the thesis in [75] and contains samples of the following six dances: *Kalamatianos*, *Siganos*, *Maleviziotis*, *Pentozalis*, *Sousta* and *Kritikos Syrtos*. Each class contains thirty instrumental song excerpts of about ten seconds length. As shown in [75], there are large overlaps between their tempo distributions. In the case of tempo-halving and doubling errors in a tempo estimation pre-processing step, these overlaps would become even larger. Thus, a similarity measure that does not rely on tempo information is necessary to achieve a good classification in that dataset. Regarding their rhythmic properties, all traditional dances from the Greek islands share the property of having a $\frac{2}{4}$ time signature ([7], page 32). Only the dance class *Kalamatianos* in D2 has a $\frac{7}{8}$ time signature. For a more detailed description of the data refer to Appendix A.3.

The dataset of Turkish music, D3, consists of six different classes of rhythm, but unlike the other two datasets, the classes are not related to specific dances. The musicological term used for the different types of rhythm in this music is *usul*. Each *usul* specifies a rhythmic pattern that defines the temporal grid for the composition, see Appendix A.4 for more details. The six *usul* in D3 have lengths from 3 up to 10: *Aksak* ($\frac{9}{8}$), *Curcuna* ($\frac{10}{8}$), *Düyek* ($\frac{8}{8}$), *Semai* ($\frac{3}{4}$), *Sofyan* ($\frac{4}{4}$), and *Türkaksagi* ($\frac{5}{8}$). According to Table 4.1, the tempo variances within each class are much bigger than in D1 and D2. This is because samples in D2 are connected to specific dance movements which puts a natural constraint to the range of tempo variations. Most of the samples in D3 are not dance music and as such, their tempo can vary in a much wider range. Thus, features for the description of the rhythmic content have to be robust to these changes. Furthermore, as all *usul* have different lengths, the recognition of the *usul* can be reduced to a recognition of its length. This is closely related to the task of time signature recognition and motivates a comparison with systems for time signature recognition. In order to acquire the samples, the teaching software *Mus2okur* [84] was used, resulting in a collection of 288 songs, distributed among the six *usul* as shown in the last row of Table 4.1. The software gives a list of songs for a chosen *usul*, which are then exported to a MIDI file. Thus, the data in D3 is available in form of symbolic descriptions, which means that their onset times can be read from the description. The MIDI files contain the description of the melody lines, usually played by only one or two instruments in unison, and the rhythmic accompaniment by a percussive instrument. As this content is separated into different voices, the rhythmic accompaniment can be excluded. This enables to focus on the relation between the melody of the composition and the underlying *usul*. To the best of our knowledge, such a study on *usul* has not been conducted before.

Table 4.1: Statistics of the tempo distributions

	D1							
CLASS	CHA	JIV	QUI	RUM	SAM	TAN	VW	WAL
MEAN	122	166	201	100	102	127	178	86
STD	5.6	14.5	11.5	11.2	18.0	4.0	2.2	4.4
N_{Songs}	111	60	82	98	86	86	65	110

	D2					
CLASS	KAL	SIG	MAL	PENT	SOUS	SYRT
MEAN	128	98	147	145	123	68
STD	8.7	4.5	8.8	10.8	8.7	5.9
N_{Songs}	30	30	30	30	30	30

	D3					
CLASS	AKS	CURC	DUY	SEM	SOF	TURK
MEAN	87	96	76	133	83	67
STD	27.7	13.6	18.8	26.0	14.9	18.8
N_{Songs}	64	57	47	22	60	38

4.2.2 Similarity Measures

Because of the scale invariance property of STM, a simple point wise distance can be applied to get a (dis)similarity measure between two STM. As shown by Foote *et al.* [49] and Holzapfel and Stylianou [73], the cosine distance outperforms the Euclidean distance. Furthermore, as described in the previous Section, measuring the angle between two STMs is to be preferred from using Euclidean distance due to the unknown normalization factor. Because of that, the rhythmic dissimilarity between songs i and j can be measured by computing the cosine distance between their mean STMs S_i^C and S_j^C

$$d_{sc}(i,j) = 1 - \frac{S_i^C \cdot S_j^C}{|S_i^C||S_j^C|} \tag{4.9}$$

In order to confirm the superiority of the cosine distance compared to the Euclidean distance, also the Euclidean distance between two mean STM, $d_{eucl}(i,j)$ will be used. For reasons of comparison, some previously proposed measures of rhythmic similarity will be used as well. As shown in [49, 73], the cosine distance denoted in (4.9) is a good measure for rhythmic similarity directly applied to periodicity spectra if the tempi do not differ widely between the pieces that are compared. Because of that, such measures can be expected to perform well on D1 with its small tempo variations while it should decrease in performance on the other datasets. The cosine measure will be denoted as $d_{cos}(P)$ when directly applied to periodicity spectra, and $d_{cos}(R)$ when directly applied to the autocorrelation sequences derived from OSS.
Recently, a dissimilarity measure based on a warping strategy was introduced [73]: periodicity

spectra as shown in Figure 4.2 are computed from OSS, and then the periodicity spectrum of one song is warped in order to be aligned with the periodicity spectrum of another song, a process referred to as Dynamic Periodicity Warping (DPW). The linearity of the warping path derived in DPW serves as a measure of rhythmic similarity: the more linear the warping path, the more similar the two pieces are considered. This dissimilarity measure will be denoted as d_{DPW}.

For D3, the note durations in the *usul* sequences can be described as a string, as for example the *Aksak* pattern shown in Figure A.4 can be described as the string `xoxxxoxox`, where `x` symbolizes the start of a note and `o` metric unit without note [155]. Note that this representation is a further simplification of the one shown in Figure A.4, because no differentiation of the intonation strength is contained. However these representations can be used for estimating the similarity between rhythms of same lengths by computing a chronotonic distance [155]. It will be evaluated in the experiments if such a distance measure between the theoretical patterns shows some correlation with the similarities estimated using the proposed measure between the MIDI samples of D3. It will be interesting to observe if samples from *usuls* with patterns that are found to be theoretically similar are confused in the classification experiments.

4.2.3 Evaluation Procedure

For a given dataset, all pairwise dissimilarities between songs are computed using the measures described in Section 4.2.2. This results in dissimilarity matrices, having values close to zero whenever two pieces are found to be similar. In order to determine the accuracy of the proposed rhythmic similarity measure, the accuracies of a k-Nearest Neighbor (kNN) classification will be determined. For this, each single song will be used as a query for which a classification into one of the available classes is desired, *i.e.*, a leave-one-out cross validation is performed using the computed dissimilarity matrix as an input. The value k_{knn} that determines the number of neighbors is varied in the interval [2...30], and the best accuracy achieved by varying k_{knn} is then reported. In order to determine if these accuracies are over-optimistic, the kNN accuracies will be compared with results achieved using a Fisher LDA classifier and an SVM classifier with a linear kernel. For SVM, the implementation included in the WEKA software [164] has been used without any parameter changes. Both LDA and SVM classifiers are evaluated using leave-one-out cross-validations, using the STM of the songs as input features.

In Section 4.3, the accuracy of the proposed STM features for the discrimination of different rhythms will be discovered. Therefore, it is necessary to evaluate the optimum set of scale coefficients for each dataset. In the first experiments, the accuracy depending on the choice of the highest included scale coefficient will be determined. In Section 4.3.4 it is evaluated if a maximum relevance feature selection as proposed in Markaki and Stylianou [108] can provide us with a consistent way to derive a compact set of features that is optimal for the classification task. For this, the relevance to the target class c of each feature x_i in a training set is computed by determining their mutual information:

$$I(x_i, \mathsf{c}) = \int \int p(x_i, \mathsf{c}) \log \left(\frac{p(x_i, \mathsf{c})}{p(x_i)p(\mathsf{c})} \right) dx_i d\mathsf{c} \tag{4.10}$$

In practice, the integration in (4.10) is problematic for continuous valued features as the scale coefficients in our case. For that reason, each feature has been discretized by using an adaptive quantization as proposed by Markaki and Stylianou [108], using $b = 5$ bins. In order to select a set of relevant features all mutual information values between the single scale coefficients and the target class have been computed. Then, a threshold has been applied to the computed mutual information, which for a value of 100% chooses all features and for a value of 0% only the one feature with the maximum relevance for the training set. Changing this threshold continuously from 0% to 100% leads to choosing a subset of features regarding their individual relevance for the classification. The influence of varying this threshold will be determined in Section 4.3.

For the data in D3, an *usul* can be expressed in a simplified way as a string, as explained in Section 4.2.1. In Section 4.3, for some *usul* their string representations will be used to estimate their similarity using a method proposed in Toussaint [155]: From the string representations chronotonic chains can be computed, by breaking down the rhythm into its smallest time unit on the *x-axis* and assigning to each element a height on the *y-axis* according to the beat-to-beat interval. This results in the chronotonic chain $[2, 1, 1, 2, 2, 2, 1]$ in case of *Aksak*. As proposed in Toussaint [155], in order to compare two such chronotonic chains, then a discrete form of the Kolmogorov Variational Distance (DKVD) can be applied. Given two chronotonic chains g and f of same length L, this distance can be computed as

$$K = \sum_{i=1}^{L} |f[i] - g[i]| \tag{4.11}$$

and is equal to the $1{-}norm$ distance between the chains. Thus, by depicting an *usul* pair as two strings of same length, their rhythmic similarity can be estimated. This method will be applied to pairs of *usul* for that samples frequently were confused in the classification experiments.

4.3 Experiments

For the proposed similarity measure d_{sc} there are mainly two critical parameters: the length of the maximum lag T_{up} considered in the autocorrelation and the numbers of coefficients C of STM in (4.9). The influence of these parameters will be explored by computing the accuracies in a grid search of these two parameters. For each dataset the optimum number for the maximum lag will be determined, and the effect of varying the number of scale coefficients will be explored. For all experiments on D3, the OSS have been computed using durational accents and the STM have been derived from tempo-variant ACF, as described in Section 4.1.3. The influence of using other accents and using the tempo-invariant ACF as features will be explored separately in Section 4.3.3.

4.3.1 Optimum upper scale and maximum lag

On both waveform datasets D1 and D2, the optimum maximum lag T_{up} found in the grid search was $8s$. The accuracies for D3 improved until a maximum lag of $14s$ is reached. It was observed that further increase does not lead to a decrease in accuracy on this dataset, as it is the case on the waveform data in D1 and D2. In Figure 4.9, the accuracies of kNN classifiers are depicted when changing the number of scale coefficients C. The optimum maximum lag was used for each dataset, for D1 and D2 the SF_OSS has been used as an input to the autocorrelation computation. It can be seen that the accuracy of the classification depends on the number of chosen scale parameters in a different way for each dataset. The highest classification accuracy was achieved for D1. More specifically, the classification accuracy increases up to 88.1% at c=170. In general, an area of almost constant accuracy is reached for $C > 80$, as can be seen from Figure 4.9. A similar behavior can be observed for D3, where the best accuracy using kNN is achieved at $C = 140$ (78.1%). On D2, a maximum is reached at $c = 30$ with an accuracy of 76.1%. Unlike for D1 and D3, when further increasing C on D2 the accuracy decreases. Similar results are obtained using the SVM classifier: while on D1 and D3 a saturation is reached just like for kNN, for D2 this does not hold. The LDA classification could not be evaluated for very large values of C, as the increasing dimensionality causes numerical problems. It is worth to note that these accuracy values are close to the accuracies achieved by human listeners on the same data (75.6%), as will be detailed in Section 4.3.5. In Table 4.2, the best accuracies for all three classifiers using the proposed features are depicted along with the value of C at which this accuracy is reached. It seems that for higher scale values on D2 the STM contain more noise than for the other two datasets. As shown in Section 4.1.4, higher scale values lead to a more accurate reconstruction at larger autocorrelation lags. Thus, regarding Figure 4.8, for D2 a stronger weighting for lags smaller than one second is optimal, while for D1 this weighting is extended to about two seconds. This behavior will be further explored in Section 4.3.4.

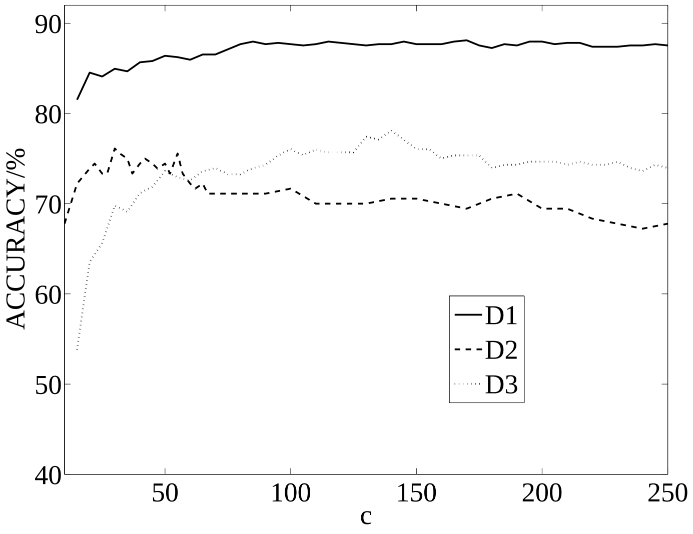


Figure 4.9: Accuracies on the three datasets for varying number of scale parameters, using SF_OSS

When using the PS_OSS instead of the SF_OSS as an input to the STM computation from D1 and D2, a similar behaviour regarding the number of scale parameters can be observed. The related accuracies are depicted in Figure 4.10 and are characterized by a similar shape as the curves shown in Figure 4.9 for these datasets. However, it must be noted that the best accuracies achieved on D2 are lower when using PS_OSS. The best accuracy of 71.1% is reached at $C = 30$ (compared to 76.1% for SF_OSS). On the other hand, a decrease is not observed for D2. On D2, the highest accuracy is slightly better when using PS_OSS (89.7% at $C = 120$) than when using SF_OSS (88.1% at $C = 170$). It is important to note that both differences

Table 4.2: Classification Accuracies at C using STM features

	kNN	SVM	LDA
D1	88.1($C = 170$)	91.7($C = 160$)	89.5($C = 120$)
D2	76.1($C = 30$)	76.1($C = 35$)	77.8($C = 25$)
D3	78.1($C = 140$)	82.3($C = 140$)	77.1($C = 40$)

are statistically not significant. The confidence intervals are 2.4% for D1 and 6.2% for D2, both at a level of confidence = 95%. Thus, based on the currently available datasets no clear conclusion about the preferable OSS can be drawn. However, for reasons of computational simplicity, the SF_OSS might be preferred.

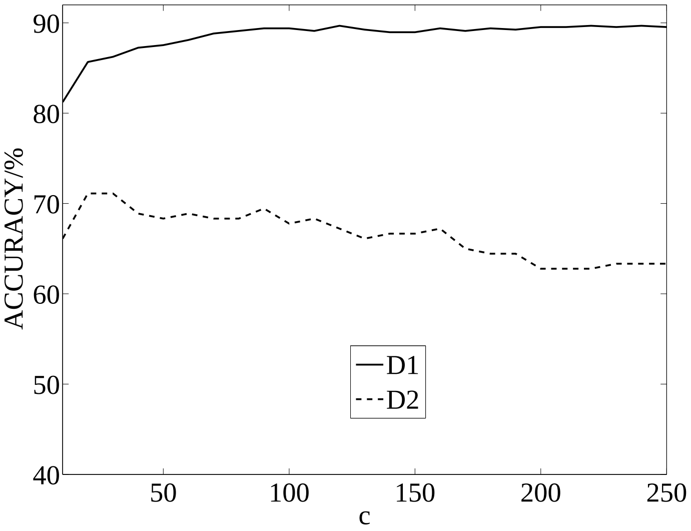


Figure 4.10: Accuracies on D1 and D2 for varying number of scale parameters, using PS_OSS

4.3.2 Comparison of distance measures

Table 4.3 shows the classification accuracies on the datasets, using the measures as described in Section 4.2.2 and kNN classification. Similar to the results presented by the author in [73], the direct cosine measures between the periodicity spectra, $d_{cos}(P)$, and between the autocorrelation sequences, $d_{cos}(R)$, work well on D1. The proposed scale method, d_{sc} achieves a slightly improved accuracy of 88.1%. However, this improvement is not significant regarding the confidence interval, which is 2.4% (level of confidence = 95%). Comparing these results to the highest accuracy, without the usage of the tempo annotations, of 85.7% [42] on the same dataset D1, the accuracy presented here using d_{sc} appears to be a satisfying proof of concept. The improvements in comparison to our previous results in [73] and [75] must be assigned to the changed sample rate of the onset strength signal which in general improved results throughout the experiments, and to the different computation of the scale transform.

For D2, Table 4.3 shows a considerable advantage for the proposed scale distance measure d_{sc}, which achieves an accuracy of 76.1% with a confidence interval of 6.2%: on this dataset it outperforms the cosine measures $d_{cos}(P)/d_{cos}(R)$ by 21.8/31.4 percentage points. This clear improvement can be assigned to the robustness to tempo changes of the scale transform.

Table 4.3: kNN-Classification Accuracies for various distance measures

	$d_{cos}(P)$	$d_{cos}(R)$	d_{DPW}	d_{eucl}	d_{sc}
D1	86.1	86.0	83.5	86.1	**88.1**
D2	54.3	44.7	60.9	73.9	**76.1**
$D3_{mel}$	53.1	56.2	50.5	75.7	**78.1**
$D3_{all}$	63.5	66.7	71.0	83.7	**86.0**

The accuracies for the dataset of Turkish MIDI files are listed in the third and fourth row of Table 4.3. The third row gives the accuracies when using the melody lines only for the onset computation as described in Section 4.1.3. Using the dissimilarity measure d_{sc} proposed in this chapter leads to the best results: an optimum accuracy of 78.1% is reached at $C = 140$, with a confidence interval of 4.8%. Direct comparisons of either periodicity spectra or autocorrelation sequences are clearly inferior due to the large changes in tempo for each *usul*. The DPW approach we presented in [73] does not lead to good results on D3. This must be assigned to the large standard deviation of the tempi in one class since DPW assumes that there are no differences in tempo larger than 20% between two songs. When tempo differences exceed this threshold, the whole procedure is becoming unreliable [73].

The fourth row of Table 4.3 (i.e., for D_{3all}) shows the accuracies that can be achieved when the tracks containing percussive accompaniment are also included in the computation of OSS. The accuracies are then in general improved, since the percussive accompaniment is typically the same for one specific *usul*. The relatively high values in the third row, $D3_{mel}$, clarify the information about the *usul* that is contained solely in the melody line of the composition. As the difference between the best accuracy in the third row and the best in the fourth row is only 7.9 percentage points, it can be concluded that this relation between the melody and the *usul* is very strong.

Comparing the measures based on the scale transform (i.e., d_{eucl} using Euclidean distance and d_{sc} using cosine distance) we see that d_{sc} indeed outperforms d_{euc}. This was expected, because of the normalization factor in (4.1) (i.e., a) is unknown, and this affects the magnitude of vectors being compared, but not their angle. Compared to d_{DPW}, the distance derived using Dynamic Periodicity Warping [73], the advantage of d_{sc} regards accuracy as well as computational: while in DPW there is the need to compute a warping path using dynamic programming, the most time consuming operation in the scale distance measure is the scale transform which is performed using a matrix multiplication.

4.3.3 Further exploring MIDI

Comparison with Scale-free ACF

Three different weighting schemes for the OSS computation from MIDI data in D3 have been evaluated in the experiments: the duration accent [123], the melodic accent [150], and the flat accent (i.e., using the same accent weight for all onsets). Using the r_c autocorrelations computed using these three accents distance matrices have been computed. Applying a kNN classifier in the same way as described in Section 4.2.3 resulted in the best accuracies for the duration accent, as documented in Table 4.4. This contradicts with the findings in Toiviainen and Eerola [154], where the melodic and flat accents were found to be preferable. Furthermore, using a selected range of autocorrelation coefficients could not further improve results on this data set, while in Toiviainen and Eerola [154] using the coefficients of longer lags and leaving out the coefficients of short lags was found superior. This must be assigned to the differences between the data sets.

DURATION	MELODY	FLAT
80.2%	68.1%	72.9%

Table 4.4: Time signature recognition accuracies when using scale free r_c representation

In Table 4.5 the confusion matrix for the best classification in Table 4.4 is shown. The biggest confusion happens between the $\frac{8}{8}$ time signature *usul* and the $\frac{4}{4}$ *usul* (*Düyek* and *Sofyan*, respectively). The pieces in the $\frac{8}{8}$-*usul* could be equivalently annotated in a $\frac{8}{4}$ time signature by changing their degree, referred to as *mertebe*, to four. The second biggest confusion happens between *Curcuna* and *Türk Aksaği*. The time signatures are related by a factor of two as well ($\frac{10}{8}$ and $\frac{5}{8}$). These types of errors have been denoted as typical as well in Toiviainen and Eerola [154]. Still, the confusion between between *Düyek* and *Sofyan* is larger. This can be attributed to the different degree of similarity of the *usul*, which can be estimated using the approach proposed in Toussaint [155]: In Table 4.6, the symbolic descriptions for the two confused *usul*-pairs are depicted as vectors of same length. From these descriptions the chronotonic chains have been derived that are depicted in Table 4.6. Note that *Sofyan* would be typically denoted as $[2, 1, 1]$ as its smallest beat-to-beat interval is a fourth note. In order to get chains of equal length, the eighth note has been chosen as smallest unit. Computing

		Predicted					
		9/8	10/8	8/8	3/4	4/4	5/8
Notated	9/8	62	0	1	0	1	0
	10/8	0	50	0	0	1	6
	8/8	1	4	24	0	18	0
	3/4	0	0	0	20	2	0
	4/4	2	0	12	0	46	0
	5/8	0	9	0	0	0	29

Table 4.5: Confusion matrix for r_c using duration accent

Symbolic Description			
Düyek:	xxoxxoxo	*Curcuna*:	xoxxoxoxox
Sofyan:	xoooxoxo	*Türk Aksaği*:	xoooxoooxo
Chronotonic Chains			
Düyek:	12212222	*Curcuna*:	2212222221
Sofyan:	44442222	*Türk Aksaği*:	4444444422
Normalized DKVD betw. Chronotonic Chains			
10/8=1.25		18/10=1.8	

Table 4.6: Computing chronotonic distances between confused *usul*

		Predicted					
		9/8	10/8	8/8	3/4	4/4	5/8
Notated	9/8	51	3	3	1	3	3
	10/8	0	52	2	0	0	3
	8/8	1	1	30	2	11	2
	3/4	3	0	3	15	1	0
	4/4	0	2	8	1	48	1
	5/8	2	4	3	0	1	28

Table 4.7: Confusion matrix for STM at $C = 140$ and maximum lag of $14s$

the Kolmogorov Variational Distances between the chronotonic chains, and normalizing by the length of the vectors it can be seen that the *usul Düyek* and *Sofyan* are more similar than the other pair. This is reflected in the higher confusion in Table 4.5. Thus, it can be concluded that the applied autocorrelation method is not only suitable for determining time signatures, but can as well capture rhythmic similarities contained in the piece.

In Table 4.7, the confusion matrix obtained from the kNN experiment depicted in Table 4.2 (accuracy of 78.1%, $C = 140$) is shown. Comparing it with the confusion matrix shown in Table 4.5 reveals very similar structure. The decrease in accuracy from 80.2% seems to be caused by some misclassification that cannot be justified by a similarity of the *usul*, as for example the $\frac{9}{8}$-time signature, which for the STM descriptor is randomly misclassified. Thus it appears that transforming autocorrelations to scale domain in the proposed way introduces some noise to the rhythm descriptors. However, the performance is only 2.1% lower than for using the scale-free autocorrelations (78.1% instead of 80.2%). This clarifies that using the tempo-variant ACF in combination with the scale transform as described in Section 4.1.3 does not lead to a significant decrease in accuracy compared to the tempo-invariant ACF. Hence, by including scale transform the currently infeasible step of beat tracking in this kind of meters is avoided and time signature estimation is made feasible even on audio signals, when presented with arbitrary types of music signals having a compound or complex meter.

Manipulating MIDI

Two more experiments have been conducted to evaluate the robustness of the proposed method. For these experiments the SVM classification that resulted in the best accuracy of 82.3% on the MIDI data has been used, which means that all scale coefficients until $c = 140$ have been used in the STM (see Table 4.2). Again, only the melody lines have been included in the OSS computations, while the percussive instruments have been left out.

The first experiments explores the influence of tempo deviations within the classes. Since for the MIDI files the tempo information is given, experiments could be conducted with the tempo of the pieces changed in a deterministic way. For this, from the data in D3 the global tempo mean value has been computed. Then, all pieces have been assigned this tempo mean plus a uniformly distributed noise. This noise has been varied in steps of 5% from 0% up to 85%. For 0% noise all pieces share the same tempo, and no scaling effects the autocorrelations. At 85% noise level noise level the global mean of about 87 bpm results in a possible tempo range from 13 to 161 bpm. In order to compensate for the noise introduced by the randomly changed tempo for each noise level the experiment has been rerun ten times, and the mean accuracies of the ten runs are reported. Computing the mean SVM-accuracy for the noise free case leads to an accuracy of 82.9%. The small difference to the accuracy of 82.3% (as shown in Table 4.2) in presence of the original tempo variance of the data proves the robustness of the proposed method to this variance. Increasing the noise level leads to an almost linear decrease in classification accuracy. However, at the largest tempo noise level of 85% the accuracy is still 73.2%. This confirms that the theoretical properties of the scale transform make the features robust to large tempo changes in practice as well.

The second experiment explores the way accuracy might get affected when dealing with real audio signals of Turkish music instead of the MIDI signals as contained in D3. For that purpose, the functionality of the *MIDI toolbox* [45] for the synthesis of an audio file from a MIDI has been used. The synthesis locates Shepard tones [142] of constant intensity wherever an onset is listed in the MIDI file. Thus, computing an OSS from the signals synthesized in this way results in almost constant onset strengths amplitudes at the locations of the note onsets. The accuracy clearly decreased to 63.5% (from 82.3%), again using SVM on STM features at $C = 140$. It was investigated if this decrease is caused by the flat characteristic of the OSS that does not allow the differentiation between strong and weak onsets. For this, the durational accent type used in the OSS computation from the MIDI files was replaced by flat accents. This means that impulses of constant height were positioned at the location of all note onsets. Indeed, removing the information about the intensity of the onset leads to the accuracy of 68.7%, and it can be concluded that the weighting of an onset according to its strength is a crucial information. Thus, it can be expected that the accuracy values obtained from real audio files of this music will be superior to the ones computed from the synthesized files, because the onsets cause varying amplitudes in the computation of OSS.

4.3.4 Mutual information based feature selection

In order to find a way to obtain an optimal set of features for classification independent of the dataset, various criteria based on the coefficient energies or the scale bandwidth [27] have been evaluated without success. We decided then to compute the mutual information, MI, between each scale coefficient and the class label as this was described in Section 4.2.3 in order to select the best features for our task from a given STM based on information theoretic criteria. This was further motivated by the fact that for D1 and D3 classification accuracies improve, when low scale coefficients are left out. Thus, for each dataset different scale coefficients appear to be relevant for classification. It was decided to use the SVM classifier, which achieved the highest accuracies in Table 4.2, and to vary the mutual information threshold as described in Section 4.2.3 on the set of features obtained for $C = 200$ for all datasets. The classification accuracies are depicted in Figure 4.11. It can be seen that from an MI threshold value of about 60% upwards for all three datasets a saturation effect is reached. These saturation levels are about the same as the best classification accuracies depicted in Table 4.2. Thus, it can be concluded that using mutual information criteria a common way to get to an optimal feature set can be defined. From Figure 4.11 it is clear that the number of samples in a dataset affects the way the accuracy changes when increasing the threshold. Increasing the threshold leads to an increasing dimensionality of the feature vector, which leads to problems especially on the smallest dataset, D2.

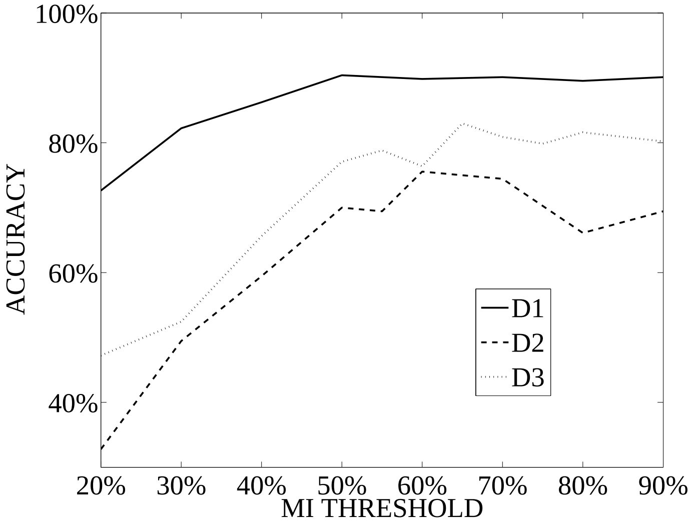


Figure 4.11: SVM classification accuracies on the three datasets for varying mutual information threshold

It is interesting to compare the compression achieved using mutual information thresholds for the three datasets. Table 4.8 shows the number of coefficients corresponding to an MI threshold value of 60%. It can be seen that for D2, a much higher compression is achieved than for D1. It was observed that for D2 scale coefficients for low scales ($c < 50$) are the most relevant, while for D1 the relevant scales were found among the whole scale range. This phenomenon is not related to the size of the datasets, but only to the different musical characteristics of the contained data. We recall from Figure 4.8 that the scale coefficients until $c = 50$ allow for a reconstruction of the autocorrelation for lags up to one second. This means that small lags are more important for this type of music than the others.

Table 4.8: Compression values for mutual information threshold of 60%

	D1	D2	D3
N_{feat}	249	27	98
Compression	34.7%	92.9%	76.5%

4.3.5 Listening Test

In order to evaluate the relation between the proposed distance measure and the way human subjects perceive rhythmic similarity on the used data, listening tests were conducted. For the first test, subjects were asked to judge the similarity measurements performed on D2 which lead to the optimum classification performance for this dataset in Section 4.3.1 ($C = 35$ for LDA). Each subject was asked to decide which of two comparison samples was rhythmically closer to a reference sample. A total amount of 25 reference samples were randomly chosen from D2 and presented to each subject. One of the comparison samples was the closest to the reference according to the proposed rhythm similarity measurement, while the other was the sample which was positioned in the middle of the ranked list of samples produced by the suggesting method as being similar to the reference sample. The subjects could decide for one of the two samples being closer, or they had the possibility to state that both comparison samples are equally close to the reference. They were informed that all music will be traditional Cretan dances, but not exactly which type of dances. Furthermore, they were asked not to restrict their judgement on the recognition of the class, but to concentrate on judging rhythmical similarity, independently of the class affiliation. They had the possibility to listen to the samples as many times as they like.

This listening test was conducted with three groups of people. The first group consisted of two experienced teachers of traditional Cretan dances. The second group consisted of 9 experienced dancers, all subjects in this group had practical experience in all style of dances present in the dataset (Cretan dances). The third group consisted of 11 university students who all had stayed for some years in the island of Crete and are familiar with the sound of Cretan music, but did not learn the traditional dances. The results for the groups are shown in Table 4.9. The right column depicts the percentage of trials in which the sample perceived more similar to the reference was indeed the sample which was estimated to be more similar by the method proposed in this thesis. The middle column depicts the percentage of the cases in which the subject did not make a decision, and the left column contains the contradictions

Table 4.9: Results of listening test for D2

	CONTRADICTION	NEUTRAL	CONSENSUS
TEACHERS	14%	12%	74%
DANCERS	16%	21%	63%
NON-DANCERS	28%	18%	54%

between proposed measure and human subject. The first row shows the results obtained for the dance teachers, and the second row for their students. It can be seen that especially for the teachers a high correlation exists between the proposed measure and the listening test results. In 74% the teachers agree with the measure, the students do so in 64% of the cases. The number of disagreements stays low for both teachers and students (14% and 16%, respectively). These results prove that apart from the objective verification of the proposed method in the classification task, the method is characterized by a high correlation of the way subjects perceive rhythmic similarity. The third row in Table 4.9 shows the results obtained with the group of non-dancers. While for this group the amount of neutral samples stayed almost the same as for the dance students, the proportion of consense decreases by 9 percentage points, while the amount of contradiction with the proposed measure increases by 12%. This difference is even larger compared to the teachers. This result shows that the perception of rhythmic similarity in this kind of music depends strongly from the familiarity of the subject with the music. It has been observed that the teachers came to a decision much faster than their students and the non-dancers. For them it was often not even necessary to listen to the samples in their whole duration, because they recognized them after a few notes. The unexperienced non-dancers needed much more time, listening several times to each sample, and trying to get to their decision. This shows that memorizing the melodic phrases plays indeed a key role in the perception of this music.
One more listening test was conducted in order to evaluate the ability of a listener to correctly categorize the dances. Six subjects were asked to classify each piece in the dataset after listening to it one time. All subjects are dancers familiar with each of the dances. A randomly chosen subset of D2 which contained 90 songs was presented to the listeners. The average correct classification by the listeners per class and overall is depicted in the first row of Table 4.10. It can be seen that some of the classes are particularly difficult and the overall accuracy is far from being perfect. The class-wise accuracies of the kNN classification on D2 as shown in Table 4.2 are depicted in the second row of Table 4.10. It is particularly interesting to observe the correlation between the classification accuracies shown in Table 4.10. Also, the type of misclassifications are the same: the class that is the most difficult to classify (*Pentozalis*) gets confused with the class *Maleviziotis* in most cases, both for the listening tests and for the automatic classification. The same holds for *Maleviziotis*, which is almost always misclassified as *Pentozalis*. However, for the automatic system there exists a confusion between *Kalamatianos* and *Sousta*, which was never observed in the listening tests. Note that the listeners were able to avoid this confusion by differentiating between the $\frac{7}{8}$ rhythm of *Kalamatianos* and the $\frac{2}{4}$ time signatures of all other dances, something in which the automatic approach does not succeed in such a degree. In general, this listening test is one more supporting result for the hypothesis

Table 4.10: Listeners' Mean Classification Accuracies, compared with accuracies of kNN classification

	Kal.	Sig.	Mal.	Pen.	Sous.	Chan.	Mean
Listeners	93.3	88.9	79.2	45.6	58.3	88.5	75.6
kNN	80.0	100.0	73.3	46.7	76.7	80.0	76.1

that the proposed system indeed detects rhythmic similarity.

4.4 Conclusions

A description of the rhythmic content of a piece of music based on the scale transform was proposed. This description is robust to large tempo variations that appear within a specific class and to large tempo overlaps between classes. Using simple distance measure and classifier techniques, the descriptor vectors can be used to classify the samples with high accuracies. The approach is computationally simple and has no need of any tempo or meter estimation which might be desirable for certain kinds of music signals. Based on mutual information criteria, a method was proposed for choosing a feature set that is optimal for the classification task. The relation between autocorrelations sequences and the Riemann Zeta function in scale domain was explored, while a discussion of the signal reconstruction by applying inverse transform enabled to gain valuable insight into the relation between variables in scale and in time domain. The inclusion of the traditional Turkish dataset provided us with a potential starting point for a detailed study of rhythmic characteristics of Turkish traditional music. The suggested measure provides a simple and efficient tool for the description and comparison of rhythm content, especially applicable to music with little or no percussive content and strong tempo variations. Its validity was confirmed in two listening tests as well.

Chapter 5

Beat Tracking

In this chapter, the potential of the phase slope onset strength signal (PS_OSS) for the usage of beat tracking especially in the context of traditional music will be evaluated. The computation of PS_OSS follows exactly the description given in Section 3.2.1. The important changes comparing the content of this chapter with the content presented by the author in [71] are related to the refined computation of the PS_OSS, and to the dataset of beat annotated Cretan dances. For this chapter, the dataset of Cretan music used in [71] has been enlarged from 20 to 41 samples in order to guarantee for the significance of the presented results. Furthermore, the dataset of western popular music has been created by combining the development data and test data from [71] to a unique dataset containing 48 samples. On these datasets, the performance of a beat tracker will be evaluated, that has been implemented according to a method described in Klapuri *et al.* [92]. The differences between the implementation used in this thesis compared to the method in Klapuri *et al.* [92] will be explained, and it will be investigated if the usage of PS_OSS for a beat tracking task can improve in some way the accuracy compared to the original implementation used in [92].

5.1 Method for beat tracking

5.1.1 Onset detection using group delay

The onset detection using group delay follows the concept explained in Section 3.2.1. The optimum value N of the analysis window of the onset detector when applied to a beat tracking task has been evaluated on a development dataset of periodic artificial signals like those depicted in Figure 3.2, with periods from 0.3s to 1s, which is related to the typical range for the tempo of musical pieces (60bpm-200bpm). This dataset, which will be referred to as BD1, is also useful in evaluating the robustness of the suggested approach against additive noise. For this purpose, a Transient to Noise Ratio (TNR) is defined in the same way as the usual Signal to Noise Ratio (SNR):

$$TNR = 10 \log_{10} \frac{\frac{1}{L} \sum_{n=0}^{L-1} x^2[n]}{\sigma_u^2} \tag{5.1}$$

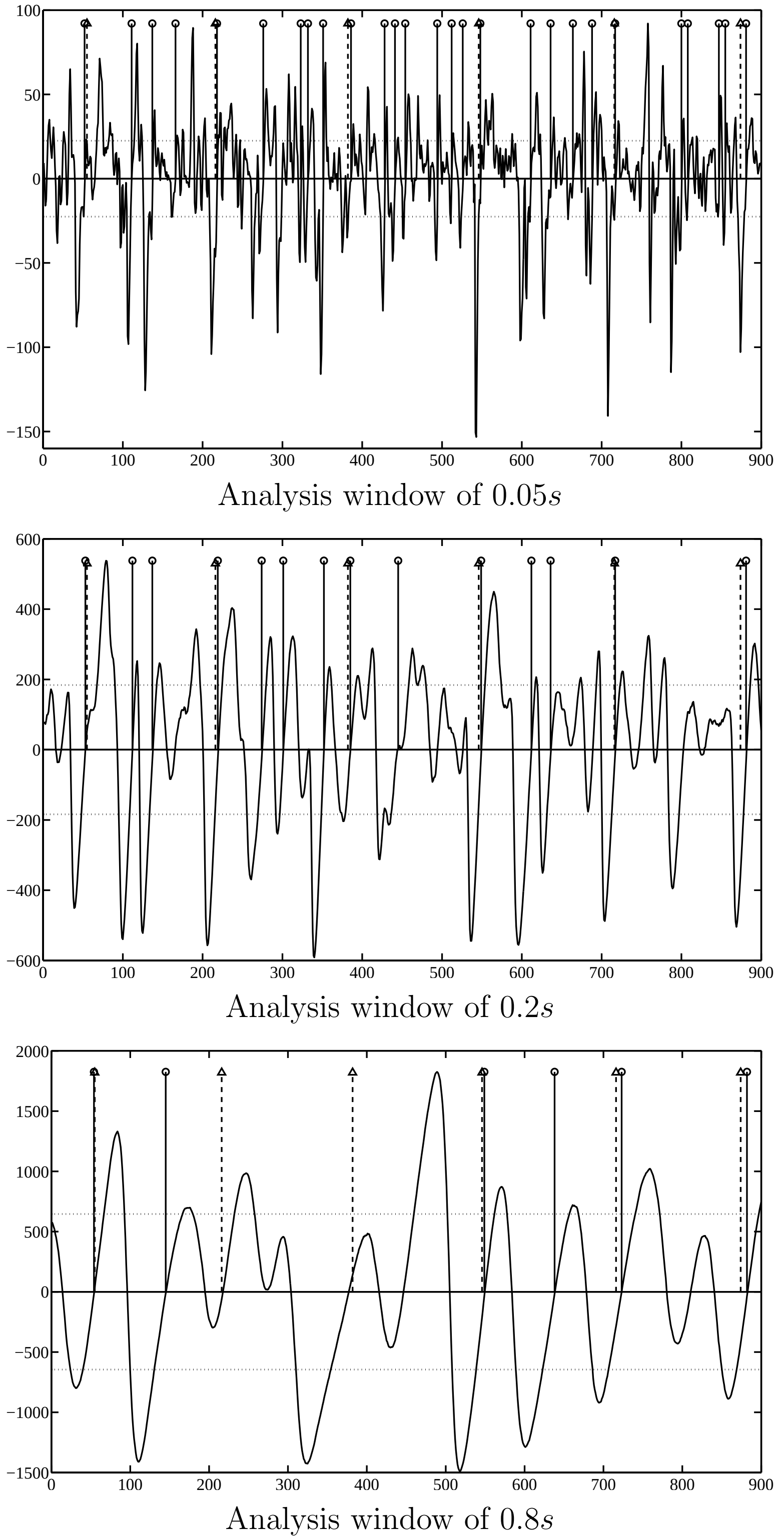
where x denotes a signal of length L and σ_u^2 denotes the variance of the noise. The artificial signals have been mixed with white noise at different TNR. For each artificial signal a corre-

sponding text file has been created containing the onset times of the impulses.
As it is indicated by Kandia and Stylianou [83], a large phase slope analysis window is appropriate for detecting major sound events in an audio signal while shorter windows may be used in case additional sound events are needed to be detected. The optimum length of the analysis window has been determined by trials and errors on the artificial development data using various TNR levels, and will be compared to other window lengths again on real music data in Section 5.3. The optimum analysis window was found to be 0.2s, thus slightly shorter than the shortest considered signal period in BD1 (i.e., 0.3 s). Figure 5.1 shows the phase slopes from a short excerpt of a music sample computed with three different analysis window lengths. It can be seen that for the short analysis window of $0.05s$ length, many spurious zeros crossings exist. On the other hand, the long analysis window of $0.8s$ leads to a small number of positive zero crossings. It can be expected that the estimation of the beat impulses (dashed impulses) from the positions of the zero crossings will be difficult. The middle figure, however, indicates a high correlation of zero crossings and beat impulses, and has been derived using an analysis window of 0.2s. This change in window length compared to $0.1s$ as applied in Chapter 3 was found crucial throughout all experiments on artificial and music signals. Except of this change, the PS_OSS has been computed with the same papameters as found optimal for onset detection in Section 3.2.1. In order to measure the correlation between the artificial signals and the positive zero crossings Equation 5.5 (see Section 5.3.2) has been used.

The way of performing a multi band analysis for the PS_OSS computation differs between our previous applied method [71] and the multiband analysis performed for optimum onset detection in Section 3.2.1. In general, dividing the spectrum into a number of bands has been shown to be meaningful for beat tracking [136, 55]. In [71], the spectrum was divided into four equally sized bands on logarithmic frequency scale. In each band, an onset strength signal using the phase slope method was computed. In order to get a single vector representation, the four band-wise onset signals, $\mathbf{c}_b[n], b = 1...4$, have been fused in the same way as in Klapuri *et al.*[92]:

$$PS\_OSS[n] = \sum_{b=1}^{4}(6-b)\mathbf{c}_b[n] \tag{5.2}$$

giving more weight to lower bands. On the other hand, the onset detection procedure described in Section 3.2.1 proposes a division into 21 bands and no weighting as in (5.2) is included. It will be examined in the experiments in Section 5.3 if this different multiband processing causes differences in the beat tracking performance.



Analysis window of 0.05s

Analysis window of 0.2s

Analysis window of 0.8s

Figure 5.1: Influence of the analysis window length on the phase slope of a music sample (x-axis: samples, y-axis: phase slope amplitude, onsets: bold peaks with circle markers, annotation: dashed peaks with triangle markers, threshold for zero crossing selection: dotted lines)

5.2 Beat tracking

For the estimation of beat times from the band-wise onset signals an algorithm based on the method proposed by Klapuri et al. in [92] has been used. The algorithm had to be adapted to the type of onset signals that are obtained using the phase slope function. This modified beat tracker will be referred to as PS/M-KLAP in the rest of the chapter. The beat tracking procedure can be divided into two parts. At first, for each time instance of a signal a beat period is determined. This can be compared with the determination of the fundamental frequency from *e.g.* a speech signal. Afterwards, given the period estimations, an optimum alignment of an impulse train having this time varying period with the signal is found. Again, a comparison with speech processing might be the determination of glottal closure instances.

5.2.1 Beat Period

For beat period estimation, Klapuri et al. suggest the computation of comb filter responses on each of the four bands separately, and summing afterwards. In PS/M-KLAP, the band wise confidence level vector $c_b(k)$, for $b = 1...21$ referring to the frequency band and k being the sample index, are simply summed without the weighting in (5.2) as proposed in Section 3.2.1. As for the onset detection task, the obtained PS_OSS has been temporally smoothed using a $51ms$ Hanning window. Afterwards, the obtained onset vector can be weighted with the sum of the spectral flux at each sample n:

$$PS\_OSS_{flux}[n] = PS\_OSS[n] \sum_{\omega} HWR(|X(\omega, n)| - |X(\omega, (n-1))|) \tag{5.3}$$

where HWR denotes a half wave rectification and $X(\omega, n)$ denotes the (short time) Fourier transform of the signal as used in the group delay computation in (3.3). This weighting was found to improve previously [71], and results with and without this weighting will be depicted in the experimental section of this chapter.

The sample autocorrelation of the vector $PS\_OSS_{flux}[n]$, or $PS\_OSS[n]$ when not applying the weighting in (5.3), is then computed in a rectangular window of $t_{beat} = 8s$ length with a step size of one second. The maximum lag considered is $4s \times f_{ons}$, which is equal to 700, since $f_{ons} = 175Hz$. The centers of the analysis windows are positioned at times $k = [1s, 2s, ..., T_N]$, where $T_N = \lfloor L/f_{ons} \rfloor$, zero padding has been applied. In the following, the beat periods β have been estimated using a *Hidden Markov Model*(HMM) as described in Klapuri *et al.*[92], where the beat period is referred to as tactus period. This results in a sequence of beat period estimations $\beta[k]$. The HMM simultaneously estimates periods at three different levels of the meter: the *tatum*, the beat and the measure period. The *tatum* refers to the smallest duration at which a period is encountered in the musical signal. The beat is related to the period. a human being would most probably tap his foot when listening to the music. The measure period is related to the positions of the bars in the musical score (see Figure 2.2 for an example). The only change in the HMM is the use of flat priors for all three estimated periods. These priors only restrict the three periods to ranges that are likely to be encountered in music. The chosen ranges are 100...700 bpm for the *tatum* period, 60...160 bpm for the beat period, and 15...90 bpm for the measure period. The exact values for these ranges have not been found to be critical. This way, the present implementation has no priors that have to be adjusted using some example dataset as in Klapuri *et al.* [92].

5.2.2 Beat Phase

In the phase estimation of the beat pulse ((27) in [92]), the computation of the likelihood of a phase $\Phi[k]$ in analysis frame k has been changed to

$$P(\hat{\mathbf{r}}_{\tilde{\mathbf{y}}\mathbf{y}}|\Phi[k]=l)=\sum_{b+1}^{21}\sum_{n=0}^{8f_{ons}}\tilde{\mathbf{y}}_k[n+l]\mathbf{c}_b[kf_{ons}+n-4f_{ons}] \tag{5.4}$$

where $\tilde{\mathbf{y}}_k$ is a reference pulse train of $t_{beat}f_{ons}+1$ samples length. Thus using the given values of $t_{beat}=8s$ and $f_{ons}=175Hz$ this pulse train has a length of 1401 samples. It has an impulse at the middle position and a period equal to $\beta[k]$. Thus, just like in the estimation of the beat period, an eight second length window has been used. Note that, as for the PS_OSS used for the beat period estimation, each band wise goodness signal $\mathbf{c}_b$ has been temporally smoothed using a Hanning window of $51ms$ length. The sum of the band wise correlations as computed in (5.4) is then used in an HMM framework as suggested in Klapuri *et al.* [92]. Again, incorporating spectral flux as in (5.3) has been examined. The influence of multiplying each c_b in (5.4) with the magnitude will be explored in the experimental Section of this chapter. Note that the accuracy of measure and tatum periods [92] have not been evaluated, as the focus is the derivation of the beat information and this information has not been annotated in the used data sets.

5.3 Experiments

This Section compares the performance of the system as suggested by Klapuri et al.[92], denoted as KLAP, with the performance of our own implementation, which uses phase slope detected onsets as input to the modified beat tracker. This system will be referred to as PS/M-KLAP.

5.3.1 Test data

Two datasets of beat annotated pieces of music have been used for evaluation. For the first dataset, two datasets previously used by the author in [71] have been combined to a single collection of 48 pieces of western popular music, which will be referred to as BT1. One of the datasets that has been merged into BT1 has been used as a training set for the MIREX 2006 Audio Beat Tracking task[1], and consists of twenty 30 second excerpts from popular music songs. Each song has been beat annotated by several listeners, who were asked to tap the beat of the piece of music. The other dataset that has been used to compile BT1 is a collection of 28 song excerpts of 30 seconds length, again all samples have been taken from popular music songs. These samples have been beat annotated by the author. The second dataset, (BT2), consists of 41 excerpts OF $30s$ length from pieces of traditional Cretan music, downloaded from the Institute of Mediterranean Studies[2]. Again, the beat for these pieces has been annotated by the author. In contrast to BT1, none of the songs contain percussive instruments, but only string instruments and vocals. In contrast to [71], the larger number of samples in the datasets will enable for a well-founded comparison between the two beat trackers.

[1] http://www.music-ir.org/mirex/2006/index.php/Audio_Beat_Tracking

[2] http://gaia.ims.forth.gr/portal/

5.3.2 Evaluation method

For the evaluations in this chapter, the two datasets have been used that are described in Section 5.3.1. For these datasets beat annotations are available. From these annotations, a unit sample sequence, $\mathbf{a}[n]$, may be obtained with pulses located at the annotated onset or beat time instance. In the same way, a unit sample sequence $\mathbf{y}[n]$ may be generated from the estimated beat pulses. The quality of a beat tracking was judged based on the function used in the MIREX 2006 Audio Beat Tracking contest[3]:

$$A_{mir} = \frac{1}{N_{ann}} \sum_{s=1}^{N_{ann}} \frac{1}{N_{max}} \sum_{m=-W}^{W} \sum_{n=1}^{N_{pt}} \mathbf{y}[n]\mathbf{a}_s[n-m] \tag{5.5}$$

where N_{pt} is the length of the two pulse trains in samples, N_{ann} is the number of different beat annotations per sound sample, N_{max} is the maximum number of impulses in the two pulse trains, $N_{max} = \max\left(\sum \mathbf{y}[n], \sum \mathbf{a}_s[n]\right)$, and W is equal to 20% of the average distance between the impulses in $\mathbf{a}_s[n]$ in samples. This function represents an estimator, of how much two pulse trains are correlated, accepting some inaccuracy regarding the placement of the beat impulses. The accuracies as computed by (5.5) will be shown, denoted as A_{mir}, in order to be able to compare with scores achieved at the MIREX contest. Furthermore, for finding the optimum length of the phase slope analysis window in Section 5.1.1, the correlation between the onset estimations and the impulse locations in the artificial data in BD1 has been determined using (5.5) as well.

Motivated by the findings in Goto and Muraoka [57], one more method will be used for the evaluation of the beat tracing. This method has been used in Klapuri *et al.* [92] as well. As mentioned in Goto and Muraoka [57], the most appropriate estimator for the performance of a beat tracking system is the length of the longest continuous correct estimated section of the song, divided by the duration of the whole song. For example, for 30s duration of a song and 12s to be the longest continuously correct beat estimation duration, the accuracy is 40%. Furthermore, the beat estimation is judged as correct when its period is half or double the period of the annotation as well. A deviation of 0.175 times the annotated period length is tolerated. Note that a beat pulse train with the same period as the annotation pulse train is considered as incorrect whenever it has a half period offset (*off-beat*). Accuracies measured with this method will be referred to as A_{cont}.

Apart from the above described beat tracking approaches KLAP and PS/M-KLAP, also the beat tracking algorithm used in Ellis and Poliner [46] will be evaluated as another state-of-the-art approach. This algorithm, just as KLAP, has been used without any changes to the original parameters, and it will be referred to as ELLIS.

5.3.3 Proof of concept

In this Section, the KLAP and PS/M-KLAP beat trackers are applied to BD1, the development set containing artificial signals. For each TNR level, computed as depicted in (5.1), the accuracies of the two beat tracking systems have been computed using (5.5) for all the signal periods in D1 (0.3s to 1s). Then the mean values and the standard errors have been computed

[3] http://www.music-ir.org/mirex2006/index.php/Audio_Beat_Tracking

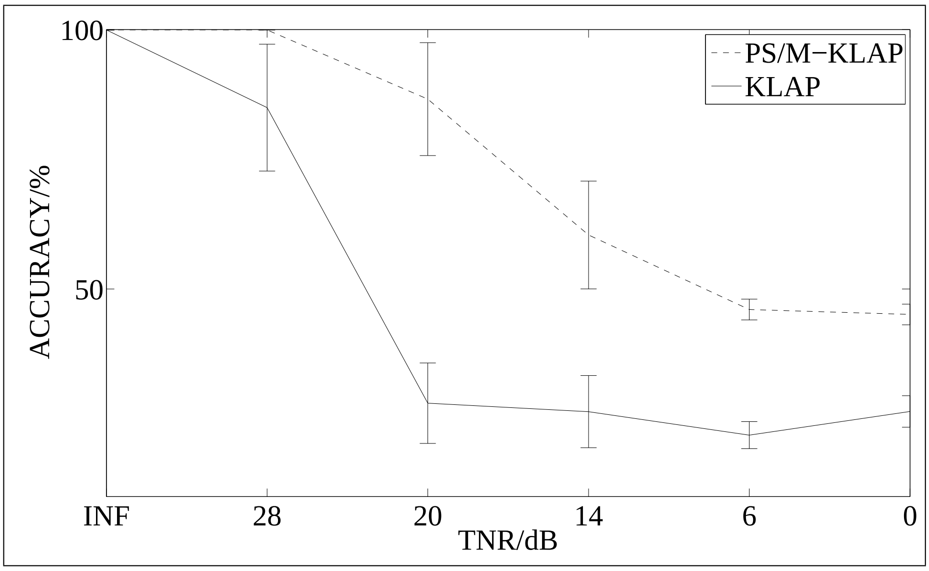


Figure 5.2: Accuracies A_{mir} of the beat tracking using the proposed method (PS/M-KLAP) and the algorithm of [92] (KLAP), on artificial signals of varying TNR

for each TNR level. The mean accuracy values along with their corresponding standard errors, shown as error bars, are depicted in Figure 5.2. Without the addition of noise both approaches estimate a beat pulse train that is perfectly correlated with the position of the impulses in the signal. When the TNR decreases, the presented approach PS/M-KLAP is persistently more accurate. This proves the hypothesis, that using the proposed approach, beat tracking will be more robust against noise which is important if an audio recording is noise corrupted. Also the presence of noise makes some of the possible percussion components found in music to soften. Based on the above results we expect the proposed approach to be also appropriate for musical signals without strong percussive components.

5.3.4 Results

The accuracies of the beat trackers applied to the music datasets BT1 and BT2 are depicted in Table 5.1 for the accuracy measures A_{cont} and A_{mir}, respectively. There are four different results for PS/M-KLAP: the first row shows the results when no spectral flux weighting is performed, the second shows the results when weighting is performed for the period estimation as depicted in (5.3), the third row depicts the accuracies when using spectral flux weighting only for the beat phase estimation in (5.4), and the fourth row shows the combination of spectral flux weighting of period and phase. The conclusions differ depending on the dataset. On BT1, the dataset containing popular music, including spectral flux weighting for the beat phase estimation clearly improves the results. It has been observed that without this weighting most errors are related to a wrong beat phase estimation. However, the same weighting leads to a strong decrease in accuracy for the Cretan music dataset BT2. Thus, it appears that for the type of signals in BT1, the spectral flux can give information about the correct phase of the beat. This is likely to be related to the different magnitude characteristics of onsets at on- and off-beats, such as a strong bass drum, which is positioned at the start of a measure and leads to

a maximum in spectral flux. On the other hand, due to the absence of percussive instruments in BT2, on this dataset no spectral flux weighting should be applied. In general, the spectral flux weighting for the beat period estimation does not lead to statistically significant changes in the beat tracking accuracies.
The comparison with the beat tracking accuracies achieved by the original beat tracking code by Anssi Klapuri reveals an interesting conclusion. These accuracies, shown in the part of Table 5.1 labeled KLAP, are about the same (A_{mir}) or better (A_{cont}) for BT1. However, for BT2 the beat tracking accuracy of KLAP is very low (28.3%/23.5%). This leads to the conclusion that the usage of PS_OSS is clearly superior to the usage of the comb filter signals in KLAP. Furthermore, the lower accuracies A_{cont} that are achieved on BT1 using PS/M-KLAP can be ascribed to the implementation of the beat tracker: Computing the comb filter signals using the code of Klapuri *et al.* [92], and using them as an input to our modified M-KLAP beat tracker leads to accuracies of 62.7%/54.1% for BT1 and 20.6%/20.3% for BT2. Comparing the results achieved with the ELLIS system also confirms the validity of the proposed method. The difference between the accuracies of the proposed method and the ELLIS system is particularly large for BT2. But also for BT1, both accuracies achieved using the ELLIS system are lower than the ones achieved using KLAP and PS/M-KLAP. Summing up, it can be concluded that the PS_OSS leads to similar accuracies for popular music when the phase alignment is supported using a spectral flux weighting, and PS_OSS is clearly superior for the beat tracking in the dataset of Cretan music.

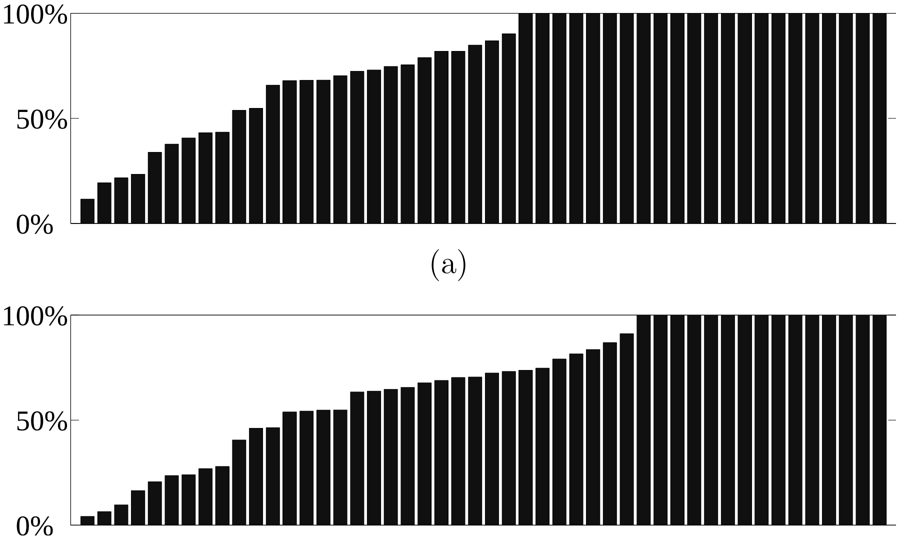
Table 5.1: Beat tracking accuracies: A_{cont}/A_{mir}, with and without spectral flux weighting for PS/M-KLAP, and for KLAP

PS/M-KLAP		
	BT1	BT2
no SF weighting	47.1/40.4	**70.8/50.8**
Period only	47.0/41.5	74.1/50.8
Phase only	**67.5/56.7**	52.9/42.3
Period & Phase	66.2/56.7	51.1/38.9
KLAP		
	BT1	BT2
	77.6/58.4	28.3/23.5
ELLIS		
	BT1	BT2
	44.7/35.3	30.3/34.9

In Table 5.2, the accuracy of the beat tracking of PS/M-KLAP are depicted when varying the length of the phase slope analysis window N from $0.1s$ to $0.3s$. The optimum window length of $0.2s$ for beat tracking has been determined in Section 5.1.1 by measuring the correlation between estimated onsets and impulse locations in the dataset of artificial signals BD1. As can be seen from Table 5.2, this finding can be justified by the accuracies obtained from music signals as well. The shown accuracies have been computing using the spectral flux weighting for the beat phase alignment for BT1 and without any spectral flux weighting for BT2, as it was found optimal in Table 5.1. Regarding the way of summing the content in the various frequency band, no improvement was observed when using a weighted summation as depicted in (5.2). For that reason, all PS_OSS have been obtained using a simple summation. A weighted summation did not lead to significant changes neither in the beat period estimation nor in the summation for the beat phase alignment in (5.4). This confirms the results obtained for the onset detection evaluations in Chapter 3, where a simple summation has been proposed as well.

Table 5.2: Beat tracking accuracies of PS/M-KLAP for varying analysis window length

PS/M-KLAP		
	BT1	BT2
0.1s	67.3/56.1	32.6/30.2
0.2s	67.5/56.7	70.8/50.8
0.3s	48.2/43.2	53.3/39.6

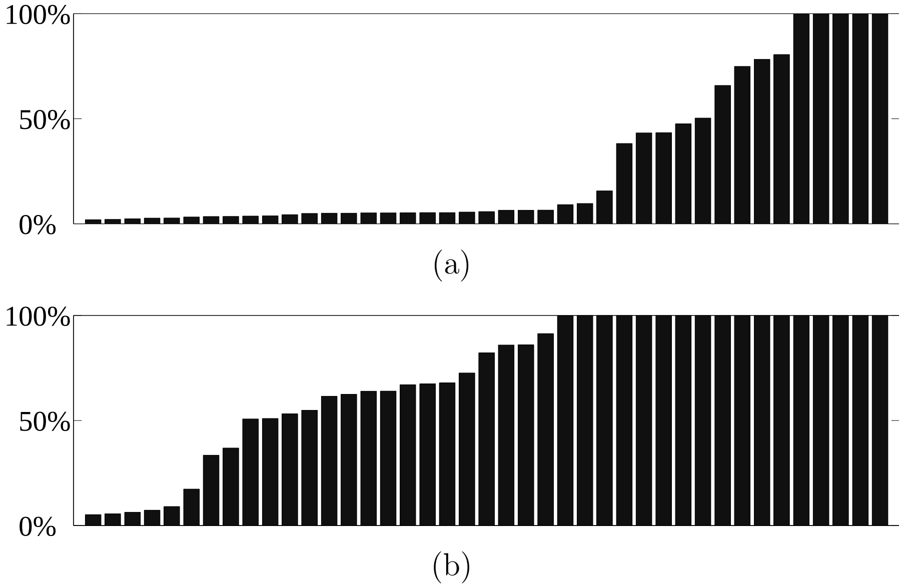


(a)

(b)

Figure 5.3: A_{cont} values for BT1 in increasing order, (a): KLAP (mean accuracy: 77.6%), (b): PS/M-KLAP (mean accuracy: 67.5%)

Let us now have a closer look at the type of errors that are encountered in the two beat trackers KLAP and PS/M-KLAP for the two datasets. In Figures 5.3 and 5.4 the accuracies A_{cont} are depicted in increasing order for BT1 and BT2, respectively. As can be seen, especially for BT2 there are a number of files for that the accuracy is very low. For that reason it was decided to have a closer look at the nature of the beat tracking errors on both datasets for both



(a)

(b)

Figure 5.4: A_{cont} values for BT2 in increasing order, (a): KLAP (mean accuracy: 28.3%), (b): PS/M-KLAP (mean accuracy: 70.8%)

beat trackers. On BT1, both beat trackers suffer from period errors and phase errors in almost the same amount. A completely different picture was obtained for BT2. On this dataset, almost only phase errors occur. This means that for both beat trackers, the beat period has been estimated correctly in almost all cases. But especially the KLAP algorithm suffers from putting the beat impulse trains to the off-beat, which means that it is characterized by 180° phase errors. These errors occur more rarely for PS/M-KLAP, where the errors are mostly caused by the phase estimation being unstable in some part of the signal. Note that this type of error is likely to be closely related to the type of music. As stated by Baud-Bovy in [6], at least the most popular Cretan dance *syrtos* is characterized by many occurrences of syncopes, *i.e.* stresses on the off-beat. As these stresses affect mainly amplitude changes in the signal, the usage of phase slope as an input of a beat tracker appears to avoid such off-beat errors.

5.4 Conclusions

In this chapter the phase slope based method to detect onsets was evaluated in a beat tracking framework. For this, a state-of-the-art approach has been adapted to the different character of the input signal (goodness signals derived from phase slope instead of comb filter outputs). It was shown that for a dataset of popular western music the proposed method is able to achieve comparable results to a state-of-the-art approach. An observed decrease in accuracy could be ascribed to problems in the beat tracker implementation. It was shown that for these type of signals, which include a reasonable amount of percussive sounds, a weighting of the phase slope signals with spectral flux is necessary to get rid of errors in the phase alignment of the beat estimation. For a dataset of traditional Cretan dances, the proposed method was clearly superior due to the better estimation of the beat phase. It appears that the off-beat times in these signals carry a larger amount of energy. This might be to some amount due to the absence of percussive instruments, and appears to have a relation with the frequent occurrence of syncopes in this kind of music. This result coincides with the observation of the author that listeners in a concert audience tend to clap the off-beat especially for slow pieces. The usage of phase slope was shown to solve this problem and it results in an applicable beat tracker for this type of musical signal. This means that when facing problems in beat tracking due to syncopes in the music signal, by using phase slope instead of magnitude derived characteristics for beat tracking off-beat errors can be avoided. This is because syncopes cause large magnitude changes on off-beats, while the phase slope without any amplitude weighting is not sensitive to these changes.

Chapter 6

Timbre Similarity

The goal in this chapter is to describe the timbre of music in a compact and salient way. Such a description should cluster sounds that are similar regarding their timbre into regions that can be discriminated using classifiers. We suggest to obtain these descriptors from a temporal/spectral description of the sound using a Non-negative matrix factorization (NMF). These NMF features will be compared to the standard features for such a task, the Mel-frequency Cepstral coefficients (MFCC), in two tasks. The first task will be the classification of a piece of music into a specific genre, such as Rock or Pop. The second task is the recognition of the lead instrument in a musical mixture signal. If the NMF features give us information about the instruments that have been mixed together in the musical sound, they have to perform better than standard MFCC in this task.

6.1 Matrix Factorization

Let's assume a real signal to be stationary within a temporal window of length t_{fft} (s). After sampling the windowed signal at a frequency f_s, its *Discrete Fourier Transform* (DFT) will provide $N_{fft} = t_{fft} f_s$ coefficients if no zero padding is used. Let $\mathbf{x}$ be an N_c dimensional column vector containing the magnitudes of the Fourier transform of the signal for frequencies up to the Nyquist Frequency, where $N_c = N_{fft}/2 + 1$. Assume that $\mathbf{x}$ has been produced by linearly combined components as:

$$\mathbf{x} = \mathbf{W}\mathbf{h} = \sum_{i=1}^{d} \mathbf{w}_i h_i \tag{6.1}$$

with $\mathbf{W}$ being an $N_c \times d$ matrix containing the description of the spectral content of the d mixture components in its columns $\mathbf{w}_i$, and $\mathbf{h}$ being a d dimensional weighting vector. Then, the problem of finding these components can be described in a Blind Source Separation [65] context. We consider the value of d in the present problem to be smaller than the number of the frequency bins, N_c, as we want to get a compact representation of the signal. Taking K observation vectors $(\mathbf{x}_1, ..., \mathbf{x}_K)$ a matrix $\mathbf{X} \in \mathbb{R}^{N_c \times K}$, containing the observations in its columns, may be constructed. This matrix is usually referred to as spectrogram, and it describes the spectral content of the signal in a temporal range denoted by t_{Timbre} in this thesis[1]. Setting the number of mixture components to a value $d \ll N_c$ we will usually not achieve equality as in (6.1) because of the time varying spectral content of the initial components throughout the spectrogram. From a mathematical point of view, every column of $\mathbf{X}$ would have to be representable as a linear combination of the columns of $\mathbf{W}$, which is unlikely to happen for a non artificial signal and $d \ll N_c$. Thus (6.1) in matrix notation becomes

$$\mathbf{X} \approx \mathbf{WH} \tag{6.2}$$

with the matrix $\mathbf{H} \in \mathbb{R}^{d \times K}$ containing the weighting vectors for time instances $1...K$ in its columns. We can pursue this approximation task with a number of error functions and assumptions on the variables.

One approach is to choose a statistical famework. In this framework $\mathbf{H}$ contains random variables (in $\mathbb{R}^d$) in its columns that are statistically independent. Then, given $\mathbf{X}$, we have to search for a matrix $\mathbf{W}^{-1}$ that minimizes the mutual information between these independent components. This approach is based on Independent Component analysis and has been presented as Independent Subspace Analysis (ISA) [24]. A necessary condition in this framework is that the distributions of the d sources that are to be estimated remain stationary throughout the length t_{Timbre} of the spectrogram under consideration. It is worth to note that the values for t_{Timbre} range from $0.25s$ up to $10s$, according to Casey [24].

Without considering a statistical framework the Non-negative Matrix Factorization (NMF) minimizes an error function like

$$D(\mathbf{X}||\mathbf{WH}) = \sum_{i,j} \left(\mathbf{X}_{i,j} \log \frac{\mathbf{X}_{i,j}}{[\mathbf{WH}]_{i,j}} - \mathbf{X}_{i,j} + [\mathbf{WH}]_{i,j} \right) \tag{6.3}$$

and constrains all the values in $\mathbf{W}, \mathbf{H}$ and $\mathbf{X}$ to be non-negative [95].

For NMF and ISA, experiments considering the influence of the length of the input spectrogram t_{timbre} and the number of components d on the Mean Squared Error (MSE)

$$MSE(\mathbf{X}||\mathbf{WH}) = \sum_{i}^{N_c} \sum_{j}^{K} (\mathbf{X}_{i,j} - [\mathbf{WH}]_{i,j})^2/(N_c K) \tag{6.4}$$

of the approximation in (6.2) have been conducted on a small number of sound samples in the author's master's thesis [69]. There, it was shown that constraining the number of observations K causes $\mathbf{X}$ to span a vector subspace of $\mathbb{R}^{N_c}$ that can be spanned by a small number of d columns of $\mathbf{W}$. In terms of musical content, due to a shorter duration t_{timbre} less different instrumental sounds will be present in the spectrogram, which causes its columns to span a more compact subspace.

In this thesis, both ISA and NMF are evaluated on a set of music samples taken from a dataset used by Li and Tzanetakis in [99]. The set consisted of twenty musical pieces of thirty seconds length each, two pieces randomly chosen from each of the ten classes contained in the dataset. The software for evaluation was taken from the MPEG-7 reference software [107]. This includes the *fastICA* algorithm [78] for the calculation of ICA. The reference software was

[1] The term *timbre* is used here since within this window the description of the spectral space of the signal will be derived

expanded by including an implementation of NMF without sparseness constraint as implemented in Hoyer [77], that minimizes the cost function shown in (6.3). The choice of this cost function has been motivated by Klabbers and Veldhuis [87], where it was found to be subjectively superior to a squared error function in measuring spectral distances. This is assigned to the property of (6.3) to emphasize differences in regions with high energy, representing therefore a weighted contrast function. The block diagram of the evaluation algorithm is shown in Figure 6.1. The power spectrum is estimated through the DFT of the signal, computed on a 40ms Hamming window with 50% overlap. The next step is a conversion from the linear frequency abscissa to a logarithmic axis. Using eight bands per octave ranging from 65.5 Hz to 8 kHz results in $N_{bands} = 56$ coefficients for each DFT window. This conversion is following the `AudioSpectrumEnvelope` descriptor (ASE) of the MPEG-7 standard. It enables a more compact description of the signal, *i.e.* it reduces dimensionality from the number of coefficients N_c on linear scale to N_{bands} . The choice of eight bands per octave has been motivated by the equal tempered musical system of western music, in which the most common tonal scales contain seven steps from the fundamental tone until its octave. Having computed the ASE vectors for a whole sample, a spectrogram representation is then obtained. This is segmented into smaller non-overlapping sub-spectrograms that represent K ASE descriptors, a step denoted as *timbre windowing* in Figure 6.1. Note that the number of observation vectors, K, defines the length of the timbre window (t_{timbre}). Varying the length of the timbre window t_{timbre} as well as the number of components d, while fixing the number of bands, $N_{bands} = 56$, we may determine the MSE of the factorizations produced by ISA and NMF. The samples of 30 seconds length

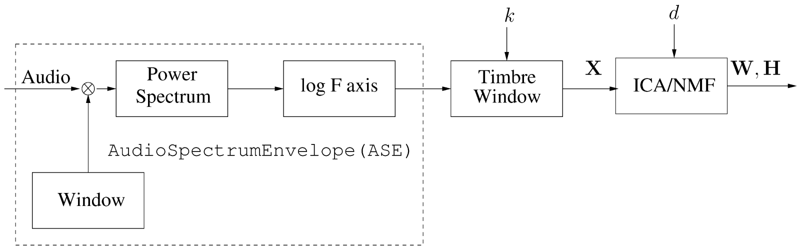


Figure 6.1: Computation of spectral bases in the MPEG-7 reference

were splitted into $N_B = [1, 2, 4, 8, 12, 16, 20, 30]$ segments of equal size. Spectrograms computed from these partitions were factorized with $d = [3, .., 30]$ components. For example, for $N_B = 4$ segments, each segment is 7.5 seconds long (segments were obtained without overlap), resulting in $K = 7500ms/20ms = 375$, where a frame rate of $20ms$ is assumed. For a given choice of splitting (*i.e.* $N_B = 4$) the corresponding MSE was computed as the sum of MSE from all segments. The number of components as well as the length of the input spectrogram influences the quality of the approximation provided by the two considered factorization methods (NMF and ISA). Increasing the number of components improved the approximation in both methods. This is, because with d increasing, the columns of $\mathbf{W}$ are more likely to construct a basis for the subspace of $\mathbb{R}^{N_{bands}}$ spanned by the columns of $\mathbf{X}$. Two example error functions averaged over the parameter d are depicted in Figure 6.2, showing that NMF is superior to ISA in the mean squared error sense for all numbers of partitions. This was consistently the case for all the songs in the set of music samples. Additionally, it can be seen that for shorter

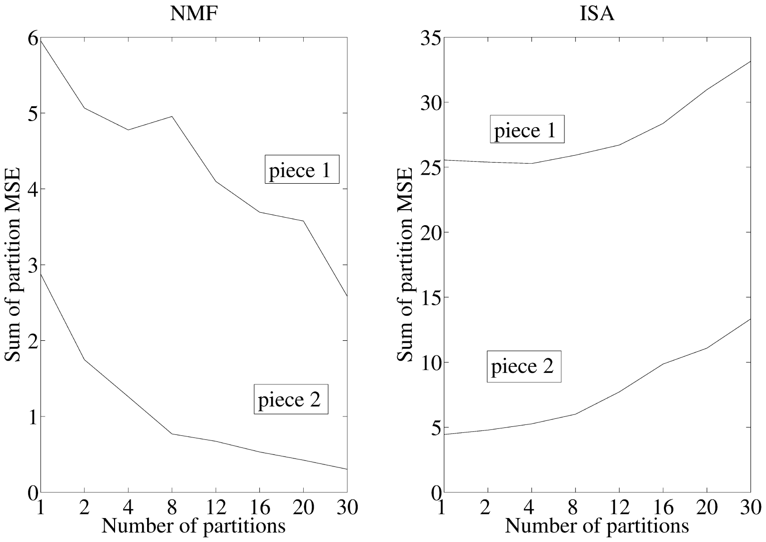


Figure 6.2: Example of error curves of NMF and ISA for two pieces of music. Approximation by NMF has generally a smaller error than approximation by ISA

spectrograms (*i.e.* more partitions) the error gets smaller for NMF while it increases for ISA. Indeed for shorter timbre windows the value of K gets closer to d and in the extreme case of $K = d$, NMF will reach a perfect result by setting $\mathbf{W} = \mathbf{X}$ while $\mathbf{H}$ being the $K \times K$ identity matrix. On the other hand, the updates in *fastICA* use sample means in order to estimate expectation values, and because of this a short timbre window leads to worse approximations (see [78] for a description of the algorithm).
We conclude that computing NMF on short spectrograms leads to more adequate spectral representations for the signals under consideration. The optimal length and number of components in the classification task will be determined in Section 6.3.4.

6.2 System Description

6.2.1 Feature Calculation

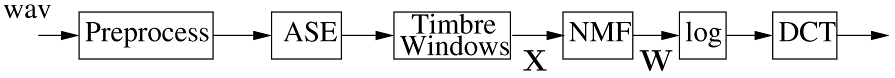


Figure 6.3: Calculation of the features used for the statistical model of musical genres

The features describing the spectral space are calculated as shown in Figure 6.3. The preprocessing steps avoid the influence of recording conditions which are not significant for classification. They include removal of mean values and normalization to an average sound pressure level of $L_{SPL} = 96dB$. The next step is the computation of the *AudioSpectrumEnvelope* descriptors (ASE), as described in Section 6.1 above. Then the timbre window is applied to segment the spectrogram of the audio signal into non-overlapping sub-spectrograms of size $N_{bands} \times K$, with $N_{bands} = 56$ and K represents the number of descriptors per sub-spectrogram. Each sub-spectrogram is then factorized using NMF providing a spectral base consisting of d vectors in the columns of matrix $\mathbf{W}$ in (6.2), with $d \ll K$. The next step transforms the energy values of the spectral bases into decibel scale, which has been shown to be crucial for an audio description task [101]. The final step of the feature calculation is a *Discrete Cosine Transform* (DCT) on the dB-scale spectral base vectors; the size of the used DCT matrix is 20×56, containing the first 20 cosine bases $\sqrt{\frac{2}{56}} \cos[\frac{(2j+1)i\pi}{2 \cdot 56}]$, $j = 0...55$, $i = 1...20$, in its rows. This helps to reduce the dimensionality of the space from 56 to 20. The resulting 20 dimensional vectors $\mathbf{v}_1, ..., \mathbf{v}_d$ represent the features of the presented system, and describe the spectral base of a sub-spectrogram in a compact way. The spectral space of the audio signal is described by the feature vectors computed from all its sub-spectrograms. Since the length of the timbre window is fixed, the number of sub-spectrograms computed from every song depends on its duration.

Psychoacoustic Model

Instead of using a logarithmic frequency axis in the *log F axis* box of Figure 6.1 the introduction of a psychoacoustic model was evaluated as well. It consists of three elements:

1. Outer ear model: At each time instance a weighting is applied to the spectrum that adapts the calculated coefficients to the actually perceived loudness of the signal. The function presented by Terhardt [147] has been used:

$$L_{TH} = \{3.64f^{-0.8} - 6.5\exp\left[-0.6(f - 3.3)^2\right] + 10^{-3}f^4\}dB \tag{6.5}$$

where L_{TH} represents the sound pressure level at hearing threshold and f denotes frequencies in kHz. It has the effect of emphasizing frequencies around 3kHz and damping low frequencies, as depicted in Figure 6.4.

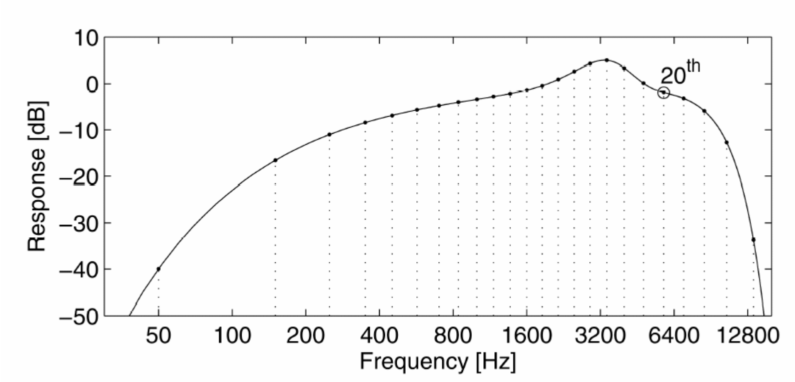


Figure 6.4: Loudness function

2. Bark scale: The linear frequency scale is converted to the *Bark* scale or critical band rate scale. This scale describes best the critical bandwidths of the human ear that lead to spectral masking when two frequencies are close enough to stimulate the same region of the basilar membrane. For an exact definition of this terminology see Zwicker and Fastl [167]. The critical bandwidths remain constant for frequencies below $500Hz$ and grow then in a non linear fashion, thus being different from the logarithmic frequency axis used in the experimental setups above. This leads to a conversion from frequencies in kHz to Bark which can be calculated as

$$z/Bark = 13\arctan(0.76f) + 3.5\arctan(f/7.5)^2 \tag{6.6}$$

Using (6.6), the lower and upper frequency limits of critical bands smaller half the sampling frequency have been calculated. Because the sampling frequency of all used data is $16kHz$, the number of critical bands to be considered is 22. The values of the power spectrum within the frequency limits of the i-th critical band, z_i, have been summed up for all bands to get the representation on the Bark scale.

3. Inner ear model: The model estimates the spread of masking between the critical bands caused by the structure of the ear's basilar membrane. The basilar membrane spreading function used to model the influence of the j-th critical band on the i-th band was derived by Schroeder in [114]:

$$\begin{aligned} 10\log_{10} B(z_i, z_j) &= 15.81 + 7.5((z_i - z_j) + 0.474) \\ &\quad -17.5(1 + ((z_i - z_j) + 0.474)^2)^{1/2} dB \end{aligned} \tag{6.7}$$

A function for a specific Bark band is steeper to the side of low frequencies which indicates that spectral masking is more present towards higher frequencies. For each of the 56 bands a function was computed using (6.7), resulting in a 22 × 22 matrix that was multiplied with the power spectrum on Bark scale. For all steps of the psychoacoustic model the implementation of Pampalk [120] has been used.

If the NMF based features used in this chapter have some connection to the characteristics that are used by humans to categorize sounds, a further improvement by this alternative preprocessing procedure may be expected.

6.2.2 Statistical Model and Classification

In this thesis, the sounds described using the method detailed in Section 6.2.1 belong to a specific class. This class is either related to its musical genre or its instrumental content. In order to construct the models for these classes we calculate the features for all audio signals of a dataset, *i.e.* the features $\mathbf{v_1}, ..., \mathbf{v_d}$ are computed for each sub-spectrogram, and then the features are stored for each class separately regardless their temporal order in the samples. This is referred to as a *bag of frames* model in Mandel and Ellis [106]. Then, a Gaussian Mixture Model (GMM), θ^i, for each class is built (i.e., with $i = 1...G$, where G denotes the number of genres or instrument classes), using a standard Expectation Maximization (EM) algorithm [9]. The EM algorithm is initialized by a deterministic procedure based on the Gaussian means algorithm presented in Hamerly and Elkan [64]. A new song is assigned to a class by applying a maximum likelihood criterion: For this, for all S feature vectors $\mathbf{v}_1, ..., \mathbf{v}_S$ collected from the sub-spectrograms of a test song the likelihoods $p(\mathbf{v}_j|\theta^i)$, with $i = 1...G$ and $j = 1...S$, are computed. Summing up the log-likelihood values for each class, the song is assigned to the class γ that has the maximum score:

$$\gamma = \underset{i}{\operatorname{argmax}} \sum_{j=1}^{S} \log p(\mathbf{v}_j|\theta^i) \tag{6.8}$$

The principle of the model training and classification is depicted in Figure 6.5. Our classification method differs from Mandel and Ellis [106] as we do not build a statistical model for the song to classify. In this way detailed information contained in the features is preserved. Design parameters of the GMM are provided in Section 6.3.4.

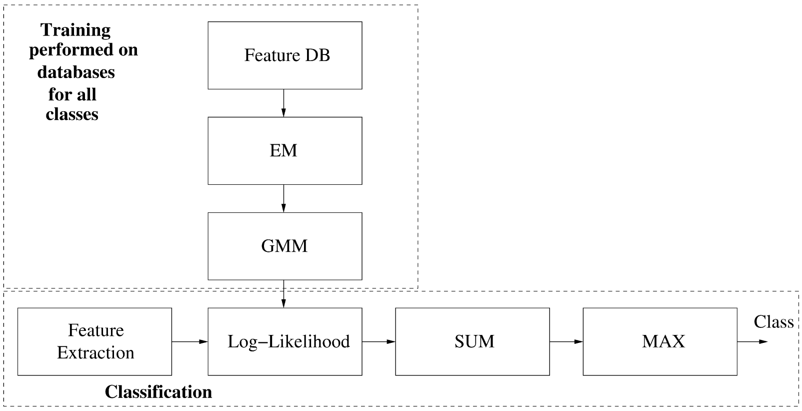


Figure 6.5: Model estimation and classification of data

6.3 Performance Evaluation

The performance of the presented system for timbre similarity is evaluated in two different ways. At first, its classification accuracy is compared with the accuracy achieved by two alternative features sets, one using MFCC, and the other using randomly chosen spectral bases. Furthermore, a stability measure is used for the evaluation as suggested in the author's master's thesis [69]. This measure is based on the distances between the statistical models built on the datasets.

6.3.1 Two alternative feature sets

In order to evaluate the performance of the proposed classification approach based on NMF it is necessary to compare with some kind of standard procedures used in many recent publications. For this purpose a *baseline* system was implemented that is as close as possible to our classification system except of the feature calculation approach. The form of the baseline system was motivated by Pachet and Aucouturier [119] which presents a frequently applied system for capturing the vertical structure of music. The model estimation and classification follow exactly the procedure depicted in Figure 6.5. However, in the baseline system MFCC are used instead of the NMF based features. Note that in contrast to [106] and [119] no model is constructed for a song to be classified. Every feature vector is considered in the same ML-classification approach as described for NMF in Section 6.2.2.

The second system to compare with differs from the NMF system only in the choice of spectral bases. These are simply d randomly chosen columns from each sub-spectrogram, which contains k columns as described in Section 6.2.1. Comparing accuracies between this system, that will be referred to as *random base* system, and NMF based system should clarify the impact of the matrix factorization in the whole classification concept.

6.3.2 A Measure of Stability

In addition to comparing the performance of the proposed classification system with those of baseline and random base system, in the author's master's thesis [69] a method to quantify the quality of the classifiers was suggested, which is based on a measure that estimates their sensitivity (or stability).

In order to judge the stability of the trained GMM, a method based on Kullback Leibler divergence (KLD) was implemented. The Kullback Leibler divergence between two distributions p_1 and p_2 is given by

$$KL(p_1||p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \tag{6.9}$$

Since there is no closed form expression for KLD in a GMM context, a possible way to get a distance measure in this case is by generating M samples $x_1, ..., x_M$ from $p_1(x)$ and then approximate KLD, by [106]:

$$KL(p_1||p_2) \approx \overline{KL}(p_1||p_2) = \frac{1}{M} \sum_{t=1}^{M} \log \frac{p_1(x_t)}{p_2(x_t)} \tag{6.10}$$

Based on (6.10) a symmetric distance measure is constructed as:

$$D_{KL}(p_1, p_2) = \overline{KL}(p_1||p_2) + \overline{KL}(p_2||p_1) \tag{6.11}$$

Let's assume that our dataset consists of G classes. Performing an n-fold cross validation, we will get a set of $n \times G$ GMMs described by their parameters $\theta_i^j, 1 \leq i \leq n, 1 \leq j \leq G$. For convenience, this set is shown as an $n \times G$ matrix in Figure 6.6. We can now determine the distances between the GMMs of different classes using (6.11) for each of the n cross validation runs separately. For example for the first run we would consider the mixture models marked by the horizontal ellipse. The minimum of these values throughout the cross validation runs gives us the least distance, D_{inter}, between two different classes. Then the distances within the classes throughout the different cross validation runs are computed, for example for the first class the mixture models marked by the vertical ellipse would be considered. The biggest value along all classes, D_{intra}, gives us a measure of how much the model differs throughout the cross validation due to diversity of the dataset. We can now define a condition measure for a specific feature set, computed by:

$$Cond_\theta = \frac{D_{inter}}{D_{intra}} \tag{6.12}$$

Obviously values for $Cond_\theta$ smaller than 1 for a specific feature set imply that a classification with this feature set might be unreliable. This is because there is a high variability between models built from a different set of data for a specific class, while at the same time there is a relatively small distance between the models for different classes. Note that using minimum and maximum values for D_{inter} and D_{intra} is a rather pessimistic approach. It penalizes a single outlier in the distances. For the intra class distance, this outlier could be the result of a single song that differed from the others in the training set and caused the model to vary strongly once it was moved from the training to the test set.

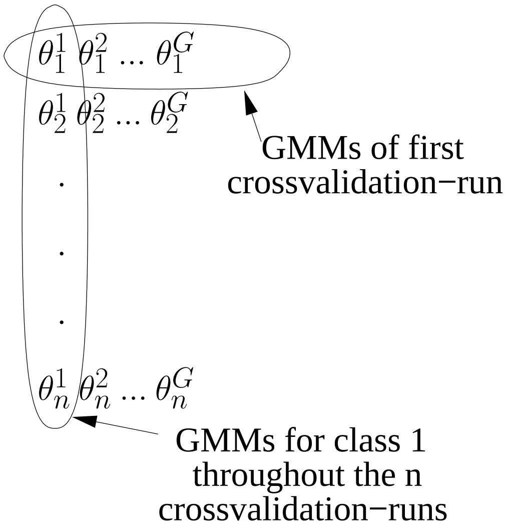

$\theta_1^1\ \theta_1^2\ ...\ \theta_1^G$
$\theta_2^1\ \theta_2^2\ ...\ \theta_2^G$
$\theta_n^1\ \theta_n^2\ ...\ \theta_n^G$
GMMs of first crossvalidation−run
GMMs for class 1 throughout the n crossvalidation−runs

Figure 6.6: Resulting GMMs from an n-fold cross validation

6.3.3 Datasets

For the experiments three different datasets have been used. All the audio files of the datasets have been converted to monaural wave files at a sampling frequency of 16000 Hz quantized with 16 bits.

The first two datasets have been widely used for the evaluation of musical genre classification systems in Western music. The first dataset (TS1) consists of ten classes[2], each containing 100 subsections of musical pieces of 30 seconds length. The dataset was collected by George Tzanetakis [99] and has been used for performance evaluation also by other researchers [15]. The second dataset (TS2) was downloaded from the website of the ISMIR contest in 2004[3], where it served as training set for the genre classification contest. The songs had been selected from the *magnatune*[4] collection. TS2 consists of six classes[5]. It contains 729 songs that are not equally distributed among the classes as they are in TS1. Also the pieces are full musical pieces and not snapshots as in TS1; therefore the lengths of the pieces in TS2 differ.

The third dataset, TS3, contains samples of traditional music from different regions of Greece and Turkey that vary regarding their instrumental content. Thus, in contrast to TS1 and TS2, the proposed timbre similarity estimation will be evaluated not in the context of musical genre classification, but rather in the context of instrumental content recognition. There are four classes in TS3, all containing polyphonic sounds as for TS1 and TS2. For the samples in the first class the main melody is played by a clarinet, in the second class by Cretan Lyra, in the third class by the Turkish wind instrument *ney*, and in the fourth class by a violin. While classes two and three contain samples from specific regions (Crete and Turkey, respectively), the other two classes contain samples that are both from Turkish and Greek traditional music. Furthermore in some of the clarinet samples in class 1 also violin is contained as accompaniment. Big

[2] Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, Rock

[3] http://ismir2004.ismir.net/genre_contest/index.htm

[4] www.magnatune.com

[5] Classical, Electronic, Jazz, Metal/Punk, Rock/Pop, World

effort has been made to avoid containing the same instrument player several times or using samples from the same album, which might cause album or artist effects. For this reason, TS3 is rather small with 20 samples per class. In all samples parts containing singing voice have been removed by hand as well as instrumental solos by other instruments.
As proposed for the MIREX 2005 evaluation[6], a 5-fold cross validation has been used. The whole dataset has been used, while stratified cross validation has not been applied. All shown classification accuracies are results of cross validations. For TS3, due to the small size of the dataset, five repetitions of 5-fold cross validations have been performed, with each of the cross validations using randomly determined train and test partitions.

6.3.4 System Parameters

For classification purposes, the optimum values for the temporal length, t_{Timbre}, of the timbre window and the number, d, of spectral base vectors to compute, should be defined. Values for t_{Timbre} from 0.25 seconds to 3 seconds have been tested. A value for d is computed by varying the values of ratio ϕ defined as:

$$\phi \leq \frac{\sum_{j=1}^{d} \sigma_j}{\sum_{i=1}^{N_{bands}} \sigma_i} \tag{6.13}$$

from 0.9 to 0.6, where σ_i denotes the i-th singular value of the *Singular Value Decomposition* (SVD) of the spectrogram to be factorized. Therefore, d provides an estimation of the minimum number of components necessary for preserving the amount of variance in the spectral basis as defined by ϕ.
These two system parameters have been defined using a subset of four classes (classical, disco, metal, rock) from the first dataset. A subset was chosen for computational efficiency and in order to avoid overfitting the system parameters to the whole dataset. The subset contains two classes that revealed to be easily classified in preliminary experiments (classic and metal), as well as two problematic classes (rock and disco). A mixture of Gaussians with five components using full covariance matrices has been built for each genre (see Section 6.2.2 for details). Figure 6.7 depicts the accuracies depending on ϕ and d. The optimum length of the timbre window is half a second while the rising accuracy for reduced values of ϕ implies that further decrease may provide even better results. However, this often leads to a value for d equal to one, especially when t_{Timbre} takes a small values. Indeed, in this case one eigenvector of the sample covariance matrix $\mathbf{X}^T\mathbf{X}$ describes a sufficient amount of the data variance (according to (6.13)). Setting d to one leads to numerical problems in the EM algorithm because some covariance matrices are close to be singular. From this we conclude that we have to assure that $d > 1$, taking therefore into account also directions of additional eigenvectors. We did experiments on the same dataset fixing t_{Timbre} to $0.5s$ and set $d = [2, 3, 4]$. We found that the classification accuracies were best for $d = 3$. This result is supported by considering the values listed in Table 6.1, which are the mean values of d determined using (6.13) to achieve the results displayed in Figure 6.7. In Table 6.1, the value of d corresponding to the best classification accuracy score ($\phi = 0.6$, t_{Timbre}=0.5s) in Figure 6.7 is close to 3. Therefore, in the following t_{Timbre} was set to $0.5s$ and d was set to 3. In this way, a meaningful representation of the signal space is achieved while the stability of the EM algorithm is assured.

[6] http://www.music-ir.org/mirex2005/index.php/Audio_Genre_Classification

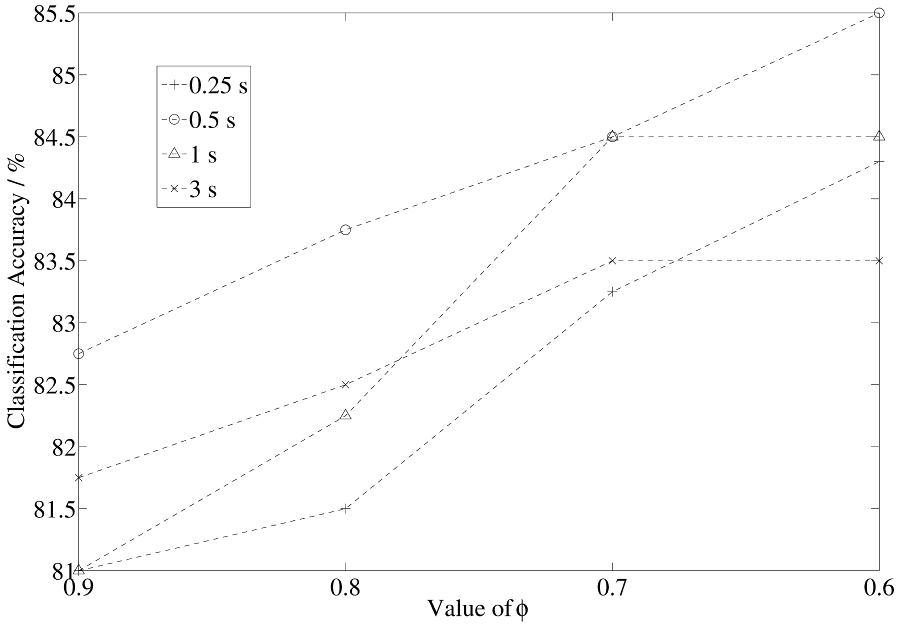


Figure 6.7: Classification accuracies for varying timbre window length and value of ϕ

Table 6.1: Mean Values for the number of spectral base vectors

		ϕ			
		0.9	0.8	0.7	0.6
$t_{Timbre}(s)$	0.25	5.38	3.60	2.62	1.96
	0.5	8.01	5.06	3.55	2.60
	1	11.30	6.95	4.76	3.41
	3	15.92	9.90	6.71	3.77

6.3.5 Classification results

Table 6.2 shows the classification accuracies on the three datasets in percentages. The rows marked with NMF contain results achieved with the system presented in Sections 6.2.1 and 6.2.2, while rows marked with MFCC contain results achieved with the baseline system as outlined in Section 6.3.1. The values in parentheses denote the number of Gaussians used in the mixture models. Full covariance matrices have been used for all experiments. We observed covariance matrices to have strong diagonals but we estimate full matrices in order to model possible covariances between the variables. For both feature sets (MFCC and NMF) the number of Gaussians had been varied in steps of five from 5 to 40. In the following Tables results that do not provide additional information have been left out to improve comprehensibility of the representation (*i.e.* for instance MFCC with 15 Gaussian components). For the fields with missing values for TS1 and TS3 training was not possible, because of the high compression performed by NMF on the training dataset. Using the bigger dataset TS2, it was possible to increase the number of components without serious estimation problems. In this case, the influence of the number of Gaussians on the classification accuracy may be observed.

The results for musical genre recognition show that our system outperforms the baseline system on both datasets TS1 and TS2. However, on TS1 the NMF based system outperforms the baseline system slightly, and the confidence interval is 2.7%, thus larger than the performance advance. But only 10 Gaussian components are necessary to reach optimum performance for the presented system, while the baseline performs best using 30 mixture components. For TS2 the performance superiority of the NMF system is clear. Also here, the proposed system achieves best results using 10 components while for the baseline system (MFCC) this is achieved using 30 components. The decline of the classification accuracy with the increased number of Gaussians may be attributed to overfitting. The dependency of the classification accuracy on the number of Gaussians for MFCC agrees with the findings in [119]. There, for 20 MFCC the best performance of the system was reached with 50 components, with slightly decreasing results when exceeding this value. Probably the lower number of components used in the baseline system for achieving the highest score can be assigned to the usage of full covariance matrices that capture correlations not extincted by the orthogonal basis of the DCT matrix used in the MFCC calculation. For the NMF features the optimum number of Gaussians is 10. This shows that more complex models do not capture significant structure in the data anymore. Thus, the usage of NMF simplified the densities of the data while keeping the significant differences between the classes. The results obtained for the instrumental content recognition on TS3 differ from the results obtained for the genre recognition datasets. From Table 6.2, it can be observed that there is no statistically significant difference between the accuracies achieved using GMM based on NMF or MFCC. Also, on this data a number of Gaussian components larger than 10 does not lead to further improvement for MFCC. Thus, it has to be concluded that there is no clear advantage for the usage of NMF features on this dataset. It is worth to point out that in the experiment a different way to train the models used for the instrument recognition has been tried as well: models have been trained on samples that contain only the lead instrument without accompaniment, while the test files contain the complex mixtures in TS3. However, also these experiments did not show any structures in the NMF features that could indicate that some of the spectral bases computed from the mixture

Table 6.2: Classification Accuracies (%) after 5-fold cross validation

	TS1	TS2	TS3
NMF(5)	71.7	75.7	76.5
NMF(10)	74.0	83.5	79.3
NMF(15)	73.9	77.7	77.5
NMF(20)	73.2	78.6	79.3
NMF(30)	-	78.5	-
NMF(40)	-	78.4	-
MFCC(10)	70.3	60.0	80.5
MFCC(20)	71.6	61.1	78.8
MFCC(30)	73.0	67.7	81.3
MFCC(40)	72.3	67.3	78.8

signals are strongly related to particular instrument sounds contained in the mixture.
The accuracies of the random base system have been extremely low for all used number of Gaussians. When comparing to the best performing systems for TS1 and TS2, *i.e.* NMF(10), the random base system with ten Gaussian components achieved accuracies of 20.2% (compared to 74.0%) and 22.8% (compared to 83.5%) on TS1 and TS2, respectively. This proves the importance of using of NMF in the computation of the spectral bases.
It is worth to note that the NMF system is trained very fast. The data reduction performed by the matrix factorization reduces a spectrogram of half a second length (25 DFT-coefficient vectors using a frame rate of 20ms) to three spectral base vectors. This yields a data compression of 88%. This is advantageous regarding training times: training a 20 component model on the first dataset took about twenty times longer using the baseline system (MFCC) instead of the NMF based system. The computation of the features for NMF took longer than computing MFCC because of the rescaled gradient descent algorithm used in NMF (about 2.3 times longer). However, summing up times for feature calculation and training, the NMF based system is still about 6 times faster than the MFCC based system. This difference in time grows non linearly with the number of Gaussians.

Even though the system suggested in this chapter captures only information about the vertical characteristics of music it also performs well in comparison with approaches incorporating more versatile feature sets that partly include *both* vertical and horizontal directions. On TS1, Li and Tzanetakis [99] report an accuracy of 71% using a feature set containing MFCC and FFT derived characteristics as well as information about beat and pitch, and Linear Discriminant Analysis as classifier. The first author of [99] presents a score of 79.5% using DWCH[7] as best performing feature and SVM as a classifier, while using GMM with three Gaussian components an accuracy of 63.5% is reported [98]. Lidy and colleagues [101] report an accuracy of 74.9% on TS1, using an SVM classifier on features describing spectral and temporal structure of a song. Pampalk and colleagues presented an accuracy on TS2 of 81% using a combination of spectral descriptors and a descriptor for modulations present in the signal, which are referred to as *fluctuation patterns* [122]. Using the training and development set of the ISMIR 2004 Audio description contest as a dataset, the system presented in [101] was reported to achieve an accuracy of 80.3%.
For sound classification approaches that are based on spectral projections and HMM, as for example [86] and [26] , no results on the presented datasets are known to the authors. Nevertheless, the approach presented in [86] has been implemented by the authors and tested on TS1, resulting in an accuracy of 50% in a five-fold cross validation. This indicates the superiority of the approach presented in this chapter to the mentioned projection based approaches, at least in the context of musical genre classification.
Another important conclusion can be drawn by comparing the results of the baseline system on TS2 with the results of [122], where MFCC have been used as an alternative feature set as well. The baseline system presented in this work does not build a statistical model of a song, but considers each MFCC vector separately by calculating its likelihood given the class models. In [122] songs have been modeled by Gaussians. This leads to an improvement in the classification accuracy of about 17% compared to our baseline system. Thus, it seems that

[7] Daubechies Wavelet Coefficient Histogram

by modeling the feature distribution for a song using GMM, results are improved, a finding confirmed in Mandel and Ellis [106] in an artist identification task. Based on the above observations it would be interesting to check if such a modeling approach will be also beneficial for the NMF based system, although such an approach is computationally quite expensive.

Confusion matrices using NMF based features are provided in Tables 6.3 and 6.4 for TS1 and TS2, respectively, using 10 Gaussians (NMF(10)). The columns contain the actual genres of the test data and rows contain the predicted classification. Apart from illustrating the above referred results and observations, Table 6.4 can be contrasted with the matrices shown in the ISMIR 2004 genre classification contest[8]. In most cases misclassifications have musical sense. For example, the genre Rock in TS1 was confused most of the time with Country, while a Disco track is quite possible to be classified as a Pop music piece. In TS2 the Rock/Pop genre was mostly misclassified as Metal/Punk. Genres which are assumed to be very different, like Metal and Classic, were never confused. The worst classification performance for the proposed system was: Rock in TS1 (57%, NMF(10)) and World in TS2 (63.3 %, NMF(10)). It is worth to note that this behavior in performance is similar to other systems as well (see ISMIR genre contest results). The low performance for these genres may be assigned to their large intra-variance of music style (at least for the analyzed data).

For TS3, the confusion matrices for NMF(10) and MFCC(30) are shown in Table 6.5. The mean values from the five cross validation repetitions are shown. Both NMF and MFCC features are not only characterized by similar accuracies, but also the confusion between the instrument classes is similar. As it was expected considering the character of the classes as outlined in Section 6.3.3, the instrument classes violin and clarinet are harder to classify due to their wide variation in musical style. On the other hand, the most compact class in terms of musical style (lyra) is classified almost without errors. Thus, the conclusion can be drawn that both MFCC and NMF features are well capable of classifying signals according to their instrumental content. However, valuable insight can only be gained if a larger dataset is available.

Table 6.3: Confusion matrix for dataset 1, using NMF based features (NMF(10))

	Bl	Cl	Co	Di	Hi	Ja	Me	Po	Re	Ro
Bl	68	1	3	0	1	4	0	1	8	3
Cl	0	94	0	0	0	4	0	0	0	0
Co	12	1	73	6	0	2	1	7	5	16
Di	3	0	10	69	8	5	4	6	2	11
Hi	0	0	0	6	69	2	1	2	12	2
Ja	1	2	0	0	1	79	0	1	1	0
Me	2	0	2	1	2	2	83	0	0	5
Po	1	0	4	10	3	1	0	79	2	2
Re	3	0	0	2	13	1	0	2	69	4
Ro	10	2	8	6	3	0	11	2	1	57

[8] http://ismir2004.ismir.net/genre_contest/results.htm

Table 6.4: Confusion matrix for dataset 2, using NMF based features (NMF(10))

	cl	el	ja	mp	rp	wo
cl	300	1	0	0	0	10
el	0	103	0	1	8	24
ja	0	0	25	0	0	0
mp	1	0	0	32	16	2
rp	6	7	0	10	69	8
wo	13	4	0	2	7	76

Table 6.5: Confusion matrices for dataset 3, using NMF(10) and MFCC(30)

NMF(10)					MFCC(30)				
	clar	lyra	ney	viol		clar	lyra	ney	viol
clar	11.2	0	2.2	4.2	clar	13.0	1.2	2.0	4.6
lyra	1.0	20.0	0	0.2	lyra	1.2	18.8	0.0	0.0
ney	5.0	0.0	17.0	0.4	ney	3.6	0.0	18.0	0.2
viol	2.8	0.0	0.8	15.2	viol	2.6	0.0	0.0	15.2

Psychoacoustic Model

The psychoacoustic processing described in Section 6.2.1 was included into the feature calculation as depicted in Figure 6.1 in the place of the simple log frequency conversion rule. All the other components of the system have been left as before and the results of the classification have been compared with the best performing NMF systems, *i.e.* NMF(10) in all cases. Classification results are shown in the first row of Table 6.6. For convenience, the best scores from Table 6.2 for log frequency rule are repeated in the third row. For TS1 and TS2, the introduction of the psychoacoustic preprocessing deteriorated the performance of the system noticeably. Experiments have been conducted in order to evaluate the influence of the individual steps of the preprocessing, *i.e.* the outer ear model, the Bark scale and the inner ear model. On TS1 using only Bark scale without inner/outer ear models performed best. On TS2, Bark scale used together with the outer ear model slightly outperformed the complete psychoacoustic model. The accuracies of these two settings are denoted in the second row of Table 6.6. For TS3, the usage of the full psychoacoustic model lead to an improvement compared to the usage of the simple log frequency conversion. However, in order to confirm the improvement of about 5% as statistically significant, the dataset would have to contain at least 200 samples, which is not possible to achieve with the music collection available to the author. Thus, it has to be concluded that neither a partial nor complete usage of the psychoacoustic preprocessing leads to significantly improved performance. If the psychoacoustic model efficiently describes the perception system, we would expect the classification results to be better than in the case of using the simple log frequency conversion rule. Therefore, either the model does not describe the perception process efficiently, or the features as input to the system have nothing to do with the cues used by humans for classifying a musical piece. Note that in Lidy and Rauber [101] the influence of the particular parts of psychacoustic preprocessing on the accuracy in a genre classification task has been analyzed. The result is the outer ear model being a crucial part of the preprocessing, which is contradictory to our results. As the psychoacoustic model used in Lidy and Rauber [101] is similar with the one used in this thesis, a reason for the bad performance of the psychoacoustic model could be the combination of this specific preprocessing with NMF. In any case, the results show that the usage of psychoacoustic models is not a guaranty for a performance improvement, and it has to be evaluated for the task and data representation at hand.

Table 6.6: Performance with and without a psychoacoustic model (%), NMF(10)

	TS1	TS2	TS3
Psychoacoustic Model	68.1	72.1	84.3
Best Psychoacoustic	72.8	77.1	84.3
Log frequency scale	74.0	83.5	79.3

6.3.6 Stability Measures

As described in Section 6.3.2, the stability of a given GMM based classifier is estimated based on distances between the models for the particular classes according to (6.12). Table 6.7 shows these condition numbers for all different configurations that had been depicted in Table 6.2. For the first two datasets, the condition numbers are always bigger for the proposed NMF based model than for the MFCC based model. Only for 5 components the NMF based features have a condition number less than 1. This can be attributed to the existence of components with large variance. Moreover, with more than ten components, the condition numbers for the NMF features are consistently bigger than one. For the baseline system all the condition numbers are smaller than one on the first two datasets. This indicates that for the NMF based features the smallest inter class distance is always bigger than the biggest intra class distance; this is not always the case for MFCC. This provides a further proof of the superiority of the proposed feature set compared to MFCC. Only for the very small data set TS3 the condition numbers for MFCC are larger than one. This indicates that this dataset results in relatively compact and well separated models for each class, independent of the feature that is applied. It has to be discovered how increasing the number of samples per class and the number of instrument classes will affect the condition measure and the accuracies achieved. Only then a conclusion can be drawn, if either MFCC or NMF are superior for the task of musical instrument recognition.

As an example, we show a graphical representation of the inter class distances for NMF(10) model on TS1 in Figure 6.8. The mean values of the inter class distances from the 5-fold cross validations have been calculated; dark areas indicate a low distances and light areas indicate higher distances. It is evident that there is a high correlation between the confusion matrix in Table 6.3 and the distances depicted in Figure 6.8 (computed using (6.11)). Similar correlations can be observed for the other two datasets as well.

Table 6.7: Condition Numbers

	Dataset 1	Dataset 2	Dataset 3
NMF(5)	0.85	0.69	1.62
NMF(10)	1.33	1.27	1.79
NMF(15)	1.62	1.29	1.78
NMF(20)	1.53	1.37	1.93
NMF(30)	-	1.20	-
NMF(40)	-	1.15	-
MFCC(10)	0.88	0.56	1.92
MFCC(20)	0.86	0.55	2.12
MFCC(30)	0.89	0.64	2.02
MFCC(40)	0.92	0.52	2.15

Note that for the NMF based features on the larger datasets TS1 and TS2 there is also a high correlation between the condition numbers in Table 6.7 and the classification accuracies in Table 6.2: The condition numbers of the NMF based system rise until a certain number of Gaussians that is bigger than the optimal in the classification accuracy sense (15 instead of 10 for TS1, 20 instead of 10 for TS2, compare with Table 6.2). Beyond this maximum the condition numbers decrease. A similar pattern may be observed for the classification score in Table 6.2. However, this structure is not clear for the MFCC based system.

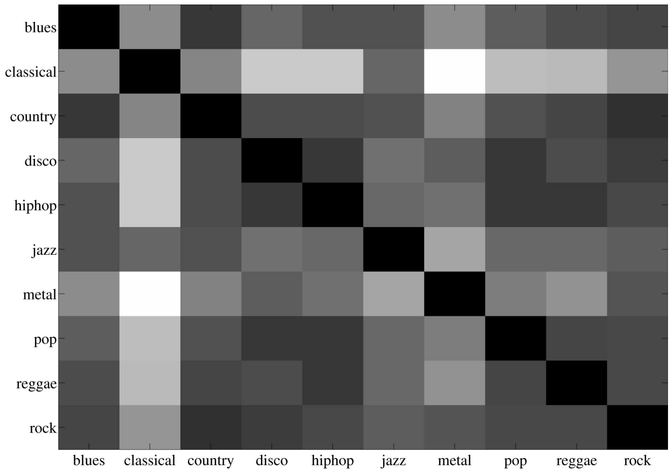


Figure 6.8: Inter class distance matrix for NMF(10) on TS1

Taking a detailed look at all the measured inter and intra class distances reveals a more informative insight into the different characteristics of the feature space modeling. Sorting all the intra class distances in increasing order gives the plots shown in Figure 6.9 for TS1, in Figure 6.10 for TS2 and in Figure 6.11 for TS3. The total number of computed distances in Figures 6.9 and 6.10 is given by $C\frac{n(n-1)}{2}$ where $n = 5$ is the number of cross validations and C is the number of classes ($C = 10$ for TS1 and $C = 6$ for TS2). For Figure 6.11 this number has to be multiplied by 5, because the cross-validation has been repeated 5 times.

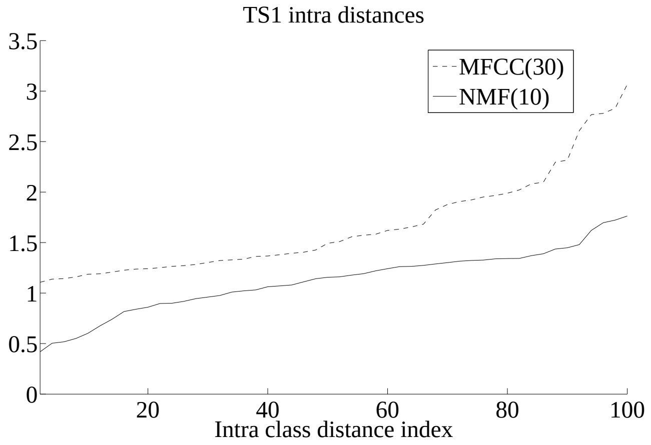


Figure 6.9: Sorted intra class distances for TS1, NMF: solid line, MFCC: dotted line

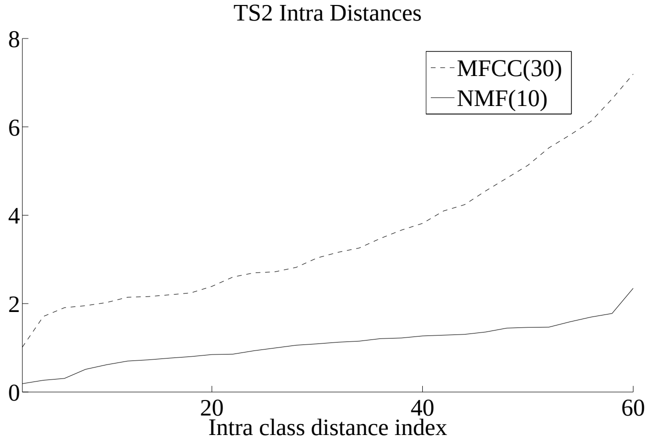


Figure 6.10: Sorted intra class distances for TS2, NMF: solid line, MFCC: dotted line

As a common difference between the two feature sets we can recognize that the intra class distances between the NMF based models are more evenly distributed. This is indicated by a less steep gradient of the corresponding curves in Figures 6.9 to 6.11. In these figures we show the intra class distances for the number of components that provided the best classification score for both features; 30 for MFCC and 10 for NMF based features. A similar behavior for both features has been observed for other numbers of components. However, on TS1 and TS2, for 5 components in the case of NMF-based features the steepness of the corresponding curve was high, which caused the condition number to be smaller than one.

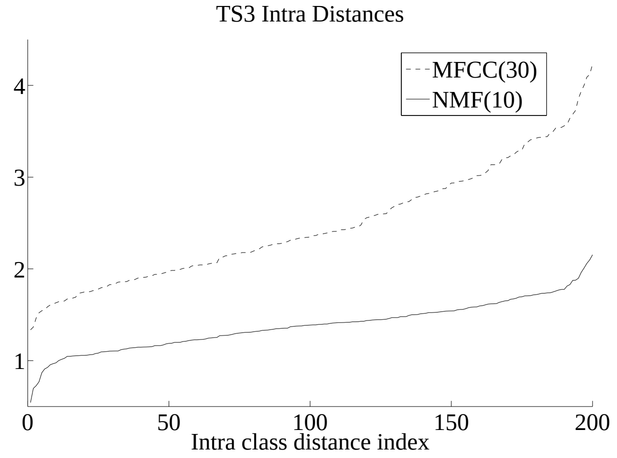


Figure 6.11: Sorted intra class distances for TS3, NMF: solid line, MFCC: dotted line

6.4 Conclusion

A feature set based on NMF of the spectrogram of a music signal has been proposed for the description of the timbre of music. It has been evaluated for the task of automatic musical genre classification and automatic instrumental content classification. Extended experiments on two widely used datasets and a newly presented dataset showed the superiority of the proposed features compared to the standard feature set of MFCC in the task of genre recognition, while for musical instrument recognition both approaches perform equally well. By using Kullback Leibler based distance measures, we were able to connect the superiority of the NMF based features in the classification task with more uniform, compared to the MFCC case, intra class distances. In addition the proposed feature extraction algorithm has the advantage of low training times of the mixture models due to the data compression and the lower number of Gaussians necessary to reach the optimum classification accuracy. Tests with a psychoacoustic preprocessing did not improve the classification accuracy. It should be evaluated in future work if the different conclusions for the instrument recognition task change when the size of the dataset is increased.

Chapter 7

Morphological Similarity: Integration

In this chapter, a preview is given for the integration of the system elements shown in Figure 1.1 to a system for morphological analysis of traditional music. Let us shortly sum up again what has been said in Chapter 1 about the morphology of the music under consideration. The traditional music of Greece often follows the logic of *parataxis*, that is it made up of small melodic phrases that are stringed together in a way that the musician considers beautiful. As mentioned in Chapter 1, these phrases appear in more or less the same form in different pieces. Thus, it would be interesting and helpful for the understanding of this music to automatically discover phrases that are similar. This is particularly interesting in the case when the amount of samples is large and, as it is usually the case for traditional music, when there is no transcription. By automatically discovering this similarity one could locate similar phrases from a large collection without the necessity of first transcribing the melodies into a score. Also music from different regions could be explored in the context of a comparative musicological study. An integration of rhythmic and melodic similarity for such a task is meaningful: It has been repeatedly confirmed by local musicians, that not only the melody is of importance for recognizing a specific dance, but also the way the instrument player puts emphasis on particular notes of the melody. On the other hand, the performed dance music classification might be improved by including melodic characteristics as well, because some of the dances (*e.g.* *Pentozalis* and *Sousta*) traditionally differ in the tonal extension of their melodic phrases.

7.1 Dataset

For the evaluation of a computational morphological analysis system a small dataset was compiled that is described in further detail in Appendix A.3.4. It contains 40 polyphonic samples of the Cretan dance *Sousta*, with the instrumentation being Cretan *lyra* and Cretan *lute* and will be referred to as MS1 in the following. For each of the 40 samples there exists a "partner" within the dataset that has been found to contain the same motif in the main melody according to an analysis by an expert. Thus the dataset contains a set of 20 morphologically related pairs that have been found in different recordings by different artists. For that reason, even though they share the same motif, they differ in terms of interpretation.

7.2 Evaluation methods

Two different tasks can be performed: First, only the 40 short samples are used to compute their mutual similarity regarding melodic and rhythmic content. The quality of the obtained similarity measure can be evaluated using the Mean Reciprocal Rank (MRR)

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i} \tag{7.1}$$

where Q is the number of queries. For our data set this means that each sample is used as a query once, *i.e.* $Q = 40$. If *e.g.* the correct partner is found on place 3 of the most similar samples, the reciprocal rank is $\frac{1}{3}$. This means that the closer the MRR is to the value 1, the better the similarity measurement.
Another test is using a sample from MS1 as a query and computing similarities for the whole duration of the piece that contains its partner motif at some time instance. If this similarity measure shows a peak at the position of the true partner, the goal of locating it in a continuous piece is achieved.

For the computation of similarity a baseline system that regards melodic content only will be used. This system was presented in Ellis and Poliner [46], and uses beat synchronous chroma features to describe the melodic content. This system was proposed for the detection of coversongs in western pop music, and it will serve as a starting point for the studies of detecting morphological similarity in traditional music. The first computational step in this approach is a beat tracking that uses a spectral flux like OSS as an input, and derives the beat time instances using dynamic programming. Then, for each beat time a 12-dimensional chroma feature is computed. These chroma features record the intensity associated with each of the 12 semi-tones of the well-tempered tonal system. In order to determine, how well two songs match, the cross-correlations between two feature matrices are computed for each possible transposition. In the following, this system will be referred to as BASE-MEL-SYS.
It will be evaluated if the usage of the rhythmic similarity measure detailed in Chapter 4 shows promising results for the given data. For this, SF_OSS will be computed with a sampling frequency of $50Hz$ as described in Chapter 4. From these OSS autocorrelations are derived in a beat synchronous approach. For that, the center of an analysis window is positioned at a time instance where the beat tracker located a beat. The length of the analysis window is set to eight beat impulses, which means that if the analysis window is centered at the $n-th$ beat, its width will be starting from beat $n-4$ and ending at beat $n+4$. The reason to decide for this width is that the used beat tracker estimated a tempo about two times higher than the correct one for all samples (tempo doubling). As the time signature of the music contained in the samples is $\frac{2}{4}$, such an analysis window length is related to two measures. This is the usual length for a melodic pattern in this dance (*Sousta*). For each analysis window, the signal inside the window is zero-padded to a length of 3 seconds, and an autocorrelation sequence is computed. The reason for the zero padding is the necessity to use patterns of constant length in our implementation of the scale transform according to (4.4). For each analysis window, a scale transform is applied to the autocorrelation sequence of the window, in the same way as described in Chapter 4. This way, for each song a feature matrix is obtained, with its number

of columns determined by the number of beat instances, and each column contains a scale transform magnitude. These matrices have been compared using the same method as applied in Ellis and Poliner [46] to the chroma features, with the exception that the correlation does not need to be computed in two dimensions, but only in the time dimension. In the following experiments this system will be referred to as RHYTHM-SYS.

As the sound files are complex mixtures, both melodic and rhythmic similarity are degraded by the other instruments contained in the mixture. Thus, a lead melody extraction using a method as the one proposed in Klapuri [89] could be included as a pre-processing at least for melodic similarity. Furthermore, instead of using chroma features, in the context of traditional music melodic histograms of a finer resolution have been found useful for the classification of melodic content [19]. In order to determine if such approaches can be adapted to the beat synchronous melody description framework, the lead melody will estimated using the algorithm presented in Klapuri [89], which was provided by the author of the paper. The parameters given as input to the algorithm are the desired number of fundamental frequency tracks to be estimated from the signal (set to 1), and the fundamental frequency range of the desired F0 tracks. This range was set to $60Hz...480Hz$, after an analysis of the available scores of the recordings. The obtained F0 tracks have been checked by resynthesizing the estimated F0 contours using a sinusoidal synthesis, and by playing these resynthesized samples in parallel to the original. In general, despite some local problems in the estimations, all melodies can be recognized from the estimation. An objective measurement of the quality of the estimations has not been performed. The next step is the computation of beat synchronous melody histograms. Motivated by the work presented in Bozkurt [19], the frequency resolution of these histograms is set higher than necessary for music using scales of the well-tempered system. This is because in Greek traditional music many modal scales are encountered which make use of tonal steps different from the half tone of the well tempered system. For example, some of these scales have their roots in the scales investigated in Bozkurt [19]. Scales like *Hidzaz* and *Kurdi* are examples for this case, and because these scales are also used in Cretan music the finer resolution of the histograms is theoretically justified. Thus, for a song a matrix is obtained with one column for a beat instance which contains the melody histogram for this beat. Again, for matching two samples the method proposed in Ellis and Poliner [46] has been used in the same way as for the chroma features. The system that uses this kind of melody histograms will be referred to as HIST-MEL-SYS.

A combination of the similarity measures derived from melody and rhythm can follow a simple procedure: the computed correlation values obtained for melody and rhythm parts can simply be added, as they have been derived in a (beat) synchronized way. Note that due to the limited size of the available data, the results shown below can be only indicative.

In Ellis and Poliner [46], the features are computed beat synchronous. This means that a beat tracking is necessary as a pre-processing step. For this, in Ellis and Poliner [46] OSS derived from amplitude are used to perform the beat tracking. However, the results in Chapter 5 indicate that for the investigated type of music a beat tracking that uses phase slope based OSS (PS_OSS, see Chapter 3) gives more accurate results. Thus, it should be evaluated as well if the accuracy of the beat tracking has some impact on the results of the matching experiments.

7.3 Experiments

7.3.1 Setup 1: Matching pairs

In the first experiment, the BASE-MEL-SYS system was applied to the data set of 40 song excerpts, MS1. Each song was used as a query and the mean reciprocal rank as defined in (7.1) was computed, which resulted in a value of $MRR_{BASE-MEL} = 0.38$, as shown in Table 7.1.
In the next experiment, scale transform based rhythmic descriptors have been computed in a beat synchronous way, as described above in Section 7.2. As can be seen in the second row of Table 7.1, the results obtained on the small sample dataset using RHYTHM-SYS are worse than the results obtained using the melody baseline system BASE-MEL-SYS. However, such a result had to be expected on the given data for the following reason: As mentioned in Section 7.1, all 40 samples are examples of the same dance *Sousta*. Furthermore, all recordings are from the same region of Crete, the municipality of Rethymnon. For that reason, all samples have a very high similarity in terms of rhythm. For that reason, the combination of rhythmic and melodic similarity measures will have to be evaluated in future on a bigger dataset that contains different kind of dances. In order to be able to obtain such a dataset, a large enough analysis of a collection of recordings has to be performed by musicologists.

In the following it will be tried if the performance in terms of the mean reciprocal rank of the BASE-MEL system can be improved by involving an estimation of the main melody from the polyphonic samples and the usage of high resolution histograms in the HIST-MEL-SYS system. In Bozkurt [19], a resolution of one Holdrian comma (Hc) is referred to as the smallest interval considered in Turkish music theory, and the authors use a resolution of $\frac{1}{3}$ Hc for their histograms. One Holdrian comma is equal to 22.6415 cents, and the octave interval can be divided into 53 Hc or 1200 cents. Various resolutions have been tried, but no clear result regarding the optimum value could be obtained on the limited sized dataset. For that reason, the resolution has been set to 2 Hc, thus 2.5 times higher than the resolution of well tempered scales (5 Hc). As can be seen from the third row in Table 7.1, the obtained mean reciprocal rank of 0.58 is improved compared to the BASE-MEL-SYS system. This improvement is present almost independently of the histogram resolution, which indicates that the sensitivity to microtonal changes is not of importance at least for the present dataset. Again, as for the rhythmic analysis, bigger and more diverse datasets have to be obtained to achieve more insight into the parameter settings necessary for a good measurement of morphological similarity.

Table 7.1: Mean reciprocal rates (MRR)

BASE-MEL-SYS	0.38
RHYTHM-SYS	0.20
HIST-MEL-SYS	0.58

7.3.2 Setup 2: Matching queries in whole songs

As described in Section 7.2, the second experimental setup is using one of the short samples contained in MS1 as a query. For this experiment 10 phrases of two measures length have been selected as depicted in the first column of Table 7.2. For example, the query file `13b42b:234` is the phrase 13*b*42*b* taken from the recording number 234 in the collection. It has been tried to locate its appearance in the file its partner in MS1 has been extracted from using the HIST-MEL-SYS method. This lead to the best pattern matching results as shown in Table 7.1. The highest correlation measures in these files are depicted in the column titled $max(R_{pos})$ in Table 7.2. In the column titled `MATCH` the success of this matching is judged. If the position connected to this highest correlation measure is exactly the position where the partner file in MS1 has been extracted from, the label `EXACT` has been assigned. If the position of the correlation maximum is related to another appearance of the same pattern in the file, it has been labeled as `CORRECT`. Finally, when a different pattern from the query pattern is located at the position of the correlation peak, the label `WRONG` has been assigned. This evaluation has been performed entirely by hand, by locating the time instance of the correlation maximum of the melody histogram in the related musical score. It can be seen that only in one case the matching gave a wrong result, while all the other 9 matches were related to an appearance of the same melodic phrase in the target file. It has to be stressed again that all the target files are different from the file that the query has been taken from. The target files used in the column titled $max(R_{pos})$ are different samples which contain at one or more time instances a melodic phrase that has been judged to be identical with the query by an analysis conducted by musicologists.

The correlation between the F0 histogram of a query sample and the histogram of the whole file it has been extracted from has been computed as well. This enables to determine how good the matching works in the perfect case, where the pattern we are looking for is indeed contained in the file exactly as found in the query. The resulting correlations are depicted in the column entitled R_{source} in Table 7.2. It can be seen that they are always larger than the correlation depicted in $max(R_{pos})$, but never equal to 1. This is likely to be caused by

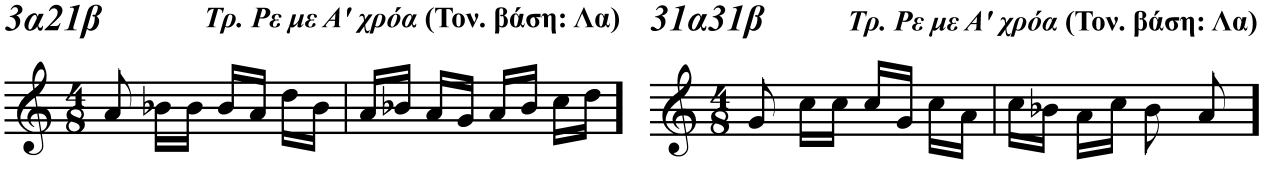
Table 7.2: Results of matching patterns from MS1 in whole song files

QUERY FILE	$max(R_{neg})$	R_{source}	$max(R_{pos})$	MATCH
(1) 13b42b:234	0.5796	0.9200	0.6403	EXACT
(2) 4a31b:217	0.3602	0.9301	0.6741	EXACT
(3) 3a3b:027	0.5059	0.9297	0.6238	CORRECT
(4) 35a35b:196	0.5482	0.9416	0.6866	CORRECT
(5) 3a21b:051	0.4511	0.8549	0.7040	EXACT
(6) 89a46b:143	0.4881	0.6571	0.5451	EXACT
(7) 31a31b:035	0.4830	0.8989	0.6351	WRONG
(8) 6a72a:167	0.5535	0.8778	0.6578	EXACT
(9) 7a6b:008	0.5073	0.8242	0.5870	EXACT
(10) 62a62b:249	0.4484	0.8333	0.5869	EXACT

slightly differing beat tracking and F0 estimation results on the small query samples and on the whole file. Furthermore, the queries have been applied also to files, where according to the annotation the phrase is not contained neither as a whole nor half of it. The correlation maxima are depicted in the column titled $max(R_{neg})$, and these values are always smaller than the correlation values computed in the other columns. This supports the assumption that the proposed method is able to separate similar phrases from those that do not share a large similarity with the query phrase.

In Figures 7.2 and 7.3, all R_{pos} vectors of the 10 queries shown in Table 7.2 are plotted. These vectors have been obtained by computing the two dimensional correlations between the query and the target histogram matrices, and the choosing the row in the correlation matrix, that contains the maximum value. In all plots, maxima have been chosen and it has be evaluated if at the related measures in the score indeed the query phrase is found. Maxima for which this is the case have been marked with dashed boxes, while maxima which are not related to the query pattern have been marked with dotted boxes. A first and important result of this analysis is that in none of the cases an occurrence of the query pattern in the investigated audio file has been missed, which means that in every case the occurrence of the pattern was related to a maximum in R_{pos}. Also the overall number of true positives (dashed boxes) is 21 while the number of false positives (dotted boxes) is only 7. However, as explained in Appendix A.3.4, these false positives do not imply that there is no similarity between the query and the target at the time instance of the false positive. The false positive only indicates that at this position the phrase played by the lead instrument does not have exactly the same label. Taking a closer look at the false positives reveals that for example all wrong detections for query (3) are phrases which contain the pattern `3a` which is also contained in the query sample (`3a3b`). A closer look has been taken at the only case, where the maximum in R_{pos} is connected to a false positive (query (7)). The query phrase and the phrases found in the dotted boxes in Figure 7.3.(7) are depicted in Figure 7.1. It is apparent that at least the first parts of the two phrases share a big amount of similarity. Thus, at least in this case, the false positive is related to a similar melodic phrase.

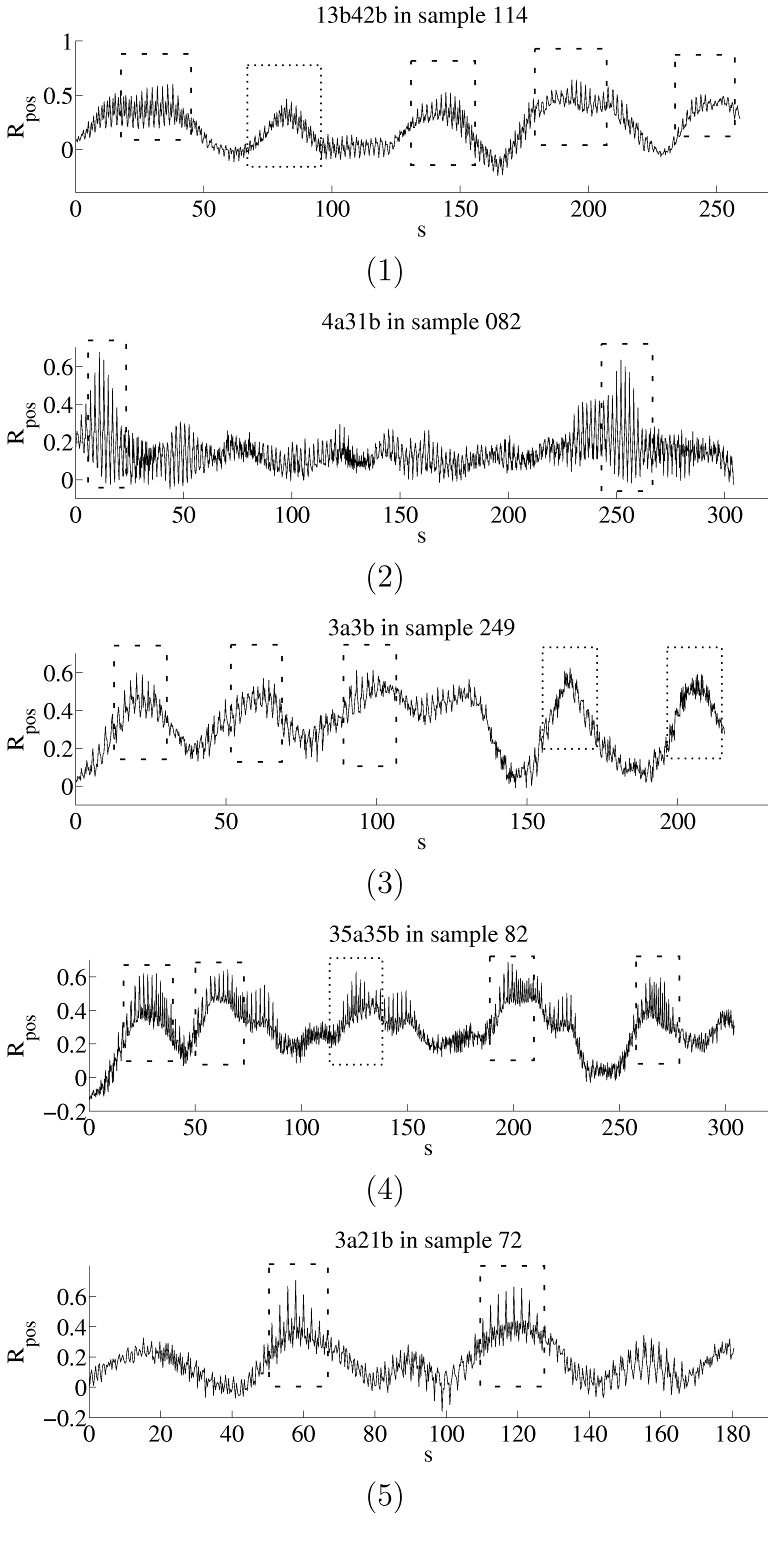
Another observation from Figures 7.2 and 7.3 is that maxima related to true positives seem to be characterized by a strong oscillation. This oscillation has been observed to have the frequency of exactly two measures. This means that the correlation shows a strong peak whenever the beginnings of the query phrase and the related phrase in the investigated file are aligned. This effect should be further investigated when a larger dataset is available, and it is possible that a detection of such oscillations, beside high correlation envelopes, further improves the result of the pattern retrieval.


3α21β
Τρ. Ρε με Α' χρόα (Τον. βάση: Λα)
31α31β
Τρ. Ρε με Α' χρόα (Τον. βάση: Λα)

(a)

(b)

Figure 7.1: Two phrases found to be similar in query (7)



(1)

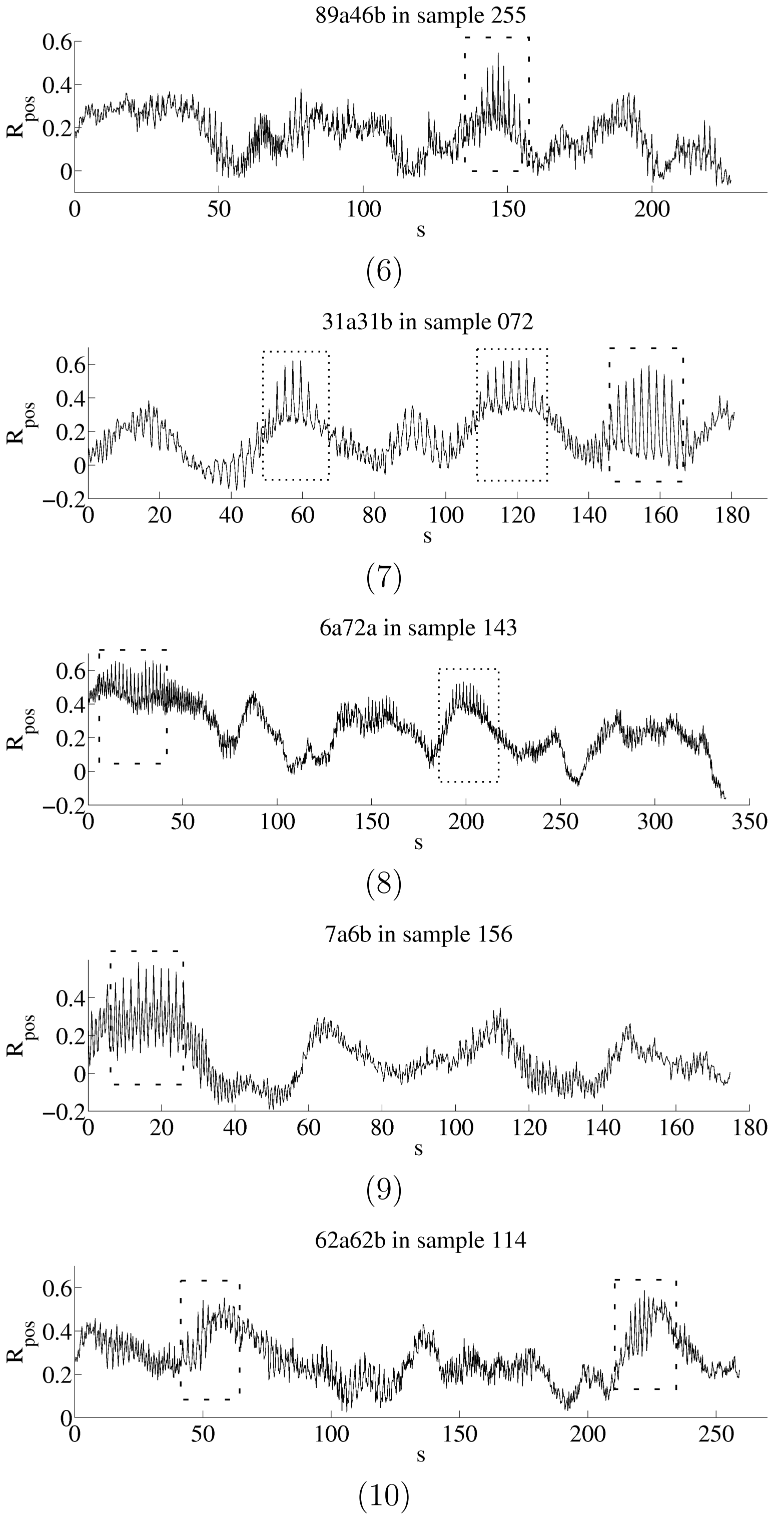
(2)

(3)

(4)

(5)

Figure 7.2: Complete R_{pos} obtained for queries 1-5 in Table 7.2, positive matches in dashed boxes, negative matches in dotted boxes



(6)

(7)

(8)

(9)

(10)

Figure 7.3: Complete R_{pos} obtained for queries 1-5 in Table 7.2, positive matches in dashed boxes, negative matches in dotted boxes

Regarding the integration of the rhythmic similarity measure into this query experiment, the approach of integration appears to be simple. As the features for melody (melody histograms) and for rhythm (scale transform magnitudes) are both computed in a beat synchronous way, the correlation values obtained for a query from these two aspects could be simply added, possibly using some weighting that favors either melody or rhythm derived correlations. However, as mentioned above in Section 7.3.1, the rhythmic similarity measure is sure to fail on this dataset which is rhythmically very homogeneous. Indeed, the computed correlations using the RHYTHM-SYS in the second experimental setup do not reveal the clear structure as the correlations depicted for the HIST-MEL-SYS. Because of that, an integration of these two aspects made no sense on the available data. However, it can be argued that such an integration is very likely to improve results on a more diverse dataset. If, for example, such a dataset would include several dances from Crete or another Greek region, it has been shown in Chapter 4 that the descriptors used in RHYTHM-SYS are suitable to discriminate different dances and to measure the rhythmic similarity between samples. It is apparent that the melody histograms from samples of different dances can be very similar or even identical, while taking the rhythmic aspect into account will help to differentiate between such samples.

Finally, the impact of the beat tracker has been evaluated. In order to determine, how large the change in the matching procedures can be if the beat tracking and hence the synchronization is optimized, all samples in MS1 and all complete samples used for the computation of R_{pos} in Table 7.2 have been beat annotated by the author. However, rerunning all experiments in the experimental setups 1 and 2 using these ground truth beat annotations did not qualitatively change the results. The original beat tracker used in Ellis and Poliner [46] lead mainly to local misalignments with the beat annotation, and it has to be concluded that these misalignments have no impact on the systems used in this chapter, at least when applied to the limited size of data that is currently available.

7.4 Conclusion

In this chapter, methods have been evaluated that help to detect morphological similarity in polyphonic recordings following the logic of *parataxis*. It has been shown that a method based on histograms of the F0 estimation of the leading melody enables for an improvement compared to a baseline system that uses chroma features. Furthermore, it has been illustrated that the proposed method is capable of spotting appearances of small melodic patterns in a whole audio file, even when both files are polyphonic mixtures and the query pattern has been derived from a different recording. Such a method can be a valuable tool for research in the field of musicology, where similar phrases in a large collection could be located without the necessity of transcription, thus leading to a large saving of time. Furthermore, the integration of melodic and rhythmic aspects have been proposed, for datasets in which different types of rhythms are contained. This integration is straight forward and as a future goal it has to be evaluated on a more diverse dataset.

Chapter 8

Conclusions

This thesis had its starting point in the investigations of musical timbre, mainly related to the classification of music regarding its genre. Then the question came up what would be necessary to classify or to find similarities in the traditional music of Greece. After extended reading of musicological studies related to the subject, it was clear that a system for the analysis of this music would have to take into account characteristics distinct from the ones encountered in western popular music. Concerning its morphology, it was clear that a system for the detection of structures in Western music, such as a *refrain*, could not lead to meaningful results on this music. Thus, the more general framework as shown in Figure 1.1 was designed. The research activity then focussed on the part of rhythmic properties of the signal. First, it was decided to take a careful look at the detection of instrument note onsets. On this field, a large onset annotated dataset was compiled. Using this dataset for evaluation, the practical use and the advantages of phase slope and fundamental frequency derived onset strength functions was demonstrated. Especially promising results were achieved when combining spectral flux and the former two characteristics in a decision fusion. An open issue is still the usage of a fundamental frequency criterion for onset detection in polyphonic mixtures.

Datasets have then been developed for the evaluation of beat tracking and rhythmic similarity tasks on traditional Cretan and also Turkish music. The beat tracking task appears to be especially demanding in the traditional music of Crete and other Greek regions. This is because the signals often do not contain percussive instruments or electric bass, which causes weaker attacks in amplitude than for popular music, and because Cretan dances are often characterized by syncopes. It has been shown that in the context of beat tracking, the usage of the phase slope for the computation of an onset strength function leads to an improved accuracy compared to two state of the art approaches. For the task of rhythmic similarity, two approaches have been proposed that aim to solve the problem of comparing the rhythmic content of two pieces of music, when their tempo differs widely. This is necessary, because interpretations of even the same piece can vary widely in tempo, and often the tempo is increased within the duration of a piece. The proposed approaches do not need any beat tracking as pre-processing step, which is too error-prone considering the beat tracking accuracies that were achieved on these type of musical signals. The first approach developed in the course of this thesis was based on a dynamic warping strategy, and the second is based on the usage of the scale transform. It could be shown that in general the scale transform based method leads to superior results in dance music classification tasks. The properties of the scale transform have been explored and a feature selection has been proposed to handle the problem of determining an optimum subset of scale coefficients. Also, for the first time a computational study of rhythmic similarity in traditional Turkish music was conducted.

Returning to the starting point of the thesis work, timbre similarity, a system based on non-negative matrix factorization was proposed for a genre classification task and applied to the classification of instrumental content of traditional music. In this classification task no advantage of the NMF system compared to MFCC features could be observed, while for genre classification improved accuracies were obtained. In order to get a final conclusion in this task, a larger dataset has to be compiled, and methods as the one presented in [67] should be evaluated in order to combine the advantages of NMF and MFCC with the separation of the lead melody in a piece.

Finally, the integration of the rhythmic similarity measure, beat tracking techniques and melodic similarity was proposed in order to estimate the degree of morphological similarity between samples of traditional music following the logic of *parataxis*. Regarding the melodic similarity an improvement was observed when a state-of-the-art system for cover song detection was modified by the usage of a fundamental frequency tracking of the main melody in the polyphonic mixture. Furthermore, a finer partitioning of the frequency axis was proposed in order to cope with the usage of modal scales. This did not result into significant changes on a dataset of Cretan music, but is likely to be important when dealing with music of other regions in Greece, Turkey or other Balkan states. This system was shown to work given a polyphonic mixture as a query. Thus it can be applied in the investigation of large datasets of field recordings in the course of musicological studies in order to simplify the process of locating morphologically related pieces of music. Furthermore, using the melodic similarity in combination with the proposed rhythmic similarity is expected to improve results when different kinds of traditional dances are considered in a dataset.

The possible applications of the developed methods are widely spread. Rhythmic similarity measures and techniques for onset detection can be applied to any kind of music signal without changes. Because the rhythmic similarity measure showed reasonable success on traditional music of Turkey, focusing future research in this direction appears to be a reasonable direction. This is the case because beside the methods proposed in this thesis also methods for the processing of the melodic content were developed recently by a research group that has a cooperation with the author's institution. Such a research direction would be of major commercial interest also for the following reason: music in the Arabic world widely follows similar principles regarding melody and rhythm as Turkish music, and tools could be applied to this music without major changes. This is particularly interesting as Arabic music addresses a much larger group of consumers than Turkish and Greek music together, and no systematic research has been conducted in this field.

In order to make such research work feasible, sufficiently large datasets of the related forms of music have to be collected and annotated. For this, a long term cooperation with experts of musicology is necessary. In such a cooperation, possible applications of computational tools can be defined, thus transferring the findings presented in this thesis into practical use by musicologists.

Interesting development task include the beat tracking for complex meters, a problem which has to be solved in order to automatically determine the temporal organization of many forms of traditional music. Regarding rhythmic similarity it will be interesting to look at possible combinations of the proposed STM based description with other rhythm descriptors such as the Rhythm Patterns, which have been shown to be applicable to traditional music in Lidy *et al.* [102].

Bibliography

[1] O. Abraham and E. M. von Hornbostel. Propositions for the transcription of exotic melodies, (in german language). In *Sonderdruck aus "Sammelbände der Internationalen Musikgesellschaft" XI, 1.* Leipzig, Internationale Musikgesellschaft, n.d., 1909.

[2] G. Amargianakis. Morphology of traditional cretan dance music. In *Proc. of 2nd Conference on Music and Dances of Crete*, 2001.

[3] B. Aning. Tempo change: Dance music interactions in some ghanaian traditions. *Institute of African Studies: Research Review*, 8(2):41–43, 1972.

[4] I. Antonopoulos, A. Pikrakis, S. Theodoridis, O. Cornelis, D. Moelants, and M. Leman. Music retrieval by rhythmic similarity applied on greek and african traditional music. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[5] H. Arel. *Türk Musikisi NazariyatıDersleri.* Husnutabiat Matbaası (reprint: 1968), Istanbul, 1930.

[6] S. Baud-Bovy. A cretan dance, (in german language). *Beitraege zur Ethnomusikologie, Studien zur Musik Suedost-Europas*, 4:175–181, 1976.

[7] S. Baud-Bovy. *An essay on the Greek folk song, (in Greek language).* Laographic Institute of Peleponese, 1984.

[8] S. Baud-Bovy. *Musical field recordings in Crete 1953-1954, (in Greek language).* Center of Minor Asiatic Studies / Musical Laographical Archives Melpos Merlie, 2006.

[9] L. Baum and J. Eagon. An inequality with applications to statistical estimation for probalistic functions of markov processes and to a model for ecology. *American Mathematical Society Bulletin*, 73:360–363, 1967.

[10] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, Sept. 2005.

[11] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Proces. Lett.*, 11(6):553–556, 2004.

[12] J. P. Bello, C. Duxbury, M. E. Davies, and M. B. Sandler. A combined phase and amplitude based approach for onset detection for audio segmentation. In *Proc. of WIAMIS*, pages 6–12, London, UK, 2003.

[13] J. P. Bello and M. Sandler. Phase-based note onset detection for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages 444–447, Hong-Kong, 2003.

[14] E. Benetos, M. Kotti, and C. Kotropoulos. Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, 2006.

[15] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and adaboost for music classification. Technical report, Kluwer Academic Publishers, 2006.

[16] W. Berry. *Structural functions in music.* Dover, New York, 1987.

[17] J. A. Bilmes. Timing is of the essence. Master's thesis, Massachusetts Institute Of Technology, 1993.

[18] C. Bohak and M. Marolt. Calculating similarity of folk song variants with melody-based features. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 597–601, 2009.

[19] B. Bozkurt. An automatic pitch analysis method for turkish maqam music. *Journal of New Music Research*, 37(1):1–13, 2008.

[20] B. Bozkurt, L. Couvreur, and T. Dutoit. Chirp group delay analysis of speech signals. *Speech Communication*, 49(3):159–176, 2007.

[21] A. S. Bregman. *Auditory Scene Analysis.* MIT Press, 1990.

[22] M. Brookes, P. A. Naylor, and J. Gudnason. A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14:456– 466, 2006.

[23] J. J. Cabrera, J. M. D.-B. nez, F. J. Escobar-Borrego, E. Gómez, F. Gómez, and J. Mora. Comparative melodic analysis of a cappella flamenco cantes. In *Proceedings of the fourth Conference on Interdisciplinary Musicology (CIM08)*, 2008.

[24] M. Casey. General sound classification and similarity in mpeg-7. *Organized Sound*, 6(2):153–164, 2001.

[25] G. Chatzidakis. Cretan music and dance. *Kritiki Stoa*, page 273ff, 1909.

[26] Y. C. Cho, S. Choi, and S. Y. Bong. Non-negative component parts of sound for classification. In *Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2003.

[27] L. Cohen. The scale representation. *IEEE Transactions on Signal Processing*, 41(12):3275–3292, 1993.

[28] N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proceedings of the 118th AES Convention*, pages 28–31, Barcelona, Spain, 2005.

[29] N. Collins. Using a pitch detector for onset detection. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 100–106, Barcelona, Spain, 2005.

[30] F. Combet, P. Jaussaud, and N. Martin. Estimation of slight speed gaps between signals via the scale transform. *Mechanical Systems and Signal Processing*, 19:239–257, 2005.

[31] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[32] G. Cooper and L. B. Meyer. *The rhythmic structure of music.* University of Chicago Press, 1960.

[33] L. Daudet, G. Richard, and P. Leveau. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 72–77, 2004.

[34] M. E. P. Davies and M. D. Plumbley. Tempo estimation and beat tracking with adaptive input selection. In *MIREX at 7th International ISMIR 2006 Conference*, 2006.

[35] M. E. P. Davies and M. D. Plumbley. Context-dependent beat tracking of musical audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1009–1020, March 2007.

[36] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-28(4):357–366, August 1980.

[37] M. Daz-Banez, G. Farigu, F. Gomez, D. Rappaport, and G. Toussaint. El compas flamenco: a phylogenetic analysis, 2004.

[38] A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[39] N. Degara-Quintela, A. Pena, and S. Torres-Guijarro. A comparison of score-level fusion rules for onset detection in music signals. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 117–121, 2009.

[40] S. Dixon. Mirex 2006 audio beat tracking evaluation: Beatroot. In *MIREX at 7th International ISMIR 2006 Conference*, 2006.

[41] S. Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects*, pages 133–137, 2006.

[42] S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2004.

[43] C. Duxbury, M. Sandler, and M. Davies. A hybrid approach to musical note onset detection. In *Proc. of the Int. Conference on Digital Audio Effects (DAFx)*, pages 33–38, Hamburg, Germany, 2002.

[44] T. Eerola and P. Toiviainen. *MIDI Toolbox: MATLAB Tools for Music Research.* University of Jyväskylä, www.jyu.fi/musica/miditoolbox/, Jyväskylä, Finland, 2004.

[45] T. Eerola and P. Toiviainen. *MIDI Toolbox: MATLAB Tools for Music Research.* University of Jyväskylä, Jyväskylä, Finland, 2004.

[46] D. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages IV–1429–1432, 2007.

[47] D. P. W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[48] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. IEEE Int. Conf. Multimedia and Expo (ICME2000)*, pages 452–455, New York, USA, 2000.

[49] J. Foote, M. D. Cooper, and U. Nam. Audio retrieval by rhythmic similarity. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 265–266, 2002.

[50] F. Fuhrmann, M. Haro, and P. Herrera. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2009.

[51] M. Gainza. Single note ornamentation transcription for the irish tin whistle based on onset detection. In *Proc. of the Int. Conference on Digital Audio Effects (DAFx)*, Naples, Italy, 2004.

[52] M. Gainza, E. Coyle, and B. Lawyor. Onset detection using comb filters. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 263–266, New York, USA, 2005.

[53] A. C. Gedik and B. Bozkurt. Automatic classification of turkish traditional art music recordings by Arel theory. In *Proc. of CIM08, 4th Conference on Interdisciplinary Musicology*, Thessaloniki, Greece, 2008.

[54] E. Gómez and J. Bonada. Automatic melodic transcription of flamenco singing. In *Proceedings of the fourth Conference on Interdisciplinary Musicology (CIM08)*, 2008.

[55] M. Goto and Y. Muraoka. Music understanding at the beat level: Real-time beat tracking for audio signals. In *Proceedings of IJCAI 95 Workshop on Computational Auditory Scene Analysis*, pages 68–75, 1995.

[56] M. Goto and Y. Muraoka. Beat tracking based on multiple-agent architecture - a real-time beat tracking system for audio signals. In *Proc. 2nd Int. Conf. Multiagent Systems*, pages 103–110, Kyoto, Japan, 1996.

[57] M. Goto and Y. Muraoka. Issues in evaluating beat tracking systems. In *Proceedings of IJCAI-97 Workshop on Issues in AI and Music*, pages 9–16, 1997.

[58] F. Gouyon and S. Dixon. Dance music classification: A tempo based approach. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 501–507, 2004.

[59] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre clasification. In *AES 25th International Conference*, 2004.

[60] F. Gouyon and P. Herrera. Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors. In *114th Convention of the Audio Engineering Society*, 2003.

[61] M. Grachten, J. L. Arcos, and R. L. de Mántaras. Melody retrieval using the implication/realization model. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2005.

[62] S. Hainsworth and M. Macleod. Onset detection in musical audio signals. In *Proc. Int. Computer Music Conference (ICMC)*, pages 163–166, Singapore, 2003.

[63] P. Hamel, S. Wood, and D. Eck. Automatic identification of instrument classes in polyphonic and poly-instrument audio. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2009.

[64] G. Hamerly and C. Elkan. Learning the k in kmeans. Technical report, Department of Computer Science and Engineering, University of California, San Diego, 2003.

[65] S. Haykin. *Adaptive Filter Theory, Fourth edition.* Prentice Hall, 2002.

[66] R. M. Hegde, H. A. Murthy, and V. R. Gadde. The modified group delay feature: A new spectral representation of speech. In *Proc. of the Interspeech-ICSLP*, pages 913–916, Korea, 2004.

[67] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2009.

[68] G. Holst-Warhaft. *Road to Rembetika: Music of a Greek sub-culture, songs of love, sorrow and hashish.* Athens, Denise Harvey, 1989.

[69] A. Holzapfel. A component based music classification approach. Master's thesis, University of Crete, Greece, 2006.

[70] A. Holzapfel and Y. Stylianou. Singer identification in rembetiko music. In *Proc. of SMC 2007, Conference on Sound and Music Computing*, Lefkada, Greece, 2007.

[71] A. Holzapfel and Y. Stylianou. Beat tracking using group delay based onset detection. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 653–658, 2008.

[72] A. Holzapfel and Y. Stylianou. Musical genre classification using non-negative matrix factorization based features. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):424–434, 2008.

[73] A. Holzapfel and Y. Stylianou. Rhythmic similarity of music based on dynamic periodicity warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages 2217–2220, 2008.

[74] A. Holzapfel and Y. Stylianou. Rhythmic similarity in traditional turkish music. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2009.

[75] A. Holzapfel and Y. Stylianou. A scale transform based method for rhythmic similarity of music. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages 317–320, 2009.

[76] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt. Three dimensions of pitched instrument onset detection. *Accepted for publication in IEEE Transactions on Audio, Speech and Language Processing*, 2009.

[77] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[78] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[79] T. Irino and R. D. Patterson. Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform. *Speech Commun.*, 36(3):181–203, 2002.

[80] ISMIR2004. Audio description contest - rhythm classification, 5th international conference on music information retrieval (ismir). http://mtg.upf.edu/ismir2004/contest/rhythmContest/.

[81] J. H. Jensen, M. G. Christensen, D. P. W. Ellis, and S. H. Jensen. A tempo-insensitive distance measure for cover song identification based on chroma features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, pages 2209–2212, 2008.

[82] Z. Juhász. Motive identification in 22 folksong corpora using dynamic time warping and self organizing maps. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 171–176, 2009.

[83] V. Kandia and Y. Stylianou. Detection of clicks based on group delay. *Canadian Acoustics*, 36(1):48–54, 2008.

[84] M. K. Karaosmanoğlu, S. M. Yılmaz, O. Tören, S. Ceran, U. Uzmen, G. Cihan, and E. Başaran. *Mus2okur*. Data-Soft Ltd., http://www.musiki.org/, Turkey, 2008.

[85] A. Kelleher, D. Fitzgerald, M. Gainza, E. Coyle, and B. Lawlor. Onset detection, music transcription and ornament detection for the traditional irish fiddle. In *Proceedings of the 118th AES Convention*, Barcelona, Spain, 2005.

[86] H.-G. Kim, N. Moreau, and T. Sikora. Audio classification based on mpeg-7 spectral basis representations. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):716–725, 2004.

[87] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9(1), 2001.

[88] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages 3089–3092, Washington, DC, USA, 1999.

[89] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 216–221, 2006.

[90] A. Klapuri. A method for visualizing the pitch content of polyphonic music signals. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 615–620, 2009.

[91] A. P. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269–282, 2004.

[92] A. P. Klapuri, A. J. Eronen, and J. T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Acoustics Speech and Signal Processing*, 14(1):342–355, 2006.

[93] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[94] A. Lacoste and D. Eck. A supervised classification algorithm for note onset detection. *EURASIP Journal on Applied Signal Processing*, 2007:153–165, 2007.

[95] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[96] F. Lerdahl and R. Jackendoff. *A generative theory of tonal music*. MIT Press Cambridge, 1983.

[97] T. Li and M. Ogihara. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3):564–574, 2006.

[98] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. 26th ACM SIGIR Conference*, 2003.

[99] T. Li and G. Tzanetakis. Factors in automatic musical genre classification of audio signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[100] T. Lidy. Evaluation of new audio features and their utilization in novel music retrieval applications. Master's thesis, Vienna University of Technology, Austria, 2006.

[101] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2005.

[102] T. Lidy, C. N. Silla, O. Cornelis, F. Gouyon, A. Rauber, C. A. A. Kaestner, and A. L. Koerich. On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-western and ethnic music collections. *Signal Processing*, 90(4).

[103] C. C. Liem and A. Hanjalic. Cover song retrieval: A comparative study of system component choices. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2009.

[104] G. List. The reliability of transcription. *Ethnomusicology*, 18(3):353–377, 1974.

[105] I. Loutzaki. Audio report: Greek folk dance music. *Yearbook for traditional music*, 26:168–179, 1994.

[106] M. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 594–599, 2005.

[107] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7.* John Wiley & sons, 2002.

[108] M. Markaki and Y. Stylianou. Dimensionality reduction of modulation frequency features for speech discrimination. In *Proceedings of InterSpeech*, 2008.

[109] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange. Polyphonic instrument recognition using spectral clustering. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2007.

[110] MIREX07. Audio onset detection. http://www.music-ir.org/mirex/2007/index.php/Audio_Onset_Detection.

[111] D. Moelants, O. Cornelis, and M. Leman. Exploring african tone scales. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 489–494, 2009.

[112] B. C. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition.* Academic Press, April 2003.

[113] E. P. Morris. *On Principles and Methods in Latin Syntax.* New York, C. Scribner's sons, 1901.

[114] M.R.Schroeder, B.S.Atal, and J.L.Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66(6):1647–1652, 1979.

[115] M. Müller, P. Grosche, and F. Wiering. Robust segmentation and annotation of folk song recordings. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 735–740, 2009.

[116] H. A. Murthy and B. Yegnanarayana. Speech processing using group delay functions. *Signal Processing*, 22(3):259–267, 1991.

[117] E. Narmour. *The Analysis and cognition of basic melodic structures : the implication-realization model.* University of Chicago Press, 1990.

[118] A. Oppenheim, R. Schafer, and J. Buck. *Discrete-Time Signal Processing.* Prentice Hall, 1998.

[119] F. Pachet and J.-J. Aucouturier. Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 1(1), 2004.

[120] E. Pampalk. A matlab toolbox to compute music similarity from audio. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2004.

[121] E. Pampalk. *Computational Models of Music Similarity and their Application to Music Information Retrieval.* PhD thesis, Vienna University of Technology, Austria, 2006.

[122] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2005.

[123] R. Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4):409–464, 1994.

[124] J. Paulus and A. Klapuri. Measuring the similarity of rhythmic patterns. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 150–156, 2002.

[125] G. Peeters. Rhythm classification using spectral rhythm patterns. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 644–647, 2005.

[126] D. Perrott and R. Gjerdingen. Scanning the dial. In *Society for Music Perception and Cognition Conference*, 1999.

[127] E. Petropoulos. *Rebetika Tragoudia, (in Greek language).* Athens, Kedros, 1983.

[128] A. D. Poularikas. *The Handbook of Formulas and Tables for Signal Processing.* CRC Press LLC, 1999.

[129] T. F. Quatieri. *Discrete-Time Speech Signal Processing.* Prentice Hall Signal Processing Series, 2002.

[130] L. R. Rabiner and B. Juang. *Fundamentals of speech recognition.* Prentice-Hall, 1993.

[131] G. F. B. Riemann. Ueber die anzahl der primzahlen unter einer gegebenen groesse. *Monatsber. Koenigl. Preuss. Akad. Wiss. Berlin*, pages 671–680, November 1859.

[132] H. Riemann. *Musik-lexikon, (in German language).* Leipzig, 1882.

[133] T. B. s. Relationships between prosodic and musical meters in the beste form of classical turkish music. *Asian Music*, 36(1), Winter/Spring 2005.

[134] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama. Specmurt analysis of polyphonic music signals. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):639–650, 2008.

[135] H. Sarris, T. Kolydas, and P. Tzevelekos. A framework of structure analysis for instrumental folk music. In *Proc. of CIM08, 4th Conference on Interdisciplinary Musicology*, Thessaloniki, Greece, 2008.

[136] E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *J. Acoust. Soc. Am.*, 103(1):588–601, 1998.

[137] A. W. Schloss. *On the Automatic Transcription of Percussive Music From Acoustic Signal to High-Level Analysis.* PhD thesis, Dept. Hearing and Speech, Stanford Univ., Stanford, CA, 1985.

[138] A. D. Sena and D. Rocchesso. A fast Mellin transform with applications in dafx. In *Proceedings of the 7th International Conference on Audio Effects (DAFx'04)*, pages 65–69, 2004.

[139] A. D. Sena and D. Rocchesso. A fast Mellin and scale transform. *EURASIP J. Appl. Signal Process.*, 2007(1):75–84, 2007.

[140] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, 2008.

[141] X. Serra. *A System for Sound Analysis/ Transformation/ Synthesis based on a Deterministic plus Stochastic Decomposition.* PhD thesis, 1989.

[142] R. N. Shepard. Circularity in judgements of relative pitch. *Journal of the Acoustical Society of America*, 36:2346–2353, 1964.

[143] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Proc. 5th International Conference on Independent Component Analysis and Blind Signal Separation*, 2004.

[144] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, 3(5):325–333, 1995.

[145] K. Steiglitz and B. Dickinson. Phase unwrapping by factorization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(6):723–726, 1982.

[146] D. Stockmann. Transcription in ethnomusicology: history, problems, methods (in german language). *Acta Musicologica*, 51(2):204–245, 1979.

[147] E. Terhardt. Calculating vitual pitch. *Hearing Research*, 1:155–182, 1979.

[148] D. Themelis. *Morphology and analysis of music, (in Greek language).* University Studio Press, Thessaloniki, 1994.

[149] I. B. Theodosopoulou. *Methodology of morphological analysis and analytic data of small rhythmic patterns of cretan folk music, (in Greek Language).* Athens: Kultura, 2004.

[150] M. T. Thomassen. Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America*, 71:1596–1605, 1982.

[151] W. F. Thompson and M. Stainton. Expectancy in bohemian folk song melodies: Evaluation of implicative principles for implicative and closural intervals. *Music Perception*, 15(3):231–252, 1998.

[152] C. C. Toh, B. Zhang, and Y. Wang. Multiple-feature fusion based onset detection for solo singing voice. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 515–520, Philadelphia, USA, 2008.

[153] P. Toiviainen and T. Eerola. Method for comparative analysis of folk music based on musical feature extraction and neural networks. In *In III International Conference on Cognitive Musicology*, pages 41–45, 2001.

[154] P. Toiviainen and T. Eerola. Autocorrelation in meter induction: The role of accent structure. *Journal of the Acoustical Society of America*, 119(2):1164–1170, 2006.

[155] G. T. Toussaint. A comparison of rhythmic similarity measures. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, 2004.

[156] I. Tsouchlarakis. *The dances of Crete - Legend, History, Tradition (in Greek language).* Center of Cretan Culture Studies, Athens, 2000.

[157] R. Typke. *Music Retrieval based on Melodic Similarity.* PhD thesis, 2007.

[158] C. Uhle and J. Herre. Estimation of tempo, micro time and time signature from percussive music. In *Proc. of the Int. Conference on Digital Audio Effects (DAFx)*, 2003.

[159] S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson. Scale transform in speech analysis. *IEEE Transactions on Speech and Audio Processing*, 7(1):40–46, 1999.

[160] P. van der Merwe. *Origins of the Popular Style: The Antecedents of Twentieth-Century Popular Music.* Oxford: Clarendon Press, 1989.

[161] P. van Kranenburg, A. Volk, F. Wiering, and R. C. Veltkamp. Musical models for folk-song melody alignment. In *Proc. of ISMIR - International Conference on Music Information Retrieval*, pages 507–512, 2009.

[162] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. Technical report, Queen Mary, University of London, 2005.

[163] W. Williams and E. Zalubas. Helicopter transmission fault detection via time-frequency, scale and spectral methods. *Mechanical systems and signal processing*, 14(4):545–559, July 2000.

[164] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[165] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(3):610–623, 1984.

[166] R. Zhou and J. D. Reiss. Music onset detection combining energy-based and pitch-based approaches. In *MIREX Onset Detection Contest in ISMIR 2007 - 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[167] E. Zwicker and H. Fastl. *Psychoacoustics - Facts and Models.* Springer, 1990.

Appendix A

Datasets

The collection of datasets is an important and necessary starting point for any experimental work. As the field of computational ethnomusicology is relatively new, there do not exist very many data sources that can be used for experiments. For that reason, in the course of my thesis work I was forced to compile a big number of data sets. While on the one hand this work is very time consuming, on the other hand it enabled me to understand the musical context I was about to examine in further detail. These datasets can provide facilitation for other researchers concerned with the same or similar subjects. Thus, in this chapter the compiled datasets are described in detail and researchers are invited to contact me if they are interested in obtaining a specific dataset.

A.1 Singer recognition data

This dataset consists of 290 songs from 21 Rembetiko singers. Rembetiko as a musical style can not be considered a traditional form of music in the strict sense. It has its roots mainly in the area of Piraeus, but also in other cities, where in the beginning of the last century existing urban musical underground culture came together with the music of refugees from Asia minor. While most songs can be assigned to specific composers, there are also songs which stem from older traditional melodies. For more details on the background of this music refer to [68, 127]. All songs contained in this dataset are polyphonic mixtures that contain, beside one or more singing voices, musical instruments that are typical for this type of music. These instruments are guitar, *bouzouki* (a plucked string instrument), *baglama* (plucked string instrument similar to bouzouki, but much smaller), and sometimes accordion or violin. Thus, the whole dataset is very homogeneous in terms of instrumental timbre. It has been used by the author in [70] for the purpose of singer recognition. The number of songs per singer ranges from eight to 18. Details of the data set are depicted in Table A.1. The numbers for musical activity list the decades in which the artist recorded music. It was tried to cover a wide range of this period with the contained pieces of music. Because of that, for some singers, as Sotiria Bellou, the singer's voice varies strongly. Note that the artist Xarmas represents a male/female duo, that throughout the given period performed together.
From each singer four songs have been hand labelled with the following labels:

- INSTR : instrumental sounds without any voice

Table A.1: Data set description

Singer	male/female	activity	songs	ID
Agathonas	m	70-now	11	S1
Batis	m	30	13	S2
Bellou	f	40-80	18	S3
Dalkas	m	30-50	14	S4
Delias	m	30-40	8	S5
Genitzaris	m	40-90	9	S6
Gkoles	m	70-now	11	S7
Glykeria	f	70-now	12	S8
Marika (Papangika)	f	20-30	18	S9
Mario	f	70-now	13	S10
Markos Bambakaris	m	30-60	15	S11
Menidiatis	m	60-now	17	S12
Nikolaidis	m	60-now	11	S13
Rita Ampatzi	f	30-50	13	S14
Roukounas	m	30-50	14	S15
Roza Eskenazi	f	30-60	15	S16
Stellakis Perpiniadis	m	30-60	18	S17
Stratos Pagiumtzis	m	30-60	16	S18
Tsaousakis	m	50-70	16	S19
Tsitsanis	m	30-70	13	S20
Xarmas	m+f	40-50	12	S21

- VOICE : voice of target singer without second voice
- MIXED : voice of target singer with second voice
- OTHER : interjections

For singer S21 all vocal frames have been labelled as VOICE, as we want to recognize this particular singer duo.
Another peculiarity of the data set is that some of the artists take part in the others' recordings. As such the artists Markos Bambakaris, Anestis Delias, Stratos Pagioumtzis, Giorgos Batis and Stellakis Perpiniadis formed a group for many years. Because of that, in many songs of the target singer, another singer, who is part of the data set, is featured as second singer. The same holds for Vasilis Tsitsanis, who wrote many songs for Bellou and Tsaousakis, and sings the second voice in some songs of Bellou. Similar relations exist for the currently performing artists Gkoles, Glykeria and Agathonas.

A.2 Onset detection data

An onset-annotated dataset of monophonic recordings as described in Table A.2 has been compiled for the evaluation of onset detection systems. Non-pitched percussive instruments, such as drums and percussions, have not been included in this dataset. The instruments contained in the dataset can be grouped according to the characteristics of their excitation:

- pitched-percussive instruments (guitar, *ud*, piano and *tanbur*)
- wind instruments (clarinet, *ney*, saxophone and trumpet)
- bowed string instruments (cello, *kemençe* and violin)

Table A.2: Main dataset details (1)

Main Set (MS)		
Instrument	Number of Onsets	Number of files
cello	150	5
clarinet	149	5
guitar	174	5
kemençe	186	5
ney	147	7
ud	211	5
piano	195	5
saxophone	148	5
tanbur	156	5
trumpet	140	5
violin	173	5
Sum	1829	57

Table A.3: Development dataset details (2)

Development Set (DS)		
Instrument	Number of Onsets	Number of files
guitar	147	7
ud	207	5
piano	117	6
violin	203	3
Sum	674	21

Effort has been made, such that each of the above classes is represented by a similar number of samples and instruments. Furthermore, besides the choice of instruments commonly used in western music, also instruments of Turkish music are included (*kemençe*, *ney*, *ud* and *tanbur*). The Turkish music examples were chosen in order to select samples that are representative for the style of performance but that do not contain many notes at which hand annotation would have been too error-prone. This restriction has been found to be necessary due to the style of performance encountered in this music, which at some point complicates the differentiation between onsets and vibrato or other effects. For annotating new samples the procedure described in Daudet *et al.* [33] was adapted: the author of this thesis and the third author of [76] did the annotations, while the fourth author of [76] corrected the results. Correcting the annotations means that it was only possible to delete annotations, and not to add new annotations. Each change in the correction process, except of a deletion, had to be discussed with the annotator. In this way cross-checked annotations were compiled for all the dataset. For the annotation the wavesurfer[1] software was used. Spectrogram, waveform and the F0 curves were used simultaneously to locate the onsets that were perceived in the sample.

Beside the data as presented in Table A.2, 21 more samples of the instruments guitar, *ud*, piano and violin were onset annotated. These files were used for our parameter evaluations and development in [76], and contain 674 onsets, see Table A.3 for details. In the overall number of 78 samples that are contained in the main dataset and in the development set, 8 samples from the dataset used in Daudet [33] are included, which contained an instrument listed in Table A.2 (one file for cello, clarinet, piano, saxophone, trumpet and violin, and two files for guitar).

A.3 Cretan music

A.3.1 General information

According to [25], Cretan songs can be separated into categories. First, there are songs *tis tavlas*, which are sung with company while sitting at a table with company, and they are characterized by a dialogue, in which one singer presents a new verse and the other people round the table answer in unison. Second, there exist historical and heroic songs. They are often sung using a free meter which has been characterized as *giusto syllabique dichrone* by Baud-Bovy in [8]. Third, there are the dances, which will also be the main focus due to their wide variety and availability in form of recordings. There exist more categories related to particular events such as weddings or funerals as well. While the second category of pieces does not follow the logic of *parataxis* (see Section 1.1), the first and the third class do. Note that while in this thesis only Cretan dances are considered, traditional dances in other parts of Greece in general follow this logic as well [6].

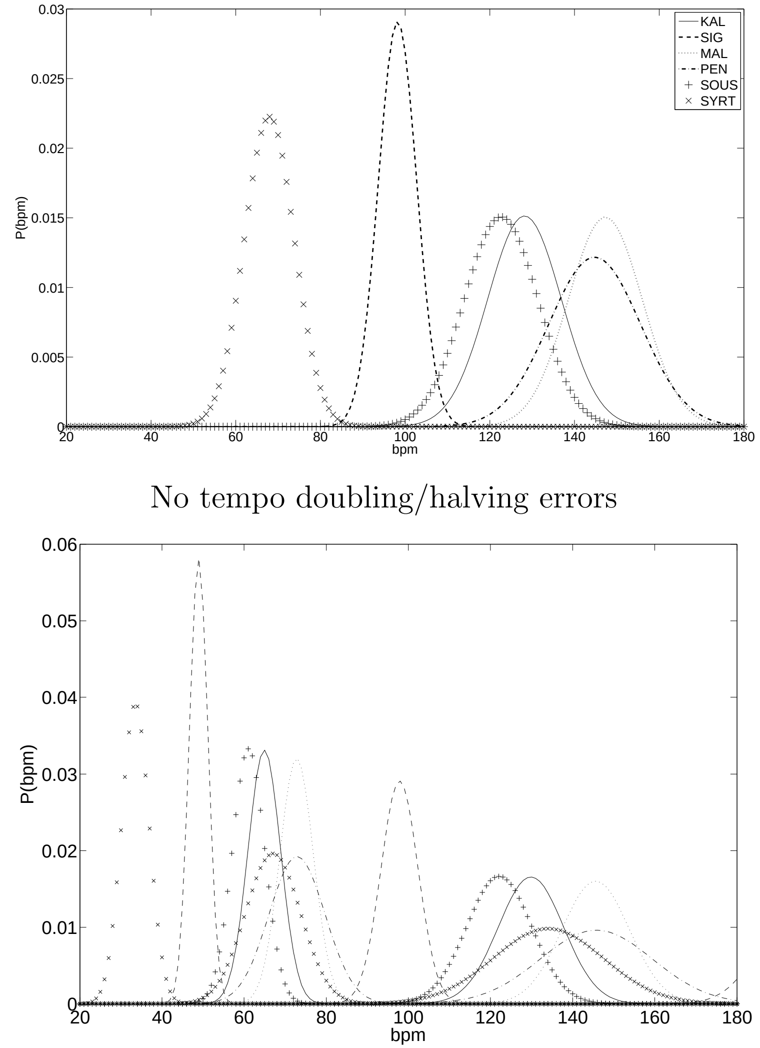
With the development of recording technology and the production costs reaching lower and lower levels, an increasing number of Cretan musicians release their interpretations of traditional pieces or own compositions having a traditional appeal. The first musicians to release their disks in the middle of the last century are the ones who are considered the masters and their playing style represents a paradigm for many players today. However, a closer look at related literature reveals that Cretan music, just as all musical traditions, is much more characterized by local particularities that have been caused by the social life in widely isolated small communities. Some examples are different styles to play the lyra, or different attitudes in singing style [8]: while in a mountain region of central Crete *lyra* players were reported to use mainly the lowest string, in Eastern Crete Baud-Bovy observed the players to use all strings.

[1] http://www.speech.kth.se/wavesurfer/

He also observes that people in Western Crete considered a person a good singer when his voice was good, while in Eastern Crete more emphasis was given to his abilities in improvisation. Furthermore, while nowadays there mainly exists a set of 6 dances that are encountered in the whole island of Crete, some decades ago each region had its characteristic dances. This fact still influences the way, people from different regions tend to interpret music either as a musician or as a dancer. In one interview a musician uttered the opinion that the music of Eastern Crete is “sweeter”, while in the Chania area in Western Crete, the same dances are in general played much faster. However, the related literature is very sparse, and currently, to the best of my knowledge, there is no institution in Crete that performs research activity that could give support in such questions. Thus a computational approach has to work with data that simplify the nature of the task, for example by categorizing dances into a specific set of classes. These classes necessarily simplify the problem but represent a first step to approach the problem of similarity in this kind of music.

A.3.2 Dance classification data

This dataset contains short excerpts of six dances commonly encountered in the island of Crete: *Kalamatianos*, *Siganos*, *Maleviziotis*, (fast) *Pentozalis*, *Sousta* and *Syrtos*. It should be pointed out that each of these classes could be divided into subclasses, because the dances differ within the same class due to different local origin. For example, in Chatzidakis [25] it is stated that according to instrument players there are about 120 different musical phrases for the dance *Maleviziotis* only. However, a study that considers all these variations cannot be conducted because the number of samples is small and ground truth on the local differences can only be obtained in the framework of a musicological study. The dance classes in this dataset can be divided into three groups according to Chatzidakis [25]: *Maleviziotis, Pentozalis* and *Sousta* are referred to as *pidichti* (*i.e.* dances connected with bouncing movements), the slower *Siganos* and *Syrtos* are *syrti* dances (syrti are connected with slow movements), and finally the *Kalamatianos* which is a dance commonly encountered in Crete but with its origin in another part of Greece. Care has to be taken about the local differences in naming these dances. The naming chosen by the author is following habits in the area of the capital of Crete, Heraklion. In other regions of Crete the same or similar dances are referred to with different names, but to the best knowledge of the author no study has been conducted to compare the different namings yet. Note that confusion is an indicator of a still vital musical tradition, as these local differences stem from habits in the regional communities. Each class contains 30 samples of about 10s length each, resulting in an amount of 180 samples in the whole dataset. In Figure A.1, the tempo annotations conducted by the authors have been modelled by Gaussians. The rate of fourth notes has been taken as tempo. Comparing with Figure 1 in Gouyon and Dixon [58], it can be seen that there are larger overlaps between their tempo distributions as for the ballroom dance data used *e.g.* in Gouyon and Dixon [58] and in Chapter 4 in this thesis. When considering that the dance *Syrtos* is often transcribed in notes of double length, this overlap gets even larger with the tempo distribution of *Syrtos* moving from the left part of the Figure to the right, creating a distribution overlapping with all dances except of *Siganos*. Furthermore, all traditional Cretan dances have a $\frac{2}{4}$ meter, only *Kalamatianos* as a dance originating from a different part of Greece has a $\frac{7}{8}$ meter. This makes the separation by considering their meter impossible as well. Also, most of the pieces contain only two kinds of string instruments, while percussive instruments are not contained in most samples, creating a very homogeneous data set considering instrumental timbre. The contained instruments are mainly the Cretan *Lyra* and the Cretan *laouto*. The first is a three-string bowed instrument used for playing the main melody. The latter is a plucked string instrument having four double strings. Its timbre and appearance bear similarity to an *ud*, but it has a brighter tone and it has movable frets. As has been observed in Baud-Bovy [8], these two instruments are the most widely used instruments in Cretan music, and the *laouto* replaced the usage of historical percussive instruments like the *davul* (in Crete referred to as *daouli*) and small bells placed on the bow of the lyra player (*gerakokoudouna*). However, it is observed that throughout the last years more players tend to use these and other instruments that had almost dissapeared from musical practice.



No tempo doubling/halving errors

Assuming tempo halving/doubling errors

Figure A.1: Tempi of the Cretan dance dataset modelled by Gaussian distributions

From a musicological point of view, the length of the excerpts, which is about 10s in average, is sufficient to classify the pieces. As detailed in Theodosopoulou [149], Cretan dances are characterized by small melodic phrases, which in the local context are referred to as *Kontilies*, which extend usually over two or four bars. As such, in the given duration at least one such element will be contained. Thus, the samples should be sufficient both for human listeners and for a computational approach, to detect present similarities in their rhythm.

A.3.3 Beat Tracking data

For the evaluation of beat tracking systems a total sum of 69 polyphonic music samples of 30 seconds length each have been beat annotated by the author. Out of these samples, 41 samples have been downloaded from Crinno data of the Institute of Mediterranean Studies[2], and contain Cretan dance music. Various dances are contained in the dataset, the pieces have been randomly chosen from the collection. None of these 41 songs contain percussive instruments, but only string instruments and vocals. Note that these samples share the musical characteristics with the data described in Sections A.3.2 and A.3.4.

The other 28 samples that have been beat annotated have been taken from the dataset TS1 (see Section 6.3.3). From TS1, the genres *classic, country* and *jazz* have been left out, and from the remaining classes the first four samples of each class have been chosen.

The beat annotation has been performed using headphones and the *audacity* software. All annotations have been acoustically checked by listening to the music on one stereo channel and to a click signal synthesized from the annotations on the other stereo channel.

A.3.4 Morphological similarity data

In the course of this thesis a small dataset of polyphonic samples has been collected that enables for a preliminary evaluation of a system for the detection of morphological similarity. For this, samples from the Crinno dataset have been used, just like for the beat tracking data in Section A.3.3. In the Crinno collection for some samples of the dance *Sousta* the lead melodies have been transcribed by musicologists and then analyzed for their morphology. All encountered phrases have been indexed, and using the list of these indexes it is feasible to locate the morphologically identical phrases in different pieces. The way to index the phrases follows the method described in Theodosopoulou [149]: the phrases have a length of either one or two bars as shown in Figures A.2 and A.3. When beginning the analysis, the first encountered two bar phrase will be titled `1a1b`. If for example the next encountered two bar phrase contains the second part of the first phrase in its second measure, while its first measure is an unknown phrase it will be titled `2a1b`, denoting the partial relation with the first pattern. In Figures A.2 and A.3 the titles of the depicted melodic phrases are denoted above the score. It is obvious that an exact partial or complete matching can be localized by using this way of indexing the phrases. However, no conclusions can be drawn about the similarity of phrases with different titles. As the amount of transcribed data is rather small (20 pieces), there are not many phrases that appear several times in various pieces. However, it has been achieved to compile a data set of 40 sound samples, each containing a complex musical mixture signal with the instruments Cretan *laouto* and *lyra* and sometimes singing voice. Each of the 40 pieces has

[2] http://gaia.ims.forth.gr/portal/

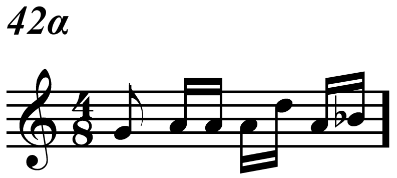


Figure A.2: Example of a one measure melodic phrase

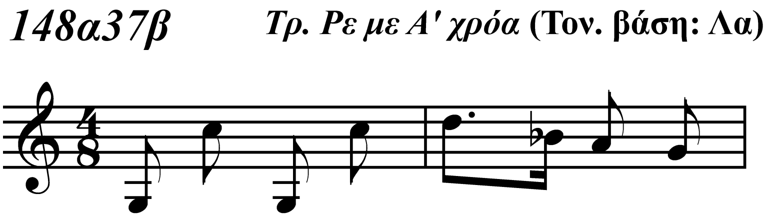


Figure A.3: Example of a two measure melodic phrase

a “partner” within the dataset that contains a similar or equal musical phrase played by the *lyra*, according to the analysis of musicologists. Thus, in this dataset exist 20 pairs of samples that contain similar phrases. Please note that according to the musicological analysis these phrases are exactly the same. However, the audio files differ because they are performed by different artists and vary due to their different playing style.

A.4 Turkish traditional music data

Compositions in Turkish traditional music follow certain schemes regarding their melodic and rhythmic content. Melodies are characterized by a modal system referred to as *makam*, and it defines a melodic texture consisting of specific tonal segments, progressions, directionality, temporal stops, tonal centers and cadences [84]. The rhythmic schemes encountered in traditional Turkish music are referred to as *usul*. An *usul* is a rhythmic pattern of certain length that defines a sequence of strong and weak intonations. An example is shown in Figure A.4: the *usul Aksak* has a length of nine beats. The notes on the upper line labelled *düm* have the

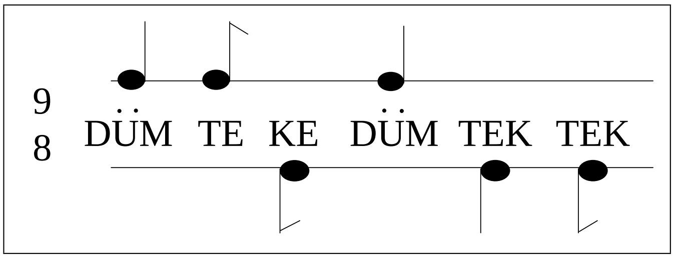


Figure A.4: Symbolic description of the *usul Aksak*

strongest intonation while the notes on the low line denote weak intonations. In vocal music practice, a student strikes his knees with his hands, the right hand for notes on the upper line and the left hand for the notes on the lower line. When acquiring an *usul*, in parallel to the hand strokes the verbal phrases denoted in Figure A.4 are pronounced, until the rhythmic flow of the pattern is well understood by the student. As soon as the student has memorized this movements (s)he starts singing the melody line of a song composed in this *usul* while (s)he continues striking the pattern. The note durations in the sequence shown in Figure A.4 can be described as the string `xoxxxoxox`, where `x` symbolizes the start of a note and `o` a metric unit without note [155].

Unlike in Toussaint [155], the length of the *usul* varies. According to H. Sadeddin Arel (1880-1955), the *usul* can be divided into minor and major *usul*. Minor *usul* have a length of up to 15 time units, while the major *usul* have up to 124 time units. As denoted in Bektaş [133], minor usul are related to small musical forms, while larger musical forms employ the major usul in most cases. Two examples of small musical forms are *Sarkı* and *Türkü*. The latter are folk songs of unknown composers, while the former are short songs based usually on four lines of text with known composer. Note that in a strict sense, *Sarkı* are elements of classical ottoman music, which share rhythmic (*usul*) and tonal (*makam*) concepts with Turkish traditional music to a great extent. Both forms have in common that a song follows a certain minor *usul* and a certain *makam*, and both forms are vocal music. The most popular songs in Turkish music are composed in these forms. Because of that, along with a system for the recognition of the *makam* as presented in Gedik and Bozkurt [53], an approach for the recognition of the *usul* represents an essential element in automatic retrieval of information from this music. Apart from that, the relation between the melody and the *usul* has not been investigated and an automatic approach like the one presented in this thesis can give valuable insight into the relation between melody and *usul*.

As mentioned, the compiled data set of traditional Turkish music consists of songs of the forms *Sarkı* and *Türkü*. They are following six different types of rhythmic schemes having lengths from 3 up to 10: *Aksak* ($\frac{9}{8}$), *Curcuna* ($\frac{10}{8}$), *Düyek* ($\frac{8}{8}$), *Semai* ($\frac{3}{4}$), *Sofyan* ($\frac{4}{4}$), and *Türk Aksaği* ($\frac{5}{8}$). As all *usul* in the data set have different length, the recognition of the *usul* can be reduced to a recognition of its length. This is closely related to the task of time signature recognition. In order to acquire the samples the teaching software *Mus2okur* [84] was used, resulting in a collection of 288 songs, distributed among the six *usul* as shown in the last row of Table A.4. The upper two rows in Table A.4, which is reproduced here from Chapter 4 for convenience, depict the mean values of the tempi in *bpm* (beats per minute) and the standard deviation of the tempi, respectively. The large standard deviations of the shown tempo values are visualized in Figure A.5, where the mean and standard deviations are shown as probability density function over the same bpm area as in Figure A.1. It is obvious that the overlaps are very large even without halving or doubling errors in a tempo estimation. This is because most of the samples in in this dataset are not dance music and as such, their tempo can vary in a much wider range, because this music is not connected to a specific dance as it is the case for the data described in Section A.3.

The *Mus2okur* software gives a list of songs for a chosen *usul*, which were then exported to a MIDI file. Thus, the data in D3 is available in form of symbolic descriptions, which means that their onset times can be read from the description. The MIDI files contain the description of the melody lines, usually played by only one or two instruments in unison, and the rhythmic

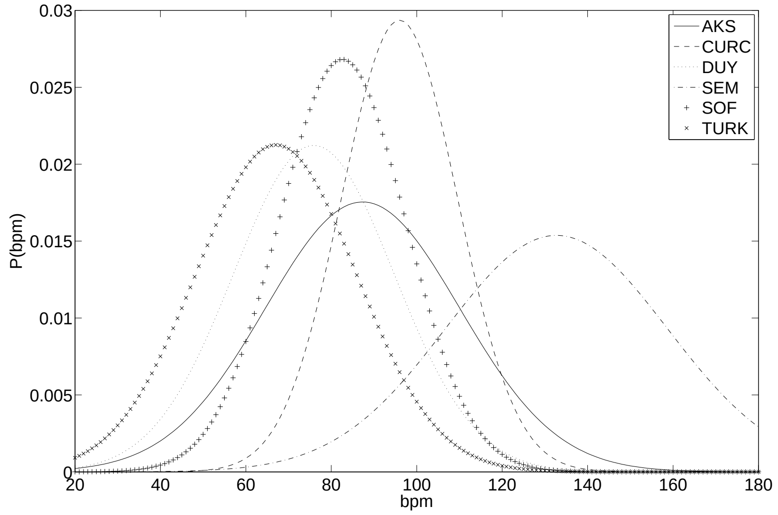


Figure A.5: Tempi of the Turkish music dataset modelled by Gaussian distributions

accompaniment by a percussive instrument. Due to the character of this music, there exists no chord accompaniment. As this content is separated into different voices, the rhythmic accompaniment can be excluded. This enables to focus on the relation between the melody of the composition and the underlying *usul*. To the the best of my knowledge, such a study on *usul* has not been conducted before.

Table A.4: Statistics of the tempo distributions in Turkish music dataset

CLASS	AKS	CURC	DUY	SEM	SOF	TURK
MEAN	98.5	98.3	70.7	131.9	81.3	73.1
STD	27.9	13.5	12.6	26.3	16.7	22.3
N_{Songs}	64	57	47	22	60	38

A.5 Instrument recognition data

This small dataset has been compiled for preliminary experiments with the timbre similarity systems presented in Chapter 6 of this thesis. It contains four classes: clarinet, Cretan *lyra*, *ney* and violin. Each class contains 20 samples of polyphonic sounds. In each class, the main melody is played by the instrument that is in the class label. It was avoided to use two or more samples of the same player or album. From each sample, sections containing singing voice have been removed by hand. The first class, clarinet contains 13 samples from various regions of the Greek mainland, 6 samples from Turkey, and one sample of a Tunesian player. The instrumental accompaniment varies widely, some samples from northern Greece contain various brass instruments (trumpet, trombone *etc.*), in some samples guitar and *ud* are part of the accompaniment. Note that also in three samples a violin is part of the accompaniment. The second class, Cretan *lyra*, is more homogeneous in terms of its local origin as well as the musical accompaniment. The accompaniment is made up of one or two Cretan *laouto* for all files and the origin of all samples is the island of Crete. The third class contains the Turkish traditional instrument *ney* as main instrument. Note that in the context of Turkish traditional music the term main instrument is not correct in all cases, as many of the samples are *heterophonic* pieces composed on a specific *makam* (see Appendix A.4 for more detail), and for that reason the instrumental accompaniment plays mainly the same melody line as the *ney*. The chosen files are mostly instrumental parts of *fasıl* (suite) compositions of Ottoman classical music, namely the first and last parts of the *fasıl*, called *peşrev* and *saz semai*, respectively. This means that this class is again, similar to Cretan *lyra*, quite homogeneous regarding its timbre. The usual accompaniment instruments in this music are *kemençe, kanun* and *tanbur*. The *kemençe* is a string instrument smaller than a violin and is played with a bow, the *kanun* is a string instrument similar to a zither, and *tanbur* is a long necked lute very different in timbre compared to *e.g.* the Cretan *laouto*.

The fourth class, violin, is more heterogeneous, as it is made up of samples from various regions. Basically, there are recordings from the Greek mainland as well as from the Greek islands and from Asia Minor. The tracks from Asia Minor originate from the common music tradition of Greek and Turkish population in the area of the western Turkish coast. However, it differs from the music contained in the *ney* class regarding instrumental timbre. Accompanying instruments in this class are guitar, accordion, *oud*, *kanun* and *santour*, which is a hammered dulcimer stemming from Persia.

Furthermore, it should be pointed out that all classes with the exception of Cretan *lyra* contain percussive instruments as well.