UNIVERSITAT POMPEU FABRA

Automatic musical instrument recognition from polyphonic music audio signals

Ferdinand Fuhrmann

TESI DOCTORAL UPF / 2012

Dissertation direction:

Dr. Xavier Serra
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain

Dissertation submitted to the Department of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfillment of the requirements for the degree of

DOCTOR PER LA UNIVERSITAT POMPEU FABRA,

with the mention of European Doctor.

Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain. http://mtg.upf.edu, http://www.upf.edu/dtic, http://www.upf.edu.

I believe that every scientist who studies music has a duty to keep his or her love to music alive.

Exploring the musical mind
J. Sloboda, 2005, p. 175

Acknowledgements

I was lucky to have been part of the Music Technology Group (MTG), where I was introduced to an intriguing and sometimes even curious field of research. The MTG has always been a warm and friendly working atmosphere, in which I could collaborate with several brilliant and stimulating minds, and moreover was able to work on one of the most fascinating things the human species ever invented, *music*. In this regard, special credit is due to Xavier Serra for leading and holding together this multi-faceted research group, and, in the end, for opening me the opportunity to dive into this 4-year adventure of a Ph.D. thesis.

It goes without saying that writing a dissertation is by no means a work of a single individual. Besides the countless collaborative thoughts and ideas that heavily influence the final outcome, apparently less important things such as draft reading or data annotation play an indispensable role in the development of a thesis. Here, the first one to mention is Perfecto Herrera, my constant advisor throughout this research, without him I still could not imagine reaching the point where I am now. He could always find a niche among his many other duties for discussing things related to my work. Due to his constant high-level input – here I have to emphasise his differentiated view on technology, his strong scientific influence, as well as his commitment to take things not too serious – this work has been shaped and became feasible from multiple perspectives.

Next, I would like to mention my teaching mates, not least for bringing back the, sometimes scarce, fun part of it. First, I would like to thank Ricard Marxer for guiding me during my first year of the signal processing course. And Martín Haro, not only for accompanying me in the following two years of teaching this course, but also for being a fantastic officemate and my intellectual partner at the endless coffee breaks. Finally, I thank Alfonso Pérez for leading the database systems classes and providing me with the necessary knowledge to overcome this slightly "out-of-context" course.

I also want to thank all the people I collaborated with in various research projects. Among the not-already-listed-above, these are Dmitry Bogdanov, Emilia Gómez, and Anna Xambó. Moreover, I thank Joan Serrà for his always keenly comments on my paper drafts, and acknowledge the proof readings of my journal article by Justin Salamon and Graham Coleman.

Several students have been involved in this thesis work, mostly fulfilling annotation duties; here, special thanks go to Ahnandi Ramesh and Tan Hakan Özaslan, for helping me setting up the instrument training collection. Moreover, Jesús Goméz Sánchez and Pratyush participated in the annotation of the evaluation collection. Finally, I want to thank Inês Salselas and Stefan Kersten for the genre annotations in the final stages of this thesis, where time was already pressuring me.

Next, Nicolas Wack and Eduard Aylon helped me a lot and deserve a special mention for their support on the provided software tools. Also Marc Vinyes, for playing the counterpart in the language exchange tandem. Moreover, I would like to thank other members of the MTG, past and present, in particular – for neutralism in alphabetical order – Enric Guaus, Piotr Holonowicz, Jordi Janer, Cyril Laurier, Hendrik Purwins, Gerard Roma, Mohamed Sordo, and all who I forgot. Finally, I want to mention Cristina Garrido for all the administrative work.

Of course, I have to thank my family for their constant support during this – mostly from my parents' perspective – sheer endless endeavour of reaching the point where I am now. For all they gave to me, I hope I will be able to give something back …

At last, very special thanks go to Kathrien, Elena, and Lelia, for being here!

This thesis has been carried out at the Music Technology Group, Universitat Pompeu Fabra (UPF) in Barcelona, Spain from September 2007 to January 2012. This work has been supported by an R+D+I scholarship from UPF, by the European Commission project PHAROS (IST-2006-045035), by the project of the Spanish Ministry of Industry, Tourism and Trade CLASSICAL PLANET (TSI-070100-2009-407), and by the project of the Spanish Ministry of Science and Innovation DRIMS (TIN-2009-14247-C02-01).

Abstract

Facing the rapidly growing amount of digital media, the need for an effective data management is challenging technology. In this context, we approach the problem of automatically recognising musical instruments from music audio signals. Information regarding the instrumentation is among the most important semantic concepts humans use to communicate musical meaning. Hence, knowledge regarding the instrumentation eases a meaningful description of a music piece, indispensable for approaching the aforementioned need with modern (music) technology.

Nonetheless, the addressed problem may sound elementary or basic, given the competence of the human auditory system. However, during at least two decades of study, while being tackled from various perspectives, the problem itself has been proven to be highly complex; no system has yet been presented that is even getting close to a human-comparable performance. Especially the problem of resolving multiple simultaneous sounding sources poses the main difficulties to the computational approaches.

In this dissertation we present a general purpose method for the automatic recognition of musical instruments from music audio signals. Unlike many related approaches, our specific conception mostly avoids laboratory constraints on the method's algorithmic design, its input data, or the targeted application context. In particular, the developed method models 12 instrumental categories, including pitched and percussive instruments as well as the human singing voice, all of them frequently adopted in Western music. To account for the assumable complex nature of the input signal, we limit the most basic process in the algorithmic chain to the recognition of a single predominant musical instrument from a short audio fragment. By applying statistical pattern recognition techniques together with properly designed, extensive datasets we predict one source from the analysed polytimbral sound and thereby prevent the method from resolving the mixture. To compensate for this restriction we further incorporate information derived from a hierarchical music analysis; we first utilise musical context to extract instrumental labels from the time-varying model decisions. Second, the method incorporates information regarding the piece's formal aspects into the recognition process. Finally, we include information from the collection level by exploiting associations between musical genres and instrumentations.

In our experiments we assess the performance of the developed method by applying a thorough evaluation methodology using real music signals only, estimating the method's accuracy, generality, scalability, robustness, and efficiency. More precisely, both the models' recognition performance and the label extraction algorithm exhibit reasonable, thus expected accuracies given the problem at hand. Furthermore, we demonstrate that the method generalises well in terms of the modelled

categories and is scalable to any kind of input data complexity, hence it provides a robust extraction of the targeted information. Moreover, we show that the information regarding the instrumentation of a Western music piece is highly redundant, thus enabling a great reduction of the data to analyse. Here, our best settings lead to a recognition performance of almost 0.7 in terms of the applied F-score from less than 50% of the input data. At last, the experiments incorporating the information on the musical genre of the analysed music pieces do not show the expected improvement in recognition performance, suggesting that a more fine-grained instrumental taxonomy is needed for exploiting this kind of information.

Resum

L'increment exponencial de la quantitat de dades digitals al nostre abast fa necessari alhora que estimula el desenvolupament de tecnologies que permetin administrar i manejar aquestes dades. En aquest context abordem el problema de reconèixer instruments musicals a partir de l'anàlisi d'enregistraments musicals (senyals d'àudio). La informació sobre la instrumentació és una de les més rellevants que els humans utilitzen per tal de comunicar significats musicals. Per tant, el coneixement relatiu a la instrumentació facilita la creació de descripcions significatives d'una peça musical, cosa indispensable per a respondre amb tecnologies musicals contemporànies a l'esmentada necessitat.

Tot i que, donada la competència del nostre sistema auditiu, el problema pot semblar elemental o molt bàsic, en les darreres dues dècades d'estudi, i a pesar d'haver estat abordat des de diferents perspectives, ha resultat ser altament complex i no existeix cap sistema que tan sols s'apropi al que els humans podem fer quan hem de discriminar instruments en una mescla musical. Poder resseguir i resoldre múltiples i simultànies línies instrumentals és especialment difícil per a qualsevol plantejament computacional.

En aquesta tesi presentem un mètode de propòsit general per al reconeixement automàtic d'instruments musicals a partir d'un senyal d'àudio. A diferència de molts enfocs relacionats, el nostre evita restriccions artificials o artificioses pel que fa al disseny algorísmic, les dades proporcionades al sistema, or el context d'aplicació. Específicament, el mètode desenvolupat modelitza 12 categories instrumentals que incloent instruments d'alçada definida, percussió i veu humana cantada, tots ells força habituals en la música occidental. Per tal de fer el problema abordable, limitem el procés a l'operació més bàsica consistent en el reconeixement de l'instrument predominant en un breu fragment d'àudio. L'aplicació de tècniques estadístiques de reconeixement de patrons, combinades amb grans conjunts de dades preparades acuradament ens permet identificar una font sonora dins d'un timbre polifònic resultant de la mescla musical, sense necessitat d'haver "desmesclat" els instruments. Per tal de compensar aquesta restricció incorporem, addicionalment, informació derivada d'una anàlisi musical jeràrquica: primer incorporem context musical a l'hora d'extraure les etiquetes dels instrument, després incorporem aspectes formals de la peça que poden ajudar al reconeixement de l'instrument, i finalment incloem informació general gràcies a l'explotació de les associacions entre gèneres musicals i instruments.

En els experiments reportats, avaluem el desemperni del mètode desenvolupat utilitzant només música "real" i calculant mesures de precisió, generalitat, escalabilitat, robustesa i eficiència. Més específicament, tan els resultats de reconeixement com l'assignació final d'etiquetes instrumentals a

un fragment de música mostren valors raonables a tenor de la dificultat del problema. A més, demostrem que el mètode es generalitzable en termes de les categories modelades, així com escalable i robust a qualsevol magnitud de complexitat de les dades d'entrada. També demostrem que la informació sobre la instrumentació de música occidental és altament redundant, cosa que facilita una gran reducció de les dades a analitzar. En aquest sentit, utilitzant menys del 50% de les dades originals podem mantenir una taxa de reconeixement (puntuació F) de gairebé 0.7. Per concloure, els experiments que incorporen informació sobre gènere musical no mostren la millora que esperàvem obtenir sobre el reconeixement dels instruments, cosa que suggereix que caldria utilitzar taxonomies de gènere més refinades que les que hem adoptat aquí.

Kurzfassung

Angesichts der immer schneller wachsenden Menge an digitalen Medien ist eine effektive Datenverwaltung für unsere moderne Gesellschaft unerlässlich. In diesem Zusammenhang widmen wir uns dem Problem der automatischen Erkennung von Musikinstrumenten aus den Audiosignalen von Musikstücken. Entsprechend der in dem jeweiligen Stück eingesetzten Instrumente verwendete Begriffe gehören zur Basis der menschlichen Kommunikation bezüglich dessen musikalischen Inhalts. Die Kenntnis der Instrumentierung einer Komposition erleichtert daher deren aussagekräftige Beschreibung – unverzichtbar für die Verwirklichung der eingangs erwähnten Datenverwaltung mittels moderner (Musik)-Technologie.

Zieht man die Fähigkeiten des menschlichen Gehörs in Betracht, erscheint das angesprochene Problem trivial. Nach mehr als zwei Jahrzehnten intensiver Auseinandersetzung mit dem Thema hat sich selbiges jedoch als hochkomplex erwiesen. Bis jetzt wurde noch kein System entwickelt welches auch nur annähernd an die Leistungen des menschlichen Gehörs herankommt. Dabei bereitet vor allem das Herauslösen von mehreren gleichzeitig klingenden Quellen aus dem Gesamtklang die größten Schwierigkeiten für artifizielle Ansätze.

In dieser Dissertation präsentieren wir eine generelle Methode für die automatische Erkennung von Musikinstrumenten aus den Audiosignalen von Musikstücken. Im Gegensatz zu vielen vergleichbaren Ansätzen vermeidet unsere spezifische Konzeption vor allem Einschränkungen in Bezug auf das algorithmischen Design der Methode, die Eingabedaten oder den speziellen Anwendungsbereich. Die entwickelte Methode modelliert 12 Musikinstrumente, harmonische und perkussive Instrumente sowie die menschliche Singstimme, welche hauptsächlich in der Musik der westlichen Welt Verwendung finden. Um der Komplexität des Eingangssignals zu entsprechen, begrenzen wir den grundlegenden Prozess der Methode auf die Erkennung des vorherrschenden Musikinstruments aus einem kurzen Audiofragment. Die Anwendung von statistischen Mustererkennungstechniken in Zusammenhang mit dementsprechend gestalteten, umfangreichen Datenbanken ermöglicht uns die Erkennung einer einzigen Quelle aus dem analysierten komplexen Gesamtklang und vermeidet dabei die Trennung des Signals in die Einzelquellen. Als Kompensation dieser Einschränkung integrieren wir zusätzliche Informationen aus einer hierarchischen Musikanalyse in den Erkennungsprozess: erstens benützen wir den musikalischen Kontext des analysierten Signals um aus der zeitlichen Abfolge der Modellprädiktionen die entsprechenden Instrumentennamen zu bestimmen. Zweitens kombiniert die Methode Informationen über strukturelle Aspekte des Musikstücks und bindet letztendlich Assoziationen zwischen musikalischen Genres und Instrumentierungen in den Algorithmus ein.

Wir evaluieren die Leistung der entwickelten Methode in unseren Experimenten durch gründliche Bewertungsverfahren, welche ausschließlich auf der Analyse von echten Musiksignalen basieren. Wir bewerten dabei die Genauigkeit, Allgemeingültigkeit, Skalierbarkeit, Robustheit und Effizienz der Methode. Im Speziellen erhalten wir sowohl für die Leistung der entwickelten Instrumentenmodelle als auch des Erkennungsalgorithmus die erwartete und angemessene Genauigkeit angesichts des vorliegenden Problems. Darüber hinaus zeigen wir, dass die Methode in Bezug auf die modellierten Kategorien verallgemeinert und auf jede Art von Komplexität der Eingabedaten skalierbar ist, daher eine robuste Extrahierung der Information ermöglicht. Im Weiteren zeigen wir, dass die Instrumentierung von Musikstücken eine redundante Information darstellt, wodurch wir den Anteil an Daten, der für die Erkennung notwendig ist, erheblich reduzieren können. Unser bestes System ermöglicht eine Erkennungsleistung von fast 0.7, anhand des angewandten F-Maßes, aus weniger als 50% der Eingabedaten. Allerdings zeigen die Ergebnisse der Experimente mit musikalischen Genres nicht die erwartete Verbesserung in der Erkennungsleistung der Methode, was darauf hindeutet, dass eine besser abgestimmte instrumentale Taxonomie für die Nutzung dieser Art von Informationen erforderlich ist.

Contents

List of Figures

List of Tables

Acronyms

Acronym	Description
ANN	Artificial neural network
ANSI	American national standards institute
ASA	Auditory scene analysis
CASA	Computational auditory scene analysis
BIC	Bayesian information criterion
CFS	Correlation-based feature selection
CL	Complete linkage
CQT	Constant Q transform
CV	Cross validation
DFT	Discrete Fourier transform
DWT	Discrete wavelet transform
GMM	Gaussian mixture model
FFT	Fast Fourier transform
HC	Hierarchical clustering
HMM	Hidden Markov model
HPCP	Harmonic pitch class profile
ICA	Independent component analysis
ISMIR	International society for music information retrieval
KL	Kullback-Leibler (divergence)
kNN	k-nearest neighbour
LDA	Linear discriminant analysis
MCP	Music content processing
MDS	Multidimensional scaling
MFCC	Mel frequency cepstral coefficient
MIDI	Musical instrument digital interface
MIR	Music information retrieval
MIREX	Music information retrieval evaluation exchange
MP	Matching pursuit
NMF	Non-negative matrix factorisation
OAM	Overlapping area matrix

continued on next page . . .

Acronym	Description
PCA	Principal component analysis
PLCA	Probabilistic latent component analysis
RA	Regression analysis
RBF	Radial basis function
ROC	Receiver operating characteristic (curve)
SL	Single linkage
SRM	Structural risk minimisation
STFT	Short-time Fourier transform
SVC	Support vector classification
SVM	Support vector machine
UPGMA	Group average linkage
WPGMA	Weighted average linkage

1

Introduction

To enjoy the music we like, we may walk in the street, move around dancing, converse with friends, drive a car, or simply relax. Meanwhile, and independent of the aforementioned, our brains perform a huge amount of complex processes to compile the auditory sensory input data into informative structures (Patel, 2007). For instance, separating the passing car in the left rear from the electric guitar solo and the driving drum pattern in your headphone, subconsciously. In everday's music listening context the human mind decodes the incoming audio stream into elementary building blocks, related to various acoustical and musical facets (Levitin, 2008). From this abstract representation musical meaning is inferred, a process that involves factors such as musical preference and knowledge, memory, lifestyle, etcetera (Hargreaves & North, 1999). Without this meaning we could not love music as we are used to do it, in some sense it would lose its value. Hence, music does not exist outside the human mind, all leftover would simply be a variation in air pressure.

One of these building blocks corresponds to the identity of sounding sources; we can only understand an acoustical scene if we are able to infer knowledge regarding the participating sound producing objects. This is evident from an evolutionary point-of-view, since specific knowledge about a particular source allows for a distinction between friend or foe, hence providing basic survival means. In a musical context this problem is termed musical instrument recognition. One may claim the simplicity of the problem, since every Western enculturated person is able to distinguish a violin from a piano. However, the task is far more complex, involving the physical properties of musical instruments, the rules imposed by the music system on the composition, as well as the perceptual and cognitive processing of the resulting sounds. Modelling the problem in a general and holistic manner still imposes a lot of difficulties to artificial systems. Besides, even humans exhibit clear limits in their abilities in distinguishing between musical instruments.

In essence, this dissertation deals with the automatic recognition of musical instruments from music audio signals. The aim is to identify the constituting instruments given an unknown piece of music. The availability of this information can facilitate index and retrieval operations for managing

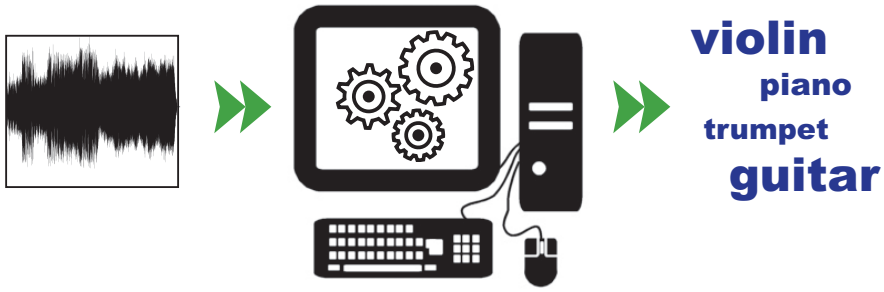


Figure 1.1: Illustration of the problem addressed in this thesis. Given an unknown musical composition the task is to identify the featured musical instruments.

big multimedia archives, enhance the quality of recommender system, can be adopted for educational purposes, or open up new directions in the development of compositional tools (see also Section 1.5).

In this context, questions regarding the involved perceptual and cognitive mechanisms may arise; or, how and to what extent can we teach a computer to perform the task? Hence, the problem is appealing from both an engineering and scientific point-of-view. Some of these questions are covered by this thesis, others remain out of the scope of this work. This section serves as an introduction to the thesis' contents and provides the corresponding contextual links. Figure 1.1 depicts an abstraction of the general problem addressed.

1.1 Motivation

Our habits in listening to music changed dramatically within the last three decades. Digitalisation of raw audio signals and the ensuing compression of the resulting data streams was developed in line with the emergence of personal home computer systems with their subsequently increasing storage capacities, together allowing for the construction of music archives immensely extending the, by then, usual dimensions. Internet technologies and the thereby initiated changes in the concept of musical proprietaries, with all involved implications for the music industry, converted music consumption and dissemination from a highly personal or within-small-groups phenomenon to a property of (on-line) societies, at least in the view of social communities. Due to the facilities of modern technology, which simplify music production and promotion processes as never before, a massive amount of new, yet unknown music is created everyday. Nowadays, music is a service, with recent digital communication devices always everywhere available, to a sheer unbounded extent. In this context, technologies for managing this huge amount of music data claim for intelligent indexing tools. From a user's perspective, an automatic separation into relevant and irrelevant items in such large archives is required. Music recommendation is more important than ever, since the enormous assortment entails an inability to select the music to listen, paradoxically (Southard, 2010).

Given these new dimensions in the availability of music and the way music is consumed, one of the big challenges of modern music technology is to provide access to these data in a meaningful way. In this respect, the precise personalisation of the recommendation process as well as the fusion of what is called *musical content* and *context*, i.e. information directly extracted from the acoustical signal and information inferred from user-assigned tags as well as collaborative listening data, respectively, will be among the future objectives (Celma & Serra, 2008). The technology thereby acts as a companion, monitoring our listening habits in connection with our activities, profiling our diverse musical preferences, supplying music on demand, providing both known and unknown music tracks with respect to the given context, and ultimately shaping and improving our music intellects, purposely! All this, however, may raise the question if it is ever possible to capture the essence of music, what keeps us listening to music, with technology? And, if yes, are such technologies really able to make us "better" music listeners? Or will they always remain artificial gadgets for nerdy technology-affine people, only representing a vast minority among all music listeners? It is, however, not intended to provide answers to these questions within this thesis, but they naturally arise in such technologically driven conceptions involving any artistic origin.

The instrumentation[1] of a musical composition is one of the main factors in the perceptual and cognitive processing of music, since it determines the piece's timbre, a fundamental dimension of sound perception (see Section 2.1.1.2). Timbre influences both the categorisation into musical styles and the emotional affect of music (Alluri & Toiviainen, 2009, and references therein), at which humans are able to deduce this information within a very short amount of time, typically several tenth of a second. Here, instrumentation shapes – together with other musical and acoustical factors – the mental inference of higher-level musical concepts. Furthermore, at longer time scales, musical instruments can exhibit very descriptive attributes, for instance in solo or voice-leading sections of a music piece. In this context, the work of McKay & Fujinaga (2010) exemplifies the importance of instrumentation in automatic classification of music. In this study the authors revealed features derived from the instrumentation of a music piece to be most descriptive, among all tested features, in terms of the piece's musical genre. Moreover, humans usually use the aforementioned semantic information to express, convey, and communicate musical meaning[2]. In short, musical instruments represent an essential part – implicitly and explicitly – in our description of music.

From an engineering point-of-view, information on the instruments featured in a given musical composition is therefore an important component in meeting the requirements of modern technology. In this regard, the aforementioned applications for indexing and retrieval, or recommendation, can only be applied in a meaningful way if the developed algorithms are able to extract those factors from the data that define why we like or do not like certain pieces of music, at which instrumentation evidently plays a major role.

From a scientific perspective, understanding and modelling the physical world together with its perception and cognition has always been the primary motivation for research. Here, questions re-

[1] The combination of instruments used by musicians to play either a certain style of music, or a particular piece within that genre (retrieved from `http://www.louisianavoices.org/edu_glossary.html`). Moreover, the new Grove dictionary of music and musicians (Sadie, 1980) suggests that the term should be considered as inseparable from the notion of orchestration, see Section 3.2, page 53.

[2] In a typical on-line collection, instrumentation constitutes, along with genre and mood related information, the most frequently used semantic dimension to describe music (Eck et al., 2008)

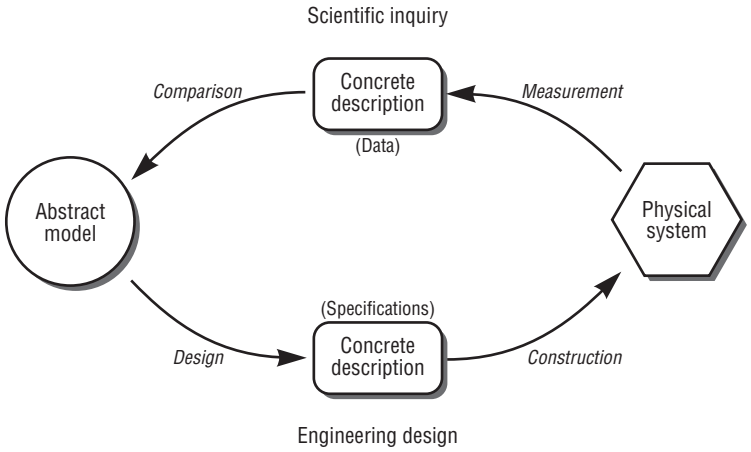


Figure 1.2: Interdependency between the science and engineering after Drexler (2009). Scientific models may influence practical realisations, while prototyped systems contribute to new or enhanced theories about the problem itself.

garding the timbral relations between musical instruments, as well as their mental processing and representations – both in isolation and within musical context – arise. Which components of a given musical instrument's sound affects its identifiability among other instruments? What perceptual and cognitive principles enable the recognition of musical instruments playing simultaneously? Furthermore, the notion of similarity may be exploited in a way of which attributes of perceived timbre influence the relations between several musical instruments.

Finally, we want to point to the interdependency of the two perspectives outlined above. Here, both domains share the basic concepts of a *physical system*, an *abstract model*, and the *concrete descriptions* of one of the aforementioned (Drexler, 2009). In particular, scientific research (*inquiry*) describes physical systems by collecting data via measurements, which results lead to the formulation of theories regarding general models of the measured systems. Engineering research (*design*), by contrast, designs concrete descriptions on the basis of a conceptual model, resulting in the construction of prototype systems. Figure 1.2 illustrates this intimately interleaved nature of science and engineering after Drexler (2009). In this regard, any scientific motivated modelling process may have practical implications on the development of a proper system applicable in an engineering scenario. On the opposite, empirical findings in the construction of engineering systems may lead to new or advanced scientific theories (Scheirer, 2000). In the context of this thesis, we hope that the development of our method for the automatic recognition of musical instruments from music audio data does not only advance modern technology for a more accurate music indexing and recommendation, but also contributes to a better understanding of human perception and cognition of perceived timbre.

1.2 Context of the thesis

This thesis is written in the context of Music Information Retrieval (MIR), an increasingly popular, interdisciplinary research area. In a very general sense, Kassler (1966) defines MIR as follows:

> "[...] the task of extracting, from a large quantity of music data, the portions of that data with respect to which some particular musicological statement is true."

This classic definition, which arose from the roots of the discipline, was entailed by the goals of the elimination of manual music transcription, the establishment of an effective input language for music, as well as the development of an economic way for printing music (Lincoln, 1967). With exception of the latter, which appears slightly out-of-date, these general aims have been actively pursued and still represent ongoing research topics inside MIR. However, the advent of digital media unfolded new, additional perspectives for the research community. In this respect, the main functionality of MIR is to provide basic means for accessing music collections. Here, the developed algorithms and systems can target the recording industry or companies aggregating and disseminating music. Furthermore, professionals such as music performers, teachers, producers, musicologists, etcetera might be addressed; or simply individuals looking for services which offer personalised tools for searching and discovering music (Casey et al., 2008).

The constant growing interest in MIR is manifested by both attendances and publication statistics of the annual International Society for Music Information Retrieval[3] (ISMIR) meeting (Downie et al., 2009), and the increasing popularity of related topics in typically not-music-focussed conventions such as IEEE's International Conference on Acoustics, Speech, and Signal Processing[4] (ICASSP), or the Digital Audio Effects[5] (DAFx) conference. The ISMIR conference in particular provides a proper platform for both research and industry, facilitating knowledge exchange and technology transfer. Moreover, the simultaneously hold Music Information Retrieval Evaluation eXchange (MIREX) competition offers an objective evaluation framework for algorithmic implementations on standardized tasks (Downie, 2008).

In general, technologies based on MIR research enable the access to music collections by supplying metadata information. Here, we can refer to any information related to a musical composition that can be annotated or extracted, but being meaningful in any way (i.e. it exhibits semantic information), with the term metadata[6] (Gouyon et al., 2008). Since it represents the main motivation for modern MIR systems, many of such systems are designed for simply providing metadata (Casey et al., 2008).

In view of the aforementioned, content-based MIR, or Music Content Processing (MCP), aims at understanding and modelling the complex interaction between humans and music by extracting

[3] http://www.ismir.net/

[4] e.g. http://www.icassp2012.com/

[5] http://www.dafx.de/

[6] Besides, metadata literally denotes *data about the data*.

information from the audio signal. Hence, the notion of *content processing*[7] refers to the analysis, description, and exploitation of information derived from the raw audio data, in contrast to the term *information retrieval* in MIR, which corresponds to the gathering of any kind of information related to music. The information provided by content processing is thought to complete the metadata derived from other sources such as knowledge deduced from community analyses or editorial metadata.

In its interdisciplinary character, MCP represents a synergy of at least the areas of signal processing, computer science, information retrieval, and cognitive sciences for both describing and exploiting musical content (Gouyon et al., 2008). In doing so, it maps the musical content to concepts related to the (Western) music system, thus providing an intuitive mean for data interaction operations. However, the extraction of this high-level, i.e. semantically meaningful, information from content is a very difficult task, beyond that of objective nature, thus requiring an explicit user modelling process. Hence, MCP systems usually try to exploit several layers of abstraction of the aforementioned semantic concepts in the description of the content, in order to meet the requirements of as many people as possible (Casey et al., 2008).

1.3 The problem – an overall viewpoint

In general, the auditory scene produced by a musical composition can be regarded as a multi-source environment, where different sound sources are temporarily active, some of them only sparsely. These sources may be of different instrumental type (therefore exhibiting different timbral sensations), may be played at various pitches and loudness, and even the spatial position of a given sound source may vary with respect to time. Often individual sources recur during a musical piece, either in a different musical context or by revisiting already established phrases. Thus, the scene can be regarded as a time-varying schedule of source activity containing both novel and repeated patterns, indicating changes in the spectral, temporal, and spatial complexity of the mixture. As an example, Figure 1.3 shows the source activity along time of two tracks taken from different musical genres.

In this context, an ideal musical instrument recognition system is able to recognise all sound-producing musical instruments inside a given mixture[8]. In practise, due to the aforementioned multi-source properties of the musical scene, time and frequency interferences between several sounding sources hinder the direct extraction of the source-specific characteristics necessary for recognition. Pre-processing must therefore be applied to minimise the interference effects for a reliable recognition.

[7] In addition, Leman (2003) denotes musical content as a 3-dimensional phenomenon which exhibits cultural dependency, represents a percept in the auditory system, and can be computationally implemented by a series of processes that emulate human knowledge structure related to music.

[8] Note the consequences this universal claim involves by considering all possible sound sources such a recognition system is confronted with. Apart from traditional acoustic instruments, which exhibit rather unique acoustical characteristics, electronic devices may produce sounds that vary to a great extent due to changes in their parameter values. Here, the question arises of whether we can model and recognise an analogue synthesiser, or a DX7 piano synthetic patch? Besides, all sounds not produced by any instrument, such as environmental or animal sounds, must be neglected by a musical instrument recognition system, even though they act as essential elements in some musical genres.

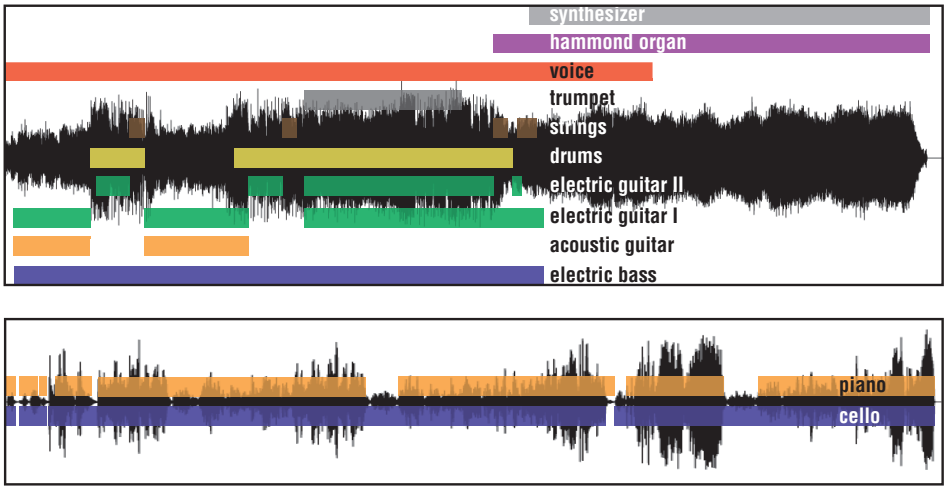


Figure 1.3: Distribution of musical instruments along time in two pieces of music. The upper track represents a rock piece whereas the lower one is a classical sonata.

1.4 Scope of the thesis

As we have seen in the previous sections, the semantic information inherent to the addressed problem of automatic musical instrument recognition from music audio signals is important with respect to the descriptive aims of MIR systems. In addition, the problem is strongly connected to other MCP tasks, in particular to the fields studying music similarity and recommendation. However, a lot of works conceptualized for automatic musical instrument recognition are not applicable in any realistic MIR context, since the restrictions imposed to the data or the method itself do not conform with the complexity of real world input data (see Chapter 3). Hence, we can deduce that the handing of real-world stimuli represents a kind-of knowledge gap inside the automatic musical instrument recognition paradigm. Moreover, the difficult endeavour of the problem and its situation at the crossroads of perception, music cognition, and engineering is challenging. Therefore, the primary objective and intended purpose in the initial development stages was the development of a method for the automatic recognition of musical instruments from real-world music signals, in connection with its integration in a nowadays MIR framework.

This focus on real-world stimuli involves three main criteria related to the functionalities of the methods to develop. First and second, the algorithms must exhibit strong data handling abilities together with a high degree of generality in terms of the modelled categories. Since the input data to the developed methods is assumed to be of a real-world nature, the complexity of the data itself and the involved variability in the properties of the musical instruments must be addressed properly (e.g. a *Clarinet* must be recognised in both classical and jazz music despite its possibly different construction types and adoptions). Third, the extracted information has to persist meaningful among several contexts, thus the abstraction of the semantic concepts in the modelling process and the thereof derived taxonomy must be carefully designed to meet as many requirements as possible.

We address the problem by modelling polyphonic timbre in terms of predominant instruments. In particular, the presented approach focuses on timbre recognition directly from polyphonies, i.e. the mixture signal itself. We construct recognition models employing the hypotheses that, first, the timbre of a given mixture is mainly influenced by the predominant source, provided its existence, and that its source-specific properties can be reliably extracted from the raw signal. Second, we hypothesise that the instrumentations of a musical composition can be approximated by the information regarding predominant instruments. In doing so we purposely avoid the adoption of any polyphonic pre-processing of the raw audio data, be it source separation, multiple pitch estimation or onset detection, since the propagation of errors may lead to even worse results compared to the information we are already able to gain without them. In order to meet requirements one and two – the data handling and generality claims – we apply a sufficient amount of representative data in the modelling process. However, given the complexity of the input data, we accept the noisy character of the approach but assume that even an imperfect inference based on these data can provide meaningful means for a rough description of the instrumentation. To address the third criteria – the preservation of meaning among several contexts – we concentrated the modelling on categories able to cover most instrumentations found in Western music, which we pragmatically define as the most frequently used in a typical collection of Western music. This guarantees a wide applicability of the developed models to different kinds of MIR problems.

In general, we are not aiming at explicitly modelling human perception nor cognition of musical instrument recognition, but we employ several related techniques in our computational implementation of the method. In this regard, we can explain many of the applied algorithmic concepts with perceptual and cognitive mechanisms. Moreover, the presented methods do not represent a holistic solution towards the problem. We rather aim at deriving an optimal solution given the scope, the context, and the methods at hand. Finally, we regard the presented methods at connecting the works studying perceptual timbre recognition and the engineering-motivated demands for intuitive music search and recommendation algorithms, where information regarding the instrumentation of music pieces is crucial.

1.5 Applications of the presented work

In this section we point towards some of the main application fields of automatic musical instrument recognition systems. From a MIR perspective, such a system can be implemented in any music indexing context, or application of a general music similarity. Tag propagation, recommender, or playlist generation systems – to name just a few – conceptually use the information regarding the instrumentation of a music piece. Furthermore, music indexing opens possibilities for educational aspects beside the pure managing abilities of big archives. Music students may browse sound archives for compositions containing a certain solo instrument; or search for the appearance of certain instruments or instrumental combinations in a musical recording.

Moreover, information regarding the instrumentation of a musical composition is necessary for other MCP algorithms acting on higher-level musical concepts. Typical artist or album classification systems can benefit from instrumental cues since these mainly exploit timbral dimensions. Moreover, the assumable subjective notions of musical genre and mood are influenced by the adoption of certain musical instruments (see Chapter 6 and the work of McKay & Fujinaga (2010)).

Music signal processing in general benefits from the information regarding the identity of the sounding sources within a music piece. From a holistic point-of-view, any information related to the musical scene, be it of low-, mid-, or high-level kind, contributes to the concept of what is called a *music understanding* system (Scheirer, 1999). Here, we want to emphasize the notion of "complete systems", processing music as a perceptual entity. As a result of their mutual interdependencies no single component of music can be analysed in isolation, both from a cognitive and a musical viewpoint. Many *transcriptive* approaches of music contrarily focus on separating the individual signals, irrespective of their perceptual and cognitive relevance (see Section 2.2 for a thorough discussion of these conceptually opposed approaches towards the computational processing of music audio signals).

Finally, new compositional tools and musical instruments working on a high-level, semantic language may take advantage of the provided information. Sample-based systems can directly select sound units according to the query in terms of a particular musical instrument (e.g. audio mosaicing or any other form of concatenative synthesis with instrumental constraints). Moreover, the concept of musical instruments may be essential for a general description of timbre in such systems.

1.6 Contributions

We regard the presented work and the resulting outcomes related to the specific problem of automatic musical instrument recognition from real-world music audio signals. The following lists our main contributions:

1. The development of a model for predominant source recognition from polyphonies and a corresponding labelling method. By directly modelling the polyphonic timbre we assure a maximum possible data complexity handling. Moreover, we provide simple labelling strategies which infer labels related to the musical instruments from the predictions of the models by analysing musical context. To our knowledge, this dissertation is the first thesis work exclusively devoted to musical instrument recognition from polyphonies.

2. The incorporation of multiple musical instruments including pitched and percussive sources, as well as the human voice in a unifying framework. This allows for a comprehensive and meaningful description of music audio data in terms of musical instruments. To our knowledge, we present one of the few systems incorporating all three aforementioned instrumental categories.

3. The quantitative and qualitative evaluation of our presented method. In comparison to other works we set a high value on the applied testing environment. We put emphasis on the development of both the training and testing collections used in the evaluation experiments, and use a great variety of different evaluation metrics to assess the performance characteristics of the method under the best possible conditions. Furthermore, we test the method with respect to its robustness against noise, as defined by the amount of participating unknown sources.

4. We contribute to the understanding of sound categories, both in isolation and in mixtures, in terms of the description of the raw acoustical data. The thesis provides several sections analysing the applicability of different audio features to the problem, both in automatic and manual processes.

5. We only use data taken from real music recordings for evaluation purposes, involving a great variety of Western musical genres and styles. To our knowledge, this represents the less restricted testing condition for a musical instrument recognition system ever applied in literature.

6. We further present and evaluate approaches for the labelling of entire pieces of music which incorporate high-level musical knowledge. Here, we both exploit inter-song structures and global properties of the music itself to develop intelligent algorithms to apply the aforementioned label inference algorithms. To our knowledge, no study in literature has addressed this problem so far, since all methods pragmatically process all data of a given musical composition, neglecting the inherent structural properties and the thereby generated redundancy in terms of instrumentation.

7. With this work we initialise both a benchmark for existing algorithms on real music data and a first baseline acting as legitimation for more complex approaches. Only if the respective methods are able to go beyond the presented performance figures, the application of heavier signal processing or machine learning algorithms is justified.

8. We provide two new datasets for the research community, fully annotated for training and testing musical instrument recognition algorithms.

1.7 Outline

This dissertation's content follows a strict sequential structure, each chapter thus represents some input for the next one. After two chapters reviewing background information and related relevant literature, the main part of the thesis starts from the frame-level analysis for automatic musical instrument recognition in Chapter 4 and ends at the collection level where we explore interactions between related musical concepts in Chapter 6. The following lists the topics involved in the respective chapters.

In Chapter 2 we present the basic scientific background from the fields of auditory perception and cognition, music signal processing, and machine learning. We start the chapter with a brief introduction to the functionalities of the human auditory system, which is followed by a more detailed analysis of the perceptual and cognitive mechanisms involved in the analysis of complex auditory scenes (Section 2.1.2). Section 2.2 introduces the basic concepts applied in the area of machine listening, an interdisciplinary field computationally modelling the processes and mechanisms of the human auditory system when exposed to sound. Here, Section 2.2.1 includes details about the involved signal processing techniques and their relation to the perceptual processes, while Section 2.2.2 refers to the notions, concepts, and algorithms adopted from the field of machine learning.

Chapter 3 covers the related work specific to the problem of automatic musical instrument recognition. We start by reviewing the general physical properties of musical instruments (Section 3.1) and assess human abilities in recognising them (Section 3.2). Section 3.3 further formulates general evaluation criteria for systems designed for the automatic recognition of musical instruments, which is followed by an assessment of the most common methodological issues involved (Section 3.4). We then examine the relevant studies in this field, concentrating on those works which developed methods for processing music audio data (Section 3.5) – in contrast to those works applying isolated sample as input for the recognition algorithm. Finally, Section 3.6 closes this chapter by discussing the main outcomes.

In Chapter 4 we present our developed method, termed *label inference*, for extracting labels in terms of musical instruments from a given music audio signal. The introductory Section 4.1 covers the main hypotheses underlying the presented approach together with their conceptual adoptions. The first part of the chapter then describes the frame-level recognition, i.e. classification, for both pitched and percussive musical instruments (Section 4.2). Here, we discuss the involved conceptual and experimental methodologies, along with all involved technical specificities. Both pitched and percussive analyses further contain an extensive analysis of the acoustical factors in terms of audio features involved in the recognition process as well as a subsequent analysis of recognition errors. The second part of the chapter describes the adoption of the developed frame-level recognition for the extraction of instrumental labels from music audio signals of any length (Section 4.3). Here, we emphasise the importance of musical context and show how a corresponding analysis leads to a robust extraction of instrumental labels from the audio data regardless its timbral complexity. In particular, we present and evaluate three conceptually different approaches for processing the output of the developed recognition models along a musical excerpt. The chapter is finally closed by comparing the developed method to state-of-the-art approaches in automatic instrument recognition and a general discussion of the obtained results (Section 4.4).

In Chapter 5 we further present a conception, termed *track-level analysis*, for an instrumentation analysis of entire pieces of music. We develop two conceptually different approaches for applying the label inference method described in the preceding chapter for extracting the instrumentation from music pieces. In the first part of this chapter we introduce an approach for locating those sections in a given music track, where robust predictions regarding the involved instruments are more likely (Section 5.1). In the second part, several methods for exploiting the recurrences, or redundancies, of instruments inside typical musical forms are presented, enabling an efficient instrumentation analysis (Section 5.2). The following Section 5.3 then assesses the performance of all introduced

track-level approaches in a common evaluation framework, where we focus on both recognition accuracy and the amount of data used for extracting the labels. At last, Section 5.4 closes this chapter by summarising its content and discussing the main outcomes.

Chapter 6 finally explores the relations between instrumentation and related musical facets. In particular, we study the associations between musical instruments and genres. In Section 6.1 we first quantify these associations by evaluating both human-assigned and automatically predicted information. In the following section we present and evaluate several automatic musical instrument recognition systems which incorporate the information regarding the musical genre of the analysed piece directly into the recognition process (Section 6.2). Section 6.3 then summarises the main ideas of the chapter and critically discusses the obtained results.

At last, Chapter 7 presents a discussion of and conclusions on the thesis's main outcomes. We first summarise the content of this thesis in Section 7.2, which is followed by a list of insights gained via the various obtained results. We then identify the main unsolved problems in the field of automatic musical instrument recognition from multi-source music audio signals and provide an outlook regarding their possible approaches (Section 7.3). Finally, Section 7.4 closes this thesis by presenting several concluding remarks.

Additionally, the Appendix provides a list of all applied audio features along with their mathematical formulations (App. A). Furthermore, a table containing the metadata information for all music pieces of the music collection used for evaluating the presented methods is added subsequently (App. B), which is followed by a list of the author's publications (App. C).

2

Background

Principles and models of human and machine sound perception

> "In order to teach machines how to listen to music, we must first understand what it is that people hear when they listen to music. And by trying to build computer machine-listening systems, we will learn a great deal about the nature of music and about human perceptual processes."
>
> (Scheirer, 2000, p. 13)

These introductory words taken from Eric Scheirer's thesis summarise best the underlying principles and purposes of machine listening systems. We regard this dissertation mainly positioned in the field of machine listening, teaching a computer to extract human-understandable information regarding the instrumentation of a given music piece. This chapter describes parts of those areas most relevant to the main directions of the thesis. In particular, we will selectively review basic concepts from the three research fields of psychoacoustics, music signal processing, and machine learning, all directly connected to the methodologies presented later in this work. The here-provided background information therefore serves as the foundation for the algorithms described in Chapters 4 - 6.

Although we have mentioned, in the previous introductory chapter, several, this thesis motivating engineering goals, we begin this chapter with a review of the most relevant processes and mechanisms of the human auditory system for processing sound in general and, more specifically, recognising sound sources. The motivation behind is that human auditory perception and cognition is, after all, our touchstone for the domain of music processing with a machine, hence the here-involved processes need some specific attention. More specifically, to develop a coherent machine understanding of music – a quite general notion which we will refer to with the term of *extracting musical meaning* – the mechanisms of the human auditory system and the thereof derived high-level understanding of music are indispensable. Here, Wiggins (2009) argues, besides referring to the so-called *semantic gap* that we introduce in Section 2.2, that

> " [...] the starting point for all music information retrieval (MIR) research needs to be perception and cognition, and particularly musical memory, for it is they that *define* Music."

In other words, music, as a construct of the human mind, is per se determined by the processes of auditory perception and cognition. With his viewpoint, Wiggins takes the matter of the importance of human auditory and cognitive processes in automatic music processing further, as yet repeatedly stated in relevant literature (e.g. Aucouturier, 2009; Ellis, 1996; Hawley, 1993; Martin et al., 1998; Pampalk et al., 2005; Scheirer, 1996). Hence, Section 2.1.1 covers the basic perceptual concepts and processes necessary for human sound source recognition. Subsequently, Section 2.1.2 takes a closer look at the handling of complex auditory scenes by the auditory system.

We then review the broad area of machine listening, a field which main research line tries to understand auditory scenes in general by means of a computer. Here, Section 2.2.1 introduces the basic principles of music signal processing with an emphasis on different signal representations used for music audio signals. In Section 2.2.2 we survey several basic concepts of machine learning and pattern recognition. We first focus on the different semantic layers for extracting information from the music audio signal in terms of audio features and derive the related musical context. The second part then addresses general aspects of learning algorithms typically applied in computational modelling.

2.1 Human auditory perception and cognition: From low-level cues to recognition models

One of the most outstanding characteristics of our species is the creation, by processing diverse information sources, of complex and abstract internal representations of the outside world, together with its transfer via communication by means of language and culture. Extracting information from the physical signal of the acoustical environment represents only one part of this multi-sensory, interactive mechanism. However, the human auditory system is able to infer, even when left in isolation, an astonishingly accurate sketch of the conditions present in the surrounding world. In this process the recognition and identification of sounding sources plays an evidently important role. Not much is yet known about the variety of complex mechanisms involved in the task of sound source recognition, but it is clear that it involves many different perceptual processes, starting from very basic, "low-level" analyses of the acoustical input to "higher-level" processes including auditory memory.

The complex nature of the problem, along with the apparent ease of its handling by the human mind, has brought some theoretical debate into literature. How the perceptual system creates meaning given the ambiguity in the sensory data itself, the loss of information at the periphery, and the potentially lacking of memory representations, all of which are assumed to be involved in sound source recognition (Lufti, 2008), is one of the essential questions raised here. In this regard, we can identify three main theoretical approaches to the problem from literature:

1. **Inferential approach.** In the 19th century, von Helmholtz (1954)[1] introduced this earliest perceptual theory, stating that the human mind adds information based on prior knowledge to the stimulus in order to make sense of the raw sensory data. Since the sensory input data is per se ambiguous and incomplete, the perceptual system performs inference from the knowledge of its likelihood, which is determined innately or originates from experience.

2. **Organisational approach.** The second theoretical approach traces back to Gestalt psychology or *gestaltism*, which believes that perception is mainly determined by the extraction of structure and order from the sensory input. Here, the notions of regularity, symmetry, and simplicity play a fundamental role in the formation of objects (see Bregman (1990) for its direct application to audition). These views originate from the assumed operational principle of the human brain's holistic, parallel, and self-organising character. Similar to Helmholtz's inferential theory, the Gestaltists consider the sensory information to be ambiguous and incomplete, at which the human mind processes these data by applying defined rules derived from the aforementioned concepts of structure and order. These two theoretical approaches therefore share several commonalities since the most likely prediction from the data is often equivalent to its organisational interpretation.

3. **Ecological approach.** This radically different theory founded by Gibson (1950) assumes that perceptual stimuli exhibit so-called *invariants* which are perceived directly without the need for any other information. Gibson emphasised the direct nature of perception, hence disregarding any form of prior knowledge involved in the respective processes. Hence, this approach relies on the ordered nature of the sensory information in opposite to the ambiguity claims encountered in the former two.

Lufti (2008) further introduces a fourth, termed **Eclectic**, approach based on principles freely borrowed from each of the three aforementioned theories. This approach has been applied in the most remarkable computational models of listening (see e.g. Ellis, 1996; Martin, 1999). In these works, the authors use an auditory-inspired sensory processing on top of which inferential, organisational, and ecological principles extract the desired information. At last, Lufti (2008) argues that the eclectic approach may be the most promising from all here-listed for an advancement of our understanding of human sound source identification.

Regarding the more specific problem of source recognition from auditory sensory data, McAdams (1993) defines a general perceptual model by identifying the following mechanisms involved in the identification of a single source. These interactive processes start with the peripheral analysis of the acoustical scene and lead to the mental descriptions of the sound source.

1. **Sensory transduction.** The first stage describes the representations of the raw auditory stimulus in the peripheral auditory system. At this level the vibrations present as air pressure differences are encoded into neural activity, which is then interpreted by higher-level perceptual processes.

[1] The German scientist (*1821, †1894) published the first major study on physical attributes of complex tones, the physiological mechanisms involved in their perception as well as the sensation of timbre in particular. Due to the extensiveness of this work and the validity of most of the presented findings up to now, von Helmholtz is often termed as one of the pioneering researcher in hearing science and his influential work is still cited as major reference.

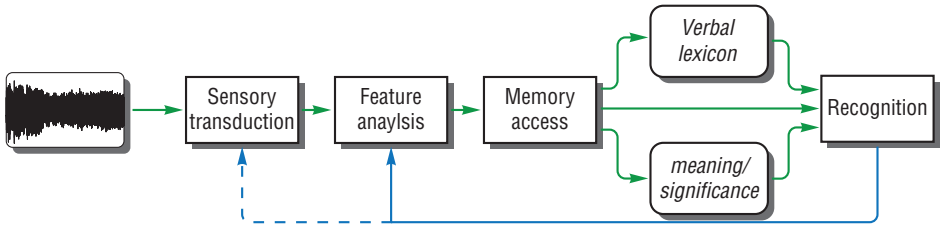


Figure 2.1: A general model of human sound source recognition after McAdams (1993).

2. **Feature analysis.** Here, the extraction of invariants, i.e. properties that stay constant despite the variation of other properties, which are the direct input representation to the actual recognition process, takes place. In particular, we can differentiate between micro-temporal properties such as structural invariants, which can be viewed – in an ecological sense – as the physical structure of the source, and transformational invariants – the specific excitation mechanism applied to the source from an ecological viewpoint. Moreover, McAdams also mentioned the extraction of macro-temporal properties related to textual or rhythmic patterns of the whole acoustic scene.

3. **Access of memory representations.** A matching procedure is performed either via a comparison process, where the nearest memory representation in terms of the used features is selected, or by a direct activation process so that the memory representations are directly accessed given a certain constellation of features in the perceptual description. Here, the memory representation exhibiting the highest activation is selected.

4. **Recognition and identification.** Finally, the verbal lexicon, in case of an already availability of language, is addressed and/or associated knowledge retrieved. At this stage, the processing is no longer purely of auditory nature. Please note that recognition and identification may take place in parallel.

The recognition process described above is by no means of a purely bottom-up kind; information originating from later stages in the processing chain influence the peripheral auditory processing and the extraction of source-specific characteristics. This top-down mechanisms of auditory organisation are accountable for the high interactivity between the different processes involved in auditory perception. Figure 2.1 illustrates this interactive process.

Before entering the very basic concepts and processes of auditory perception, let us consider some basic theoretical issues regarding the actual recognition process. In particular, we adopt a viewpoint similar to Martin (1999, p. 11 et seq.) and Minsky (1988), who viewed recognition as a *process* in a classification context. Recognition is thus taking place at different levels of abstraction in a categorical space, a given sounding source may therefore be described at different layers of information granularity. Thus, each recognition level enables the listener to draw specific judgements exhibiting a certain information content, or entropy, about the sounding object. Moving towards less abstracted categorical levels will reveal more specific details about the object under analysis, at the expense of a higher information need to classify the object into the respective categories. More

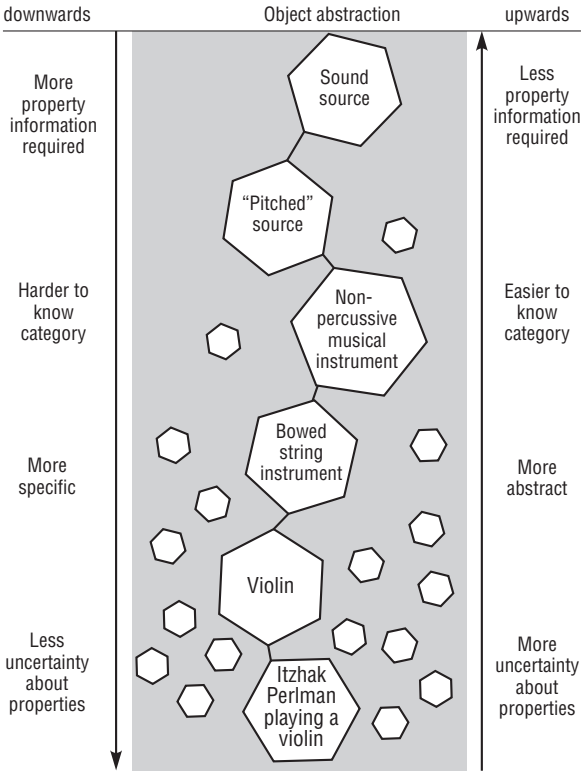


Figure 2.2: Recognition as classification in a category-abstraction space after Martin (1999). Recognition is regarded as a process which starts at a basic level of abstraction and evolves downwards towards a more specific description of the object, depending on the information needs. Hence, the different layers of abstraction represent the information granularity, at which less property information is needed for a recognition at higher levels of abstraction while accordingly more object-specific data is necessary for a more detailed recognition. The columns at the left and right margins indicate the changes involved when moving in the corresponding direction of abstraction. The not addressed, small hexagons indicate other possible categories in the recognition process such as "my favourite instrument", "brown wooden thing", or "seen live one year ago", etcetera.

abstract levels accordingly require less source-specific information for recognition, but less details about the sounding source are revealed. In this context, recognition is regarded as a process that starts at a certain lower level of abstraction and may be continued according to the required granularity of the extracted information. Thus the process moves down the hierarchy and refines the prediction strength by accumulating more sensory data. Figure 2.2 depicts the underlying ideas, synthesised from drawings by Martin (1999). Minsky (1988) particularly argues that there is a privileged, entry-level category representation for reasoning and recognition that occurs at an intermediate level of abstraction, which has also been suggested by the experiments of Rosch (1978) on category emergence and prototype establishment.

The next section describes, on one side, the main mechanisms involved in auditory processing and, on the other side, its functional building blocks, from the sensory input to the recognition of the source of a single stimulus. Hence, it discusses the perceptual processes and concepts involved in the recognition of an arbitrary sound source based only on the sound it emits.

2.1.1 Basic principles of human auditory perception

This section covers a review of several important processes involved in human auditory perception. In particular, we first discuss the basic low-level mechanisms, which are common to all auditory processing. At higher levels of the processing chain we focus more specifically on the mechanisms necessary for sound source recognition. Hence, this section should serve as an overview of all related perceptual processes and will build the basis for the understanding of the developed computational approaches later in the thesis. However, we do not claim, in any respect, completeness regarding the concerned concepts.

2.1.1.1 The peripheral auditory system

In general, the human auditory periphery can be regarded as a connected system consisting of successive stages, each with an input and output (Moore, 2005a). Some of these devices behave in a somehow linear (e.g. middle ear), while others in a highly non-linear manner (e.g. inner ear). The following reconstructs the paths of an arriving sound through the different stages of the peripheral auditory system. It should be noted that many of the described mechanisms were studied by experimentation on animals or human cadavers, hence their real functionality in a living human organism may differ from the experimental results. Moreover, many of the involved processes, mostly the higher-level mechanisms, are still experimentally unexplored, thus the examination of their behaviour and functionality is largely of speculative nature.

First, the pinna modifies the incoming sound by means of directive filtering, which is mostly used for determining the location of the sound-emitting source. The sound then travels through the outer ear canal at which end it causes the eardrum to vibrate. Compensating for the impedance mismatch between outer and inner ear, the middle ear then transforms the oscillation pattern to the oval window, the membrane in the opening of the cochlea, the main part of the inner ear. Both outer and middle ear again apply a filter to the sound, emphasising mid frequencies in the range of 0.5 to 5 kHz, important for speech perception, while suppressing very low and high ones.

The cochlea itself represents a conical tube of helical shape, which is filled with almost incompressible fluids. Along its length it is divided by two membranes, one of which is the Basilar membrane. A vibrating oval window applies the respective pressure differences to the fluid, causing the Basilar membrane to oscillate. Since the mechanical properties of the Basilar membrane vary along its length, this transformation process acts as an effective frequency-to-place mapping; the location of the maximum displacement only depends on the stimulus frequency. Hence, for complex sounds, the Basilar membrane acts like a Fourier analyser, separating the individual frequency components of the sound into distinct vibration maxima along its length (Plomp, 1964; von Helmholtz, 1954). This Fourier behaviour is however no longer valid for close-in-frequency components, mostly due to the limited frequency resolution of the Basilar membrane; the patterns of vibration interfere, causing a more complex movement of the membrane. We will later revisit this phenomenon by reviewing the frequency selectivity and masking properties of the human auditory system.

The displacements of the Basilar membrane directly activate the outer hair cells and indirectly excite the inner hair cells, which create action potentials in the auditory nerve. The functionality of the

outer hair cells is believed to actively influence the mechanisms of the cochlea, controlling sensitivity and fine tuning. It is further assumed that the outer hair cells are partly top-down controlled, since many of the nerve fibres connecting the brain's auditory system with the cochlea contact with the outer hair cells. Here, Moore (2005a) remarks the following:

> "It appears that even the earliest stages in the analysis of auditory signals are partly under the control of higher centers."

The aforementioned frequency-to-place mapping characteristics of the Basilar Membrane is preserved as a place representation in the auditory nerve. High frequencies are encoded in peripheral parts of the nerve bundle while the inner parts are used for transmitting low-frequency information. Hence, the properties of the receptor array in the cochlea represented as frequency, or tonotopic map are preserved up to the brain. Besides, this tonotopic representation is believed to play a fundamental role in the perception of pitch (see the next section).

The physical properties of the incoming sound are directly translated to the neurons' firing characteristics. First, the stimulus intensity is encoded in the firing rate of the activated neurons. Second, the fluctuation patterns of the nerve fibres are time-locked to the stimulating waveform. The frequencies of the incoming sound components are additionally encoded in the temporal properties of the neurons' firing, which occur phase-locked, i.e. roughly at the same phase of the component's waveform.

Regarding the aforementioned frequency selectivity and masking properties of the human auditory system, von Helmholtz (1954) already assumed that the behaviour of the peripheral auditory system can be modelled by a filter bank consisting of overlapping bandpass, i.e. auditory filters. The auditory system separately processes components of an input sound that fall in different auditory filters, while components falling in the same filter are analysed jointly. This defines some of the masking properties of the auditory system; concurrent components can mask each other, depending on their intensity and the frequency ratios (Moore, 1995). Experimentally determined masking patterns reveal the shape of the auditory filters with respect to their masking properties. Moreover, the form of theses auditory filters along the frequency axis also determines human abilities to resolve components of a complex tone. Remarkable here is that only the first 5 to 8 partials of harmonic sounds, as produced by most musical instruments, are processed separately (Plomp, 1964; Plomp & Mimpen, 1968), the rest is perceived as groups with respective group properties (Charbonneau, 1981).

The subjective masking strength of a given stimulus strongly depends on the stimulus's kind along with the context at hand. Here, we can differentiate between informational masking, which occurs if the same kinds of stimuli are involved in the masking process, e.g. masking speech with speech, and energetic masking, attributed to no contextual dependencies between the participating sounds, e.g. masking with noise (Yost, 2008). The former is evidently more difficult to process for the human auditory system since it is assumed that the brain performs a kind of segregation of the two informationally similar sounds.

2.1.1.2 The basic dimensions of sound

On top of the peripheral processing, the auditory system performs a computational analysis of its input, concurrently extracting basic perceptual attributes from the incoming neural fluctuation patterns (Levitin, 2008). In this context, literature usually emphasises the difference between the physical and perceptual qualities of a sound (Licklider, 1951; Scheirer, 2000). Physical sound properties can be measured by means of scientific instrumentation, perceptual qualities are however defined by human perception and thus highly subjective. Anyhow, we can identify the physical correlates of these perceptual attributes, linking the physics with the perceptual sensation. Here, some relations can be found quite easily, e.g. the frequency-pitch relation, while others exhibit a more complex relationship, e.g. the physical correlate of timbre sensation. Besides, the human auditory system extracts these perceptual dimensions in a time span between 100 and 900 ms, depending on the respective attribute (Kölsch & Siebel, 2005).

In what follows we review the most important perceptual dimensions of sound. These include three of the primary perceptual attributes of sound, namely loudness, pitch, and timbre. The auditory system extracts these attributes, among others, in parallel, i.e. independently from each other, and in a both bottom-up and top-down controlled manner (Levitin, 2008).

Loudness. Corresponds to the subjective sensation of the sound's magnitude. The American National Standards Institute (ANSI) defines it as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud". Loudness sensation is highly subjective, hence difficult to quantify. Following several perceptual experiments, Stevens (1957) suggested the perceived loudness to be proportional to the sound's intensity raised to the power of 0.3. Thus, the loudness represents a compressive function of the physical dimension of intensity. Moreover, Moore (1989) notes that the perceived loudness is related to the sound's acoustic energy it exhibits at the position of the listener, on the duration of the stimulus (up to a certain length loudness increases with duration), and on the sound's spectral content.

Pitch. Pitch is a perceptual dimension that describes an aspect of *what* is heard. The American National Standards Institute (ANSI) formally defines it as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale". In contrast to musical pitch, however, pitch sensation is highly subjective, hence difficult to measure by scientific means. Since it forms the basic element of musical melodies as well as speech intonation, pitch represent a musical and perceptual key concept.

For a simple periodic sound pitch roughly correlates to its fundamental frequency. For complex, harmonic sounds the pitch is merely defined by the lower harmonics than by the fundamental (Moore, 2005b; Schouten, 1970). This has been derived from studying the perceptual phenomenon called *missing fundamental* (e.g. Ohm, 1873), stating that the pitch of a given sound is not determined by the presence/absence of its fundamental frequency.

Hearing research derived two main theories for the perception of pitch. The first, so-called *place theory* assumes that the auditory system determines the pitch of a sound by the location of the excitation in the cochlea's receptor array. On the contrary, the *temporal theory* attributes the phase-locking mechanism of the auditory neurons to determine the pitch of an incoming sound. In recent years

many researchers, however, believe that the actual pitch perception is based on principles borrowed from both aforementioned theoretical approaches (Moore, 2005b).

Music psychology developed several models of pitch perception, among which the most famous is probably the 2-dimensional representation proposed by Shepard (1964). His helical model differentiates between the dimension pitch *chroma* and pitch *height*. It reflects the circular characteristics of perceived pitch proximity and similarity, as observed in psycho-acoustical experiments. Here, *chroma* represents the pitch in the 12-stage chromatic scale of Western music, while *height* refers to its octave belongingness.

There have been attempts to estimate a quantisation of pitch in terms of a perceptual scale based on psycho-acoustical evidence. Stevens & Volkmann (1940) constructed a mapping of frequency values in Hertz to values of units of subjective pitch, entitled *mel*[2], in tabulated form. The authors evaluated comparative judgements of listeners on distance estimations of pitches, thereby assessing the dependency of perceived pitch on frequency. The parametric representations of this scale (see e.g. Fant, 1974) represents an approximation of the aforementioned experimental data. As it roughly approximates the non-linear way human pitch perception changes as a function of frequency, this scale has been incorporated into the Mel Frequency Cepstral Coefficients (MFCCs) to measure the shape of a sound's frequency spectrum, one of the most important descriptors for the perceptual sensation of timbre (e.g. Jensen et al., 2009; Logan, 2000; Nielsen et al., 2007; Rabiner & Juang, 1993).

Timbre. In this work, the concept of perceptual timbre obviously plays the most important role of the here-considered basic dimensions of sound. It however exhibits the most complex relationship of the sound's physical attributes to its perception. In this regard, its formal definition by the American National Standards Institute (ANSI) leaves a rather big room for interpretation[3]:

> "[Timbre represents] that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

Bregman (1990) is further emphasising this rather imprecise conception by writing, with respect to the definition of timbre[4],

> "This is, of course, no definition at all [...] We do not know how to define timbre, but it is not loudness and it is not pitch."

[2]Besides, the name *mel* was literally derived from the word *melody*.

[3]An extensive list of various definitions of timbre throughout literature can be found at http://acousticslab.org/psychoacoustics/PMFiles/Timbre.htm, substantiating the interpretative character of this perceptual concept.

[4]Due to its non-existing physical correlate, the adoption of the term "timbre" in scientific research brought much debate into the discipline of hearing science. Timbre, as a purely perceptual quality, lacks any direct relation to a sound's physical parameter, hence a quantification in a scientific sense is impossible. In this regard, especially Martin (1999) criticises that timbre " *... is empty of scientific meaning, and should be expunged from the vocabulary of hearing science*". In this thesis, we will however frequently apply the term in order to refer to the corresponding perceptual sensation elicited by any sound stimulus. Moreover, timbre exhibits a strong interrelation to the musical concept of instrumentation, which represents an important consideration in the subsequent chapters.

This difficulty in defining the concept is somehow rooted in the multidimensional character of timbre – unlike loudness and pitch, which are unidimensional quantities. Furthermore, the perceptual mechanisms behind the sensation of timbre are yet not clear. Handel (1995) suggests two possible answers to the general perception of timbre; either timbre is perceived in terms of the actions required to generate the sound event, which would coincide with the ecological notion of the production, or transformational invariants. This would allow us to recognise the object despite large changes in other acoustical properties. The second possible perspective refers to the separate perception of the underlying dimensions. In this case, the perceptual system learns the particular connections of the different features to the respective auditory objects.

A lot of work has gone into the identification of the underlying dimensions of timbre perception. Here, most works applied the technique of Multidimensional Scaling (MDS), evaluating perceptual timbre similarities. One of the early researchers using this technique in timbre research, Grey (1977) writes, explaining the underlying hypotheses of MDS studies:

> "[The researcher] may start with the perceptual judgements of similarity among a diverse set of (naturalistic) stimuli, and then explore the various factors which contributed to the subjective distance relationships. These factors may be physical parameters of the stimuli, which then would lead to a psychophysical model; yet, multidimensional scaling techniques may also uncover any other factors involved in judgement strategies."

Up to now, many researchers studied the perceptual dimensions of timbre via MDS, using different sets of stimuli and experimental conditions (e.g. Caclin et al., 2005; Grey, 1977, 1978; Iverson & Krumhansl, 1991; Kendall & Carterette, 1993; Lakatos, 2000; McAdams et al., 1995). These works usually use short isolated sound stimuli, originating from natural or specially synthesised sources[5]. The participants of the experiment are then asked to rate the similarity/dissimilarity of all tone pairs from the set of stimuli. Upon these ratings, the MDS algorithm produces a geometrical model of the "timbre space", wherein the different stimuli are represented as points and the respective distances refer to their dissimilarities. The space-spanning dimensions are later interpreted in terms of acoustic, perceptual, or conceptual attributes and often related to computational descriptions of the respective sounds.

Remarkably, the only dimension revealed in all of these studies relates to the brightness of the stimuli, which is attributed to the concept of the spectral centroid. Other attributes found in these works refer to the attack and decay transients (e.g. Lakatos, 2000), the (time-varying) spectral shape (e.g. McAdams et al., 1995) – here, the perceptual important amplitude and frequency modulation may play an important role – or the spectrum's fine structure (e.g. Caclin et al., 2005). Hence, from this observation it is evident that the perceptual space revealed in MDS studies strongly depends on the input stimuli. In this regard, the subjects' similarity ratings seem to strongly depend on the respective context (Hajda et al., 1997).

[5] McAdams et al. (1995), for instance, synthesised special artificial sounds to simulate timbres falling "between" natural musical instruments, in order to further substantiate the validity of their results.

Therefore, these studies have met some criticism; first, due to the limited number of both stimulus pairs and categories, together with the special conditions of the stimulus presentation applied in the respective works, it seems hard to generalise the obtained results. In this context, the aforementioned studies completely neglected the contextual components of timbre perception, which generates, by no means, a realistic testing scenario. Furthermore, the technique only reveals continuous dimensions, though timbre perception is assumed to be at least partially influenced by categorical attributes, e.g. the hammer noise of the piano or the breathy characteristics of the flute. But one of the main critics, however, arises from the conception of MDS studies itself. In this regard, Scheirer (2000) writes, criticising the lacking re-evaluation of the identified dimensions with computational models of timbre perception:

> "The testable prediction that is made (often implicitly) by such a research model is that it is these particular properties that are really used by a listener to identify objects from their sounds. It is incumbent upon those researchers who wish to assert the continued utility of the multidimensional-scaling paradigm for timbre research to conduct such computational studies to confirm that these properties contain sufficient information to support the behaviors imputed to them."

To correct for the downsides of some of the aforementioned MDS studies, several authors responded to the potentially misleading insights obtained from too-constrained experimental settings. In particular, both the acoustical properties of the stimuli related to the revealed dimensions and the influence of musical context were subject to consideration.

Since most earlier MDS studies used very short sounds, i.e. mostly discarding the steady-state of the stimulus, Iverson & Krumhansl (1991) evaluated the influence of the stimulus's length and its respective sub-parts on the similarity ratings. The authors found high correlations between the results obtained from the attack part, the steady-state part, and the entire stimulus. This suggests that the cues important for stimulus similarity, and thus presumably also for source recognition and identification, are encoded independently of the traditional note segmentation. Each here-analysed part of the signal, i.e. attack and steady-state, separately provides important acoustical information for similarity rating and hence source recognition.

Kendall (1986) revealed the significance on musical context on the categorisation abilities of humans using sounds from musical instruments, which are regarded as a direct representation of timbre (see also Chapter 3). The study compares the performance of a whole-phrase to a single-note context, at which the former indicates phrases form complete folk-songs and the latter thereof extracted single notes. Furthermore, the author explores the effect of transients and steady-state on the performance in the respective context by editing the various stimuli. Results suggest that transient components are neither sufficient nor necessary for the categorisation of the instruments in the whole-phrase context. Moreover, transient-alone stimuli led to the same results than full notes and steady-state-alone settings in the isolated context. In general, Kendall identified the whole-phrase context to yield statistically significant superior categorisation performance than the isolated-note context, emphasising the importance of musical context in this kind of perceptual mechanisms.

Sandell (1996) performed a musical instrument identification experiment in which he tested subjects' abilities in dependence on the number of notes presented from a recorded arpeggio. Here, results indicate that the more notes are presented, the higher the identification performance of the subject, hence emphasising the role of simple musical context for source identification (see also Chapter 4 for its ubiquitous presence in our algorithmic implementation). In this context, Grey (1978) notes with respect to the simplistic harmonic and rhythmic contextual settings used in this early experiment, though already foreseeing the importance of musical context on the perception of timbre,

> "I hoped to begin to understand the effects of context on timbre perception. I believe that studies using musical contexts will have a greater relevance to normal perceptual experience than those which merely concentrate on tones in isolation"

Furthermore, Grey also concluded that attacks are of minor importance compared to steady state for timbre discrimination in a musical context.

To validate these results obtained from perceptual examination, Essid et al. (2005) performed an automatic instrument recognition experiment with separated attacks and steady-states. Their first note, however, relates to the non-triviality of extracting the attack portion of a sound even from monophonic audio signal. The performed experiments show that in short isolated frames (45 ms and 75 ms), the attack provides on average better estimates than the steady-state alone. A system mixing both attacks and steady-state frames, again on a short time basis, then yielded nearly the same performance as the attack-only system. However, systems using a much larger decision window (465 ms and 1815 ms), not considering the distinction between attacks and steady-states, performed by far best, yet another indication for the important role of musical context even for automatic recognition systems.

The perception of polyphonic timbre was by far less studied in literature. Noticeable here are the works performed by Alluri & Toiviainen (2009; in Press), exploring perceptual and acoustical correlates of polyphonic timbre. In the first study, the authors performed MDS, correlation, and regression analysis (RA) of similarity ratings obtained from Western listeners on polyphonic stimuli taken from Indian music. Revealed acoustic dimensions include *activity*, *brightness*, and *fullness* of the sound. Here, the sub-band flux, measuring the sound's spectral difference in 1/3 octave bands, represents the most important computational description of the timbral dimensions, highly correlated to both the activity and fullness factor. The brightness dimension however does not reveal such a evident correlation with one of the applied audio features. Surprisingly, the MFCCs showed no significant correlation with the identified perceptual factors, suggesting a re-evaluation of the dimensions in subsequent computationally modelling experiments (see Scheirer's criticism above on the MDS paradigm).

In a follow-up study, Alluri & Toiviainen (in Press) followed the same experimental methodology but using listeners from both Western and Indian culture, hence estimating the cross-cultural dependencies of the perception of polyphonic timbre. The results suggest that familiarity with a given culture, e.g. Indian listeners rating stimuli taken from Indian music, leads to a finer estimation of the

dimension of the perceptual timbre space, here the authors found a value of 3 dimensions for both settings. Cross-cultural ratings, however, revealed only 2 dimensions in the respective perceptual space. Moreover, the interpretation of the identified dimensions coincide with the ones obtained from the first study, both for intra- and cross-cultural testing (again, the dimensions *activity* and *brightness* were the most explanatory in the different tests). Finally, one of the major insights of these works is the overlapping of the dimensions identified in experiments using monotimbral data and the here-obtained ones. This suggests that the timbre perception of multi-source sound mixtures is based on the analysis of the compound signal constituting of the involved sources. Source recognition can therefore be seen as independent process, that happens concurrently or subsequently to the initial timbre perception, i.e. the mixture is segregated and the individual sources recognised successively.

2.1.1.3 Higher-level processing

Auditory learning. According to von Helmholtz (1954), information obtained from the raw sensory input is ambiguous and therefore complemented by cues taken from prior knowledge. Much effort has been taken to identify the role of this prior knowledge, but not much has been gained beyond the peculiarities of the individual studies (Lufti, 2008). Here, the difficulties arise from the fact that recognition takes place at different levels of abstraction (see above), as well as the subjective nature of the prior knowledge.

There is a high consensus among researchers that auditory knowledge is acquired in an implicit manner. It is believed that humans are highly sensitive to the stimuli's contingent frequencies, i.e. probabilities, which form the basis for anticipatory behaviour regarding the future. Hence, the perceptual system learns properties of auditory objects and events by mere exposure (Reber, 1967). In this context, the exposure allows for both the acquisition of an abstract representation of theses objects and the formation of predictive expectations (Cont, 2008; Hazan, 2010).

Many works studied the implicit learning mechanisms inherent to human auditory perception. Saffran et al. (1999) showed that humans already perform such learning schemes at the age of 8 month by testing both adult and infant listeners in a grammar acquisition experiment using note triplets. Loui & Wessel (2006) confirmed these results by using tonal sequences derived from non-Western scales in the same experimental context. Both studies showed that subjects were able to learn the exposed grammar by recognising thereof generated melodies. Similarly, Tillmann & McAdams (2004) added timbral information to the tone triplets used in the aforementioned studies in order to estimate the influence of factor dependencies on the implicit learning capabilities of humans. The authors used timbral distances related to the statistical regularities of the tones by using different musical instruments (i.e. high timbral similarity corresponds to intra-word transition, while low similarity relates to inter-word transitions). Results revealed that subjects do significantly better in learning the code words provided the respective timbral cues. This emphasises the importance of timbral information in learning and recognition from music.

Moreover, Krumhansl (1991) showed that judgements about tonal "fit" are highly consistent among subjects, indicating the learnt nature of these predictions. Participants of the experiments were asked to rate how well different tones fit within in a given tonal context, established by either a melody line

or a harmonic progression. In this context, Serrà et al. (2008) pointed out that when analysing the statistical distribution of automatically extracted tonal information in terms of pitch class profiles (PCP) from large music collections, a remarkable analogy to the "tonal hierarchies" of Krumhansl could be observed. Shepard & Jordan (1984) reported a similar effect regarding the statistical learning of musical scales; in their experiment subjects mapped heard scales exhibiting equidistant notes to mental traces of acquired musical scales (i.e. major/minor), reporting perceived differences in the interval sizes. Finally, Bigand et al. (2003) showed a comparable behaviour in a harmony context, where listeners were able to identify spurious, "wrong" tones in a functional tonic chord in a more accurate way than in a functional sub-dominant, which in general is less probable in the tested musical context. See also the recent works by Hazan (2010) and Cont (2008) for a more detailed review.

The implicit character of learning is also manifested in the fact that for some experimental tasks, music experts do significantly better than novices (Crummer et al., 1994; Kendall, 1986). Moreover, explicit training of subjects leads to an improvement in performance compared to untrained subjects (Jordan, 2007; Sandell, 1996).

The acquired knowledge forms the basis for creating, mostly subconsciously, expectations regarding the acoustical environment [6]. Many authors regard the process of evaluating these expectations with the actual sensory information as a basic means for survival in a continuously sounding world. Literature from research on music processing developed several theories about the nature and functionality of this mutual process (e.g. Huron, 2006; Meyer, 1956; Narmour, 1990). Meyer (1956) was one of the first acknowledging expectations to be the main source for the perceived emotional qualities of music. Narmour (1990) expanded this theory, further constructing a computational model for melodic perception. Finally, Huron (2006) takes it to the next level by stating that music perception per se is a result of successively evaluating expectations by the auditory system. Moreover, Huron notes that composers purposely guide listeners' expectations by establishing predictability or creating surprise in their works.

Similarity and Categorisation. Given a proper representation, or cue abstraction (Deliege, 2001), of the sensory information related to the auditory event to identify, how does the auditory system retrieve the relevant information from memory? As part of the above-introduced general model of the auditory recognition process, the concepts of similarity, categorisation, and contextual information play an important role. Following Cambouropoulos (2009), the concepts of similarity and categorisation are strongly linked. In a famous work, Rosch (1978) studied how the perceptual system groups similar entities into categories along with the resulting category prototypes. The emerging categories represent partitions of the world and are both informative and predictive, i.e. the knowledge about an object's category belongingness enables the retrieval of its attributes or features.

Literature derived three main theories of categorisation based on different assumptions on their mental representation. The *classical*, or container theory assumes that categorisation is defined by a set of rules derived from attributes which define the respective categories. The *prototype* theory uses a model which estimated probability given the input data results in the respective category decision.

[6] In his influential work, Huron (2006) introduces four kinds of musical expectations. The veridical, schematic, and dynamic-adaptive expectations corresponding, respectively, to the episodic, semantic, and short-term memory are of subconscious kind. Conscious expectations of reflection and prediction constitute the forth one.

Finally, the *exemplar* theory relies on a set of examples that resembles the mental representation of the given category (see Guaus (2009) for a more detailed discussion).

On the basis of the performed categorisation, recognition and identification is accomplished. In this context, identification describes the process of assigning a class label to an observation. The specific taxonomy or ontology defines the respective verbal descriptions, or labels of the categories. Moreover, the retrieved associated knowledge positions the auditory object in the context at hand and enables the evaluation of its significance. In conclusion, Cambouropoulos (2009) notes, regarding the highly contextual, thus complex nature of the entire categorisation process:

> "It is not simply the case that one starts with an accurate description of entities and properties, then finds pairwise similarities between them and, finally, groups the most similar ones together into categories. It seems more plausible that as humans organize their knowledge of the world, they alter their representations of entities concurrently with emerging categorizations and similarity judgments. Different contexts may render different properties of objects/events more diagnostic concurrently with giving rise to certain similarity relationships and categorisations. If context changes, it affects similarity, categorisation and the way the objects/events themselves are perceived."

2.1.2 Understanding auditory scenes

In general, the acoustical environment does not provide the sound sources in isolation. The thereof obtained auditory sensory information rather involves multiple sound sources, presumably overlapping both in time and frequency. The ability of human perception to resolve this acoustical mixture forms the basis for the analysis of the acoustic scene. However, the perceptual mechanisms behind are still not well understood (Carlyon, 2004). Since music represents, in general, a multi-source acoustical environment (see Section 1.3), the here-described data properties indeed represent the main complexity involved in this thesis.

Cherry (1953) coined the problem as the *cocktail party problem* by exemplifying a conversational situation where several voices, overlapping in time, are embedded in a natural acoustical environment including other stationary or dynamic sound sources. The listener, however, is able to focus on the targeted speech stream and transform the acoustical data into semantic information. In particular, Cherry writes:

> "One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others...we may call it the cocktail party problem."

Bregman (1990, p.29) draws an analogy to vision to emphasise the complexity of the problem. He writes

> "Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?"

And also Levitin (2008) uses a metaphor for describing the cocktail party problem,

> "Imagine that you stretch a pillowcase tightly across the opening of a bucket, and different people throw Ping-Pong balls at it from different distances. Each person can throw as many Ping-Pong balls as he likes, and as often as he likes. Your job is to figure out, just by looking at how the pillowcase moves up and down, how many people there are, who they are, and whether they are walking toward you, away from you, or are standing still. This is analogous to what the auditory system has to contend with in making identifications of auditory objects in the world, using only the movement of the eardrum as a guide."

In this context, automatic recognition of musical instruments from polytimbral music represents a special variant of the cocktail party problem. Though in opposite to the classic example of different human voices in a noisy environment, the involved sources are by no means independent in music. Usually composers adopt the musical instruments following distinct rules, which, depending on the current praxis in the respective period, may include voice-leading constraints, harmonically-rooted specification, timbral conceptions of the composer to no rules at all. Hence, the musical instruments act together to form the harmonical, timbral, and emotional affection – to name just a few – of music.

Bregman (1990) describes the processes necessary for decoding the information provided by the acoustical scene into understanding. This influential work, naming the field Auditory Scene Analysis (ASA), provides a theoretical framework for research in the field along with numerous experimental evidence for the described auditory principles. The underlying approach towards auditory perception is strongly influenced by the organisational principles of Gestalt psychology (see the very beginning of this section), though Bregman argues that besides this bottom-up – he calls it *primitive* – processing, top-down mechanisms – termed *schema-based* – must be involved in general auditory perception.

ASA assumes the auditory system to order the incoming neural information in a primitive, low-level manner, grouping and segmenting the data, composed of frequency, level, and time information, into so-called auditory objects. Here the mechanisms follow *gestaltism* by applying the rules of closure, similarity, proximity, symmetry, continuity, and common fate (Wertheimer, 1923). Hence, the acoustical signal is transduced and transformed into grouped representations according to principles of perceptual organization. The thereby occurring uncertainties in the interpretation of the raw neural codes, resulting from the ambiguity of the sensory data, are resolved by learned pref-

erence rules, which are continuously updated by stimulus information regarding register, timbre, duration, dynamics, etcetera (Temperley, 2004).

This behaviour seems evident since our perception is, for instance, highly sensitive to common onsets and modulations of different components across frequency, or to the frequency relation of partials of harmonic sounds. Moreover, natural sounds vary slowly in time, hence proximity and continuity play an important role in auditory perception. Here, Bregman introduces the notion of "old-plus-new", stating that an unknown auditory scene is first analysed in terms of the already-known; what is left is then attributed to a "new" object. Also, the inferential character of the auditory system, as demonstrated in the auditive restoration phenomenon as shown by Warren (1970), may be partially explained by these rules.

In general, the auditory system is not able to analyse the properties of a sound event until its consistent components are integrated as a group and segregated from those of other sound events. Hence, auditory perception has to perform a kind of separation of meaningful events in both frequency and time, a process that is commonly known as *stream segregation* or *auditory streaming* (Bregman, 1990). In this context, the inherent limitations of the human brain in processing information are controlling the amount of concurrent streams[7]. Moreover, the temporal ordering of the auditory objects and events plays an important role (Hazan, 2010). Most acoustical cues are somehow correlated across time insofar that they become redundant and substitutable to a certain extent. This property can partially explain the effects of auditory restoration of masked sound events and therefore enables robust source recognition in noise (Handel, 1995).

Those cues involved in the streaming process can be of different kinds. They may be of low-level nature as described by the gestalt principles or higher-level concepts such as timbre, pitch, loudness, harmony, etcetera. At this stage, top-down processing is heavily involved in the formation of these auditory streams. Given the sensory data, the most likely – across senses – constellation of auditory objects will form the respective streams. Moreover, depending on the listening experience, those cues which lead to the best performance are selected to control the formation process of the auditory streams. Furthermore, selectivity, adaptation, and attention interactively control the process (Carlyon, 2004; Yost, 2008). In this context, an auditory stream may – but not necessarily has to – correspond to a single acoustic source.

It has been shown that this ability to form auditory streams from complex acoustical mixtures is already partially present in newborns (Winkler et al., 2003). It therefore seems that most of the low-level processes of the auditory systems are innate (Crawley et al., 2002), while the ability and power of the schema-based control evolves with experience.

In the context of this thesis, streaming-by-timbre takes a special role. It describes the process of auditory streaming based on timbral cues, hence it can be understood in the sense of how the perceptual system segregates sound sources according to their sounding characteristics (Bregman, 1990). Research put some effort in studying this perceptual mechanism, mainly driven by the question of what factors influence the separability of sound sources (e.g. Singh, 1987; Wessel, 1979). Here, mostly musical instruments were adopted to create different timbre sensations. It has been shown

[7] We will address these limitations in more detail in Section 3.2.2, see also the works by Miller (1956) for a quite general, and Huron (1989) for a music-specific assessment of human information processing capabilities.

that especially both the static and dynamic spectral characteristics of the sound are decisive for the streaming abilities of concurrent timbres. These properties correspond to the formant areas and small spectral variations inherent to musical instruments (Reuter, 2009) (see also Section 3.2.2).

Once the auditory system has segregated the sensory data, the perceptual streams are analysed. Here, source recognition is based on the extraction of features describing the sounding object. In this context, Martin (1999) notes that humans have to extract source-invariant information already from the incoming, thus unresolved, audio stream to reliable segregate and categorize. It should be kept in mind that the process of feature extraction, together with the auditory attention mechanisms, is involved in both segregation of concurrent sources and the subsequent analysis after segregation (Yost, 2008).

As already stated above, the process of auditory streaming is assumed to be based on both primitive, i.e. bottom-up, and schema-based mechanisms. On the one side, low-level processes successively transform the input into elementary symbolic attributes of the sensory stimulus. Here, the above-described mechanisms of perceptual organisation take place (Bregman, 1990). This process also conforms with the theory of visual perception by Marr (1982), who viewed the perceptual process as a successive series of computational stages. Hence, the perceptual system performs a successively abstraction of the input data, creating several levels of data representation. Each of these levels encodes a different kind of information, at which higher levels contain a more semantic description of the stimulus. The stages are assumed to be rather independent, each stage can therefore be modelled separably, at which the concrete processing can be accomplished locally, i.e. is not influenced by other stages. Finally, the combination of all models of all stages yields the complete system.

On the other side, top-down processes take control of the various stages in the data processing chain. Here, the auditory system compares, at each level in the hierarchy, a mental representation of the acoustical environment to the actual sensory data. This mental representations are created by both short-term and long-term prior knowledge regarding the data. According to the resulting match, the perceptual system adapts both its low-level sensory processing and the mental representation. This top-down control is most likely accountable for perceptual phenomena such as completion or residual pitch. See the works of Slaney (1995) and Ellis (1996) for more detailed evidence of the involved schema-based processes in audition.

2.2 Machine Listening

Machine listening represents the area of research that teaches computers to generate an abstract representation of a sound signal. Hence, it involves the automatic analysis and description of the given auditory scene for extracting meaningful information. Since we assume that the meaning of the information is defined by the human mind[8], the performance of machine listening systems should always be evaluated against human abilities on the corresponding task at hand. However, engineering-

[8] As already mentioned earlier it is human perception and cognition that define music (Wiggins, 2009).

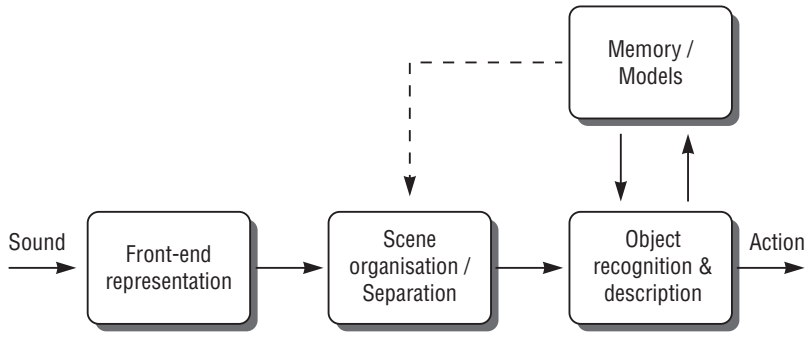


Figure 2.3: Processes involved in machine listening after (Ellis, 2010).

motivated criteria often define the evaluation context, since many applications, although inspired by human behaviour, are neither interested in the perceptual abilities of humans nor on mimicking them (Scheirer, 2000).

Tzanetakis (2002) identified the following stages involved in machine listening – he uses the term *computer audition* – and connects them to the most relevant research areas:

1. **Representation.** Refers to the transformation of the time-domain acoustical signal into a compact, informative description by simulating the processes of the auditory periphery. Here, Tzanetakis exemplifies the time-frequency transformations usually applied in machine listening systems for a proper representation of the frequency content of a given signal. The most important area of research referring to these signal transformations is *signal processing*.

2. **Analysis.** An understanding of the acoustical environment is obtained from the given representation. The processes applied here may include similarity estimation, categorisation, or recognition, which include abstract knowledge representations and learning mechanisms for both humans and machines. The main research area here is the field of *machine learning*.

3. **Interaction.** The user is actively involved in the process of the presentation and control of the extracted information from the signal. Here, ideas and concepts from *human-computer interaction* have major influence.

In this line, Ellis (2010) describes the processes involved in the representation and analysis stages of machine listening as follows; a front-end processing transforms the signal into a proper representation for the analysis, on top of which an organisation, or scene analysis algorithm extracts the relevant objects. Then, recognition and description takes place by consulting memory representations, which store information on the objects, and moreover act as an adaptive, top-down control for the scene analysis component. Figure 2.3 illustrates the processes involved.

In the context of this thesis, which addresses the problem of automatic recognition of musical instruments from music audio signals, the key process involved represents the scene analysis stage. Here, literature in machine listening has developed two conceptually different approaches to resolve

a complex mixture signal and extract the relevant objects. Scheirer introduced the terminology relating to these different viewpoints on the problem and inspired many subsequent studies by his influential works (e.g. Scheirer, 1996, 1999, 2000). In particular, the author defines the following two general methodologies:

1. **Separation for understanding.** This approach assumes that a successful extraction of a music understanding requires a central representation of the input data. It borrows the idea of structuralism stating that cognitive processes are based on symbolic models of the sensory data. In particular, the specific symbolic model of audition is rooted in music theory and its fundamental concept of the score. The central representation is correspondingly a transcription of the sensory data into a score-like description. In the machine listening field, this entire conception is often termed the *transcriptive model*. Hence, a full understanding of the acoustical environment requires a piano-roll-like representation of the input signal according to music theoretical entities, which can be used to separate the mixture into the concurrent sources. Typical systems either apply the cues obtained from the transcription to segregate the entire signal into the sources or to directly synthesise the source signals. The isolated signals can then be analysed separately in terms of the extracted features with respect to the desired information.

2. **Understanding without separation.** In relation to human perception mechanisms this approach assumes that the lack of any structural representation of the sensory data leads to an iterative abstraction of meaningful information directly from the input signal. Here, such systems apply continuous transformations to the input signal until the desired information is accessible. A completed transformation stage and its abstractions define the input to the next, higher-level stage of meaning extraction. Typical implementations usually use simple signal processing and pattern recognition techniques to directly infer judgements about the musical qualities of the stimulus. This *signal understanding* approach can be regarded as a class of sensor interpretation problems known from general artificial intelligence; the goal is to abstract the signal into a symbolic stream so that the most meaningful elements are exposed, while other agencies can operate on deeper qualities of the source (Hawley, 1993). To emphasise the conceptual advantages of this approach, Scheirer (2000) writes

   > "In a separation-less approach, the required action is one of making feature judgements from partial evidence, a problem that is treated frequently in the pattern recognition and artificial intelligence literature. Rather than having to invent a answer, the system can delay decision making, work probabilistically, or otherwise avoid the problematic situation until a solution presents itself."

Due to the apparently appealing challenge of constructing automatic music processing systems based on the transcriptive model there is a vast amount of related approaches in literature. In recent years, researchers paid specific attention to the problems of source separation (e.g. Casey, 1998; Smaragdis et al., 2009; Vincent et al., 2010; Virtanen, 2006; Weintraub, 1986) and automatic music transcription (e.g. Abdallah & Plumbley, 2004; Goto, 2004; Klapuri, 2003; Moorer, 1975; Smaragdis & Brown, 2003). Alternative approaches towards music understanding however claimed the invalidity of the approach with respect to the human processing of music signals and enforced the signal

understanding approach (e.g. Ellis, 1996; Herrera et al., 2000; Scheirer, 1996, 1999). In this regard, the authors argue that a simultaneous separation of the audio signal into the concurrent sources cannot account for components of the signal that are masked or shared by different sources. Hence, the process directly involves an information loss that is not present in signal understanding systems. Moreover, most listeners do not transform the sensory data into a score-like representation. On the contrary, the organism produces various output mechanisms related to the perceived qualities of music such as foot-tapping, emotional responses, or high-level judgements about musical genre or style (Levitin, 2008). Besides, Martin et al. (1998) argue that music transcription should only be viewed as an engineering problem, possibly of interest for practical applications, rather than as a prerequisite for music understanding. In this context, Scheirer (2000) writes

> "if useful analyses can be obtained [...] that do not depend on transcription or sound separation, then for many purposes there is no need to attempt separation at all."

Finally, Ellis (1996) notes, quite pessimistically concerning the limits of the transcriptive model

> "The idea of a machine that can convert a recording of a symphony into the printed parts for an orchestra, or a MIDI encoding for storage and resynthesis, remains something of a phantasy."

Since we approach the problem addressed in this work without applying automatic music transcription and musical source separation techniques, this thesis is positioned in the signal understanding field. In this respect, the developed methodologies include inferring the characteristics of the objects to recognise, i.e. the musical instruments, directly from the mixture signal without any form of polyphonic pre-processing (e.g. multi-pitch estimation, onset detection, transient reduction, etcetera).

2.2.1 Music signal processing

This section covers several basic concepts from signal processing necessary for machine listening approaches. We here concentrate on the area of audio signal representations which usually comprises the front-end processing stage of an automatic music processing systems. In what follows we shortly review the most important representations of audio signals as applied in related literature. Similar to the previous section, we however do not claim completeness in any respect.

2.2.1.1 Audio signal representations

In this section we survey the most common representations of audio signals as applied in automatic music processing systems, due to their important role in the recognition process (see Figures 2.1 and 2.3). The most used signal representation is probably based on the Fourier decomposition of

the sound signal due to the similarities to the analysis performed by the Basilar membrane (see Section 2.1.1). In the context of this thesis, especially the Fourier Transform equivalent for finite time-sampled signals, the *Discrete Fourier Transform* (DFT), is applied extensively to transform the input sound into a Fourier representation. The DFT represents a specific case of the additive expansion or decomposition models, which can be generally described by a weighted sum over a set of particular expansion functions. Here, the expansion functions correspond to the pre-defined, frequency-localised complex sinusoidal bases. One of the big advantage of such additive decomposition models over conceptually different signal representations lies in their implementation of the superposition principle; as a direct implication a transformation applied to the mixture signal equals to the weighted sum of the transformations applied to the respective decomposition functions.

The DFT is frequency-, but not time-localised, hence providing no temporal information regarding the applied sinusoidal decomposition. To overcome this shortcoming, the input signal is represented as a sequence of short segments, or frames, on top of which the DFT is performed. Hence, the analysis is shifted along the time axis using a fixed step, or hop size. This process can be regarded as the application of a time-localised window function to the signal prior to the Fourier analysis. Moreover, the specific formulation of including a special window additionally to the sinusoidal into the decomposition function is known as *Gabor expansion*, the resulting expansion functions are called *Gabor atoms*. This time-frequency representation is usually termed *Short-Time Fourier Transform* (STFT). An in-depth study of the STFT and its various interpretations is given by Goodwin (1997).

Typical higher-level signal representation for music processing use the STFT as starting point. Here, sinusoidal modelling techniques have been particularly applied widely across the field, due to their usefulness for the analysis of harmonic sounds. The Sinusoid Transform Coder introduced by McAulay & Quatieri (1986) extracts distinct sinusoidal tracks from the STFT, hence regarding the mixture signal as a sum of partial tracks. The system picks spectral peaks from each STFT frame, the entire mixture signal is therefore represented as a time-varying set of triplets including amplitude, frequency, and phase information of the respective estimated partials. By using a birth-death tracking algorithm the system extracts continuous frequency tracks, which correspond to the sinusoidal components of the analysed sound. Serra (1989) extended this methodology by explicitly considering transient and noise components in the signal model. The author suggested a "deterministic-plus-stochastic" decomposition of the signal, where harmonic sounds are modelled via sinusoidal tracks and the remainder of the spectrum by an autoregressive noise model.

In the context of music signal processing, the *constant Q transform* (CQT) represents a popular alternative to the standard DFT. It has been introduced to avoid specific shortcomings observable with the DFT and to conform the inherent properties of the Western tonal music system (Brown, 1991). In particular, the CQT adapts its frequency resolution to the one of musical scales, while applying complex sinusoids as expansion functions; the subdivision of the octave into intervals of equal frequency ratios in the equal-tempered tuning system results in a logarithmically spacing of the successive notes, hence the CQT offers the corresponding logarithmic frequency resolution. This is in opposite to the standard DFT formulation, which provides a linear spacing of its bins along the frequency axis. More precisely, when viewed from a filter bank perspective[9], this logarithmically

[9]In signal processing, the DFT is often regarded as a bank of band-pass filters. Here, each frequency bin represents a single filter with a constant-length prototype impulse response.

frequency spacing results in a constant *frequency-to-bandwidth* ratio of the filters. This, in turn, leads to a good frequency resolution at lower frequencies together with a good time resolution for higher frequencies. According to the uncertainty principle, which is inherent to any kind of time-frequency decomposition (Burred, 2009), low frequencies thus exhibit bad temporal resolution, while high frequencies provide bad frequency resolution. These frequency-dependent properties of the CQT are however in line with some general characteristics of music, since, inside this duality, high frequencies usually offer strong temporal information while low frequencies only vary slowly in time.

Furthermore, the Wavelet Transform offers a more general multi-resolution frequency transform. It provides the facility to use a large variety of expansion functions, such as Haar or Daubechies wavelets (Mallat, 1999), hence the transform is not necessarily limited to complex sinusoids such as the aforesaid. Since its frequency resolution can be related to the characteristics of the human auditory system, the Wavelet Transform equivalent for sampled signals, the *Discrete Wavelet Transform* (DWT), has been applied for auditory modelling (Moore, 1989). In principle, it performs an octave-band decomposition of the signal, hence providing good frequency resolution for low frequency components and high temporal resolution in the upper regions of the spectrum.

Another frequently used approach is the signal's decomposition via adaptive models using an overcomplete dictionary of time-frequency localised atoms. The main characteristic of decomposition methods using overcomplete dictionaries is their inability to reconstruct the time signal from the derived time-frequency representation. Such models select those atoms from the dictionary which best match the analysed signal. The most common dictionaries consist of, e.g., Gabor atoms or damped sinusoids. Some examples of overcomplete decomposition algorithms include the *Basis Pursuit* (Chen et al., 1999) or *Matching Pursuit* (MP) (Mallat & Zhang, 1993). The latter iteratively subtracts the best match of the dictionary from the signal until some stopping criterion has been reached and has been applied for automatic music processing (e.g. Leveau et al., 2008).

Finally, we review those signal representation which model the auditory periphery processing. Such representations are inherent to computational models of ASA in the field of Computational Auditory Scene Analysis (CASA). In general, these models transform the acoustical signal into a pattern of nerve firing activity. First, the signal is filtered according to the outer- and middle-ear frequency transfer characteristics. Next, such models apply a filter bank consisting of overlapping gammatone filters, simulating the frequency analysis performed by the cochlea. At last, a inner hair cell transduction model is used to account for the compression, rectification, and phase locking properties at this stage of the auditory processing. The resulting time-frequency representation is termed *Cochleagram* (e.g. Brown & Cooke, 1994; Cooke, 1993; Godsmark & Brown, 1999). Often, authors apply an additional autocorrelation analysis to the cochleagram, resulting in the 3-dimensional *Correlogram*, used for the analysis of harmonic sounds (e.g. Ellis, 1996; Martin, 1999; Wu et al., 2003).

2.2.2 Machine learning and pattern recognition

Literature provides many different formulations regarding the definition of machine learning. Following Langley (1996), we suggest a formal, thus quite general attempt.

> "[Machine learning is] a science of the artificial. The field's main objects of study are artefacts, specifically algorithms that improve their performance with experience."

Algorithms from machine learning and especially pattern recognition have been extensively applied in automatic music processing systems. This is partially due to the aim of both pattern recognition and the human cognitive system to determine a robust linkage between observations and labels for describing the current environment. In this regard, Duda et al. (2001) phrases the following:

> "[Pattern recognition is] the act of taking in raw data and taking an action based on the category of the pattern"

Usually, an observation is represented as a n-dimensional feature vector, describing the properties of the observation. This vector, or pattern, represents a point in a multi-dimensional space, in which a machine learning algorithm models the inherent structure of the data in either a supervised or unsupervised manner. The resulting model is able to present evidence for a given unseen observation, according to the learnt criteria. In the following, we first take a closer look at different audio features involved in the hierarchical semantic layers used to describe music, and subsequently review some relevant learning algorithms typically applied in automatic music processing systems.

2.2.2.1 Audio features

In a very broad sense, a feature denotes a quantity or quality describing an object of the world. Thus, it serves as a synonym for attribute or description of the object. Conceptually, it can be regarded as an abstraction in a compact description of a particular information. Hence, it facilitates the handling of noisy data, allows for compression, or can be used to suppress unnecessary details, thus enabling a robust analysis (Martin, 1999)

Music Content Processing (MCP) typically differentiates between hierarchically structured description layers corresponding to broad feature categories – an analogy to the perspective of a hierarchical ordering of information in human perceptual and cognitive system (Martin, 1999; Minsky, 1988). In this regard, a representation addressing these general description layers can be derived, which is depicted in Figure 2.4, showing a graphical illustration of the different levels of abstractions addressed by MCP systems, synthesised from drawings of Celma & Serra (2008).

In a machine listening context, and following the music understanding approach as introduced above, the raw audio signal subsequently passes the three layers of abstractions, processed by the respective transformations. First, such systems derive low-level features from the data which are combined to so-called mid-level descriptors. From these descriptors high-level, human-understandable[10] information regarding the audio signal can be extracted. We can therefore group the extractable

[10]Here, the term *human-understandable* refers to the general case of listeners, hence musical novices which are unfamiliar with most low- and mid-level musical concepts. Human experts, however, may be able to extract meaningful information yet from the extracted mid-level representation.

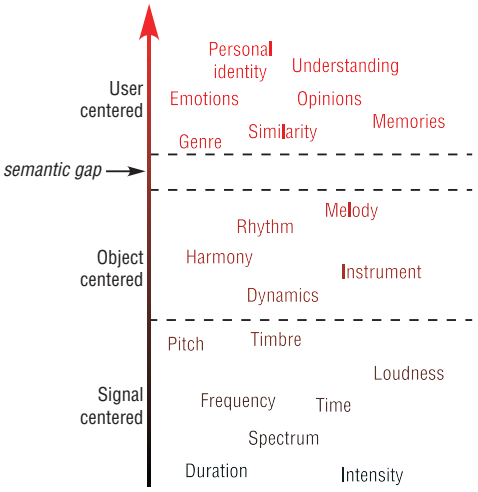


Figure 2.4: Different description layers usually addressed by MCP systems, after Celma & Serra (2008).

descriptors according to those 3 categories. The first category represents *low-level features*, describing the acoustic information by a numeric representation. Typical features at this level include descriptions of the spectral content, pitch, vibrato/tremolo, or temporal aspects of the signal. Those features form the class of *signal-centered* descriptions of the data. The next higher level corresponds to the mid-level description of the signal, thus including tonality, melody, rhythm, or instruments, to name just a few. Here, typical descriptors include the Harmonic Pitch Class Profile (HPCP) or the Beat Histogram. We relate the term *object-centered* to this category of descriptors. Finally, music semantics such as genre, mood, or similarity assessments, hence contributing to the "understanding" of music, are grouped into high-level descriptions of the music; this information is usually regarded as *user-centered* descriptors.

From Figure 2.4 we can also identify a conceptual and methodological problem, inherent to many MIR algorithms, entitled *semantic gap*. It is manifested in a ceiling of machine performance when addressing the extraction of high-level musical concepts such as genre or mood. In particular, the semantic gap arises from loose or misleading connections between low- and mid-level descriptors of the acoustical data and high-level descriptions of the associated semantic concepts, be it in music classification or similarity assessment (Aucouturier & Pachet, 2004; Celma & Serra, 2008). However, it can be identified as methodological problem, namely treating a perceptual construct such as music as pure, independent in it, data corpus, hence ignoring its inherent social, emotional, or embodiment qualities. Moreover, there is a high consensus in literature that methods working in a purely bottom-up manner are too narrow to bridge the semantic gap. Therefore, Gouyon et al. (2008) argues that the step from the mid- to the high-level description of music has to include a user model. See Casey et al. (2008) and particularly Wiggins (2009) for a thorough discussion on this phenomenon.

The here-considered audio features represent static descriptions of musical qualities. The description in terms of an HPCP vector, or pitch value, for instance, refer to, respectively, one single estimate of the tonality, or one single value of the pitch for a given point in time. Temporal information is,

Time scale	Dimension	Content
Short-term	Timbre	Quality of the produced sound
	Orchestration	Sources of sound production
	Acoustics	Quality of the recorded sound
Mid-term	Rhythm	Patterns of sound onsets
	Melody	Sequences of notes
	Harmony	Sequences of chords
Long-term	Structure	Organization of the musical work

Table 2.1: Dependencies of various musical dimensions and their time scale, after (Orio, 2006).

however, indispensable for the perception of musical qualities (Huron, 2006; Levitin, 2008). In this regard, the auditory system extracts different musical attributes at different time scales, as indicated by insights obtained from neural experimentation (Kölsch & Siebel, 2005). Moreover, Casey & Slaney (2006) explicitly show that including temporal information is necessary for addressing the modelling of several higher-level musical aspects. To account for these effects MCP systems usually extract the low-level features on a frame-by-frame basis – frame sizes of around 50 ms are typically applied – and, depending on the context and the modelled concept, accumulate this information over longer time scales to extract the higher-level information. Hence, different musical facets, or concepts, need different integration times, and can therefore be grouped according to their time-scale. Table 2.1 shows an overview of the linkage between several musical dimensions and their time-scale after Orio (2006).

2.2.2.2 Learning algorithms

Pattern recognition provides a vast amount of conceptually different learning algorithms. Typical methods include association learning, reinforcement learning, numeric prediction, clustering, or classification. Figure 2.5 shows a hierarchical conceptual organisation of various approaches in pattern recognition after Jain et al. (2000).

The figure illustrates the differences between supervised and unsupervised learning as well as generative and discriminative concepts. In this respect, unsupervised learning refers to techniques where the distribution of categories emerges from the data itself, without prior knowledge concerning the class membership of the instances. Contrary, supervised learning approaches rely on prior information on the instances' label or cost assignment. Such algorithms learn the relations between the observations' properties of the different pre-defined categories. Moreover, generative learning concepts refer to algorithms that model, for each class separately, the class conditional densities, i.e. likelihoods. On the other hand, discriminative approaches focus on the discrimination between classes and directly model decision function, or posterior probabilities.

Since in this thesis we mainly apply algorithms for categorization and classification, we here shortly review several methods typically found in related literature. Among unsupervised learning methods, clustering represents the most utilised approach. This technique includes k-means clustering,

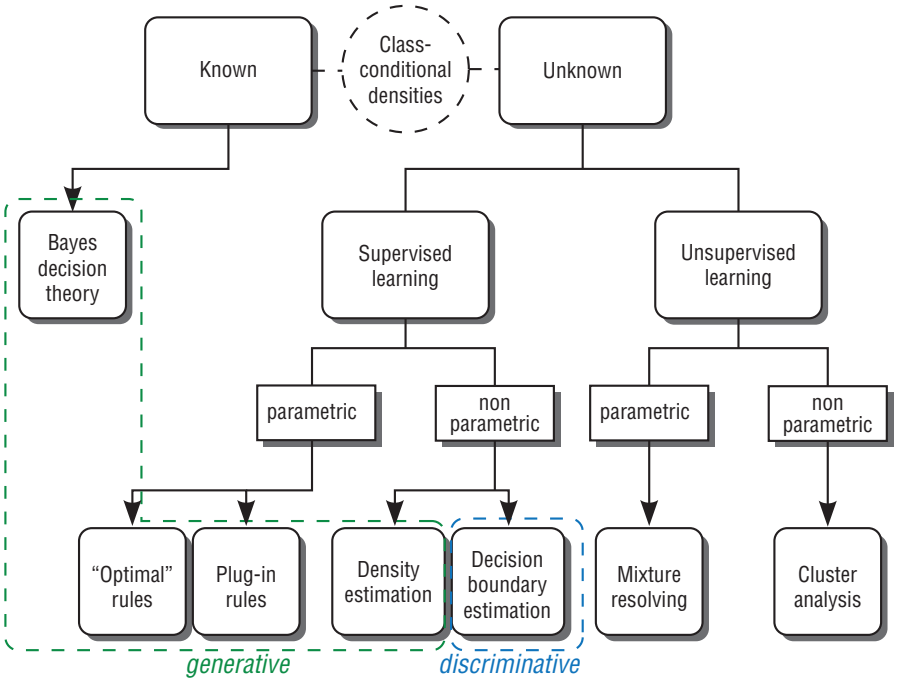


Figure 2.5: Various approaches in statistical pattern recognition after (Jain et al., 2000).

single Gaussian, or Gaussian Mixture Models (GMMs). Recently more advanced techniques such as Independent Component Analysis (ICA), Non-negative Matrix Factorisation (NMF), or Probabilistic Latent Component Analysis (PLCA) became popular. Regarding the supervised techniques, a variety of algorithms have been applied. Here, methods such as Naïve Bayes classifiers or Decision Trees, simple Nearest Neighbour (NN), Artificial Neural Networks (ANN), or Support Vector Machines (SVM), which will be described in detail in Section 4.2.1, have been the most popular. Moreover, several systems use ensembles of combined classifiers by applying techniques such as boosting or bagging. Finally, state models such as Hidden Markov Models (HMM) incorporating temporal information via transition probabilities represent another popular technique for the modelling of frame-wise extracted features in automatic music processing systems.

2.3 Summary

In this chapter we provided the background information behind the methods developed and applied in the remainder of this work. In particular, we entered those research fields mostly related to this thesis, namely psychoacoustics, signal processing, and machine learning, reviewing some of their basic notions and concepts. First, we put special emphasis on the mechanisms involved in human auditory processing since we regard it the touchstone for addressing the problem at hand. Here,

we discussed the most essential psychoacoustic processes and concepts, including the controversial notion of perceived timbre as well as the statistical nature of our internal learning processes. We then reviewed human mechanisms to process and resolve multi-source environments, which form the foundation for the analysis of polyphonic, multitimbral music in terms of source recognition.

The second part of the chapter concentrated on the area of machine listening, which combines the aforementioned fields of music signal processing and machine learning. Due to the important role of the signal representation in the process of automatic source recognition, we first discussed, from a signal processing point-of-view, different audio signal representations as applied in related literature. We then explored the different semantic layers for extracting information from music audio signals and subsequently reviewed some of the basic concepts machine learning offers for the categorisation and classification of music.

3

Recognition of musical instruments

A state of the art review of human and machine competence

Historically, the task of classifying musical instruments deserved quite a lot of attention in hearing research. From von Helmholtz (1954) onwards researchers utilised the instruments' acoustical and perceptual attributes in order to understand the processes underlying timbral categorization operations as performed by the human mind. In this regard, musical instruments provide a representation of timbral subspaces for experimental purposes, as they exhibit a kind-of objectively defined taxonomy with a natural grouping into different categories, which can be related via timbre; the description of the acoustical properties of musical instruments offer basic means to directly assess timbral qualities of the sound. Moreover, musical instruments allow for the control of the sound's basic dimensions aside from timbre, i.e. pitch, loudness, duration. These properties made instrumental tones popular for estimating the perceptual dimensions of timbre (see Section 2.1.1.2). The resulting dimensions found in these studies are assumed to be involved in timbral decision tasks, hence the respective acoustical correlates may play decisive roles in the specific problem of categorisation among different musical instruments.

With the availability of modern computer systems computational modelling of perceptual phenomena became feasible. The first attempts toward automatic musical instrument recognition mostly focused on studying basic methodologies for computational modelling (e.g. Cemgil & Gürgen, 1997; Kaminsky & Materka, 1995). Hence, these experiments were conducted on rather aseptic data – mostly monotimbral material recorded under laboratory conditions – along with a limited set of instrumental categories. The developed systems therefore exhibited by no means completeness in the sense of covering a great variety of musical instruments or applicability to different types of input data, but provided significant insights into the nature and value of different types of acoustical features and classification methodologies, thus paving the way for more enhanced systems. Nevertheless, some of the first approaches offered a high degree of complexity and generalisation power in terms of the applied concepts, see for instance the influential work of Martin (1999). In recent

years, along with increasing computational power, more complex systems were developed, focussing on a larger variety of instrumental categories even in complex musical contexts.

The basic problem, underlying all musical instrument identification systems – including the human mind, is the extraction of the invariants specific to the considered categories as the foundation of the classification process (see Section 2.1). Thus, the information that discriminates one category from all the (modelled) others has to be encoded without ambiguities from the input data. Computational realisations of such systems therefore usually extract features from the raw audio signal. Monotimbral data offers a direct access to the acoustical properties of the corresponding musical instruments, hence making them ideally suited for the aforementioned perceptual studies. Real music, however, is predominantly composed in polytimbral, and presumably polyphonic[1] form, complicating the automatic recognition of musical instruments (and sound sources in general) from this kind of data. Since the different sources constituting the mixture overlap both in time and frequency, the extraction of the acoustical invariants related to the respective sounding objects is not trivial. Thus systems dealing with recognition from polyphonies demand for more complex architectures, involving heavier algorithms for pre-processing the raw data, or need additional a priori knowledge to perform the task.

This chapter is thought to be an introduction into the field of automatic musical instrument recognition, hence covering all relevant areas related to the topic. It is organised as follows; to begin with, we examine the main characteristics of musical instruments in terms of their acoustical properties and show how they group together by reviewing well-established taxonomies of instruments (Section 3.1). This is followed by the examination of human capabilities in recognising musical instruments from both mono- and polytimbral contexts (Section 3.2). We then postulate requirements for any musical instrument recognition system as a guidance for comparing their general performance (Section 3.3), and discuss the basic methodology common to most systems (Section 3.4). Section 3.5 finally presents the review of the relevant literature, a subsequent discussion in Section 3.6 then closes this chapter.

3.1 Properties of musical instrument sounds

3.1.1 Physical properties

Any musical instrument can be regarded as a vibrating system, which oscillates, when set into excitation by imposing a force, at distinct frequencies with certain strength (Fletcher & Rossing, 1998). Furthermore, the underlying sound producing mechanism can be regarded as a two-component,

[1] Polyphony connotes the rhythmical independence of simultaneous parts, or voices, of a musical composition with respect to each other. Contrary, Homophony denotes the movement of multiple voices with the same rhythmic pattern along time. In consequence, monophonic music consists from just one voice, but note that a single voice can be played by multiple sources. We therefore want to emphasise the subtle differences between the two terms monophonic and monotimbral in connection with music.

interactive process; the first part being the actual sounding source, e.g. a string of the violin, which resulting complex tone is further shaped by a filter, the so-called resonator, e.g. the wooden body of the violin (Handel, 1995). When excited, the source produces an oscillation pattern which consists of individual components, termed *partials*, generated by its different vibration modes. The resulting frequencies and corresponding amplitudes of the partials are defined by the resonance properties of the respective vibration mode – the resonance frequency and its damping factor. Both are defined by the physical and geometrical characteristics of the sounding source. These frequencies may be located at quasi integer multiples of a fundamental frequency, as characteristically produced by periodic signals. The resulting spectrum is said to be harmonic[2], a typical property of instruments stimulating a strong sensation of pitch ("pitched" instruments). In contrast, the partials of aperiodic sounds are rather spread across the whole frequency range, generating an inharmonic, noise-like tone, observable with most percussive sound sources[3] ("unpitched" instruments). The damping influences the time-varying strength of the partial, where a light damping exhibits high vibration amplitudes in a narrow frequency region around the corresponding resonance frequency together a slow response to temporal changes of the source, and vice-versa for a heavily damped mode.

This complex vibration pattern is then imposed to the resonator which acts as a filter, reshaping the amplitudes of the individual frequency components. Since coupled to the source, the instrument's body vibrates accordingly in different modes, at which distinct frequency regions are activated by the oscillation of the source. Which frequencies to what extent being affected again depends on the physical and geometrical properties of the resonator. For many instruments several distinct frequency regions are amplified, creating so-called *formants*, or formant areas. Being an effect of the acoustic properties of the static resonator, their frequency location does not depend on the actual pitch of the excitation pattern produced by the sounding source. As a consequence, formants are paradoxically one of the reasons for the dependency of timbre on the pitch of many musical instruments (see below). For some instruments the resonance of the filter even influences the geometrical properties of the source, hence generating a direct interaction with the source vibration pattern. Figure 3.1 shows a simplified illustration of this source-filter production scheme of instrumental sounds. It can be seen that the process is equivalent to a multiplication of the source's spectral excitation pattern with the resonator's transfer function in the frequency domain. The depicted abstraction of the resulting representation of amplitudes versus frequencies – the dashed line in Figure 3.1 – is usually denoted as *spectral envelope*.

Besides their distinct spectral distributions, tones produced by musical instruments exhibit strong temporal patterns as well. The most evident are related to the sound's temporal envelope, which can be roughly divided into three different parts; the attack, sustain, and release (Figure 3.2). In addition to the attack and release phases, which are featured in all natural sounds – a consequence of the excitation of the vibration modes – some instrumental sounds exhibit a strong sustain part, an implication of the specific excitation method; strucked or plucked sources obviously cannot be sustained anyway, hence their sounds enter the release directly after the attack phase of the tone (e.g. piano or guitar). Other instruments in opposite offer sustained parts of finite duration (e.g. blown instruments) as well as possibly infinite duration (e.g. bowed string instruments). Besides these macro-temporal properties, micro-temporal processes related to the spectral components addition-

[2]Accordingly, the frequency components (partials) of these spectra are usually termed *harmonics*.

[3]There exist some in-between instruments which are able to produce a clear pitch sensation but do not exhibit a harmonic spectrum, e.g. bells.

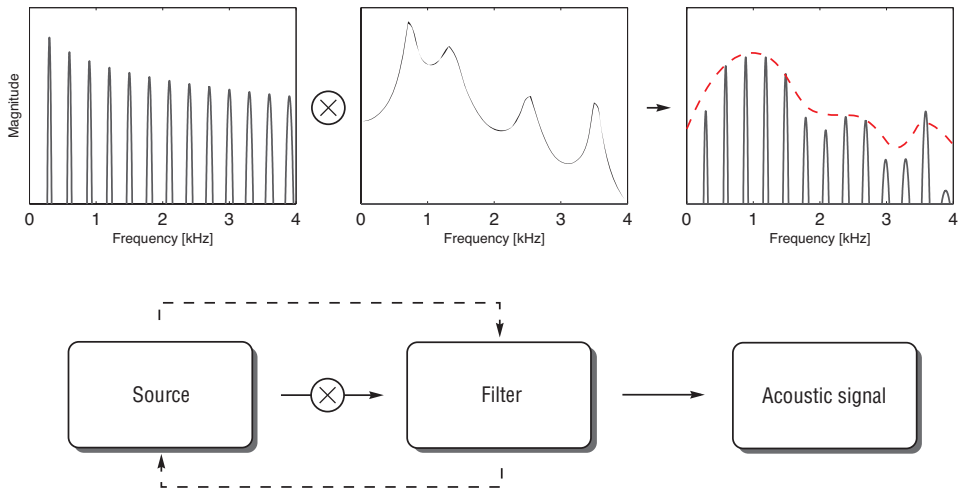


Figure 3.1: Source-filter representation of instrumental sound production. The process can be regarded as a multiplication of the excitation spectrum with the resonator's transfer function. The coupling of source and filter causes an interaction of the two components, depicted as forward and feedback loops. The dashed line in the upper right plot denotes the resulting spectral envelope of the sound.

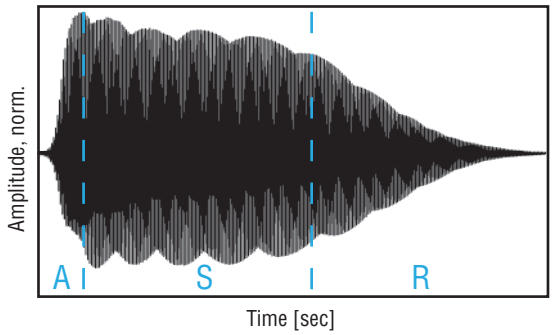


Figure 3.2: Temporal envelope of a short clarinet tone. Attack (A), Sustain (S), and Release (R) phases are marked.

ally shape the perception of instrumental sounds. Since the partials' temporal behaviour is influenced by the damping factors of the respective resonance modes, each components behaves differently with respect to changes of the source along time. Moreover, pitch-independent transients during the attack phase and noise signals, artefacts of the excitation method (e.g. b(l)owing), are part of the sound and consequently influence its temporal behaviour.

By considering these temporal aspects the concept of the spectral envelope can be extended by adding a temporal dimension, resulting in the *spectro-temporal envelope* (Burred, 2009; McAdams & Cunible, 1992). There is a great consensus among hearing researchers that this representation is best uniting the different timbral dimensions, since it captures most of the acoustical correlates identified in the corresponding studies reviewed in Section 2.1.1.2. Figure 3.3 shows an example of the spectro-temporal distribution of a single instrument tone played by a violin.

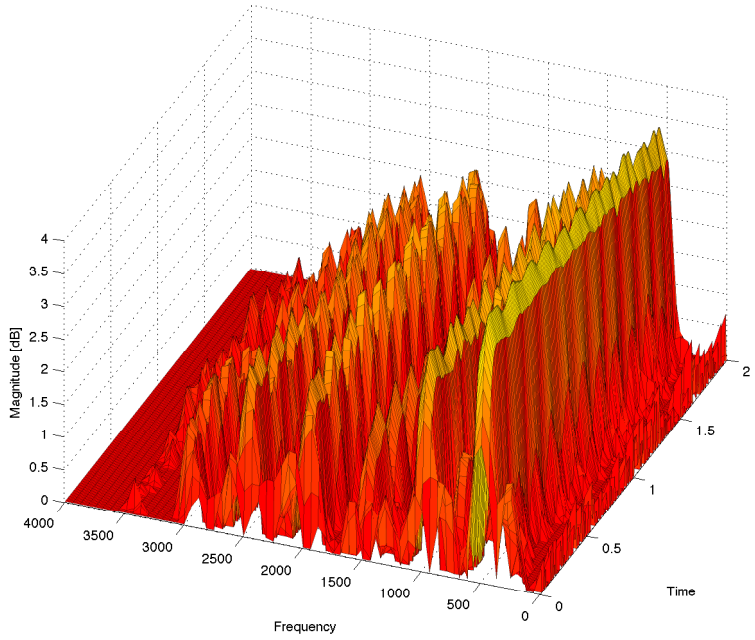


Figure 3.3: Spectro-temporal distribution of a violin tone played with vibrato. Both the spectral and temporal characteristics can be easily seen. Note, for instance, the anti-formant at the frequency position of the third partial.

3.1.2 Perceptual qualities

In general, the timbral sensation of a specific musical instrument's tone is a result of several characteristics, or variables, of the heard sound. First, spectral cues derived from the amplitude and frequency ratios of the individual partials constitute the basis for timbral decisions. In particular, they result from the product of the spectral characteristics of the vibrating source and the resonances introduced by the filter of the instrument's body. With respect to the latter, the absolute location in terms of frequency of the main formants as well as the frequency relation of the respective components having maximum amplitude between different formant areas seem to have a major influence on the timbral sensation of the tone (Reuter (2003) referring to Schumann (1929)). Those spectrally related cues correspond to the spectral shape dimension identified in the aforementioned MDS studies (e.g. brightness). Time-varying characteristics further influence the timbre of an instrument's tone, since the individual spectral components do not follow similar temporal trajectories along its duration (see above). Moreover, transients as well as noise components exhibit strong discriminative power between tones of different musical instruments. Even with pitched, i.e. harmonic, components removed from the signal, the remaining "noise" part showed high recognition rates in experimental studies (Livshin & Rodet, 2006). The corresponding dimensions revealed by the timbre similarity experiments are the attack characteristics as well as the temporal variation of the spectrum (e.g. the spectral flux as identified by McAdams et al. (1995)).

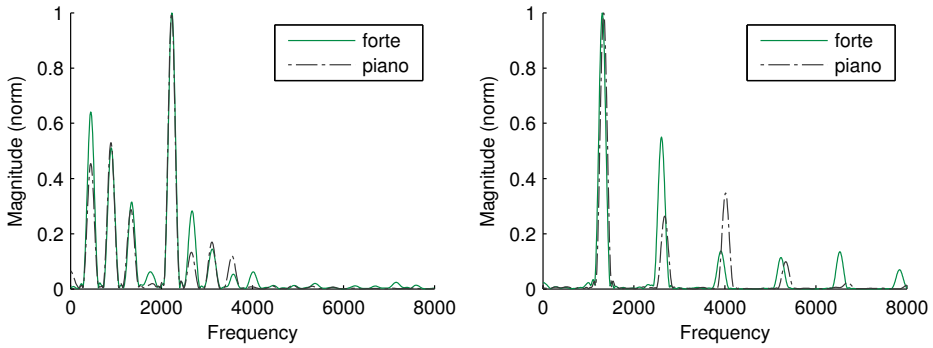


Figure 3.4: Influence of dynamics and pitch on perceived timbre. Part (a) shows the spectra of a low-pitched violin tone played with *piano* and *forte* dynamics, while (b) depicts the spectra of a high-pitched violin tone played with the same two dynamics. Note the differences in the harmonics' relative magnitudes between the two figures due to the different pitches played, and within each plot due to the different dynamics applied. All spectra are normalised to emphasise the relative differences in the partials' magnitudes.

Moreover, most instruments show dependencies of the timbral sensation on articulation and pitch. Often changes in register are accompanied by strong changes in timbre. Due to differences in the playing method (e.g. "overblowing" techniques are used with many wind instruments to change the register) or the excitation source (e.g. different strings are played at different registers of the piano or string instruments) the resulting timbre is evidently altered to a great extent. However, intra-register factors play a distinctive, even though subordinate, role in the timbral sensation of an instrument's tone. First, the strength of the excitation is directly affecting the amplitudes of the source's partials, here a stronger excitation produces a richer spectrum by enhancing higher harmonics, generating an overall brighter sound. Hence, depending on the place and intensity of the excitation, different modes of both the source and the resonator are activated, the latter producing different formant areas along the frequency spectrum. Furthermore, the formants might affect different partials at different pitches played, resulting in slightly modified spectral envelopes. Figure 3.4 exemplifies these dependencies for two pitches played by a violin with different dynamics, i.e. excitation strengths. To summarise, Handel (1995, p. 428) wrote:

> "Each note of an instrument [...] engages different sets of source and filter vibration modes so that we should not expect a unique "signature" or acoustical property that can characterize an instrument, voice, or event across its typical range. The changing source filter coupling precludes a single acoustic correlate of timbre."

Evidence from various experimental studies supports these indications; in a psychoacoustic study Marozeau et al. (2003) showed that despite the observed intra-register dependency of timbre on the fundamental frequency (intervals of 3 and 11 semitones were used in those experiments), the different timbres of the same musical instrument stay comparable. The authors demonstrated that the perceptual intra-instrument dissimilarities were significantly smaller than the cross-instrument ones. Moreover the hypothesis of a general non-instrument-specific dependency of timbre on fun-

damental frequency had to be rejected for intervals smaller than one octave, i.e. certain instruments' timbres are more affected by changes in fundamental frequency than others. Furthermore, Handel & Erickson (2004) suggested that humans use the timbral transformation characteristics across the playing range of a particular instrument for identification purposes, since even human expert listeners seemed to be unable to ignore timbral changes across different pitches of the same instrument (here, intervals of one and two octaves were used). The authors further argued that these transformation properties exist for both category and instrument-family level and are heavily involved in the, presumably hierarchical, recognition process.

Nevertheless, by using an automatic instrument recognition algorithm Jensen et al. (2009) showed that even a transposition of testing instances by more than 5 semitones with respect to the training samples degrades the recognition performance significantly. These results seem odd in comparison to the perceptual evidence coming from the aforementioned studies. However, this low threshold of 5 semitones may be explained by the transformation process the authors applied to generate the different pitches for their experiment. By shifting the sound's spectrum for generating the transposition, the timbre is altered since the formant areas are shifted as well, hence resulting in weaker identification performance of the system.

In general, the information provided by the different timbral cues is highly redundant, hence in real situations the human mind may assign weights dependent on the context. In essence, those variables that gives the most confident estimate in the current acoustical situation are chosen for label inference (see Section 2.1.2).

Finally, it should be noted that an instrument's historical usage and development play a fundamental role for its present sound characteristics. Orchestral instruments have always been continuously modified and improved along centuries, hence to conform to the current composition methods and performance practice at hand. They therefore exhibit highly adaptation to the conventions imposed by the Western music system, and reflect many properties of human auditory perception.

3.1.3 Taxonomic aspects

In general, a taxonomy characterises a field of (abstract) knowledge by describing, classifying, and representing its elements in a coherent structure. For musical instruments, a certain taxonomy has to reflect organology, *"…the science of musical instruments including their classification and development throughout history and cultures as well as the technical study of how they produce sound"*[4]. Historically, many different taxonomic schemes have been proposed, based on the instruments' geometric aspects, material of construction (e.g. wood and brass instruments), playing method (e.g. blown or bowed instruments), or excitation method (e.g. struck or plucked instruments). The most well-known, however, was certainly defined by von Hornbostel & Sachs (1961), considering the sound production source of the musical instruments. In particular, this taxonomy groups the instruments into the basic classes *aerophones* (the instruments' sounds are generated by the vibration of an air column), *chordophones* (strings are set into oscillation to produce a sound), *idiophones* (these instru-

[4]Retrieved from http://www.music.vt.edu/musicdictionary/

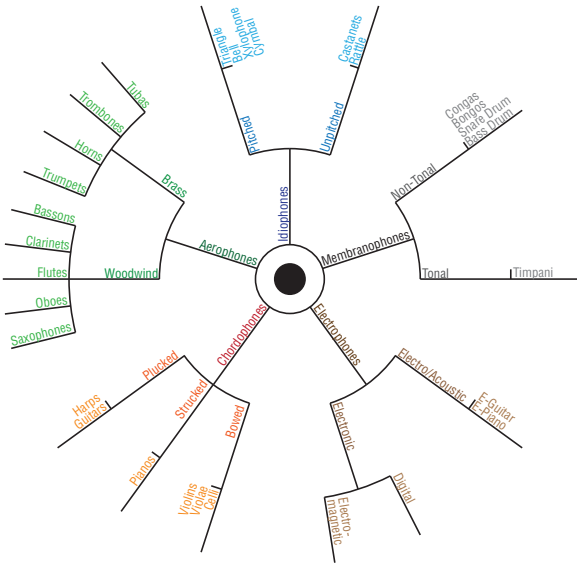


Figure 3.5: A simplified taxonomy of musical instruments, enhancing the classic scheme of von Hornbostel & Sachs (1961) by the category of the *electrophones*.

ments are excited by imposing the force on its body), and *membranophones* (vibrating membranes act as the sound source).

Since the development of electronic sound generators the classical taxonomies, including mostly orchestral instruments, have to be expanded by the category of the Electrophones. For instance, Olson (1967) provides an additional category entitled *Electric instruments*, a diverse class grouping together instruments like electric guitar, music box, metronome, and siren. Hence, it remains unclear how to subgroup this extremely varied category, nevertheless it seems to be possible to roughly divide them into Electric/Acoustic (e.g. the electric guitar) and Electronic instruments, whereas the latter can be further separated into instruments using electromagnetic (e.g. analogue synthesizers) or digital sound producing methods (e.g. digital or sample-based synthesis systems). Figure 3.5 shows a tree-like structure of an enhanced taxonomy including the most prominent musical instruments based on the classic scheme described above.

3.1.4 The singing voice as musical instrument

As *singing voice* we consider all sounds that are "*…produced by the voice organ and arranged in adequate musical sounding sequences*" (Sundberg, 1987). This definition covers a rich amount of timbral modifications of the voice's acoustical signal, a variety that goes far beyond the possibilities of most other traditional musical instruments in altering the timbre of their sounds. Despite its evidently different musical role compared to other instruments, in the context of this thesis, however, the singing voice is regarded consistently with respect to the latter, in a sense that it contributes to the mixture in the same way any other active source does.

In general, the sound production mechanism of the human voice can be described in a similar way to other musical instruments using the source-filter abstraction (Bonada, 2008); the vocal folds, or cords, act as the voice source, which are set into oscillation by the air flow produced by the lungs. The fundamental frequency of the generated sound mainly depends on the tension, length, and mass of the folds. The filter, or resonator, consisting of the mouth and nose cavities, then shapes the source signal according to its formant areas. Finally, the resulting sound is radiated through the air via the lips.

However, the singing voice provides a much greater flexibility in terms of variation of the sound spectrum than other musical instruments. First, the voice source is able to produce harmonic, inharmonic, and in-between sounds, enabling typically vocalisation styles such as pure singing, whispering, and growling as well as any intermediate expressive mode. Second, the geometric properties of the resonator are dynamic, hence formant areas can be created "on demand" by altering the mouth and nose cavities. A well-known example is the singing formant of (male) opera singers who use a distinct configuration of the resonator to produce a formant in the range around 3 kHz, allowing for a better audibility in the context of orchestral accompaniment.

3.2 Human abilities in recognising musical instruments

Recognising musical instruments is an elementary, supposable subconscious, process performed by the human mind in everyday's music listening. But contrary to the common perception that instrument identification is an easy task, several studies have shown clear limitations in the recognition abilities of subjects (Martin, 1999). Moreover, humans tend to overestimate their performance in comparative experimental settings, most noticeable when assessing the performance of automatic recognition systems.

Almost all experiments examining human recognition abilities of musical instruments have been performed on monophonic audio data, in order to exclude any perceptual or cognitive mechanism not related to the recognition process itself. It is assumed that sound-source recognition from more complex stimuli involves more sophisticated processing of the brain which acts as a kind-of preprocessing for the actual recognition (see Section 2.1.2). Regarding the polytimbral context, only very little research has been conducted for estimating human abilities to identify concurrent instrumental tones, and most of the existing sparse works concentrated on the laboratory condition of recognition from tone pairs. However, we can spot some more general, thus related to source recognition, aspects of human brain processing of complex auditory stimuli in the respective literature. Due to these conceptual differences, the following section is divided into two corresponding parts, separating experimental findings derived form studies using monotimbral and polyphonic/polytimbral stimuli, respectively.

3.2.1 Evidence from monophonic studies

Since the amount of conducted studies examining the ability of subjects in discriminating between sounds from different musical instrument is sparse, a qualitative comparison among them is difficult. Though differences in the used experimental settings and methodologies often resulted in heterogeneous conclusions. However, some commonalities between the respective results can be identified which are presented here. In doing so, we mostly concentrate on the more recent experiments carried out by Martin (1999), Srinivasan et al. (2002), and Jordan (2007), the latter being the most exhaustive, accounting for factors such as register, dynamic, and attack-type differences in the presented stimuli. What follows are the most important observations derived from literature:

1. The maximum recognition performance achived by expert human listeners, including professional musicians, was 90% of accuracy in a 9 instrument, forced-choice identification task (Srinivasan et al., 2002). Adding more categories degrades performance subsequently, reported values include 47% and 46% of accuracy for recognising, respectively, 12 out of 12 (Jordan, 2007) and 14 out of 27 instruments (Martin, 1999).

2. Confusions between instruments of the same instrument family (e.g. stings) are more likely to happen than confusions between instruments of different instrument families. Inside a family, regular and coinciding confusions between certain instruments were found across studies (e.g. French horn with Trombone, or Oboe with English horn), most probable resulting from either overlapping formant areas or similar spectral fluctuations (Reuter, 1997). Hence, performance in terms of recognition accuracy increases significantly when evaluated at the instrument family level. Authors could observe an increase of 5 and 46 pp (!) for the 9 (Srinivasan et al., 2002) and 14 (Martin, 1999) instrument recognition experiments, respectively.

3. Subjects extensively use musical context for timbral decisions. Experiments on solo phrases showed better performance figures than studies using isolated sounds. Martin (1999) reported recognition accuracies of 67% on a 19 out of 27 recognition task, supporting previously found evidence (Kendall, 1986).

4. Prior exposure to the sound sources improves accuracy. Hence, musical training is beneficial for the recognition performance. In his experiments, Jordan (2007) found a significant difference in the performance of identifying musical instruments between the groups of professional and hobby musicians. Moreover, results reported for untrained listeners showed an absolute difference in recognition accuracy of up to 21 pp when compared to the accuracies obtained by testing trained musicians (Kendall, 1986).

5. Features derived from the attack portion of the signal are decisive for timbral decisions on isolated note samples. Jordan (2007) found significant differences in recognition accuracy of subjects when comparing isolated sounds with attacks replaced by a constant fade-in to the unmodified versions. However, the influence of the attack is by far less important than the influence of the register the instrument is played in[5]. Comparisons of different registers showed p values smaller 10^{-3}, an indication of the importance of the formant areas in the recognition

[5]Besides, alterations in the dynamics of the stimuli (the study in question examined the dynamical forms of *piano* and *forte*) revealed no effect on the recognition accuracy (Jordan, 2007).

process. Since the identification performance significantly dropped for high pitched sounds (fundamental frequencies ranging from 250 to 2100 Hz, depending on the instrument), the author argued that the degradation of the recognition accuracy can be explained by the absence of the first formant; due to the high fundamental no partial falls into the frequency range of the first formant. Moreover, there is evidence that, in musical context, features derived from the attack phase are irrelevant and replaced by the analysis of the steady-state part of the sound (Kendall, 1986). To conclude, Grey (1978) hypothesised:

> "In that spectral differences are more continuous throughout the presentation of tones, the extension of the context [...] may amplify such differences, giving the listener more of a chance to store and compare spectral envelopes. [...] Musical patterns may not let the listener take such care to store and model for comparison the fine temporal details [i.e. the attacks], since information is continuously being presented."

3.2.2 Evidence from polyphonic studies

In a quite general regard, the perceptual and cognitive capacities of the human mind are limited. Experiments on subjects' channel capacities, i.e. the amount of information they are able to capture, showed that these limits exist in almost all areas of cognitive processing with a rather constant magnitude. In this context, Miller (1956) presented the "magical number 7", a numeric quantity corresponding to the information capacities of various, but supposedly unrelated, cognitive processes. He identified, across the respective studies, quantities ranging from 4 to 10 categories (or roughly 7 $\pm$ 2) the human mind is able to ambiguously process. Above this threshold, subjects are more likely to produce errors in the respective tasks. In particular, Miller (1956) reported studies assessing subjects' abilities in absolute judgement (i.e. judging the order of magnitude of a certain set of stimuli), the size of their attention span (i.e. the quantity allowing for a simultaneous focus), and the size of their immediate memory (i.e. the number of symbols to remember instantaneously).

The results suggest that the amount of information a human can process in a given task seem to be quite low, at least lower than expected. These limitations certainly play a functional role in our understanding of music as well. However, with respect to the stimuli used in the aforementioned work, music is different in many respects; among others it provides massive contextual information as well as meaning, and both short- and long-term memory is involved (see Section 2.1.1.3). Nevertheless, we can find noticeable analogies when reviewing literature studying human perceptual and cognitive abilities in polyphonies.

But first, let us consider the related field of speech perception and cognition. Here, Broadbent (1958) reported that inside a multi-speaker context, subjects were only able to attend to one single speaker, not even able to correctly report on the spoken language of the concurrent speakers. That is, in the cocktail party situation (see Section 2.1.2), attention mechanisms seem to be employed to capture and convert the acoustical information of a single source into meaning, or switch between several speakers. Sloboda & Edworthy (1981) noted that in addition social conventions, restricting the number of voices in a typical conversational situation to one, may have an influence on this massive restriction of the human brain.

In case of music, literature reveals a slightly different picture. Huron (1989) conducted a study determining human abilities in estimating the number of concurrent voices[6] in polyphonic, but monotimbral music. Subjects had to continuously determine, along a fugal composition of Baroque composer J. S. Bach, the number of active voices. Obtained results examined their abilities in estimating voice entries and exits as well as their accuracy in spotting the amount of present voices. In general, musicians showed a slightly more accurate performance than non-musicians, indicating the presence of mental images of timbral densities inherent to musicians. Moreover, a threshold of 3 concurrent voices could be observed, below which subjects responses reflected accurately the amount of present voices. If the number of concurrent voices exceeded this value of 3, subjects showed both slower and more inaccurate responses. But even more remarkable, subjects reported that below the threshold they rather counted the number of voices whereas above they were only able to estimate their amount[7].

However, highly elaborated Baroque contrapuntal works exhibit up to 6 different, independently from each other composed, voices. Here, harmony can play an additional role in a sense that it provides contextual information to fuse the individual voices (Sloboda & Edworthy, 1981). In the same work the authors committed that listeners are unable to actively attend to more than one voice at a time, a link to the experimental findings from the speech domain.

Kendall & Carterette (1993) conducted one of the first studies examining subjects' instrument identification abilities in polytimbral contexts. In the experiment listeners were asked to both estimate the perceived blend of, and recognise the two different instruments constituting a dyad tone. Several musical contexts were employed (isolated tones and musical phrase, both in unison, major third, and harmonic relation) to asses subjects' abilities on 10 different instrumental combinations from the brass and wind families, in a forced-choice task. In general, an inverse relation of blend and identifiability was observed. An MDS analysis of the similarity rating of dyad pairs revealed the qualities *nasality* and *brilliance* – in contrast to usually found attributes such as *sharpness* or *brightness* resulting from studies using single tones – as the primary two dimensions, which were re-encountered by analysing the listeners' blending ratings via MDS. This indicates that the perceptual qualities of polytimbral sounds are directly related to the separability of the respective constituting sources. In particular, identification abilities of sound combinations were found to be correlated to both contrast in stable spectral properties and time-varying spectral fluctuation patterns.

Similarly, Sandell (1995) examined the main factors for this kind of timbre blending. In his experiments the author identified two main features of major importance; first the absolute difference in spectral centroids of the tones, and second the position in terms of frequency of their compound spectral centroid. Moreover, the tested intervals unison and minor third suggested no dependency of the blending abilities on the fundamental frequency of the respective tones, thus emphasising the stable spectral, i.e. formant, and time-varying characteristics identified by Kendall & Carterette (1993).

[6] With the term *voice* we refer to a single "line" of sound, more or less continuous, that maintains a separate identity in a sound field or musical texture (Huron, 1989).

[7] Given the fact that the author observed a beneficial influence of timbre on the estimation, i.e. a difference in timbre improves the accuracy, we may speculate that in context of distinct timbres and the above presented evidences from information theory, the threshold can be raised to 5, which would be perfectly in line with the experiments presented by Miller (1956) on the attention span.

In the more general scenario of polyphonic, multitimbral music, the human mind is assumed to resolve the problem of source identification by performing streaming-by-timbre, hence grouping the different sound objects into separate streams, from which decisions regarding the timbral nature of the sources are inferred (see Section 2.1.2). Hence, the ability to stream different timbres seems to depend on the aforementioned blending tendencies of the involved sounds. Reuter (1997; 2009) identified strong analogies between streaming and identification/blending abilities of concurrent instrumental timbres. He determined two properties of musical instruments to be crucial for the ability to stream, hence identify, multiple sources; the first relates to the formant areas of the instruments, the second – in absence of characteristic formants – corresponds to their spectral fluctuations. Both enable the identification of concurrent timbres as well as their segregation into different streams by the human mind. In a follow-up experiment the author showed that artificially manipulating formant areas of musical tones directly affects their segregation tendencies in multi-source contexts (Reuter, 2003). In this regard, it seems most probable that the underlying operations for perceptual streaming are of primitive nature (i.e. low-level processes) which is further controlled, adapted, and complemented by high-level contextual and top-down processes (Bregman, 1990). See also the work of Crawley et al. (2002) for more evidence on the primitive nature of perceptual grouping.

It should be emphasised that these mutual properties of individual instruments have been utilised over centuries by composers to blend or separate timbres; from the Baroque period onwards specific combinations of instruments were used to create artificial, blended timbres. Hence, there exist simple rules in the praxis of orchestration[8] which pairs of instruments tend to blend and which not. As expected, these rules are largely based on the parameters identified above. Finally, the concept of the orchestra as an entity purposely includes the coexistence of contrasting families of timbres (Kendall & Carterette, 1993).

At last, the number of not controllable parameters seems to complicate extensive experiments studying human capacities when listening to real music. It is not clear how attention mechanisms, temporal encoding, their interaction, and musical meaning itself influence the performance on various tasks. Hence, conclusions with respect to the more general case of polytimbral music, derived from the aforementioned studies, are at best of speculative nature.

3.3 Requirements to recognition systems

In his thesis Martin (1999, p. 23 et seq.) postulated six criteria for evaluating and comparing sound-source recognition system. Due to their universality, we strictly follow them here, emphasising their implications on the field of automatic musical instrument recognition:

[8] The art of arranging a composition for performance by an instrumental ensemble; retrieved from http://www.music.vt.edu/musicdictionary/.

1. **Generalisation abilities.** Generalisation in terms of modelled categories in the sense that regardless of the instruments' construction type, pitch, loudness, duration, performer, or the given musical and acoustical context, the recognition accuracy of the system should be stable. Hence, a successful recognition system has to capture the categories' invariants independent of the variability of the aforementioned parameters.

2. **Data handling.** The ability of the recognition system in dealing with real-world data, which exhibit a continuous degree of temporal and timbral complexity. Similar to the first criterion, recognition performance should not be affected by the variability of the real-world data. For instance, systems designed for monotimbral data act poor in this respect, since they may fail to produce reliable predictions when input a polyphonic sound. It should be noted that those systems nevertheless might be useful in certain contexts, but this fact has to be taken into account when comparing systems.

3. **Scalability.** Scalability in terms of modelled categories; a recognition system should exhibit enough flexibility in a way that new categories can be easily learned. Furthermore, Martin introduces the notion of *competence of the approach* to evaluate systems which limited knowledge. It addresses the system's capabilities of incorporating additional categories and the thereby generated impact on its performance.

4. **Robustness.** With increasing amount of noise the system's performance should degrade gracefully. In this context we can identify manifold definitions for noise, e.g. the number of concurrent or unknown sources, the degree of reverberation, etcetera, which should affect the recognition accuracy to an adequate, hence reasonable, amount.

5. **Adaptivity.** Adaptivity in terms of the employed learning strategy in the sense that both labelled and unlabelled data are incorporated in the learning process. Learning, as such defined by the human mind, is a life-time process and includes supervised training by teachers as well as flexible unsupervised processes for new input data. Hence, computational systems should use semi-supervised learning algorithms and keep updating their repositories continuously to guarantee the best possible abstraction of the categories' invariances.

6. **Real-time processing.** There is strong evidence that the essential qualities of music are defined via time-varying processes (Huron, 2006). Martin argues that any music processing system aiming at understanding the musical content is therefore required to mimic these real-time aspects. However, the author admits that this would bear too many limits for computational systems, hence he proposed to add the term *in principle* to the real-time requirement. Hence the criterion is reduced to the sequential processing of the input data.

At last, in case of an equal performance of competing systems regarding all of the aforementioned criteria, Ockham's razor, or *lex parsimoniae*, should be applied, stating that the approach making the fewest assumptions is to favour (Martin, 1999).

3.4 Methodological issues

As an introduction to the status-quo of the related research, we first point towards common modalities and shared methodologies among all developed approaches designed for identifying musical instruments from audio signals. To begin with, this section starts with presenting several important methodological issues inherent to classification paradigms in order to provide the appropriate context necessary to assess the pros and cons of the different works discussed in Section 3.5. This is followed by a review of the general architecture of an automatic musical instrument recognition system.

3.4.1 Conceptual aspects

When comparing different classification systems and their performance, it is of major importance to consider the pre-conditions the respective systems were designed under. These pre-conditions may result from the intended purpose (e.g. a system designed for classical music only) and thereby imposed system requirements, from the availability of resources (e.g. adequate data), or computation facilities. In case of musical instrument recognition systems several parameters reflecting those pre-conditions can be identified. The two most crucial parameters are certainly the type of data used for evaluation and the number of categories covered by the developed recognition models. Other factors – maybe less obvious but nevertheless of high relevance – include the variability of the used data (e.g. the number of distinct musical genres covered), the number of independent data sources, or any prior knowledge input to the system.

In practice, we can identify four main types of data that are used for building and evaluating systems for the automatic recognition of musical instruments. Most early approaches, but also studies having a stronger focus on perceptual aspects, frequently applied sample libraries of in isolation recorded instrumental tones, among which the most popular being the MUMS[9], IOWA[10], IRCAM's studio online (SOL), and RWC (Goto et al., 2003) collections. These sample libraries offer a rich amount of different categories, thus allowing to investigate and reveal the complex perceptual and acoustical correlates between instances of a wide range of musical instruments. On the other hand, the generalisation and data handling capabilities of systems developed with this kind of data are generally poor, since the data is not reflecting the complexity of real world stimuli (see Section 3.3). Recognition performance of such systems usually degrades dramatically when applied to data of a different type (Eronen, 2001; Martin, 1999), even though a different sample library is used (Livshin & Rodet, 2003).

Monotimbral music audio data, often termed *solo recordings*, are usually applied to put the systems in a more ecological context, as these data guarantee more "naturalness" such as reverberated signals, noisy ambient backgrounds, different recording conditions as well as musical aspects related to articulation and playing styles. Moreover, a quasi "clean" access to the sources' parameters under real

[9] http://www.music.mcgill.ca/resources/mums/html/MUMS_dvd.htm
[10] http://theremin.music.uiowa.edu

conditions is possible, hence enabling a modelling of the instruments' timbres inside musical context. However, a direct translation of the developed models to more complex signals is not straightforward, since such systems require "perfect" source separation a priori, which output is then used for the actual classification process. Since the former is nearly impossible to achieve, at least from nowadays perspectives, the whole thought experiment is to question.

To simulate real music signals researchers often revert to artificially created polytimbral data, either by MIDI-directed or undirected, i.e. quasi-random, synthesis of isolated notes taken from sample libraries. Since the acquisition of labelled polytimbral music is difficult, time consuming and sometimes even costly, synthesising data offers a simple strategy to mimic the complexity of real music. However, these data are only partially reflecting the properties of music, lacking effects, reverberation, compression and other aspects of the mixing and mastering applied in the production process of music. In general, these factors alter the spectro-temporal properties of sounds and accordingly those properties of the musical mixture signal to a great extent. Moreover, in the case of a quasi-random mixing of the data, all sort of musical context is neglected, since different sources are by no means independent in music. These generated sounds thus do not represent the intended approximation of the targeted real-world conditions.

Therefore, designing and testing an instrument recognition system with real music recordings is the only remaining option in order to meet the requirements 1, 2, and 4 presented in Section 3.3. Moreover, evaluation itself should be performed on a varied set of music audio data, covering different musical genres, in order to reliably estimate generalisation and data handling capabilities. Paradoxically, only few works tested their approaches on such a varied set of data.

The number of incorporated categories has been identified as the second influential parameter for evaluating and comparing systems designed for the automatic recognition of musical instruments. A classification system, by definition, should cover the whole universe in terms of categories that it attempts to describe. However, in computational modelling the amount of classes is primary controlled by the scope of the study. Hence, systems accounting for an applicability in a real-world engineering context (e.g. a query-by-example system) obviously incorporate different categories, both in number and kind, than, for example, systems designed for examining the perceptual separability of instances of the Wind instrument family. Moreover, restrictions in data size, model complexity, or processing power control the amount of incorporated categories, further narrowing the respective systems' generalisation and scalability characteristics. The limitations in the number of categories lead to a reduced categorical space wherein both training and evaluation is usually performed. Thus, a direct comparison of different systems is evidently not possible due to the differences in the dimensions of the respective evaluation spaces.

The conclusion, however, that fewer categories lead to easier recognition problems is not always valid; distinguishing between Oboe and English Horn is by far more difficult than, for instance, constructing a recognition system for Violin, Piano, Flute and Trumpet. Hence, the taxonomic specificity applied in the classification system has to be taken into account when judging the complexity of the system. In general, we can recapitulate that there is a certain trade-off between the number of applied categories and the resulting recognition performance (see also the comparative analysis of different perceptual identification experiments performed by Srinivasan et al. (2002)). That is,

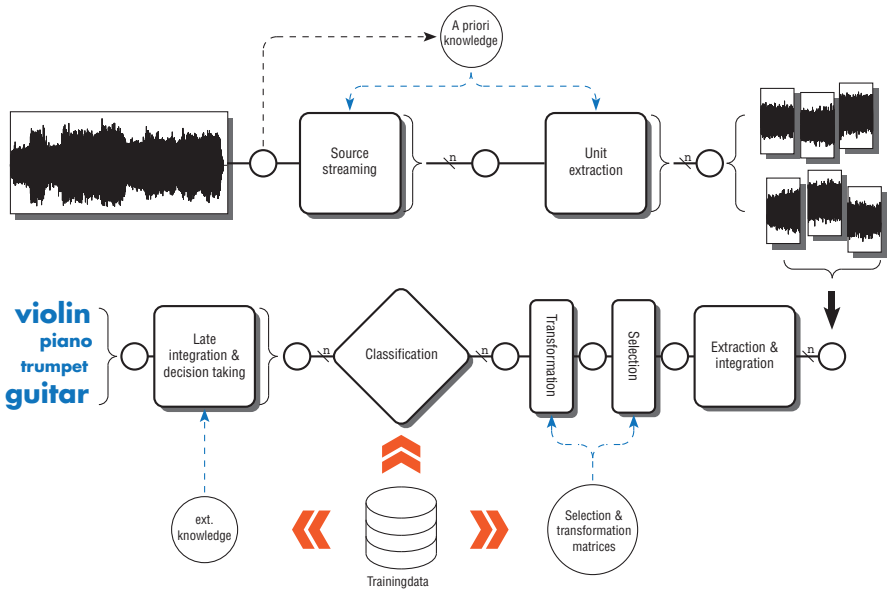


Figure 3.6: General architecture of an instrument recognition system. The signal is first pre-processed to extract the basic acoustical units on which the further processing operates. Then, features are extracted, selected, and transformed to form the input to the actual classification step. The resulting model decisions are further corrected by post-processing strategies from which resulting representation the labels are extracted. Note that depending on the respective approach, some components might only be partially active or even "short-cut". Most of the approaches reviewed in Section 3.5 can be explained by using this scheme.

the fewer categories are learned, the less confusions the system should produce, generating higher recognition accuracies. Contrary, a system incorporating many categories will exhibit a significant lower identification performance, since the great amount of categories lead to higher confusion rates.

3.4.2 Algorithmic design

In what follows we examine a general approach for constructing a musical instrument recognition system. Here, we take an engineering point-of-view and describe the building blocks of an artificial system, abstracted from the approaches presented in the literature. The result is a modular system architecture, where one may virtually plug different components together to accomplish the system that best fits the requirements at hand. We want to emphasise the universality of the scheme, hence it reflects the architecture of almost all systems that can be found in literature. Indeed, depending on the specific approach, certain blocks are only partially included or even missing at all. Figure 3.6 illustrates this general scheme with all the involved modules for the design of an automatic musical instrument recognition system.

1. **Pre-processing.** First, any prior information concerning, for instance, the number of sources, fundamental frequencies or onset times of the present notes is input to the system. This can be realized via a manual insertion or by a corresponding algorithmic implementation, which acts directly on the acoustic signal. Based on the information provided by the previous module the audio is then segregated into timbral streams which itself are segmented into acoustical units. These units constitute the fundamental blocks on which the further processing operates.

2. **Feature processing.** The acoustical units are transformed into a vector of features describing the properties of the underlying sound. The features are initially chosen based on the assumed characteristics of the categories' invariances. This low-level information is typically derived by framing the audio into small chunks, from which short-term features are extracted and integrated over the length of the unit by applying statistical measures. What follows is either a transformation or a filtering of the generated feature vector according to a previously performed analysis of the training data. Both processes decrease the redundancy of the information captured by the features and thereby reduce the complexity of the forthcoming algorithms.

3. **Classification.** A previously trained model is applied to predict the class probabilities on the input for each acoustical unit.

4. **Post-processing.** The classifier output is re-weighted by either globally estimated (e.g. structural or timbral information) or local contextual information (e.g. classifier decisions of neighbouring units). From the resulting data the corresponding labels are finally extracted.

3.5 State of the art in automatic musical instrument recognition

This section covers a literature survey of approaches dealing with the identification of pitched and percussive instruments from music audio data. Our main focus lies on methods developed for recognising pitched instruments, the respective review is therefore by far more extensive than the corresponding one dealing with unpitched instruments. We nevertheless discuss some of the more recent approaches towards the recognition of percussive instruments from polyphonies, additionally providing a reference to an already existing literature overview. Due to the fact that these two groups of musical instruments exhibit major differences in their sound characteristics (see Section 3.1), resulting in partially great conceptual differences in the respective recognition approaches, they are often regarded as separate problems. Here, we will stick to this distinction and present the respective approaches separately.

3.5.1 Pitched instruments

Since the approach taken in this thesis is mainly motivated by an engineering point-of-view, i.e. the design of an instrument recognition system for the analysis of real-world music audio signals, which implies the handling of any music audio data at hand, the subsequent review focuses on related work studying approaches in musical context, be it mono- or polytimbral[11]. Consequently, timbral studies, which are primarily motivated to identify perceptual differences in the timbral sensation of musical instruments, are not taken into considerations. Since a review of these kind of approaches is not provided here, we refer the interested reader to the comprehensive overview presented by Herrera et al. (2006). The following is again subdivided into two parts, covering, respectively, the approaches designed for monotimbral and polytimbral music.

3.5.1.1 Monotimbral studies

Martin (1999) published the first large scale study examining the automatic recognition of musical instruments. In this influential work the author evaluated the accuracy of his developed recognition model and compared it to the performance of human subjects on the same task. Among others, a corpus of music audio data comprising monophonic recordings of orchestral instruments was used for evaluation in a 15 out of 27 recognition task. The author constructed a hierarchical classification system on features calculated from a perceptually motivated representation of the audio. Context-dependent feature selection and search strategies were additionally applied to adapt the system to the complexity of the problem. Finally, a maximum likelihood decision was performed based on an univariate Gaussian prototype for every instrument to obtain the class membership of an unknown instance. Although his algorithm showed good performance on the experimental problems, results revealed that for all requirements listed in Section 3.3, the computer was outperformed by the human subjects.

Two years later Eronen conducted a similar study. In this work an even larger corpus of audio data was analysed, including three sample libraries (MUMS, SOL and IOWA), output sounds of a synthesizer and monophonic recordings taken from compact audio discs. The system used MFCCs in combination with different other spectral features of the audio signal to train and test GMMs for the 31 target instruments. Feature selection algorithms as well as hierarchical classification schemes were applied to reduce dimensionality and enhance the performance of the system, respectively. Reported results showed similar performance compared to the work of Martin (1999), identifying the MFCCs as the best performing features in the recognition task.

In a follow-up work, Eronen (2003) used Independent Component Analysis (ICA) to transform feature vectors of concatenated MFCCs and their derivatives. Via pre-trained basis functions the testing data was mapped into a space where the mutual information of the dimensions should be minimized. Furthermore, discriminatively trained HMMs were applied to capture the temporal behaviour of the instruments' timbre. Finally, classification was done by selecting the model which provided the highest probability. The algorithm was tested on the same data as described in Eronen (2001) and results showed that the transformation improves the performance consistently whereas

[11] In the remainder we will refer with the term *music audio data* to any kind of audio data exhibiting any form of musical context, in opposite to the term *out-of-context data* denoting all data lacking of musical context

the use of the discriminative trained HMMs is only beneficial for systems using a low number of components in the instrument prototypes.

Essid et al. (2006b) evaluated SVM classifiers together with several feature selection algorithms and methods, plus a set of proposed low-level audio features. The system was applied to a large corpus of monophonic recordings from 10 classical instruments and evaluated against a baseline approach using GMM prototype models. Classification decisions were derived by performing a voting among the classifiers' predictions along a given decision length. Results showed that SVMs outperformed the baseline system for all tested parameter variants and that both pair-wise feature selection and pair-wise classification strategies were beneficial for the recognition accuracy. Moreover, longer decision length always improved recognition performance, indicating the importance of the musical context (i.e. the integration of information along several instances in time) for recognition, as similarly observed in perceptual studies (see Section 2.1.1.2).

Joder et al. (2009) studied both early integration of audio features and late integration of classifier decisions in combination with SVMs for instrument recognition from monotimbral recordings. Early integration denotes the statistical evaluation of short-term features inside a texture window prior to classification, while late integration refers to the combination of classifier decisions on several texture windows for decision making (e.g. fusion of decisions or HMM). In this work the same data was used as applied by Essid et al. (2006b). Reported results showed only slight improvements over a baseline SVM system when early and late integration were combined. Interestingly, best early integration resulted from taking the mean of the short-term feature values along a segment length corresponding to the "basic" acoustical unit, a musical note. The authors concluded that early integration is mainly for smoothing feature values, i.e. removal of features' outlier values, while late integration should roughly capture temporal aspects of the music.

Finally, Yu & Slotine (2009) proposed a pattern matching approach for classifying monophonic instrument phrases. The technique, coming from image processing, treats spectrograms of musical sounds as texture images. No specific acoustic features were used since in the learning stage of the system only sample blocks of different scales were taken from the training spectrograms. These blocks are then convolved with the test spectrogram at each time-frequency point and the minimum was stored at the corresponding position of a feature vector. This process was repeated for all blocks and the final classification was obtained by applying a simple kNN rule in the feature space. Given 85.5% of average accuracy on a seven instruments plus drums classification task, the authors suggested the technique as a promising tool for the separation of musical instruments in polyphonic mixtures.

3.5.1.2 Polytimbral studies

As already stated above, a direct translation of the models reviewed in the previous section to more complex data is not straightforward. Heavy signal processing is often required to adapt the data in a way that the recognition approaches can be applied. In the course of our literature review of polytimbral recognition approaches we identified three main classes of studied methodologies; first, *pure pattern recognition* systems try to adapt to the more complex input by releasing the constraints on either the data or the categories itself. Recognition is usually performed directly from the poly-

phonic signal, identifying a single dominant instrument or a certain combination of instruments from the mixture. Second, *enhanced pattern recognition* aims at combining signal-processing front-ends with pattern recognition algorithms, introducing source separation or multi-pitch estimation prior to the classification step. The pre-processing should minimise the influence of the source interference on the extracted features, which are input to the recognition algorithm. Finally, the class of *template matching* algorithms derives class memberships by evaluating distances to abstracted representations of the categories. Here, global optimisation methods are often applied to avoid erroneous pre-processing resulting from, for instance, source separation.

Before presenting the works in detail, Table 3.1 lists all the reviewed approaches together with their main properties with respect to the applied data, recognition algorithm, and evaluation results. It can be seen that more recent studies already incorporate a sufficient number of categories (up to 25 different instruments) in real-world complex mixtures (polyphonies up to 10 concurrent instruments), obtaining acceptable performance figures. A direct comparison between them, however, is not possible due to the different data sources used (note the dominance of personal collections) and differences in the applied categories. Moreover, only 3 studies tested their approaches on a sufficient variety of musical styles, giving insights into their generalisation and data handling capabilities.

Pure pattern recognition. Many studies dealing with instrument recognition from polytimbral audio data tried to directly apply the knowledge derived from the monophonic scenario. Although some extensions had to be incorporated, the methodology and techniques remained the same in the majority of cases. For instance, Simmermacher et al. (2006) approached the identification of four classical instruments (Flute, Piano, Trumpet, and Violin) in solo passages from concerti and sonatas by applying a classifier trained on isolated note samples (IOWA collection). The authors assumed that in the test scenario, where the soloing instrument is accompanied by various other instruments, the extracted features remain descriptive with respect to the target instrument, since it predominates the mixture. Both perceptually motivated features and features from the MPEG-7 standard were used in combination with classical MFCCs, at which feature selection was performed to reduce dimensionality. Results showed an average classification accuracy of maximum 94% depending on the respective set of audio features applied.

Essid et al. (2006a) presented a rather unconventional approach for identifying musical instruments in polytimbral music. Unlike focussing on the individual instruments present in the mixture, the signal was classified according to its overall timbre, which results from the individual sounds of the concurrent instruments. The authors derived a suitable taxonomy by hierarchically clustering the training data prior to the actual classification process. The obtained categories were labelled according to the featured instrumentation and statistical models were built for the respective classes. The approach seems promising for data containing a limited set of target instruments – the study used instrumental jazz music – since it avoids all kind of preprocessing usually involved in the source segregation of polytimbral data. On the other hand, it is to question whether the method can be applied to more varied types of music with a greater number of instrumental combinations.

Little & Pardo (2008) used a weakly-labelled data set to learn target instruments directly from polytimbral mixtures. Here, weakly-labelled refers to the fact that in a given training file the target is not assumed to be continuously present. Using 4 instruments from the IOWA sample collection, ar-

	Data & experimental settings					Algorithmic specifications				Evaluation		
Author	Poly.	Cat.	Type	Coll.	Genre	Class.	A priori	PreP.	PostP.	#Files	Metric	Score
Simmermacher et al. (2006)	4	4	real	pers.	C	SVM	×	×	×	10	Acc.	0.94
Essid et al. (2006a)*	4	12	real	pers.	J	SVM	×	×	✓	n.s.	Acc.	0.53
Little & Pardo (2008)	3	4	art. mix	IOWA	–	SVM	×	×	✓	20	Acc.	0.78
Kobayashi (2009)*	n.s.	10	real	pers.	P,R,J,W	LDA/RA	×	×	×	50	Acc.	0.88
Fuhrmann & Herrera (2010)*	10	12	real	pers.	C,P,R,J,W,E	SVM	×	×	✓	66	F	0.66
Eggink & Brown (2003)	2	5	real	pers.	C	GMM	×	✓	×	1	Acc.	1.0
Eggink & Brown (2004)	n.s.	5	real	pers.	C	GMM	×	✓	×	90	Acc.	0.86
Livshin & Rodet (2004)	2	7	real	pers.	C	LDA/kNN	×	✓	×	108	Acc.	n.s.
Kitahara et al. (2006)	3	4	syn. MIDI	RWC	C	HMM	×	✓	✓	n.s.	Acc.	0.83
Kitahara et al. (2007)	4	5	syn. MIDI	RWC	n.s.	Gauss.	✓	✓	✓	3	Acc.	0.71
Heittola et al. (2009)	6	19	art. mix	RWC	–	GMM	✓	✓	✓	100	F	0.59
Pei & Hsu (2009)	3	5	real	pers.	C	SVM	✓	✓	✓	200	Acc.	0.85
Barbedo & Tzanetakis (2011)*	7	25	real	pers.	C,P,R,J	DS	×	✓	✓	100	F	0.73
Cont et al. (2007)	2	2	real mix	pers.	n.s.	NMF	×	×	×	4	Acc.	n.s.
Leveau et al. (2007)	4	7	real mix	pers.	n.s.	MP	×	×	✓	100	Acc.	0.17
Burred et al. (2010)	4	5	art. mix	RWC	–	prob. dist.	×	✓	×	100	Acc.	0.56

Table 3.1: Comparative view on the approaches for recognising pitched instruments from polytimbral data. Asterisks indicate works which include percussive instruments in the recognition process. Synonyms of the header denote, among others, polyphonic density (Poly.), number of categories (Cat.), type of data used (Type), the name of the data collection (Coll.), the classification method (Class.), imposed a priori knowledge (A priori), any form of pre-processing (PreP.) and post-processing (PostP.), and the number of entire tracks for evaluation (#Files). Abbreviations for the evaluation metric refer to Accuracy (Acc.) and F-measure (F). Furthermore, the legend for musical genres include Classical (C), Pop (P), Rock (R), Jazz (J), World (W), and Electronic (E). The three main blocks represent the grouping into pure and enhanced pattern recognition, and template matching with respect to the recognition approach.

tificial mixtures of a maximum polyphony of 3 were created for training and testing, in a random manner at different mixing levels of target to background, in order to estimate the capabilities of the approach. Then, classifiers were constructed using instances taken from these training files. The produced models showed superior performance compared to models trained on isolated notes only, which indicates that sound mixtures exhibit many spectro-temporal characteristics different from isolated sounds. Reported results included a recognition accuracy of 78%, in comparison to 55% of the model trained with isolated tones.

An evolutionary method was applied by Kobayashi (2009) to generate an instrument detector. The approach used genetic algorithms and feature selection along with a Linear Discriminant or Regression Analysis (LDA/RA) to automatically generate the feature set and classification mapping from a set of supplied basis functions. Moreover, foreground/background separation is applied to the stereo signal to separate monaurally from binaurally recorded instruments (e.g. voice versus string sections). The separated data is further transformed via the wavelet transform into a time-pitch representation by applying mother wavelets corresponding to a semitone band-pass filter. Ten broad instrumental categories were annotated in 100 music pieces taken from commercial recordings, which were cut into 1 second extracts, shuffled and split into train and test set. The author reported excellent results in terms of recognition accuracy (88% on average), despite the absence of a clear separation of training and testing data.

Another complete system for labelling musical excerpts in terms of musical instruments was presented by Fuhrmann & Herrera (2010), virtually combining the approaches of Little & Pardo (2008) and Joder et al. (2009). Hence, statistical models (SVMs) for 12 instruments were trained by early integrated low-level features which were extracted from weakly-labelled polytimbral music audio data, whereas a late integration of the classifier decisions via contextual analysis of the music provided the final labels. Two separate classifiers for pitched and percussive instruments were employed, and several strategies for the late integration examined. Moreover, the applied dataset was purposely designed for containing both music pieces from various genres (even rather atypical styles such as electronic music were used) and unknown, i.e. not trained, categories to estimate the performance of the system under realistic conditions. Reported results of an F-measure of 0.66 for around 240 excerpts extracted from 66 tracks indicate the potential of the approach as well as some clear limitations which cannot be overcome without the application of more enhanced signal processing techniques.

Enhanced pattern recognition. The studies presented here addressed the problem of source interference from polytimbral audio by incorporating additional knowledge about the source signals in the recognition process. Pitch and onset information were often used to determine the parts of the signal which are unaffected by the interference. Furthermore, some authors applied source-separation to pre-process the mix and apply pattern recognition techniques on the obtained source signals.

Eggink & Brown published two studies dealing with instrument classification from polyphonic mixtures. In their first work the authors applied the missing feature approach to instrument recognition from polyphonies in order to handle feature values corrupted by interfering frequency components (Eggink & Brown, 2003). One composition of a Duet was analysed by first estimating the fundamental frequency using a harmonic sieve, which eliminates frequency regions not exclusively be-

longing to the target source. Hence, all source interference was excluded prior to the classification step. A statistical model (GMM) trained on solo performances and isolated tones for the five classes Cello, Clarinet, Flute, Oboe, and Violin was then applied to obtain the class membership for each instance. Results indicate that the models were able to recognise the instruments from the masked signal, although the testing conditions were quite limited (only 1 excerpt from 1 recording was used).

In a subsequent study Eggink & Brown (2004) studied the identification performance of a slightly modified recognition system for the same five instruments in a richer polyphonic context. Classical sonatas and concerti were analysed by recognising the soloing instrument. The fundamental frequency estimation algorithm based on the harmonic sieve was applied to locate the partials of the predominant instrument. Then, a statistical prototype (GMM) was trained with low-level features extracted from the spectral peak data on isolated notes and monophonic recordings. The models were created for every instrument and every fundamental frequency to account for the pitch dependency of the instruments' timbre. Finally, an unknown frame was classified according to the model which returned the highest probability, integrating the decisions along the whole excerpt. Evaluation on 90 classical pieces resulted in an average recognition accuracy of 86%.

Livshin & Rodet (2004) performed identification from duet compositions in addition to a conventional monophonic study within a real time framework. Their approach estimated the frequency components of the respective instruments, which were then input to a subtraction algorithm to isolate the two sources. From each source, features, which had been selected from the monophonic dataset by repeated LDA, were extracted and classified with a kNN rule. The performance of the duet system showed promising recognition accuracy, although the evaluation scenario was quite restricted.

A complete probabilistic approach for instrument identification in polyphonies was presented by Kitahara et al. (2006). The system used a probabilistic fundamental frequency estimation module based on the work of Goto (2004) for detecting melody and base lines in complex audio. Additionally to the note probability an instrument probability was derived by computing the harmonic structure for every possible fundamental frequency and extracting 28 features to train 15-state HMMs. To derive the final estimate the values for note and instrument probability were then multiplied and a maximum likelihood decision returned the instrument for each time-frequency point. The resulting representation was further post-processed by an additional HMM with limited transition probabilities to derive the most probable instruments given the observed probabilities. An average recognition accuracy of 83% on a 4 instrument identification task was reported from experiments using music audio data generated with the RWC instrument samples, but limited to a polyphony of three. Furthermore, neither drums nor vocal samples were used to test the robustness of the system.

Furthermore, Kitahara et al. (2007) presented a method to recognise musical instruments from artificial music audio data by eliminating unreliable feature data caused by the source interference. The authors developed a weighting method that estimates to what degree a given feature is influenced by overlapping frequency components of concurrent sources. LDA was used to minimise within-class and maximize between-class variance, thus enhancing features which discriminate best the categories. The features were extracted from the harmonic structures of the corresponding instruments using annotated fundamental frequencies and onset times. For evaluation the authors constructed

a dataset with artificial mixtures up to a polyphony of four, generated from the RWC instrumental library. Pitch-dependent Gaussian prototypes were trained for all instruments, and recognition was derived by taking a maximum a posteriori decision. By additionally analysing musical context the resulting class hypotheses were corrected and performance improved. Results of 71% of average accuracy were reported for the maximum tested polyphony of four.

Heittola et al. (2009) built a recognition system integrating informed source separation prior to the classification process. First, a polyphonic pitch estimator provided the values of the concurrent fundamental frequencies in all frames of a polyphonic mixture. The pitch information was then used to initialise a Non-negative Matrix Factorisation (NMF) separation algorithm which output streams corresponding to the individual instruments. Features were extracted from the generated source spectrograms and finally evaluated by pre-trained GMM models of the instruments. Polyphonic mixtures of 4 seconds length with a constant number of simultaneous instruments were generated for training and testing in a quasi-random manner using the samples from the RWC library. Reported results for the 19 instrument recognition problem included an F-measure of 0.59 for a polyphony of 6. Given these excellent performance figures, the approach, however, seems to be preliminary since the number of sources is needed as input parameter and a constant number of sources along the excerpt is assumed.

Fuzzy clustering algorithms were applied by Pei & Hsu (2009) to group feature vectors according to the dominant instruments in a given piece of music. The features were derived by averaging short-term values along beat-defined texture windows. From each resulting cluster the most confident members were taken for classification using a SVM model trained on monophonic recordings of 5 instruments. Results showed an average accuracy of 85%, according to the authors a comparable quantity with respect to literature. The presented algorithm requires the number of concurrent instruments beforehand to work properly, since this parameter defines the number of the final instrumental labels.

Finally, Barbedo & Tzanetakis (2011) developed a simple strategy for instrument recognition from polyphonies by extensively using voting and majority rules. The core system classifies isolated individual partials according to the instrumental categories. By focussing on isolated partials only, the authors purposely excluded ambiguous data caused by source interference. Hence, the systems is working on a pre-processing which estimates the number of sources and the corresponding fundamental frequencies for each frame. For a given fundamental frequency partials are then found by peak picking in the neighbourhood of their estimated positions and isolated by a filtering process. Then, features are extracted and pairwise classification for each instrument combination performed. A first majority vote among all pairs' decisions determines the instrument for the respective partial, a second one identifies the instrument of the given fundamental frequency of the considered partials. This is repeated for all simultaneous sources in a given frame. Finally, all instruments present in more than 5% of the total amount of frames are taken as labels for the entire signal. Experimental results for 25 instrument on music taken from several musical genres showed excellent recognition performance (F-measure of 0.73), although the authors admitted that accuracy dropped significantly when analysing music containing heavy percussive elements. This seems reasonable since the broadband spectra of these instruments are likely to mask partials from pitched instruments.

Template matching. The last group of approaches covers methods based on evaluating predefined templates related to the musical instruments on an unknown mixture signal. Similar to percussive instrument detection, templates can be constructed, whose match quality gives an estimate regarding the presence of the respective musical instruments. The match of a given instrument is usually determined by a predefined distance metric, calculated between the template and the signal. Some approaches rely on a single template per instrument, where classification is derived by evaluating a single distance measure, whereas others construct multiple instances per instrument and decompose the signal via an optimization method involving all templates simultaneously.

Cont et al. (2007) used a NMF decomposition system to simultaneously estimate the pitches and instruments of a given polyphonic recording. To capture the instrument specific information the authors used the modulation spectrum as input to the NMF algorithm. Templates for each note of each instrument were constructed in the training process as single basis function in the resulting classification matrix. Prediction was then performed by matching an unknown input to the training matrix, using additionally sparsity constraints to limit the solution space of the NMF decomposition. The authors evaluated the system both subjectively and objectively, whereas the latter was rather limited. Two mixtures of two different monophonic recordings were used but no average performance figure given. The authors, however, argued that given the difficulty of the addressed task, the results were satisfactory.

Sparse coding algorithms can further be considered as template matching processes. In particular, dictionary based algorithms such as the MP algorithm match templates from a given dictionary to the signal. Leveau et al. (2007) applied dictionaries containing harmonic atoms corresponding to the individual pitches of different musical instruments to decompose a polyphonic mixture. The dictionaries were trained with isolated notes and refined by further adapting them with monophonic phrases. When decomposing a mixture signal the selected atoms indicate which instruments at which pitches are present at a given time in the mix. At each time instance, the resulting atoms are then grouped into ensemble classes, which salience depend on the salience of the containing atoms. To derive labels for an entire segment of music, a probabilistic voting algorithm, which first maps the saliences of the ensembles onto log-likelihoods and then sums the resulting values for each ensemble, was applied to obtain the most likely ensemble. Evaluation was performed on a dataset consisting of artificial signals which were generated by mixing monophonic phrases, extracted from commercial recordings. Results of the evaluation on instrument recognition performance only showed satisfactory results for rather small polyphonies (i.e. ≤ 3 concurrent sources), indicating that the technique is not robust enough to process real music audio data containing more difficult source signals.

Finally, Burred et al. (2010) used source separation prior to a template matching algorithm in order to apply prototypical, spectro-temporal envelopes for classification. These timbre models were derived by applying a Principal Component Analysis (PCA) on the spectral envelopes of all training data of the respective instruments. The separation algorithm combined onset and partial tracking information to isolate individual notes in a polyphonic mixture. The evaluation mixtures consisted of simultaneous, quasi-random sequences of isolates notes from two octaves of the respective instruments. The extracted notes were then directly matched to the timbre models of five different musical instruments. Classification was finally derived by evaluating the probabilistic distances based on Gaussian processes to all models and choosing the model which provided the smallest distance. Ac-

curacy for instrument recognition yielded 56% in mixtures of a polyphony of three, for all correctly detected onsets.

3.5.2 Percussive instruments

Since the focus of our presented instrument recognition approach lies on pitched instruments – our algorithm roughly estimates the presence of the drumkit in a music signal – we only shortly review some recent approaches to the classification of percussive instruments from polytimbral music. Apart from the obvious differences in their spectral characteristics, percussive instruments generally carry more energy in the mixture (e.g. consider the presence of the drumkit in pop or rock music) than pitched sources. Therefore, the application of proper onset detection algorithms allows for a more robust localisation of the percussive events in time, as compared to pitched instruments. Furthermore, percussive sounds exhibiting rather stable characteristics along time (i.e. the sound of a Bass Drum will not change dramatically in a single piece of music) and their number is usually quite limited inside a given musical composition. Due to these properties the problem of recognising percussive instruments from polyphonies gained some attention in the MIR research community. Since an extensive overview of the relevant approaches is not provided here, we refer to the comprehensive review presented by Haro (2008).

Gillet & Richard (2008) constructed an algorithm for drum transcription combining information from the original polyphonic music signal and an automatically enhanced drum-track. In their framework the authors evaluated two drum enhancement algorithms for cancelling pitched components; the first used information provided by binaural cues, the second applied an eigenvalue decomposition for a band-wise separation. The basic transcription approach consisted of an onset detection stage, from which detected events a feature vector was extracted, both from the original and enhanced track. After feature selection a pre-trained classification model (SVM with normalised RBF kernels) was applied to predict the instruments in the respective events. To combine the information of the two tracks, the authors evaluated early as well as late fusion strategies, which refer, respectively, to the combination of the feature vectors prior to classification and the combination of classifiers' decisions. Evaluation was performed on the publicly available ENST collection (Gillet & Richard, 2006), which provides a full annotation of almost all percussive events as well as separated drum and accompaniment recordings of the featured tracks. Besides evaluating the separation accuracy of the respective algorithms, the authors reported the classification accuracies obtained by the system for three instruments (Bass Drum, Snare Drum, and Hi-Hat). First, only a slight improvement in recognition performance could be observed when comparing the results from the enhanced to the original track. However, the late fusion of the classifier decisions improved the results significantly, indicating that the two signals cover complementary information which can be exploited for percussion detection.

Alongside their system for pitched instrument recognition, Fuhrmann et al. (2009a) used a similar approach for percussive instrument classification. Here, the same methodology was applied as described above; an onset detection algorithm detected percussive events in polytimbral music, from which frame-wise extracted acoustic features were derived. These features were integrated along

a texture window placed at the respective onset and classified by a pre-trained recognition model (again, SVMs were used). Besides reporting similar identification performance in comparison to Gillet & Richard (2008), the authors additionally evaluated the importance of temporal aspects in the feature integration step. Since it has been shown that temporal characteristics of timbre are essential for human recognition, three levels of temporal encoding were tested on their influence on the recognition performance from polyphonic music. Experimental results showed that a coarse level of encoding (i.e. using statistics on the derivatives of the respective features) is beneficial for accuracy, whereas a fine-grained temporal description of the feature evolution in time is not improving recognition performance, indicating that these characteristics are difficult to extract from polyphonies given the assumable source interference.

Finally, Paulus & Klapuri (2009) approached the problem of transcribing drum events from polytimbral music by applying a neural network of connected HMMs for time-located instrument recognition. With this approach the authors argued to overcome the shortcomings usually encountered when performing segmentation and recognition of the audio separately, as implemented by the systems reviewed above. The authors additionally compared a modelling strategy of instrument combinations to the common strategy modelling the individual sources independently. In their approach the audio was first analysed by a sinusoidal-plus-noise model, which separated pitched components from noisy portions of the signal. The tonal information was discarded and features were extracted from the residual. MFCCs and their first derivatives were applied for training the individual signal models (4-state left-to-right HMMs), based on the information obtained from the annotation data. Once the models had been trained the connection between them was implemented by concatenating the individual transition matrices and incorporating inter-model transition probabilities, all deduced from the training data. In the recognition step, the connected models are then applied and the Viterbi algorithm used to decode the obtained sequence. Results on the public available ENST dataset, including 8 different percussive categories, showed superior performance for the individual instrument modelling approach in combination with a model adaptation algorithm, which adapts the trained models to the audio under analysis. Reported evaluation scores yielded an F-measure of 0.82 and 0.75 for isolated drums and full mixture signals, respectively.

3.6 Discussion and conclusions

In this chapter we identified the automatic recognition of musical instruments as a very active field in the research community of MCP, which has produced a great amount of high-quality works – as well as many noisy studies, too – related to the task. Many conceptually different approaches have been presented to tackle the problem, incorporating knowledge derived from human perception and cognition studies, and recent techniques from machine learning or signal processing research.

Perceptual studies have revealed the basic components of the timbral sensation of instrumental sounds. Additionally, several mutual attribute have been identified which cause confusions among certain groups of instruments or effect in a blending of their timbres when simultaneous active. This

blending properties also hinder their segregation and consequentially the individual recognition of the instruments. Hence, the physical characteristics of the instruments determine a kind-of upper limit for musical instrument recognition, even for humans.

The fact that some of the developed systems for monophonic input sounds score close to the performance of human recognition indicates that the problem itself can be regarded as solved to a certain extent. Since this kind of data allow for the best insights into the nature of the recognition task, results suggest that machines are able to extract the timbre identifying properties of musical instruments' sounds and use them to build reliable recognition systems. In particular, studies on feature applicability for musical instrument recognition showed that mainly the robust estimation of the spectral envelope enables the successful recognition, and modelling the temporal evolution of the sound improves results subsequently (Agostini et al., 2003; Lagrange et al., 2010; Nielsen et al., 2007). Moreover, the applied methods from machine learning allow for the handling of complex group dependencies in hierarchical representations and for reliable intra-family separation of musical instruments.

However, multi-source signals still cause a lot of problems for automatic recognition systems. In connection to the aforementioned it seems that the processing of the complex acoustical scene prior to the actual identification step is of major importance. It is assumed that the human mind uses a complex combination of attention mechanisms, auditory restoration and virtual pitch to perform a streaming-by-timbre to segregate the individual perceptual streams and determine their timbres. This process is presumably of mutual nature, thus recognition and segregation accompany each other. Since hearing research is far away from understanding these complex operations, a computational modelling seems to be – at least from the current signal processing point-of-view – nearly impossible. Up to now there exist no artificial system that can handle the interferences between concurrent sources in an auditory scene in such a way that a robust source recognition is possible.

Nevertheless, some approaches towards the automatic recognition of musical instruments from polyphonies assumed slightly simplified conditions to accomplish recognition systems that work on even complex data. Focussing on parts of the signal where no or only slight source interference can be observed allows for a robust extraction of the instruments' invariances. In this way several systems have been constructed that reach acceptable recognition performances even in complex polytimbral environments.

Comparing the different approaches from literature remains a very difficult, nearly impossible, task. Since most of the studies used their own dataset for training and testing, a direct comparison is not possible (see also Table 3.1). Moreover, even if the number of classes and the source complexity is the same, the employed music audio data may be extremely different. Furthermore, most works still impose restrictions on the nature of their data and algorithms, which further complicates any general comparison. Thus, the reported evaluation figures can only be partially used to assess the recognition performance of the respective algorithms in a more general way. To conclude, the best way to objectively estimate the performance of a musical instrument recognition algorithm is to perform its evaluation on a rich variety of natural data, i.e. real music. In this context the number of classes, the amount of noise, and the data complexity reach a realistic level on which the method must perform. Only if tested at this scale, the real capacities of the approaches can be identified!

4

Label inference

From frame-level recognition to contextual label extraction

What remains evident from the literature review presented in the previous chapter is that hardly any of the examined approaches does not impose restrictions to the employed training and evaluation data, or to the algorithmic processing itself. Most methods applied narrow taxonomies in terms of the modelled musical instruments, tested with a limited polyphonic complexity, i.e. number of concurrent sources, or evaluated with an inadequate data diversity in terms of musical genres. Moreover, many studies used artificially created data lacking any kind of musical context for evaluation. As a consequence, almost all of these approaches cannot be applied and exploited in systems of a broader purpose, e.g. typical MIR applications such as search and retrieval or recommender systems. Besides, the heavy restrictions involve only scant advances to models of listening, or more general, machine listening systems. Hence, those approaches do not contribute to research in a scientific sense. From this viewpoint, and as already pointed out in Section 1.4, the primary objective of this thesis was to design a method without the aforementioned shortcomings in connection with its embedding into a typical MIR framework.

In this chapter we present our method for assigning instrumental labels to an audio excerpt of any finite length. Here, we want to note the subtle difference we take when using the possibly ambiguous terms *classification* and *labelling*. The former is used in connection with the raw frame-based estimates predicted by the classification module, while the latter connotes attaching a semantic label to the entire analysed signal. Hence, whenever referring to classification we reside on a frame level, while labelling comprises the integration of the signal's entire temporal dimension. Consequentially, the term *label inference* denotes the extraction of semantic information in terms of labels, or tags, from a series of frame-based estimates, output by the classifier.

Conceptually, this chapter is divided into two parts which cover, respectively, the aforementioned classification and labelling stages of the presented method. Before that, we first present the the-

oretical methodology underlying the overall design process (Section 4.1). Then, we present the developed approach towards musical instrument classification in Section 4.2, which is further subdivided into the sections covering the pitched (Section 4.2.3) and percussive instruments (Section 4.2.4). Here, we illustrate the respective taxonomic choices, the applied data, the experimental settings, and the results of the corresponding classification problem together with a thorough analysis of the involved acoustical descriptions in terms of audio features and the resulting prediction errors. Section 4.3 covers the strategies examined for integrating the frame-based classification output to derive instrumental labels given an unknown music excerpt; we first introduce the underlying conception (Section 4.3.1) and the constructed evaluation dataset (Section 4.3.2), followed by a brief discussion of the applied evaluation methodology (Section 4.3.4) and all obtained results (Section 4.3.5). Furthermore, Section 4.3.6 contains an analysis of the resulting labelling errors. Finally, this chapter is closed by a comparison of the presented method's performance to other state-of-the-art approaches (Section 4.4.1) and a general discussion in Section 4.4.2.

4.1 Concepts

Prior to examining the methodological details, we want to illustrate our main assumptions that led to the development of the presented method. These assumptions, or hypotheses, refer to the basic extraction and modelling approaches of the musical instruments' sound characteristics from music audio signals and are subsequently validated in the remainder of this chapter. They can be stated as follows:

1. The perceptual characteristics, or timbre, of a certain musical instrument can be extracted from polytimbral music data, provided a certain amount of predominance[1] of the target source.
2. Musical context provides basic means for label inference, as it is similarly utilised by the human mind.
3. This extracted information enables a meaningful modelling of musical instruments in connection with MCP/MIR.

ad 1. Our approach towards extracting the instrument's characteristics from the audio data relies on a statistical pattern recognition scheme. This choice is perceptually and cognitively plausible from the viewpoint of how the human mind organises knowledge (Section 2.1.1.3); moreover, the approach is widely used in related literature (Essid et al., 2006b; Gillet & Richard, 2008; Heittola et al., 2009; Kitahara et al., 2007; Martin, 1999). In this framework we generate statistical models of musical instruments and apply these models for prediction. In our particular conception, modelling itself is performed directly on the presumably polytimbral data without any form of preprocessing. In doing so we purposely avoid source separation and related techniques of polyphonic

[1] In this thesis, we pragmatically define the predominance of an instrument as being perceptually clearly audible in, and outstanding from, the context of other instruments playing simultaneously.

music transcription, since their applicability cannot be fully guaranteed in the context addressed by the developed method[2]. We furthermore want to examine the potential of the presented approach as a general method towards instrument recognition from musical compositions, concentrating on its own peculiarities and specificities. In consequence, we limit the method to the modelling and recognition of predominant sources from the music audio signal since we assume that the main characteristics of these instruments, encoded in their spectro-temporal envelope, are preserved. Thus, the polyphonic mixture sound is mainly affected by the spectral characteristics of the predominant instrument.

ad 2. From the perceptual point-of-view it seems evident that musical context provides important cues for sound source recognition (Grey, 1978; Kendall, 1986; Martin, 1999). However, only few approaches towards musical instrument recognition in polytimbral environments incorporate this general property of music. Moreover, there is a broad consensus among researchers that the temporal dimension provides necessary and complementary information for retrieval (Casey & Slaney, 2006). Here, we exploit the property of stationary sources in music to reliably extract the labels from a series of classifier predictions. We assume that musical instruments are played continuously for a certain amount of time, thus their predominance along time can be used as a robust cue for the label inference. Moreover, this approach enables the recognition of multiple predominant instruments in a given musical composition.

ad 3. Statistical modelling requires, in general, a sampling of the target population. Since in the majority of cases measuring all elements of the target population is impossible, the population is approximated by a representative sample (Lohr, 2009). In this regard, representativeness denotes the ability to model the characteristics of the population from the sample. Thus, the sample used for training a statistical model has to reflect the properties of and their variabilities inside the target population. In the context of this thesis, we can regard the above-defined problem as recognition from noisy data, since the accompaniment of a predominant source can be simply considered as noise. Here, the results obtained by Little & Pardo (2008) suggest that introducing "noise" in the training process of musical instrument classifiers improves the robustness and thus the recognition performance from polytimbral testing data. Hence, a meaningful modelling of musical instruments from polyphonies is possible, if, and only if, the training data reflects the variability of the sampled population. To guarantee this variety we emphasise the construction of the collections used to train the classification models, comprising a great variety in musical genres and styles, recording and production conditions, performers, articulation styles, etcetera. Moreover, the restriction of modelling predominant instruments only is not impairing the applicability of the method to various kinds of data; we can further assume that most of the targeted data, i.e. Western music compositions of any kind, exhibit enough predominant information related to musical instruments from which sufficient instrumental information can be gained. Finally, a meaningful modelling of the extracted information implies that the used taxonomy reflects the system's context. Thus, depending on the problem at hand, a too fine-grained taxonomy can result in a model too complex with respect to the observed data, causing a general performance loss. On the other hand, a too coarse taxonomy may not satisfy the user's information need, thus results in useless output. We therefore decided

[2]There is, however, recent evidence that an incorporation of polyphonic pre-processing techniques is beneficial for recognition systems under certain constraints, see e.g. (Barbedo & Tzanetakis, 2011; Haro & Herrera, 2009).

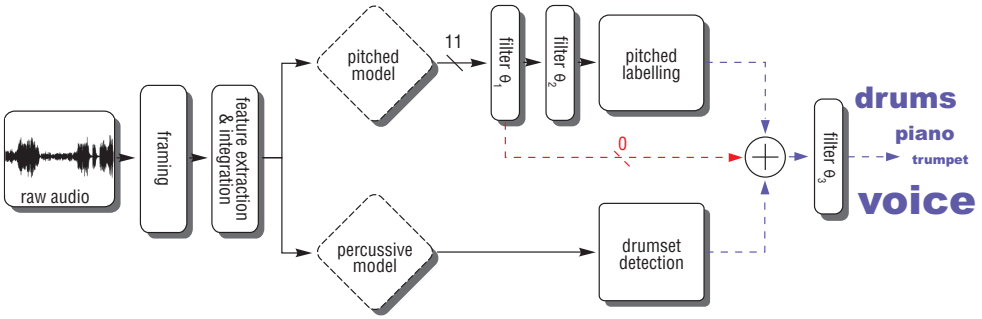


Figure 4.1: Block diagram of the presented label inference method. The audio signal is first chopped into chunks from which short-term features are extracted and integrated. The signal is then split in two separate paths representing the recognition of pitched instruments and the Drumset detection. Both branches apply a classification model to the feature vectors, at which the models' time-varying output is subsequently used for label inference. See text for more details.

on an abstract representation in the hierarchy of musical instruments[3], valid across multiple use cases and understandable by Everyman, incorporating pitched and percussive categories, as well as the human singing voice. All three carry important semantic information which is necessary for a sufficient description of the musical content in terms of instruments. Furthermore, pitched and percussive instruments are modelled separately, due to their evidently different acoustic characteristics, whereas the human singing voice is regarded as a pitched instrument and consequentially modelled in conjunction with the latter (see also Section 3.1).

Before entering the theoretical and experimental playground behind our method, Figure 4.1 shows a schematic illustration of the label inference process. Note the two separate branches in the classification and labelling stage, corresponding, respectively, to the pitched and percussive analysis.

4.2 Classification

4.2.1 Method

The most basic process executed by the presented method is the determination of the main instrument, for both pitched and percussive categories, within a short time-scale. For this purpose we employ a pattern recognition approach by following the typical notions of training and prediction; in the former a statistical model is trained by applying the training collection, the latter uses the trained model to predict class assignments for unseen input data. In both stages the basic methods of feature extraction, feature selection, and classification are involved. Figure 4.2 shows a conceptual illustration of this train/test setup, which can be summarised as follows; first, the signal of a

[3]This applied coarse taxonomy of musical instruments can be regarded as the entry-level for reasoning and recognition at an intermediate level of abstraction, as introduced by Minsky (1988) (see Section 2.1).

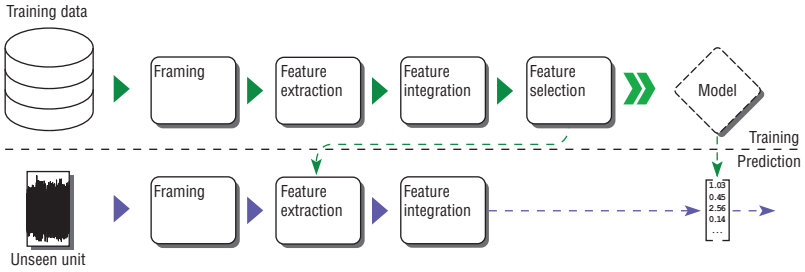


Figure 4.2: Illustration of the pattern recognition train/test process as employed in the presented method.

single acoustical unit[4] is partitioned into very short chunks, which are transformed into a low-level feature representation of the raw audio. The time series of extracted audio feature vectors along the unit is integrated into a single vector by taking statistical measures of the individual features. All so-processed units of the training data are then passed to a feature selection algorithm to reduce the dimensionality of the feature space. The resulting lower-dimensional representation of the data is finally applied to train the classification model, which is further used for the prediction on unseen instances. The next sections cover these structural building blocks of our classification system in detail. Parts of the approaches described in this section have been previously published by Fuhrmann et al. (2009a).

4.2.1.1 Audio features

Given an acoustical unit, the signal is weighted by an equal loudness filter (Robinson & Dadson, 1956), incorporating the resonances of the outer ear and the transfer function of the middle ear, and framed into short fragments of audio, using a window size of 46 ms and an overlap of 50%. For calculating the FFT a Blackman-Harris-92dB window function is applied to weight the data accordingly. A total of 92 commonly used acoustical features, describing the temporal, timbral, and pitch related properties of the signal are extracted. These features can be roughly classified as follows[5]:

Local energies. A great part of these features is based on the psycho-acoustical Bark scale (Zwicker & Terhardt, 1980), an implementation of the frequency resolution of the cochlea's Basilar membrane in terms of critical bands. Additionally to these 26 energy band values[6], we derive four broader energy bands, dividing the spectrum into the regions corresponding to the frequency limits of 20, 150, 800, 4 000, and 20 000 Hz. Finally, we introduce a global estimate of the signal's energy derived from its magnitude spectrum (Peeters, 2004).

Cepstral coefficients. We obtain MFCCs (Logan, 2000) by calculating the cepstrum of log-compressed energy bands derived from the Mel scale, a psychoacoustic measure of perceived

[4] In this thesis the term *acoustical unit* denotes the quantity of audio data, or length of the audio, the recognition models use to perform a single prediction.

[5] A complete mathematical formulation of all described features can be found in the Appendix.

[6] We expand the originally proposed 24 bands by replacing the lowest two by four corresponding bands, covering the frequencies between 20, 50, 100, 150, and 200 Hz. For convenience we provide a table containing the complete list of all 26 bands numbered with the applied indexing scheme together with the respective frequency ranges in the Appendix.

pitch height (see Section 2.1.1.2). In our implementation we extract the first 13 coefficients from 40 Mel-scaled frequency bands in a frequency range from 0 to 11 000 Hz. These features are used to estimate the spectral envelope of a signal, since the cepstrum calculation involves a source-filter deconvolution due to the logarithmic compression of the magnitudes (Schwarz, 1998).

Spectral contrast and valleys coefficients. A shape-based description of spectral peak energies in different frequency bands is used for capturing the spectral envelope characteristics of the signal under analysis (Akkermans et al., 2009). We calculate 6 coefficients for both contrast and valleys features, using the frequency intervals between 20, 330, 704, 1 200, 2 300, 4 700, and 11 000 Hz. To the best of our knowledge this feature has not been used in the context of automatic musical instrument recognition so far.

Linear prediction coefficients. These features are further used to describe the spectral envelope of the signal (Schwarz, 1998). Linear predictive coding aims at extrapolating a signal's sample value by linearly combining the values of previous samples, at which the coefficients represent the weights in this linear combinations. Since the coefficients can be regarded as the poles of an corresponding all-pole filter, they also refer to the local maxima of the estimated description of the spectral envelope. Here, we derive 11 coefficients from the linear predictive analysis.

Spectral. Various features are extracted from the signal to describe its spectral nature. Many of them are common statistical descriptions of the magnitude spectrum, including the centroid, spread, skewness, and kurtosis – all 4 calculated both on the basis of FFT bin and Bark band energies – spectral decrease, flatness, crest, flux, and roll-off factors (Peeters, 2004), high-frequency content of the spectrum (Gouyon, 2005), spectral strongpeak (Gouyon & Herrera, 2001), spectral dissonance (Plomp & Levelt, 1965), and spectral complexity (Streich, 2006).

Pitch. Based on the output of a monophonic pitch estimator, we derive several features describing the pitch and harmonic content of the signal. In particular, we calculate the pitch confidence (Brossier, 2006) and its derived harmonic features inharmonicity, odd-to-even harmonic energy ratio, and the three tristimuli (Peeters, 2004), which all use the pitch extracted by the monophonic estimator as input for the respective calculations. Additionally, we compute the pitch salience feature as defined by Boersma (1993).

Temporal. We calculate the zero crossing rate as an estimate of the "noisiness" of the signal (Peeters, 2004). This feature simply counts the sign changes of the time signal, hence periodic signals generally exhibit a lower value than noisy sounds.

4.2.1.2 Temporal integration

The framing process results in a time series of feature vectors along the unit, which are integrated by statistical measures of the individual features' distribution. This is motivated by the fact that humans use information accumulated from longer time scales to infer information regarding the instrumentation in a music listening context. Here, we apply the results obtained from a previous work, where we studied the effect of temporal encoding in the integration process on the classification accuracy for both pitched and percussive instruments in polytimbral contexts (Fuhrmann

et al., 2009a). We tested three levels of temporal granularity in this integration phase, showing that temporal information is important, but its extraction is limited due to the complex nature of the input signal. Hence, in this thesis we use simple temporal dependencies of feature vectors that are incorporated by considering their first difference values. That is, the difference between consecutive vectors is calculated and stacked upon the instantaneous values, thus doubling the size of the vector. Then, mean and variance statistics are taken from the resulting representation along time.

4.2.1.3 Feature selection

Creating highly dimensional feature spaces usually leads to redundancy and inconsistency in terms of individual features (Jain et al., 2000). To reduce the dimensionality of the data along with the models' complexity we apply a feature selection algorithm. We use the Correlation-based Feature Selection (CFS) method (Hall, 2000), which searches the feature space for the best subset of features, taking the correlation of the features with the class and the intercorrelation of the features inside the subset into account. More precisely, the goodness Γ of a feature subset S containing k features is defined as follows:

$$\Gamma_S = \frac{k\overline{\varrho_{\mathrm{fc}}}}{\sqrt{k + k(k-1)\overline{\varrho_{\mathrm{ff}}}}}, \tag{4.1}$$

where $\overline{\varrho_{\mathrm{cf}}}$ denotes the average feature-class correlation and $\overline{\varrho_{\mathrm{ff}}}$ the average feature-feature intercorrelation. If the problem at hand is classification, i.e. with discrete class assignments, the numerical input variables have to be discretised and the degree of association between different variables is given by the symmetrical uncertainty (Press et al., 1992)

$$U(X,Y) = 2 \times [\frac{H(Y) + H(X) - H(X,Y)}{H(X) + H(Y)}], \tag{4.2}$$

where $H(X)$ denotes the entropy of X. To derive the resulting subset of features in a reasonable amount of computation time (in general, evaluating all 2^k possible feature subsets is not feasible, with k being the total number of features), the method utilises a *Best First* search algorithm (Witten & Frank, 2005), implementing a greedy hill climbing strategy, to efficiently perform the search problem. This feature selection technique has been used widely in related works (e.g. Haro & Herrera, 2009; Herrera et al., 2003; Livshin & Rodet, 2006; Peeters, 2003).

If not stated differently, we perform feature selection in a 10-Fold procedure, i.e. we divide the data of each category into 10 folds of equal size, combine them into 10 different datasets each consisting of 9 of each categories' generated folds, and apply 10 feature selections. Thus each fold of each category participates in exactly 9 feature selections. This results in 10 lists of selected features from which we finally keep those features, which appear in at least 8 of the 10 generated lists. This procedure guarantees a more reliable and compact estimate of the most discriminative dimensions of the feature space, as the resulting set of features is independent of the algorithm's specific initialisation and search conditions.

4.2.1.4 Statistical modelling

The statistical models of the musical instruments applied in this work are implemented via SVMs (Vapnik, 1999). SVMs belong to the general class of learning methods building on kernels, or kernel machines. The principal idea behind is to transform a non-linear estimation problem into a linear one by using a kernel function. This function projects the data from the initially low-dimensional input space into a higher-dimensional feature space where linear estimation methods can be applied. Furthermore, the SVM is regarded as a discriminative classifier, hence applying a discriminative learning scheme (see Section 2.2.2.2), since it directly models the decision function between 2 classes and is not relying on prior estimated class-conditional probabilities.

Support Vector Classification (SVC) applies the principle of Structural Risk Minimisation (SRM) for finding the optimal decision boundary as introduced by Vapnik (1999). In short, SRM tries to minimise the actual risk, i.e. the expected test error, for a trained learning machine by implementing an upper bound on this risk, which is given by the learning method's performance on the training data and its capacity, i.e. the ability to learn from any data without error. Hence, SRM finds the set of decision functions which balances best the trade-off between the maximal accuracy on the actual training data and minimal overfitting to these particular data (Burges, 1998).

Given the training data pairs $\{\mathbf{x}_i, y_i\}, i = 1 \dots l, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$, the SVC finds the linear decision boundary that best separates the training instances $\mathbf{x}_i$ according to the binary class assignments y_i. Hence, the objective is to determine the parameters of the optimal hyperplane in the d-dimensional space denoted by

$$\mathbf{w}^T\mathbf{x} + b = 0. \tag{4.3}$$

Thus, a decision function can be derived which assigns to any $\mathbf{x}_i$ the class membership as follows,

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b > 0, \quad \text{for} \quad y_i = +1, \\ \mathbf{w}^T\mathbf{x}_i + b < 0, \quad \text{for} \quad y_i = -1. \end{aligned} \tag{4.4}$$

Hence, for constructing a SVC system one has to determine the proper values for $\mathbf{w}$ and b that define the decision boundary. However, for certain problems many possible $\mathbf{w}$s and bs may be identified, leading to the non-existence of a unique solution along with the risk of a low generalisation ability of the resulting classifier. To overcome the aforementioned limitations the idea of the maximal margin is introduced; instead of looking for a hyperplane that only separates the data, the aim is to determine the hyperplane which additionally maximises the distance to the closest point of either class. This concept of the maximal margin guarantees both better generalisation properties of the classifier and the uniqueness of the solution (Friedman et al., 2001). Hence, the parallel hyperplanes that define the maximal margin, which represents the optimal decision boundary, are, after scaling, given by

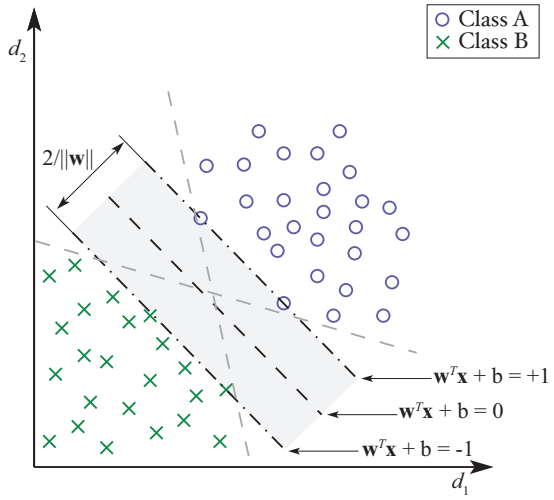


Figure 4.3: Principles of the support vector classification. The optimal decision boundary is represented by the dashed line, the corresponding hyperplanes framing the margin are dash-dotted. Note the dashed light grey hyperplanes which indicate possible hyperplanes separating the data but not fulfilling the maximum margin constraint.

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b &> +1, \\ \mathbf{w}^T\mathbf{x}_i + b &< -1, \end{aligned} \tag{4.5}$$

together with the corresponding width of the margin

$$\frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{\mathbf{w}^T\mathbf{w}}}. \tag{4.6}$$

Maximising this distance is equivalent to minimising its reciprocal, here resulting in a convex optimisation problem, i.e. a quadratic criterion together with linear inequality constraints (Friedman et al., 2001), that can be formulated as follows,

$$\begin{aligned} \hat{\mathbf{w}}, \hat{b} &= \arg\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to } \; & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1 \ldots l, \end{aligned} \tag{4.7}$$

where the inequalities of Eq. (4.5) are merged by multiplying each with the corresponding value of y_i to form the inequality constraint. Since the decision function is linear, SVC belongs to the class of linear classification; moreover, as the objective is to maximise the margin, a SVM generally denotes a maximum margin classifier. Figure 4.3 illustrates these geometric considerations for a separable problem in two dimensions.

Real world problems, however, only exhibit in rare cases the property of a perfectly linear separability of the data. To relax the constrain of a perfect separation, SVC allows for a certain amount of training error, creating a "soft" margin. Hence a penalty term is introduced in Eq. (4.7) to compensate for these errors while keeping the problem still linear. Moreover, many problems cannot be solved by a direct application of linear estimation methods, thus the data is mapped from the input space into a higher-dimensional feature space, where a general linear solution is more probable, via a kernel function $\phi(\cdot)$ (see above). This leads to the formulation of the standard SVM optimisation problem as presented by Cortes & Vapnik (1995),

$$\begin{aligned} \hat{\mathbf{w}}, \hat{b}, \hat{\xi} &= \arg\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to } \; & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, i = 1 \ldots l, \xi_i > 0, \end{aligned} \tag{4.8}$$

introducing, besides the kernel function $\phi(\cdot)$, the cost, or regularisation parameter C and the slack variable ξ.

Since the projected data may exhibit a very high, possibly infinite, dimensionality, the dual problem is used to derive a solution for $\mathbf{w}$. Solving the dual problem is simpler than than solving the corresponding primal, and can be achieved with standard optimisation techniques (Friedman et al., 2001). Here, the Lagrangian Dual simplifies the optimisation to a great extent since the dimensionality of the problem, which can be infinite, is reduced to l. As a result, $\mathbf{w}$ is defined as a weighted linear combination of the training instances. The weights, which are derived from the solution of the dual problem, correspond to the scalar Langrange multipliers α_i. Hence, the optimal $\hat{\mathbf{w}}$ can be written as

$$\hat{\mathbf{w}} = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i). \tag{4.9}$$

In case of the evaluation of the decision function, $\hat{\mathbf{w}}$ is substituted into the original decision function $\mathbf{w}^T\mathbf{x}+b$. However, the obtained equation exhibits the calculation of an inner product in the feature space, which is difficult to achieve due to the high dimensionality of the data. Hence, special kernel functions (symmetric and positive definite $\phi(\cdot)$) are applied that allow for a calculation of this inner product directly in the low-dimensional input space. The resulting relation can then be elegantly written as a kernel evaluation in the input space

$$\hat{\mathbf{w}}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i)^T\phi(\mathbf{x}) + b = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{4.10}$$

By using this so-called *kernel trick* the high-dimensional vector $\mathbf{w}$ is never explicitly used in both the calculation and the evaluation of the decision function. The complex calculation of the inner product

in the high-dimensional feature space is rather replaced by the kernel evaluation in the input space. Moreover, many of the α_is are zero, reducing the terms in the summation of Eq. (4.10). The training instances $\mathbf{x}_i$ with a corresponding $\alpha_i \neq 0$ are therefore called the support vectors[7]. Thus SVMs handle both forms of the so-called *curse of dimensionality*; moderate complexity is guaranteed while overfitting is avoided by using only the most decisive instances in the construction process of the decision function (Burges, 1998).

Additionally, many application scenarios need an estimate of the "class belongingness" of the testing instances rather than a categorical label of ± 1. Hence, the output of a SVM has to be transformed into a probabilistic estimate, i.e. a real number between 0 and 1, by using methods such as the one proposed by Platt (1999). Here the instances' posterior can be approximated via a mapping of the classifier's output into probabilities using a sigmoid function.

As indicated above, SVMs are inherently binary classifiers. Thus, in a multi-class problem, individual binary classifiers are combined into a single classification system. Basically, there exist two distinct approaches, related to the nature of the classification problem (Manning et al., 2009), to combine multiple categories in a SVM architecture. In an *any-of* situation a given instance can belong to several classes simultaneously or none at all (one-vs-all architecture), while a *one-of* classification problem assumes that instances are only affiliated with a single category (one-vs-one architecture). Hence, the specific choice of the architecture depends on the mutual exclusiveness of the classes.

A K-class one-vs-all classification system comprises K independent binary classifiers, each one modelling the target category and its respective complement, i.e. the "rest" class. That is, evaluation of one category is not influencing the decisions on all other classes. Thus a single prediction includes the application of K classifiers to one single data instance.

In case of a one-vs-one scheme the classification system is built form $K(K-1)/2$ individual models, with K being the number of classes. Here, category membership and probabilistic estimates for all target classes of the given instance have to be derived from the combined raw output of the binary classifiers. Several strategies such as voting (the class which scores the most binary votes wins) or maximum likelihood decision (the class exhibiting the highest single probability value wins) output the class label of the instance under analysis. However, in many situations class-wise probabilistic estimates are desired for subsequent processing. Then methods for combining the class probabilities, termed *pair-wise coupling*, can be applied (Hastie & Tibshirani, 1998; Wu et al., 2004).

In all subsequent experiments we use the SVM implementation provided by LIBSVM[8]. The library provides two different versions of the classifier (C-SVC and nu-SVC) together with 4 different kernel functions (linear, polynomial, RBF, and sigmoid kernel). Moreover, in case of an one-vs-one architecture, pairwise coupling of the individual probabilistic estimates is applied to obtain the class-wise values using the method presented by Wu et al. (2004).

[7] Those instances falling on the margin hyperplanes are furthermore used together with the corresponding α to derive the constant b.

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4.2.2 Evaluation methodology

Evaluating statistical pattern recognition systems refers to assessing its error on the target population. Here, the error rate P_e generally captures this performance of such systems. In practice, however, P_e cannot be determined directly due to finite sample sizes and unknown data distributions (Jain et al., 2000). As a result of these common limitations, P_e is usually approximated by the error on the used sample of the target population. Typically, a split procedure is followed, which divides the available instances into a training and test set, assuming mutual statistical independence. Then, a classifier is constructed using the training samples and the system's error rate is estimated via the percentage of misclassified test samples. Often, a single evaluation process provides only poor insights into the generalisation abilities of the system, hence reflecting real world conditions only weakly. However, given a single dataset, there exist numerous methods for partitioning the data into training and testing parts for near-optimal performance estimation, e.g. holdout, leave-one-out, or rotation methods (Duda et al., 2001).

In our experiments we apply a 10-Fold CV procedure from which the average accuracy $\overline{A}$ is obtained; the data is divided into 10 Folds, at which 9 parts are used for training and one for testing the respective model in a rotation scheme, averaging the accuracies A of the 10 different testing runs as performance estimate. Here, A refers to the fraction of correctly predicted evaluation instances. It is calculated by comparing the class estimates obtained from a maximum likelihood decision on the probabilistic output of the respective model to the ground truth labels of the instances. To further account for the initialisation and randomisation mechanisms in the fold generation process, we perform 10 independent runs of the CV and average the resulting accuracies obtaining a robust estimate of the overall recognition performance of the developed classification system.

Due to their particular conception – generalising from the sample to the target population – most pattern recognition systems are highly sensitive to the distribution of samples among the respective categories. Under the constraint of minimising the amount of wrongly predicted samples, such systems usually favour predictions for the majority classes. Common solutions for avoiding these class-specific biases include adjusting the costs for misclassification in the respective categories, or the artificial sampling of either the majority (down) or the minority (up) class (Akbani et al., 2004). To avoid any bias towards more frequent categories we always use a balanced dataset, i.e. the same amount of instances in all classes, in all upcoming classification experiments. In the case of an imbalance we therefore limit the amount of instances per category to the amount of instances the class with the fewest total amount exhibits. All categories comprising more than this value are randomly downsampled to guarantee a flat class distribution.

Moreover, we introduce additional class-specific metrics for assessing the performance of the system in recognising the individual categories. Here, we use standard metrics of precision and recall, as formally defined by Baeza-Yates & Ribeiro-Neto (1999),

$$P = \frac{|Retrieved \cap Relevant|}{|Retrieved|}, \quad \text{and} \quad R = \frac{|Retrieved \cap Relevant|}{|Relevant|}, \tag{4.11}$$

where $|\cdot|$ denotes the cardinality operator.

In a classification context both metrics can be rewritten using the notions of true and false positives plus negatives, i.e. tp, fp, tn, fn. Then, P is defined as $\frac{tp}{tp+fp}$ and R similarly as $\frac{tp}{tp+fn}$. Furthermore, we apply the balanced F-score, or F-measure, to connect the aforementioned:

$$F = 2\frac{PR}{P+R}. \tag{4.12}$$

Finally, we estimate the performance of the classification system under the null hypothesis of no learnt discrimination with respect to the modelled categories, i.e. a random assignment of the labels. $\overline{A}_{null}$ is consequentially defined as $1/K$, with K being the number of classes.

4.2.3 Pitched Instruments

The evaluation procedure for pitched instrument classification basically follows the concepts described in the previous section. In what follows we give insights into more specific issues of the specific recognition problem. In particular, we present the chosen taxonomy, introduce the used dataset, and provide details about all conducted experiments. Finally we present the obtained results for the proposed classification system along with a thorough analysis of the involved audio features and the resulting recognition errors.

4.2.3.1 Taxonomy

Since we aim at imposing as few restrictions or limitations as possible on the presented method, the applied taxonomy has to be able to reflect the instrumentations typically found in various genres of Western music. More precisely, the specific taxonomic choice should allow for a sufficient description of a better part of Western music in terms of instrumentation, which can be used in a MIR context. Hence, we agree on an abstract representation which covers those pitched instruments most frequently found in Western music. In particular, we model the musical instruments *Cello*, *Clarinet*, *Flute*, *acoustic* and *electric Guitar*, *Hammond Organ*, *Piano*, *Saxophone*, *Trumpet*, and *Violin*. Additionally we include the *singing Voice*, since its presence or absence in a given musical composition can carry important semantic information.

It should be noted that the chosen taxonomy allows for a great variety of musical instruments even inside a given category (consider, for instance, the *acoustic Guitar* category containing instruments such as the concert guitar, 6 and 12 steel string acoustic guitars, lap steel guitar, etcetera), which was done thoroughly on purpose. This agrees to a consistent and clear semantic label output, understandable by Everyman, as well as keeping the complexity of the model at a low level. Furthermore, in a perceptual and cognitive context the proposed taxonomy could serve as an intermediate level of abstraction in the hierarchical model the human brain uses to store and retrieve sensory information regarding musical instrument categories (see Section 2.1).

4.2.3.2 Classification data

Statistical modelling techniques demand for quality and representativeness of the used training data in order to produce successful recognition models. In the case of noisy data sufficient data instances are needed to model both the target categories' characteristics as well as their invariance with respect to the noise (see Section 4.1).

In order to construct a representative collection we collected audio excerpts from more than 2 000 distinct recordings. These data include music from the actual and various decades from the past century, thus differing in audio quality to a great extent. It further covers a great variability in the musical instruments' types, performers, articulations, as well as general recording and production styles. Moreover, each training file of a given category was taken from a different recording, hence avoiding the influence of any album effect (Mandel & Ellis, 2005). In addition, we tried to maximise the distribution spread of musical genres inside the collection to prevent the extraction of information related to genre characteristics.

We paid two students to obtain the data for the aforementioned 11 pitched instruments from the pre-selected music tracks, with the objective of extracting excerpts containing a continuous presence of a single predominant target instrument. Hence, assigning more than one instrument to a given excerpt was not allowed. In total, approximately 2 500 audio excerpts were accumulated, all lasting between 5 and 30 seconds. The so-derived initial class assignments were then double-checked by a human expert and, in case of doubt, re-determined by a group of experienced listeners.

Figure 4.4 shows the distribution of labels inside the training collection with respect to the modelled pitched musical instruments and genres. As can be seen we neither were able to balance the total amount of instances across categories nor come up with a flat genre distribution for each class. Nevertheless, we think that the collection well reflects the frequency of the modelled musical instruments in the respective musical genres, i.e. one will always find more electric guitars in rock than in classical music.

4.2.3.3 Parameter estimation

In this section we present and evaluate the stages to be examined in the design process of the classification system. Here, most of the experiments are related to parameter estimation procedures. In particular, we first determine, in terms of classification accuracy, the best-performing length of the acoustical unit on which the classifier performs a single decision ("time scale"). Next, we study the influence of the amount of audio instances taken from a single trainings excerpt on the classification performance ("data sampling"). We then estimate the best subset of audio features ("feature selection") and finally determine the optimal parameter settings for the statistical models ("SVM parameters").

Given the nature of the classification task, all pitched instrument classification experiments reported in this section apply a one-vs-one SVM architecture. Since the problem at hand is the recognition of a single predominant instrument from the mixture, this choice is plausible. Moreover, in all experiments prior to the final parameter estimation via the grid search procedure, we use standard

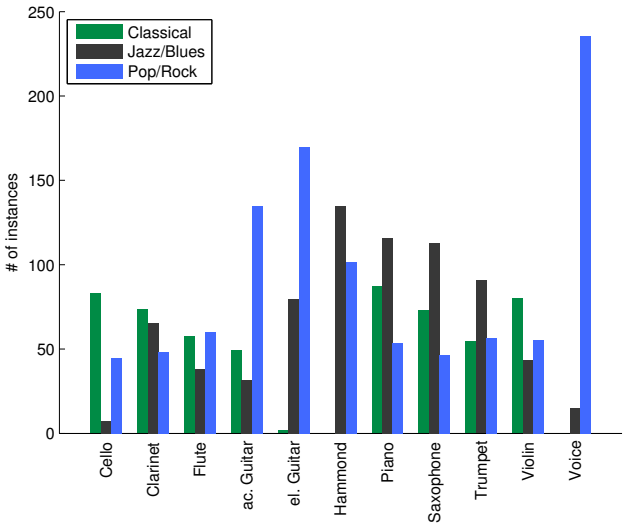


Figure 4.4: Distribution of pitched musical instruments inside the music collection used for extracting the training and evaluation data of the instrumental models.

parameter settings of the used classifier as proposed by the library (Hsu et al., 2003).

Time scale. The objective of the experiment is to define the length of the acoustical unit, on which a single prediction of the model is performed. Many approaches in literature use the entity of a musical note as a dynamic length for the basic acoustical unit (Burred et al., 2010; Joder et al., 2009; Lagrange et al., 2010). This makes sense from the perceptual and cognitive point-of-view, since onset detection and harmonic grouping seems to be very basic operations of the auditory system, naturally grouping the incoming audio stream into objects (see also Section 2.1.2). An accurate segregation for complex signals is, however, almost impossible from nowadays signal processing point-of-view (Liu & Li, 2009; Martin et al., 1998). Moreover, experiments with subjects showed that the human mind accumulates the information extracted from several of these basic units for timbral decisions (e.g. Kendall, 1986; Martin, 1999). The same effect could be observed in a modelling experiment performed by Jensen et al. (2009); here, the incorporation of several notes of a given phrase played by a single instrument in a single classification decision does not affect the performance of the recognition system. Since the variation in pitch of a series of consecutive notes may not exhibit those magnitudes which affects the timbre of the particular instrument (Huron, 2001; Saffran et al., 1999; Temperley, 2007) (see also Section 3.1), these finding seem plausible. In our experiments we nevertheless evaluate classification frames ranging from the time scale of a musical note to the one of musical phrases. However, given the polyphonic nature of our recognition problem we expect longer frames to perform superior.

To compare the performance of the models on various lengths we build multiple datasets, each containing instances of a fixed length. Here, we extract one instance of a given length, i.e. an acoustical unit, at a random position from each audio training file. We then perform 10×10-Fold CV to estimate the model's average accuracy in predicting the correct labels with respect to the annotated data. Since the class distribution of the data is skewed (see the previous Section and Figure 4.4), we

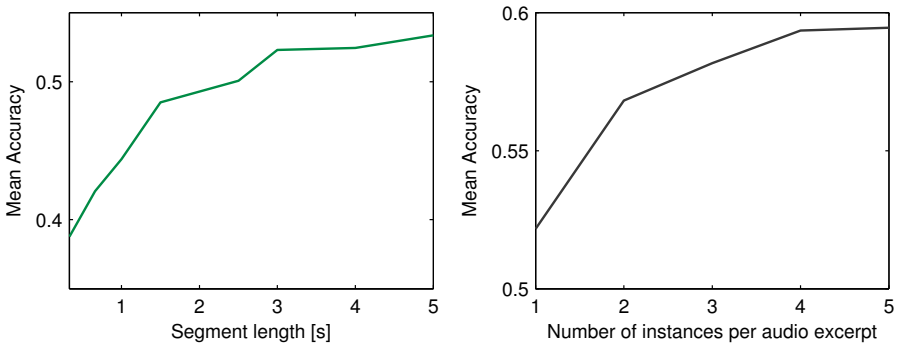


Figure 4.5: Results of the time scale and data size experiments for pitched instruments. Part (a) refers to the recognition performance with respect to the length of the audio instances, while (b) depicts the mean accuracy for different number of instances taken from the same training excerpt.

take, for each run of 10-Fold CV, a different flattened sample from the data. In each classification turn we furthermore apply feature selection.

Figure 4.5a shows the results obtained for the time scale experiment. As expected, the recognition performance improves with larger time scales as this probably results in a more reliable extraction of the instrument's invariant features; insensitivity to outliers in terms of feature values as well as to corrupted or noisy signal parts increases by incorporating more data in the feature integration process. According to these results, we use a length of 3 seconds for the audio instances in all upcoming classification experiments.

Data sampling. Here we study the effect of data size on the recognition performance of the classification system. In general, increasing the amount of data results in better generalisation abilities of the model, which leads to an improved recognition performance, assuming independence of the samples. In our particular case, an increase in data size refers to the extraction of multiple instances from a single audio excerpt in the dataset, hence violating the assumption of the independence of the respective samples. However, the underlying hypothesis is that the assumable greater variety in pitches, articulations, and musical context of the target instrument along a single training excerpt improves the recognition performance of the system. We therefore test the influence of the number of instances taken from a single excerpt in the dataset on the system's recognition performance.

We employed the same experimental methodology and setup as described in the aforementioned time scale experiment by constructing multiple datasets, each containing a different number of fixed-length instances randomly taken from each audio file in the training set. Furthermore, we kept instances of the same file in the same fold of the CV procedure, in order to guarantee a maximum independence of training and testing set. The increased variety in musical context and articulation styles together with the dependency of most pitched instruments' timbre on fundamental frequency (see Marozeau et al., 2003) should result in a positive effect on the recognition performance when increasing the number of instances taken from each audio file in the training dataset, although this effect might be of limited nature.

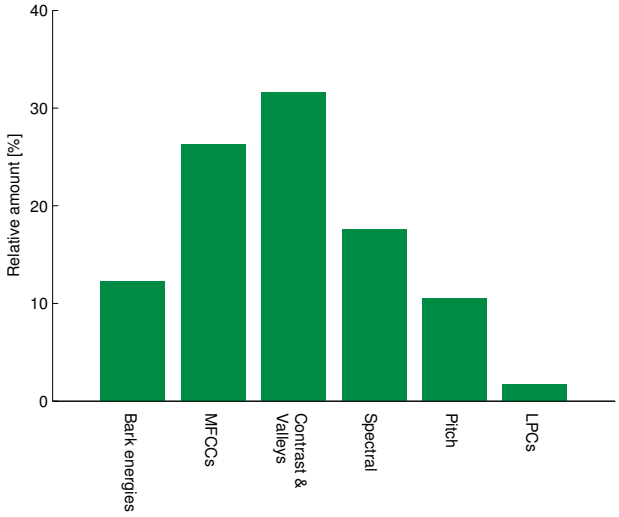


Figure 4.6: Selected features for pitched instruments grouped into categories, representing the acoustical facets they describe.

Figure 4.5b depicts the mean accuracy $\overline{A}$, resulting from the 10×10-Fold CV, for different number of instances extracted from a single audio excerpt. As can be seen, the identification performance of the model can be increased to a certain extent by augmenting the size of the used data. It seems that the ceiling of the mean accuracy results from the limited instrument's variation inside the audio training file. In consequence, we use the values of three instances per audio files for the pitched models in all subsequent classification tasks.

Feature selection. Table 4.1 lists all features selected by the performed 10-Fold CV feature selection procedure. In addition, Figure 4.6 shows this final set of selected features grouped with respect to the acoustical facets they describe. In total, we can reduce the dimensionality of the data by approximately 85% by selecting 59 out of 368 low-level audio features. For pitched instruments the description of the spectral envelope seems to be of major importance – MFCCs and spectral contrast and valleys features cover approximately 60% of the selected features. But also pitch and harmonic-related features, which are derived from algorithms designed for monophonic music processing, along with basic spectral statistics seem to be important. We note that the selected features roughly resemble those that had been identified in different monophonic classification studies (e.g. Agostini et al., 2003; Nielsen et al., 2007). This confirms our main hypothesis that with the chosen methodology an extraction of the instrument's relevant information from polytimbral music audio signals is possible, given a certain amount of predominance of the target.

SVM parameters. In general, the performance of a SVM in a given classification context is highly sensitive to the respective parameter settings (Hsu et al., 2003). The applied SVM library requires several parameters for both classifier and kernel to be estimated a priori. Determining the parameter values of the classifier in a given problem is usually arranged by applying a grid search procedure, at which an exhaustive parameter search is performed by considering all predefined combinations of parameter values. As proposed by Hsu et al. (2003), we estimate the respective classifier's regular-

Feature	Statistic	Index
Barkbands	mean	3
Barkbands	var	7, 8, 12, 23
Barkbands	dvar	4, 6, 8
MFCC	mean	2-6, 9-11
MFCC	var	2, 3, 6, 12
MFCC	dmean	6, 7
MFCC	dvar	1, 2
Spectral contrast	mean	0, 2-4
Spectral contrast	var	2-5
Spectral contrast	dmean	5
Spectral contrast	dvar	3, 5
Spectral valleys	mean	0
Spectral valleys	var	5
Spectral valleys	dmean	2, 5
Spectral valleys	dvar	3-5
LPC	mean	10
Tristimulus	mean	0
Tristimulus	var	0, 1
Barkbands spread	dmean	–
Barkbands skewness	mean	–
Spectral strongpeak	mean	–
Spectral spread	mean	–
Spectral spread	dmean	–
Spectral rolloff	mean	–
Spectral dissonance	dmean	–
Spectral dissonance	dvar	–
Spectral crest	mean	–
Spectral crest	var	–
Pitch salience	mean	–
Pitch confidence	mean	–
Pitch confidence	dmean	–

Table 4.1: Selected features for the pitched model. Legend for the statistics: mean (mean), variance (var), mean of difference (dmean), variance of difference (dvar).

isation parameters C and ν, and the kernel parameters γ and d for the RBF and polynomial kernel[9]. We furthermore perform the grid search in a two-stage process by first partitioning and searching the relevant parameter space loosely for each combination of classifier and kernel types. Once an optimal setup has been found, we use a finer division to obtain the final parameter values using a 10 $\times$ 10-Fold CV scheme. For illustration purpose, Figure 4.7 shows the parameter space spanned by the classifier's cost parameter ν and the RBF kernel's parameter γ, evaluated by the mean accuracy $\overline{A}$ on the entire dataset.

[9]As already mentioned before, the regularisation parameter controls the trade off between allowing training errors and forcing rigid margins. The kernel parameter γ determines the width of the RBF's Gaussian as well as the inner product coefficient in the polynomial kernel. The parameter d finally represents the degree of the polynomial kernel function.

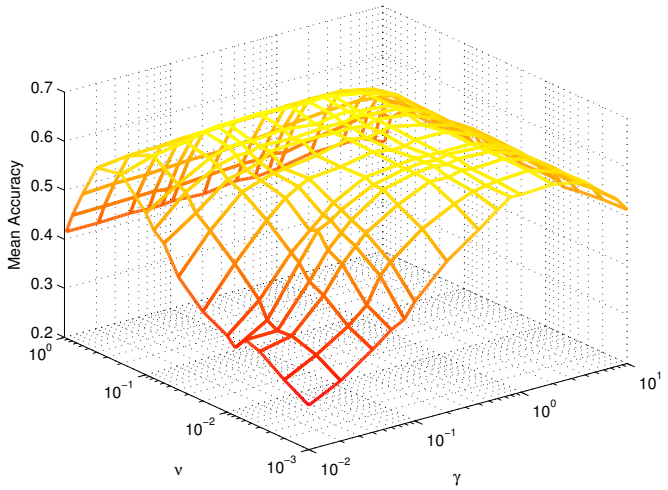


Figure 4.7: Mean accuracy of the pitched model with respect to the SVM parameters. Here, the classifier's regularisation parameter ν and the RBF kernel's γ are depicted.

$\overline{A}_{null}$	C4.5	NB	10NN	MLP	SVM*
9.1%	33%	37.1%	57.4%	57.9%	63%±0.64pp

Table 4.2: Recognition accuracy of the pitched classifier in comparison to various other classification algorithms; a Decision Tree (C4.5), Naïve Bayes (NB), Nearest Neighbour (NN), and Artificial Neural Network (MLP). Due to the complexity of the data, simple approaches like the C4.5 perform worse than more enhanced ones, e.g. the MLP. However, the proposed SVM architecture is superior, demonstrating the power of its underlying concepts. The asterisk denotes mean accuracy across 10 independent runs of 10 Fold CV.

4.2.3.4 General Results

Table 4.2 shows the result obtained from the 10×10-Fold CV on the full dataset. To illustrate the power of the SVM on this kind of complex data, the recognition accuracy of other classification methods, usually found in related machine learning applications, is added. It can be seen that relatively simple methods such as Decision Trees (C4.5) or Naïve Bayes (NB) more or less fail to learn the class specificities, while more enhanced algorithms such as the artificial neural network (MLP) are coming close with respect to the recognition performance. We used the WEKA library (Hall et al., 2009) to estimate the recognition accuracies of the additional classifiers. We mostly apply standard parameter settings in a single 10-Fold CV experiment. Moreover, Figure 4.8 shows the mean precision, recall, and F-measures for the individual instrumental categories, together with the corresponding standard deviations. Additionally, we perform a single run of a 10-Fold CV and construct the confusion matrix from all testing instances in the respective folds. Table 4.3 shows the result.

In the following we qualitatively assess the performance of the developed model for pitched instrument recognition on the basis of the presented quantitative results. The objective is to interpret the model's functionality in terms of the acoustical properties of the respective audio samples and the thereof derived description in terms of audio features. This further helps for understanding the acoustical dimensions primarily involved in the recognition task as well as the extracted character-

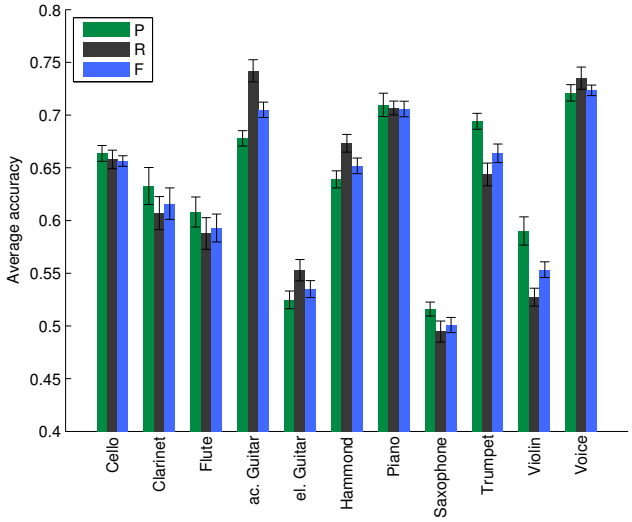


Figure 4.8: Performance of the pitched model on individual categories. Mean values across 10 independent runs of 10 Fold CV are shown, error bars denote the corresponding standard deviations.

	Cello	*Clarinet*	*Flute*	*ac. Guitar*	*el. Guitar*	*Hammond*	*Piano*	*Saxophone*	*Trumpet*	*Violin*	*Voice*
Cello	**269**	7	15	25	22	11	8	7	7	31	5
Clarinet	7	**240**	39	8	5	5	19	41	17	23	3
Flute	19	33	**228**	13	6	32	25	13	9	11	18
ac. Guitar	20	5	5	**304**	16	5	16	10	5	8	13
el. Guitar	6	5	3	21	**223**	39	19	23	12	28	28
Hammond	8	2	26	9	38	**277**	9	4	2	13	19
Piano	12	13	15	18	18	18	**288**	13	1	6	5
Saxophone	11	36	14	12	25	11	16	**211**	39	18	14
Trumpet	7	27	7	12	15	15	3	48	**261**	9	3
Violin	37	20	22	12	33	11	4	23	13	**220**	12
Voice	11	3	14	13	28	14	3	9	6	9	**297**

Table 4.3: Confusion matrix of the pitched model. The vertical dimension represents the ground truth annotation, while the horizontal dimension represents the predicted labels of the respective instances.

istics of the individual instruments in the polytimbral context. Most of our analysis is based on the distribution of instances in the confusion matrix shown in Table 4.3, thereby taking into account both correct and confused instances as well as their differences. In doing so we identify and subsequently compare the most prominent acoustical facets, captured by the audio features involved in the developed model's decision process, to the intrinsic acoustical properties of the respective musical instruments (Meyer, 2009; Olson, 1967). In particular, we first provide an analysis in terms of the most decisive features by looking at the recognition task at a whole as well as focussing on individual instrumental categories. This is followed by a qualitative analysis of the prediction errors.

4.2.3.5 Feature analysis

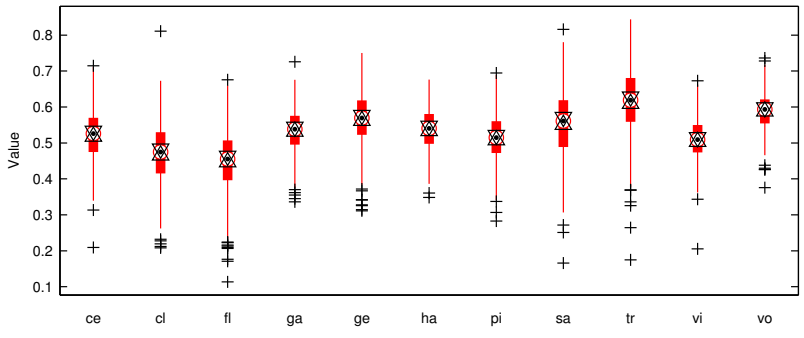
Here we estimate the most important acoustical facets integrated in the classification model by determining the amount of information a single audio feature carries within the current recognition task. Even if the decision functions among the individual instruments in the audio feature space is assumable of a highly non-linear kind, determining the most crucial audio features may give insights into the basic acoustical analogies the model applies to discriminate between (groups of) individual instruments. Hence, in this first experiment we evaluate the cumulative degree of association of an individual attribute to all target classes in order to qualitatively asses its informativeness for discriminating among the 11 categories. In particular, we first normalize each of the 59 involved features (see Table 4.1) similar to the SVM model and subsequently compute its χ^2 statistic with respect to the classes. The general idea behind this non-parametric hypothesis testing is to compare the observed to the expected frequencies of two variables of a random sample to evaluate the null hypothesis of no mutual association via contingency tables; for large sample sizes a large value indicates large deviations of the observations from the expectations so as to reject the null hypothesis. Given X, a discrete random variable, with x_i possible outcomes, $i = 1 \ldots m$, and n independent observations grouped by $K = 1 \ldots k$ classes, then the χ^2 statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{i,j} - nP(K=i)P(X=x_j))^2}{nP(K=i)P(X=x_j)}, \tag{4.13}$$

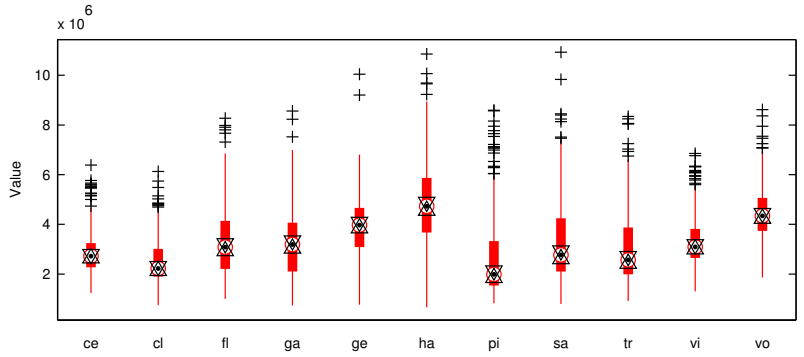
where k and m denote, respectively, the number of classes and possible outcomes of a given feature, $n_{i,j}$ the observation frequency of outcome x_j given class i, $P(K=i)$ the prior probability of class i, $P(X=x_j)$ the probability of outcome x_j.

Since the attributes to evaluate are of numeric kind, all features are discretised using the method presented by Fayyad & Irani (1993) prior to the evaluation. We then rank all features according to their calculated χ^2 value. Figure 4.9 shows Box plots of the 5 top-ranked features, assumed to carry the most discriminative power among all features in the classification task at hand. Note that non-overlapping comparison intervals between categories correspond to a statistically significant difference in sample medians at a significance level of 5% (i.e. $p < 0.05$). Here, the comparison interval endpoints are given by the centres of the triangular markers.

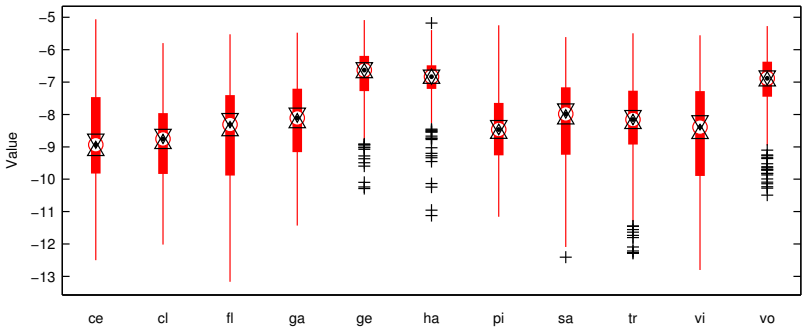
It can be seen from the resulting figures that each of the 5 features carries information for discriminating groups of instruments, but none of them is able to significantly separate one particular instrument from the rest. However, we are able to deduce some general acoustical characteristics that separate groups of instruments from this depicted information; for instance, *Flute* and *Trumpet* are well discriminated by the pitch salience feature (Figure 4.9a), since the sound of the former is noisy due to the blowing technique while the one of the latter is the brightest of all modelled instruments. Moreover, *electric Guitar*, *Hammond organ*, and the *singing Voice* are separated from all other instruments via a measure of the spectral spread (Figure 4.9b), indicating that these sounds carry a higher amount of high frequency components, most probably due to the applied distortion effects in case of the former two and the unvoiced sibilants in case of the singing voice. Similarly, the 0^{th} coeffi-



(a) Mean of pitch salience feature.



(b) Mean of spectral spread feature.

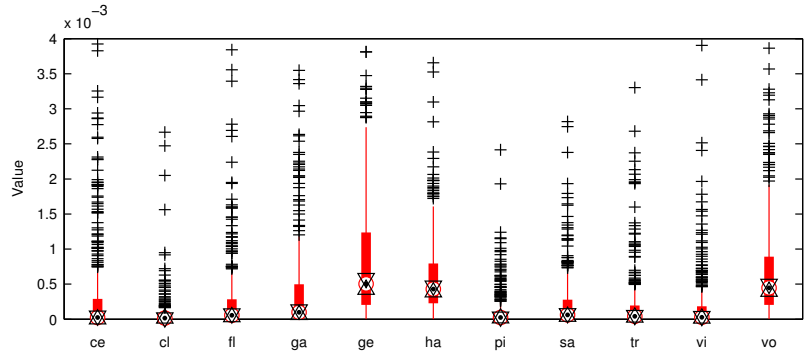


(c) Mean of 0^{th} spectral valley coefficient.

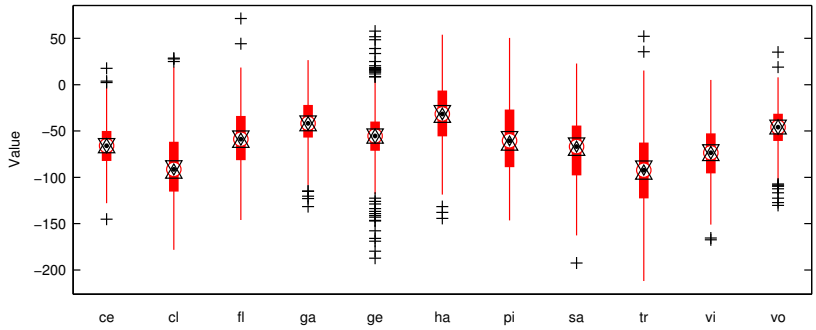
Figure 4.9: Box plots of the 5 top-ranked features for pitched instrument recognition. See second part of the figure for a detailed caption.

cient of the spectral valley feature discriminates the same groups of instruments[10] (Figure 4.9c). The 3^{rd} Bark band energy however exhibits similar separation abilities (Figure 4.9d), indicating that the magnitude of frequency components between 150 and 200 Hz are important acoustic properties for discriminating *electric Guitar*, *Hammond organ*, and the *singing Voice* in this context. Finally,

[10]Unfortunately this feature cannot be interpreted directly in terms of the acoustical properties it captures, since the applied PCA linearly combines the information from each band by applying a transformation matrix calculated from the data itself.



(d) Mean of 3rd Bark energy band.



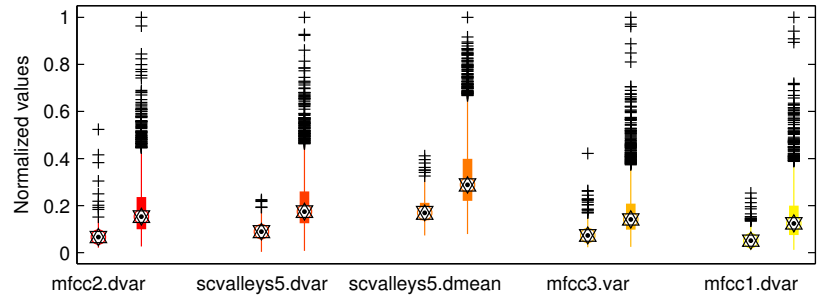
(e) Mean of 2nd MFCC coefficient.

Figure 4.9: Box plots of the 5 top-ranked features for pitched instrument recognition. Despite the assumable non-linear feature dependencies applied by the classification model for category decisions, several discriminative properties with respect to groups of instruments can be observed from the depicted features. Legend for the instrumental categories plotted on the abscissa: *Cello* (ce), *Clarinet* (cl), *Flute* (fl), *Acoustic Guitar* (ga), *Electric Guitar* (ge), *Hammond* organ (ha), *Piano* (pi), *Saxophone* (sa), *Trumpet* (tr), *Violin* (vi), and *singing Voice* (vo).

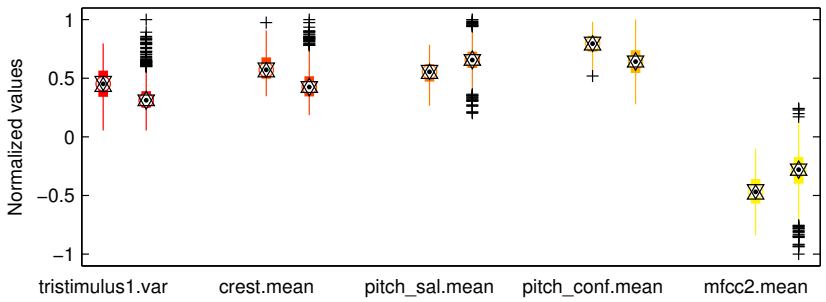
the equal position of some boxes in Figure 4.9e, for instance the boxes corresponding to *Clarinet* and *Trumpet*, may explain the mutual confusion that can be observed between these instruments in Table 4.3.

In a further experiment, we assess, for a single instrument, the informativeness of the individual audio features. That is, our aim is to identify the most discriminative features used by the developed model for separating a given instrument from all others. In doing so we build, for each musical instrument, a binary dataset from the instances falling on the diagonal of Table 4.3, grouping the instances of the respective instrument against the rest. Next, we compute the χ^2 statistic between all features and the respective class in order to determine the dimensions captured by the features the model uses for discrimination between the individual categories. We then rank the features according to the obtained values. In other words, we only analyse these data which are perfectly recognised by the trained model[11], avoiding any confused instances. In the course of the following analysis we therefore also determine those acoustic characteristics of the individual instruments which enable

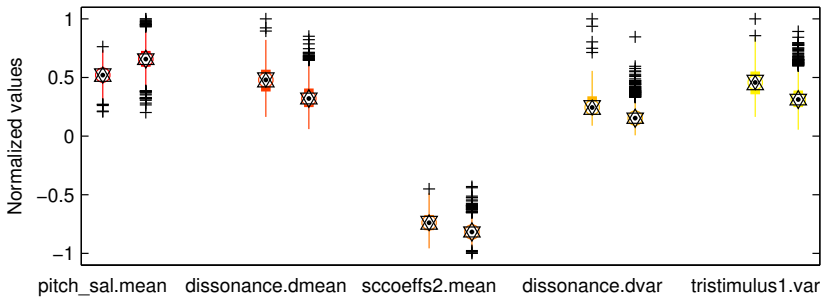
[11]Although the instances in Table 4.3 are classified by slightly different models due to the CV procedure, we hypothesize that the conclusion drawn from the forthcoming analyses does not cause any loss of generality. The small value of the standard deviation – obtained by averaging the results of 10 different CV – in Table 4.2 is further suggesting this hypothesis.



(a) Cello.



(b) Clarinet.

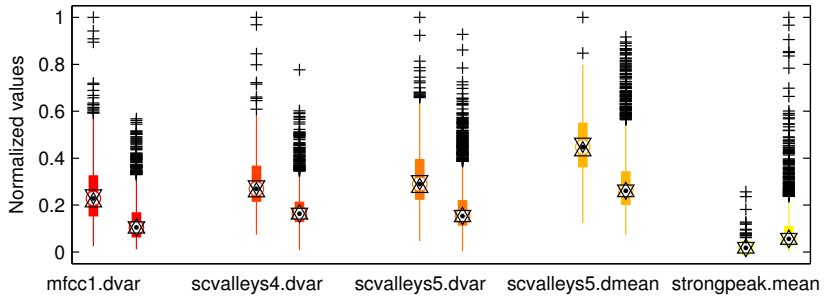


(c) Flute.

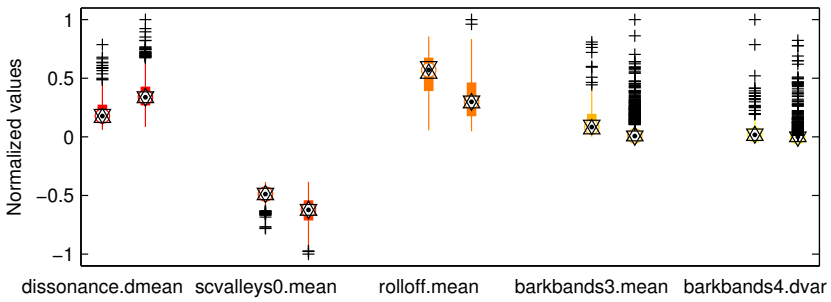
Figure 4.10: Box plots of the 5 top-ranked features for individual pitched instrument recognition. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. See last part of the figure for a detailed caption.

a successful discrimination among them. Figures 4.10(a)-(k) show the obtained Box plots for the respective 5 top-ranked features.

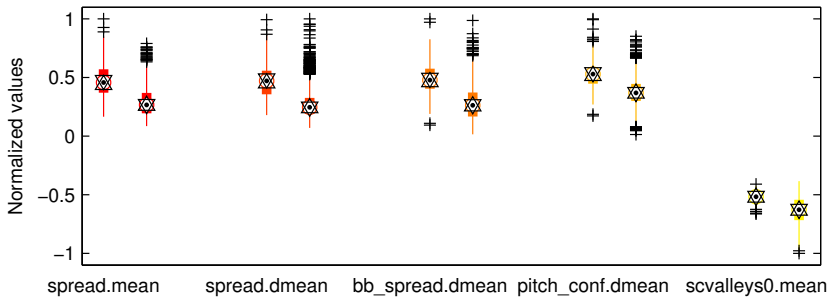
Similarly, we examine, for a given instrument, which of the applied audio features are accountable for misclassification. That is, we take all instances of a single ground truth category, as shown in one single entire row in the confusion matrix of Table 4.3, and group them into correctly and incorrectly recognised instances. Again, the χ^2 statistic is calculated for all features in each binary scenario and the resulting values are ranked. We hypothesise that those features ranked as most informative are



(d) Acoustic guitar.



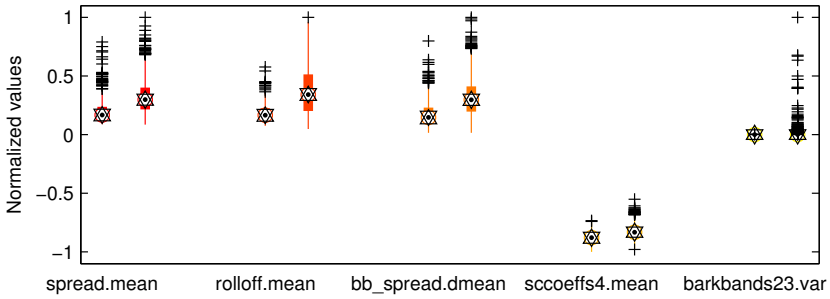
(e) Electric guitar.



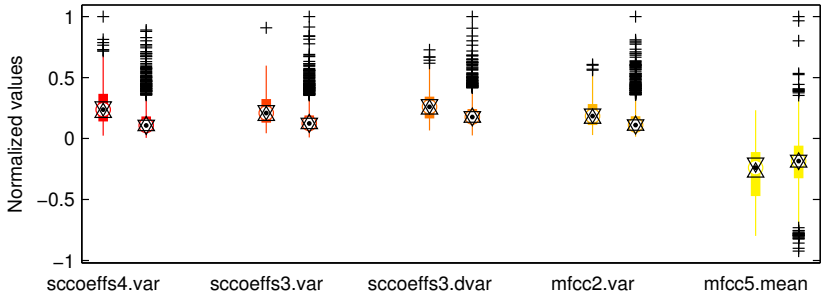
(f) Hammond organ.

Figure 4.10: Box plots of the 5 top-ranked features for individual pitched instrument recognition. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. See last part of the figure for a detailed caption.

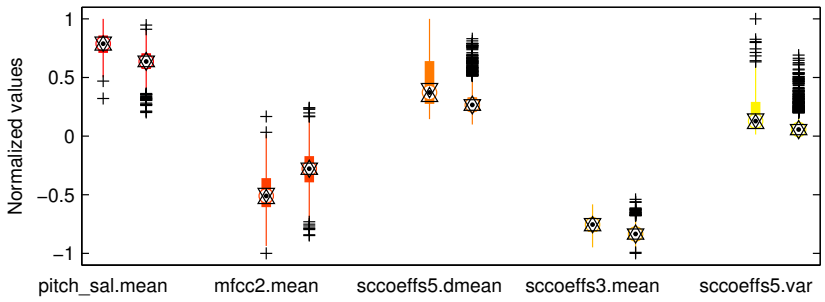
accountable for the main confusion of the particular instrument. The here identified features will, to a certain extent, resemble those found in the experiment described above, but also reveal additional features only attributable to the instrument-specific confusions. In other words, the aim here is to identify the saxophone qualities of a *Clarinet* instance recognised as *Saxophone*, rather than the general qualities of *Saxophone* tones which separates the instrument from all others, as performed in the previous experiment. Figures 4.11(a)-(k) show the obtained Box plots for the respective 5 top-ranked features. In what follows we discuss the outcomes of both aforementioned experiments for each instrument separately and relate the respective features to the acoustic characteristics of the



(g) Piano.



(h) Saxophone.

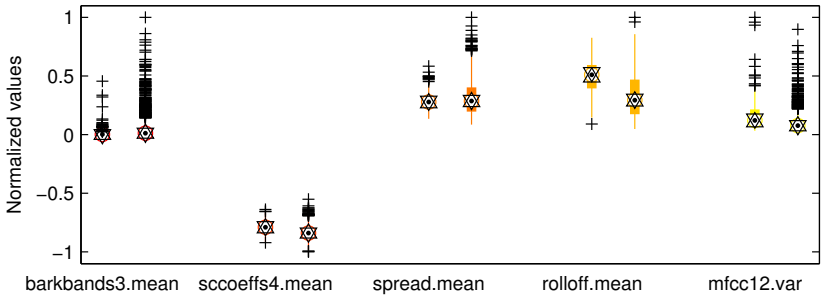


(i) Trumpet.

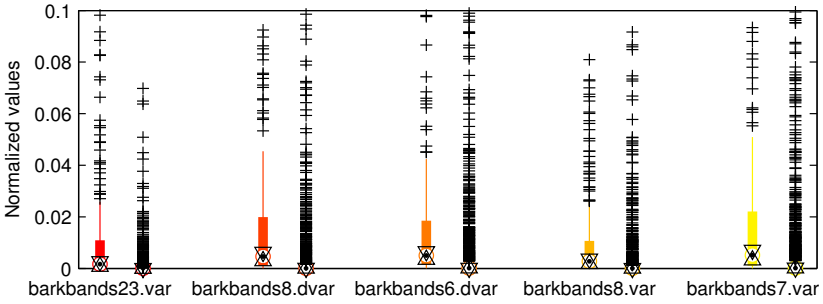
Figure 4.10: Box plots of the 5 top-ranked features for individual pitched instrument recognition. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. See last part of the figure for a detailed caption.

particular instruments. At last we provide a summary of the obtained insights in Table 4.4.

Cello. In Figure 4.10a, the *Cello* is most significantly defined in terms of audio features, compared to all other instruments, by the description of its spectral envelope. The instrument's intrinsic characteristics are encoded in the 2nd and 3rd MFCC coefficients most probably accounting for the strong body resonances in the spectral envelope. The spectral slope properties of low-pitched sounds, more common for this instrument, are further described by 1st MFCC coefficient. Also the 5th spectral valleys coefficient seem to play an important role. Analogously, the 4th and 5th spectral contrast



(j) Violin.

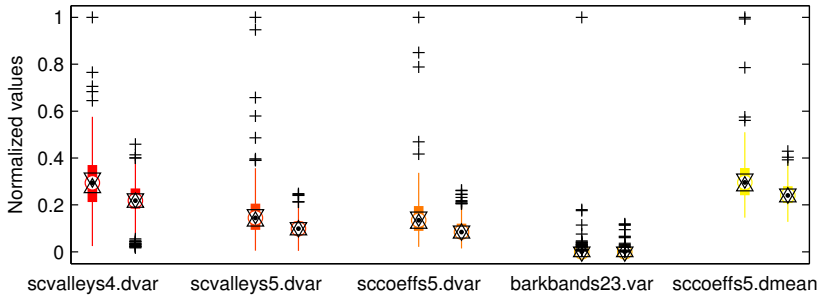


(k) Singing Voice.

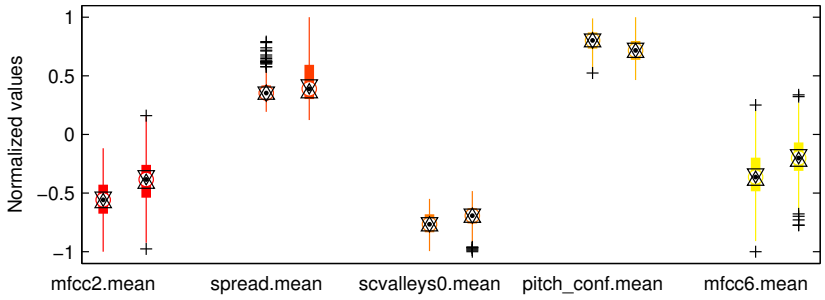
Figure 4.10: Box plots of the 5 top-ranked features for individual pitched instrument recognition. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. The figure shows those features mostly correlated with the respective binary classes, only consisting of instances correctly predicted by the developed recognition model. Not surprisingly, many of the depicted features can also be found in Figure 4.9. Legend for the feature statistics: mean value of instantaneous values (*mean*), variance of instantaneous values (*var*), mean value of first difference values (*dmean*), and variance of first difference values (*dvar*).

and valleys coefficients in Figure 4.11a may explain the confusions with *acoustic Guitar* (see also Figure 4.11d). Moreover, the appearance of the 23rd Bark energy band in the figure could indicate the confusions of some distorted *Cello* instances with the *electric Guitar*, since those high frequency components are rather atypical for "natural" cello sounds.

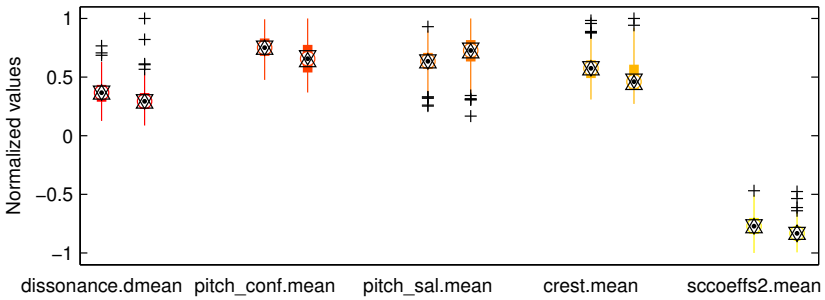
Clarinet. Remarkably, the most prominent property of the *Clarinet* – the attenuation of the even harmonics for low- and mid-register tones – is described by the top ranked feature in Figure 4.10b; the 2nd tristimulus (here ambiguously denoted *tristimulus1* due to 0-based indexing) describes the relative strength of the 2nd, 3rd, and 4th harmonic. Similarly, the spectral crest feature seems to account for the lacking harmonics since it relates the spectrum's maximum to its average energy value. A source of both recognition and confusion are features accounting for pitch strength (*pitch salience* and *pitch confidence*), since strong clarinet tones exhibit rich harmonics in the upper part of the spectrum while very soft tones can produce spectra consisting of only 4 harmonics. The aforementioned features directly account for the relative harmonics' strength since they derive their value from an autocorrelation of the signal. Moreover, the appearance of the 2nd MFCC coefficient in Figures 4.11b and 4.10i may be an indicator for the mutual confusions between the instruments



(a) Cello.



(b) Clarinet.

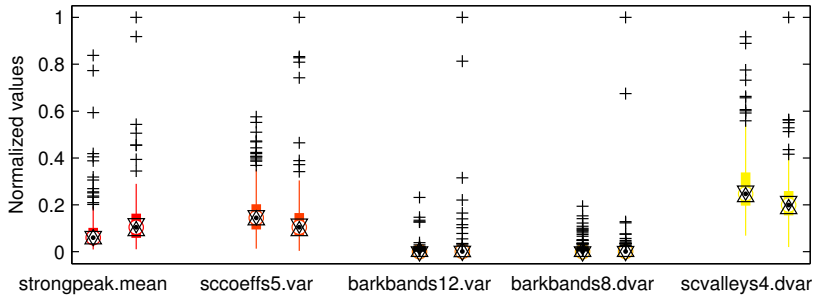


(c) Flute.

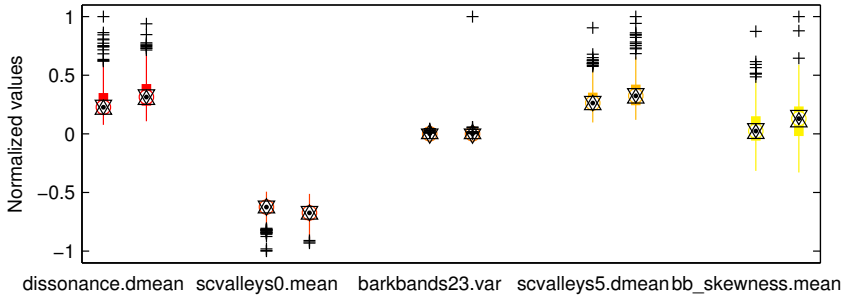
Figure 4.11: Box plots of the 5 top-ranked features for individual pitched instrument confusions. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. See last part of the figure for a detailed caption.

Clarinet and *Trumpet*. Further evidence for this assumption can be derived from the relative position of the Clarinet and Trumpet boxes in Figure 4.9e, showing the distribution of the 2nd MFCC coefficient's mean statistic with respect to all categories.

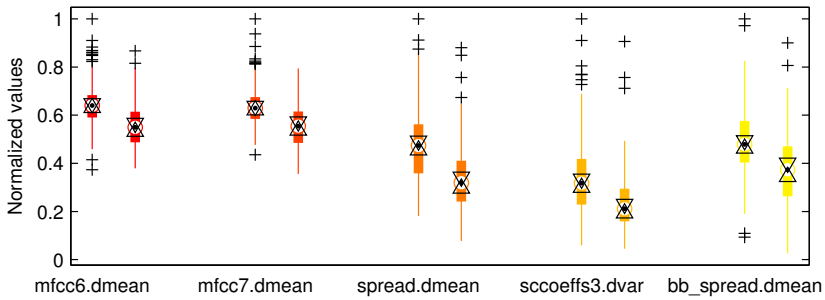
Flute. The *Flute* is separated by the description of the pitch strength and the roughness of the tone. This can be related to the uniform overtone structure attributable to the flute's sound for almost all pitches as well as the strong noise components incorporated in the signal due to the blowing (Figure 4.10c). Basically, this also applies for the confusions associated with flute sounds (Figure 4.11c),



(d) Acoustic guitar.



(e) Electric guitar.

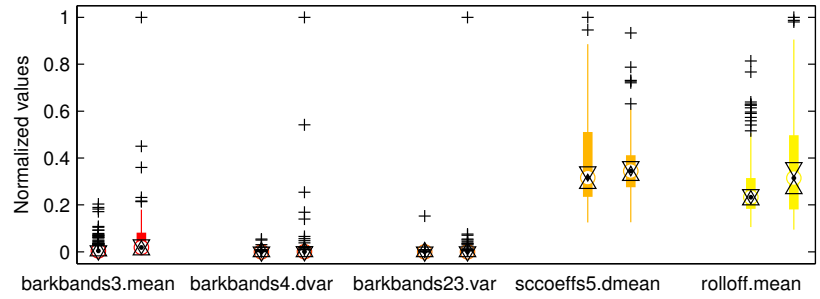


(f) Hammond organ.

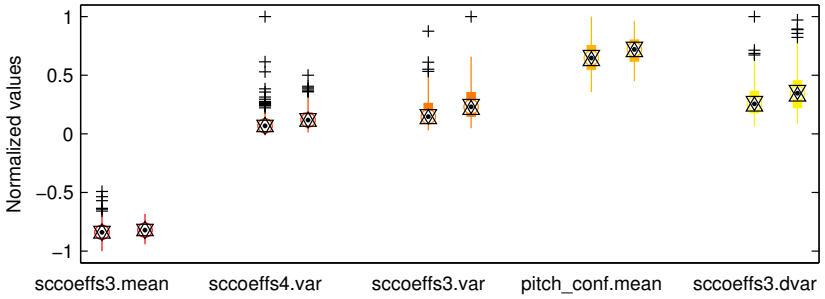
Figure 4.11: Box plots of the 5 top-ranked features for individual pitched instrument confusions. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. See last part of the figure for a detailed caption.

where the same features indicate the most common sources of errors. Remarkably, due to the absence of a pronounced formant structure in the flute's tones, features directly describing the spectral envelope (e.g. MFCCs) are not listed in the respective Box plots.

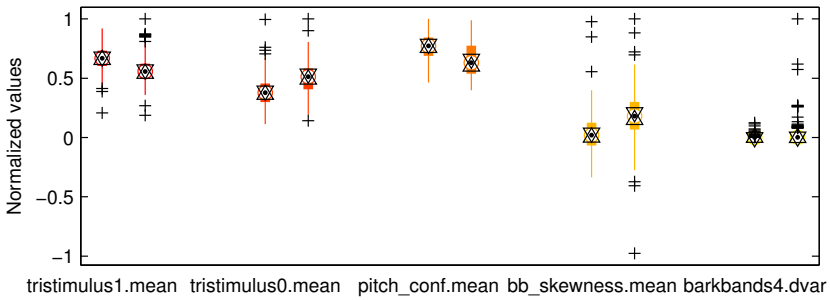
Acoustic Guitar. Figure 4.10d shows those features best separating *acoustic Guitars*, including the slope of the sound's spectrum as represented by the 1st MFCC coefficient as well as additional spectral envelope descriptions via the 4th and 5th spectral contrast and spectral valleys coefficients, most probably to distinguish the *acoustic Guitar* from other stringed instruments, e.g. *Violin* and *Cello*.



(g) Piano.



(h) Saxophone.

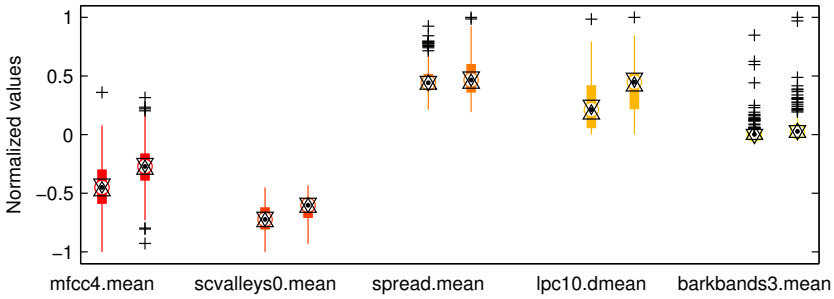


(i) Trumpet.

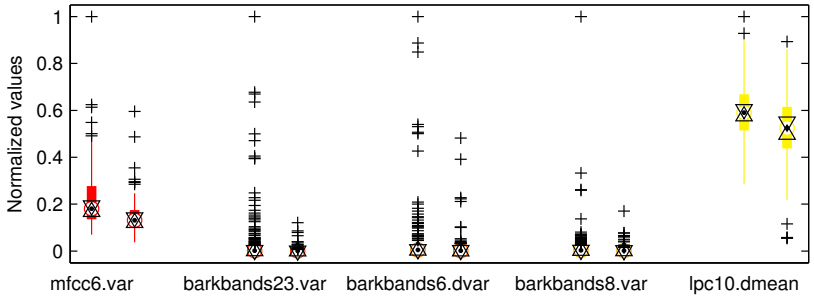
Figure 4.11: Box plots of the 5 top-ranked features for individual pitched instrument confusions. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. See last part of the figure for a detailed caption.

These latter features also appear in Figure 4.11d, showing those features primarily involved in confusing *acoustic Guitar* sounds; here, the 12$^{\text{th}}$ Bark band seems to describe the instrument's formant around 1.5 kHz, while the variance in the change of the 8$^{\text{th}}$ band may cause the confusions with the *singing Voice* (see also Figure 4.10k).

Electric Guitar. On the contrary, the *electric Guitar* behaves differently, since the instrument does not exhibit prominent body resonances like the *acoustic Guitar*, but is frequently played with artificial sound effects – the most prominent the distortion. Thus, features capturing these effects seem



(j) Violin.



(k) Singing Voice.

Figure 4.11: Box plots of the 5 top-ranked features for individual pitched instrument confusions. The two boxes for each feature correspond, respectively, to the target instrument and the "rest" class. The figure shows those features most accountable for the main confusions of a given instrument since we compare the feature values of the correctly labelled instances to the incorrectly labelled ones of each category. Legend for the feature statistics: mean value of instantaneous values (*mean*), variance of instantaneous values (*var*), mean value of first difference values (*dmean*), and variance of first difference values (*dvar*). See text for more details.

to have high discriminative power; *dissonance* directly accounts for the non-linearities involved with distorted sounds, whereas *rolloff* captures the enriched high frequencies, being significantly higher than for other instruments (Figure 4.10e). Moreover, energies between 150 and 300 Hz are descriptive (3rd and 4th Bark energy band) as well as the 0th coefficient of spectral valleys, which can also be deduced by looking at Figures 4.9(c) and (d). On the other hand, the aforementioned dissonance feature also provides the greatest source for confusions, most probably with other distorted instruments such as the *Hammond organ* (Figure 4.11e). Also noticeable is the appearance of the variance of the 23th Bark energy band, which seems to cause the confusions with the *singing Voice* (see also Figure 4.10k).

Hammond. The *Hammond organ* is best characterised in our model by spectral distribution features. Probably due to the absence of any intrinsic spectral shape property – the timbre of the instrument can be modified by mixing the generated harmonic components at different amplitudes via the *drawbars* – the distribution of the frequencies around the spectrum's mean carries most information to recognise the instrument (Figure 4.10f). This can also be observed from Figure 4.9b, where the Box representing the *Hammond organ* instances takes the most extreme position. Other

discriminative features are the pitch strength and the 0^{th} spectral valleys coefficient. Source for confusions are the change in higher-order MFCC coefficients, as well as features capturing the spectral spread (Figure 4.11f).

Piano. The characteristics of the *Piano* are primary described by features addressing the distribution of the components in the generated spectrum (Figure 4.10g), likely due to the rapid decrease of the amplitudes of higher-order partials, a result from the strucked excitation of the strings. Here, the *Piano* shows significantly lower values when compared to all other instruments (e.g. Figure 4.9b). Accordingly, Figure 4.11g determines spectral energy between 150 and 300 Hz (3^{rd} and 4^{th} Bark energy bands) most important for confusions of Piano sounds, most probably with *electric Guitars* (e.g. see Figure 4.10e). Moreover, the *rolloff*, which accounts for high frequency properties of the sound may explain the frequent confusions with *Hammond organ* and again with *electric Guitar*.

Saxophone. Figure 4.10h shows the features top-ranked for the *Saxophone* category. Apparently, the 3^{rd} and 4^{th} spectral contrast coefficients account for the instrument's distinct resonance structure which separates it best from all other instruments. The 2^{nd} MFCC coefficient seems to describe similar aspects of the instrument's sound, while the 5^{th} coefficient capture the higher frequency modulations of the spectral envelope. Recognition errors are most importantly assigned to the 3^{rd} spectral contrast coefficient (Figure 4.11h), which explains the high confusion rate with *Trumpets* (note the co-occurrence of the feature also in Figure 4.10i). Moreover, the pitch strength feature points towards confusions with the *Clarinet*.

Trumpet. Since the sound of the *Trumpet* is characterised by a rich overtone spectrum along with little noise components, salience of the pitch is the top-ranked feature in Figure 4.10h. Moreover, the 2^{nd} MFCC coefficient and the spectral contrast and valley coefficients describe the formant structure of the sound. These features may also be accountable for the strong confusion of *Trumpet* sounds with *Saxophone*. Since the *Saxophone* combines both brass and woodwind characteristics, a confusion on the basis of the formant properties of the sound is not so far off. Also, the amplitude of the first 4 harmonics as captured by *trisimulus0* and *tristimulus1* seems to be important for the misclassification of *Trumpet* sounds (Figure 4.11i). Finally, the pitch strength may be again accountable for the prominent confusions with the *Clarinet*.

Violin. The features important for discriminating sounds from the *Violin* are depicted in Figure 4.10j; here, both the spread feature and the 3^{rd} Bark energy band are probably used to distinguish the instrument from *electric Guitar*, *Hammond organ*, and the *singing Voice* (Figures 4.9(b) and (d)). Moreover, the 4^{th} spectral contrast coefficient seems to model the instrument's formant regions. On the other side, Figure 4.11j shows these features most important for confusing *Violin* sounds with other instruments. The re-occurrence of the 3^{rd} Bark energy band and the spread feature points toward confusions with *electric Guitar* and the *singing Voice*. Finally, the 4^{th} MFCC coefficient, which captures the general formant structure of the sound, may be addressable to cause various confusions, e.g. with *Clarinet* or *Saxophone*.

Singing Voice. Lastly, the *singing Voice* is best characterised in our model by various Bark energy band features between 400 and 800 Hz (Figure 4.10k). This emphasizes the importance of the first

formant in distinguishing the *singing Voice* from other instruments, whilst the second formant region is primarily used to differentiate different human voices. Moreover, the shown features account for both variances in the instantaneous and first difference values of the energy bands, which may refer to the vibrato used by professional singers. Since the used data set does not comprise audio recordings of opera, the characteristic singing formant of opera singers is not captured by the model. Moreover, the high order Bark energy (23$^{\text{rd}}$ band) in Figure 4.10k seems to address unvoiced sibilants, which can reach up to 12 kHz. Similarly, confusions of *singing Voice* sounds can be attributed to the same features (Figure 4.10k); here, the region of the first formant area is important as well as the high frequency content extracted by the 23$^{\text{rd}}$ Bark energy band, which most probably produces the confusions with the *electric Guitar*. Additionally, the 6$^{\text{th}}$ MFCC coefficient, capturing higher order modulations in the spectral envelope, thus referring to the fine-grained formant structure of the sound, may partially explain the confusions with instruments like *acoustic Guitar*.

Summary. Table 4.4 shows a summary of the instrument-wise feature analysis. Here, the left half contains the analysis for the individual instrument recognition, while the right half refers to the confusion analysis (see above for more details). We grouped the identified features into broad categories, representing the acoustical facets they capture, at which we assign a group to a particular instrument if we find more than one feature of the given group in the respective 5 top-ranked features of Figures 4.10 and 4.11. In particular, *Bark* denotes local spectral energies as typically described by the Bark energy bands, while *Env.* corresponds to all features accounting for the spectral envelope of the signal, such as MFCCs or spectral contrast and valleys. Furthermore, we group all features describing statistical properties of the spectrum into the *Spec.* category, whilst *Pitch* finally addresses the features capturing pitch-related characteristics of the analysed sound. It can be seen from the table that the developed recognition model uses the audio features referring to those properties of the musical instruments which describe their intrinsic acoustical characteristics for discriminating the respective categories. In particular, spectral envelope descriptions are the most important for those instruments exhibiting strong body resonances (e.g. *Cello*, *Violin*, or *acoustic Guitar*), while pitch-based features are associated with blown instruments such as *Clarinet* or *Trumpet*. Moreover, the recognition model relates instruments applying artificial audio effects to descriptions of their spectral statistics, noticeable here are the *electric Guitar* and the *Hammond organ*, both frequently using the distortion effect which influences the spectral characteristics of those instruments' sounds to a great extent. Remarkably, the *singing Voice* is mostly characterised by local spectral energies in the frequency region of the first formant. Not surprisingly, similar feature-instrument combinations can be found in the confusion analysis, at which the spectral envelope description, being the most decisive timbral characteristic of instrument sounds, causes the most confusions throughout the instrumental categories. But also other, more specific, descriptions cause frequent inter-instrument confusions, see, for instance, the pitch category for *Clarinet* and *Trumpet* or the spectral features for *electric Guitar* and *Hammond organ*, accounting most probably for the applied distortion effects.

4.2.3.6 Error analysis

In this section we perform a qualitative analysis of the recognition errors by perceptually evaluating the wrongly predicted instances from Table 4.3. We thereby group the respective instances of a given confusion pair according to the observed perceptual correlates. Our aim is to find perceptual

	Recognition				Confusion			
	Bark	Env.	Spec.	Pitch	Bark	Env.	Spec.	Pitch
Cello		✓				✓		
Clarinet				✓		✓		
Flute			✓	✓			✓	✓
ac. Guitar		✓			✓	✓		
el. Guitar	✓		✓			✓	✓	
Hammond			✓			✓	✓	
Piano			✓		✓			
Saxophone		✓				✓		
Trumpet		✓						✓
Violin		✓	✓			✓		
Voice	✓				✓	✓		

Table 4.4: Summary of the feature analysis for pitched instruments. The left half of the table shows those features important for the recognition of the instruments by the developed model, while the right half contains the ones most probably involved in the confusions of the particular instruments. Here, *Bark* denotes local energies in the spectrum as captured by the Bark bands, *Env.* corresponds to those features describing the spectral envelope, e.g. MFCCs or spectral contrast and valleys features. Furthermore, *Spec.* refers to features accounting for statistical characteristics of the spectrum such as the spectral spread or skewness, and *Pitch* contains all features related to the pitch properties of the signal, e.g. salience or tristimuli.

regularities in the confusions between particular instruments and adjust the training data according to these found criteria. As already mentioned in the previous section, the confusion matrix in Table 4.3 contains many instrument pairs with strong mutual confusions, thus we expect to find those aforementioned regularities for some instrumental combinations.

Our first observation is, for each instrument across its respective confusions, a constant amount of "noise". That is, there exist a certain amount of instances, which confusion cannot be attributed by any perceptual explanation. This amount of "noise" instances lies between 3 and 10, depending on the confusion rate of the given instrumental pair, and is quite evenly distributed across the confusion matrix. Moreover, we identify a significant number of instances in which signal the confusion instrument is clearly audible, representing kind-of "correct confusions" (e.g. an instance labelled as *Flute* but labelled with *acoustic Guitar*, wherein the accompaniment *acoustic Guitar* takes a prominent part). The obvious reason for such instances is wrong annotation, which is natural for a dataset of this size. Moreover, such instances can also contain two similarly predominant instruments. Since only one label is attached to each training instance, the model may use the not-annotated sound for classification. Or, those instances are artefacts of the data generation process – the random extraction of the acoustical units from the audio file in the training corpus. Although the target instrument is supposed to be continuously playing, in a typical 30 second excerpt containing a single predominant instrument plus accompaniment, it can be expected that small sections of the signal happen to be without the target. If those sections are extracted by the random data generation process, the instance is labelled with the wrong instrument.

Furthermore, some of the found groups of errors can be identified intuitively by considering the sound producing mechanisms of the respective instruments (e.g. *Cello* instances recognised as *Violin*), some by musical attributes (e.g. *Saxophone* instances recognised as *Clarinet* due to the soft play-

ing style of the instrument), while others can hardly be grouped by perceptual explanation (e.g. *Violin* instances recognised as *Saxophone*).

We further observe, for all instruments, that unison lines played by two different instruments are often sources for confusion. Here, either the instance is labelled with one instrument, but the model predicts the other one. Or, the unison generates a complete different timbre ("fusing of timbre", see Section 3.2.2) which is then recognised as none of the participating instruments. Moreover, we identify several additional factors attributable to the main confusions as produced by the model. Some of these produced errors are perceptually obvious, while others are difficult to discover, even for a well-trained listener. In the following we describe the sources of the most significant regularities in the confusions determined during this perceptual error analysis, and indicate the corrections we applied to the training dataset.

Sound production. Well-established confusions, e.g. between *Cello* and *Violin*, can be explained by the similar sound producing mechanism. Also confusions between *acoustic Guitar* and *Cello*, both string instruments, or between *Clarinet* and *Saxophone*, both Wind instruments, can be attributed to the sound producing mechanism. Most of these cases also pose difficulties to experienced listeners in a perceptual discrimination task. Here, factors such as register and dynamics play an important role.

Register. Since different registers of the same instrument may exhibit very different timbral sensations (see Section 3.1), instruments are more likely to be confused when played in specific tone ranges. This happens, for instance, in the upper register of *Clarinet* and *Flute*, where the perceptual difference between tones of these two instruments can only be determined by the amount of noise in the signal (Meyer, 2009). But also high-pitched sounds from the *Piano* are frequently confused with *Clarinet* or *Flute*, probably due to the missing modelling of the hammer sound. Another example is the confusion between *Cello* and *Violin*, and vice versa, as high-pitched *Cello* tones sound similar to *Violin*, while the *Violin* can be easily confused with the *Cello* in the low register. Indeed, many confusions between those instruments in our model can be attributed to the pitch range of the respective sound, a fact that similarly happens to humans (Jordan, 2007; Martin, 1999).

Dynamics. Analogously, dynamic changes in the sound of a particular instrument may have a significant effect on its perceived timbre. For instance, we can address parts of the *Saxophone*'s confusions with *Clarinet* to the low dynamics of the respective *Saxophone* sound, since the sounds of the two instruments become perceptually very similar. Also *Trumpet*, when played with low dynamics, is often confused with *Clarinet* in our model.

Since most of the above-described phenomena result in "natural" confusions, i.e. the sounds of two instruments get perceptually hard to discriminate, we do not adapt the data accordingly. Also one has to question if it is in general possible to account for this subtle differences at this level of granularity, i.e. the extraction of instrumental characteristics directly from a polytimbral mixture signal.

Distortion. Instances of several instruments use a distortion effect, causing regular confusions with *electric Guitar* and *Hammond organ*. These latter two are frequently played with distortion so

that this particular effect becomes part of the instrument's sound characteristics. Since the training data of these two instruments include many distorted samples, other instruments applying the distortion effect can be easily confused; for many instances the audio effect causes the sounds to become perceptually very similar to distorted sounds of *electric Guitar* or *Hammond organ*. We accordingly remove all distorted instances from the training data of all instruments except the two aforementioned in order to model the instruments' characteristics rather than the audio effect.

Recording Condition. We observe a correlation between confused instances exhibiting old recording conditions and the instrument *Clarinet*. Probably due to the high amount of *Clarinet* samples taken from sources with such recording conditions, instances from other categories showing the same recording style are frequently confused with *Clarinet* in the model. Moreover, the overall sound characteristics of these instances, i.e. missing lower and upper frequency components in the signal's spectrum, seem to corrupt perceptual discrimination abilities between the respective instruments (e.g. *Clarinet* and *Piano* sound more similar under this recording conditions). We therefore remove most of such *Clarinet* samples from the training dataset and replace them with proper instances.

4.2.3.7 Discussion

Given the results presented in Section 4.2.3.4 and the insights provided by the respective feature and error analysis, we first can confirm the main hypotheses postulated in the beginning of this Chapter (Section 4.1). That is, given a certain amount of predominance, the spectral envelope and its coarse temporal evolution of the target instrument is preserved, which enables the extraction of the instrument's characteristics for computational modelling. We also show that longer time scales of several seconds are needed for a robust recognition, most probably due to the complex nature of the data the features describing the instrumental characteristics are extracted from. Since in polytimbral data masking of the target or interference between several concurrent sources frequently occur, more confident decisions can be derived by integrating the data of longer time spans. Moreover, similar observations were reported from perceptual recognition experiments, where humans performed significantly better at longer time scales (Kendall, 1986; Martin, 1999).

The figures in Table 4.2 demonstrate that the resulting recognition performance is far from random, indicating a successful extraction of the instrument-specific characteristics. Moreover, the applied SVM architecture is suitable for modelling the complex relationships between categories in terms of audio features. Here, the model's ability to handle highly non-linear data together with its generalisation abilities seems to be a key aspects for its superiority against the other classification methods shown in the table.

More evidence for the successful extraction of the instrument-specific invariants can be found in the nature and importance of the applied audio features, as analysed in Section 4.2.3.5. In general, the features selected in the construction process of the recognition model resemble those identified in automatic recognition studies performed with monophonic data (e.g. Agostini et al., 2003; Nielsen et al., 2007). Furthermore, the acoustical facets captured by these features correspond to those acoustical characteristics known to be decisive between musical instruments' timbres from musical

acoustics research (Meyer, 2009) (see also Figures 4.9 and 4.10). Finally, the most prominent confusions identified in a perceptual analysis of the recognition errors coincide with those usually found in analogue experiments with human subjects. This suggests that similar features as applied by the developed model could also be used by the human auditory system to discriminate between different instrumental categories.

However, the limitation in recognition accuracy of around 65% indicates that certain acoustical or perceptual attributes of the instruments' timbres are not captured by the applied audio features and are therefore not modelled in the current system. The perceptual analysis of the errors suggest that additional features are needed to account for the persistent confusions, which can be observed in Table 4.3. Here, more fine grained description of the spectral envelope characteristics would enable the discrimination between instruments from the same instrumental family (e.g. *Cello* versus *Violin*). In addition, the description of the attack portion of the sounds may help in the extraction of intrinsic characteristics not directly manifested in the spectral envelope. This can reduce confusions between blown instrument such as *Clarinet* and *Trumpet*, or string instruments like *Cello* and *acoustic Guitar*. Moreover, a better modelling of noise transients would improve the recognition performance, for instance the noise introduced by the hammer mechanics of the *Piano*, or the breathy sound as produced by the *Flute*. Most of those aforementioned characteristics are known to improve recognition accuracy (e.g. Lagrange et al., 2010), but cannot be directly extracted from the raw polytimbral audio signal.

In conclusion, the developed model shows a robust recognition performance on a complex task – the direct recognition of predominant pitched musical instruments from polytimbral music audio data – but leaves much headroom for improvement.

4.2.4 Percussive Instruments

In our method we focus on the detection of a single percussive instrument, the Drumkit. This choice is motivated by its predominance in almost all genres of Western music, except for classical compositions. We therefore assume that its presence or absence in a given musical context carries important semantic meaning. Moreover, adding less frequently used percussive instruments (e.g. Bongos, Congas, Timpani, etc.) would complicate the model and may not increase the overall information.

In what follows we present our approach towards the detection of the Drumkit in Western music; it is based on the modelling of the overall timbre of the Drumkit, without focusing on its individual components. In previous works we used an instrument-based method for detecting the presence of the Drumkit (Fuhrmann et al., 2009a; Fuhrmann & Herrera, 2010), accomplished via individual percussive instrument recognition (*Bass Drum*, *Snare Drum*, and *Hi-Hat*), as developed by Haro (2008). Onset detection was applied to locate the percussive events in the music signal and pre-trained SVMs predicted the presence or absence of each individual instruments. The so-found information was then aggregated by a simple majority vote among these decisions along the entire audio to indicate the presence of the Drumkit.

A quantitative comparison of the two approaches – which is not included in this thesis – showed no significant differences with respect to the recognition accuracy, but clearly favoured the timbre-based over the instrument-based approach in terms of computational complexity; additional to the performed onset detection, the latter applies the recognition models far more frequently to the audio signal, since each onset has to be evaluated for all three instruments. The former evaluates a single model sequentially, similar to the process for pitched instrument recognition. What follows are the methodological details of our timbre-based method for approaching the problem of Drumset detection.

Conceptually, we assume that the timbral properties of music with and without drums differ significantly. This is reasonable since the different percussive instruments of the Drumset exhibit distinct spectral energy patterns, e.g. pulsed low-frequency excitation for the Bass Drum, compared to the other instruments the Drumset usually plays along. The problem can therefore be regarded as a binary pattern recognition task, as described in Section 4.2.1. Moreover, the following shares several commonalities with the process of the pitched instrument recognition.

4.2.4.1 Classification data

As data corpus the same collection as for the pitched instruments is used. That is, we labelled these excerpts according to the presence or absence of the Drumkit. In the case of ambiguity, i.e. both classes inside a single excerpt, the excerpt was skipped. In total, we accumulate more than 1.100 excerpts per category, i.e. *Drums* and *no-Drums*.

4.2.4.2 Parameter estimation

The parameter estimation experiments described here are similar to those performed for the pitched instruments, as described in Section 4.2.3.3. We therefore only briefly review the underlying concepts and present the respective results.

Time scale. Here, we estimate the length of the audio instance, i.e. the acoustical unit, required for a robust recognition of the Drumkit's timbre. Again, to evaluate the problem, we construct multiple collections for different extraction length, at which we randomly extract one single audio instance from a given excerpt, and measure the respective recognition accuracy. Evidence from perceptual experiments suggests that pure timbral categorizations are done at short time scales (several few 100 ms) and serve by this means as cues for higher-level organization tasks related to genre and mood (Alluri & Toiviainen, 2009; Kölsch & Siebel, 2005). In contrast to the pitched instruments, where an increase in recognition performance with increasing extraction length is observed, we therefore expect the performance of the percussive model to be quasi-independent of the audio length the information is taken from. Figure 4.12a shows the obtained results. We observe a slight increase in recognition performance with longer time scales, which may result from the improved outlier removal in terms of feature values for longer windows. According to these results we use a length of 3 seconds[12] for the percussive acoustical units in all subsequent experiments.

[12]Note that this value is the same as for the pitched instrument recognition, thus enabling the prediction of both models from the same basic feature extraction process, which simplifies the whole recognition system to a great extent.

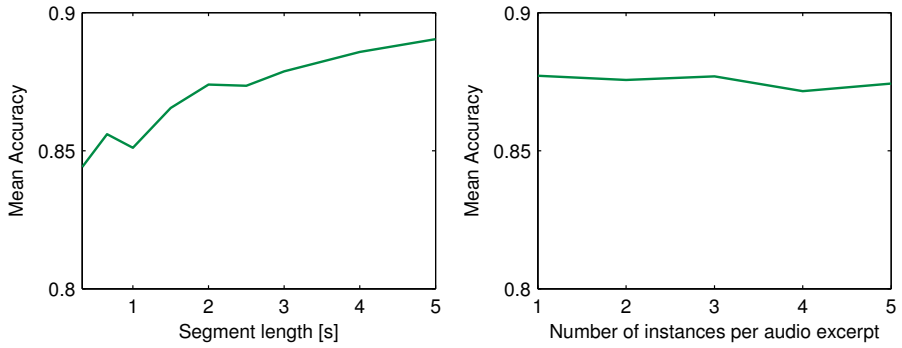


Figure 4.12: Results of the time scale and data size experiments for percussive timbre recognition. Part (a) refers to the recognition performance with respect to the length of the audio instances, while (b) depicts the mean accuracy for different number of instances taken from the same training excerpt.

Data size. Similar to the pitched instruments, we estimate the influence of multiple instances taken from a single audio excerpt on the recognition performance. Since the timbre of the Drumkit should be rather constant across a single excerpt, we expect no influence of the datasize. Again, we build multiple datasets in which we alter the amount of instances taken from a single excerpt and compare the respective mean accuracies. Figure 4.12b depicts the experimental results, at which no dependency of the recognition performance on the data size can be observed. Given those results we use one single instance from each audio excerpt for percussive models in all upcoming experiments.

Feature selection. Table 4.5 lists the features resulting from the selection process described in Section 4.2.1.3. In addition, Figure 4.13 shows the relative amount of features selected with respect to the acoustical facets they describe. In total, we reduce the initial feature set comprising 368 low-level audio features to 43, a reduction of approximately 90%. More precisely, we observe the relative importance of local energy distribution, represented by the Bark band energies, for the Drumkit's timbre recognition. Particularly only very low and high bands were selected, indicating the discriminative character of these frequency regions. This confirms the intuition that the presence of drums mainly adds significant components in both extrema of the audio spectrum, primary due to the sounds of the Bass Drum and the Cymbals[13].

SVM parameters. Here, we follow the same methodology as described for the pitched instruments. That is, we perform a two-stage grid search procedure to optimize the parameter settings for the SVM model. Again for illustration purpose, Figure 4.14 shows the mean accuracy with respect to the classifier's cost parameter C and the RBF kernel's γ parameter evaluated for the entire dataset.

[13]The frequency regions where the Snare Drum and the Tom-toms are usually located seem to be very dense due to overlapping frequency components of other instruments, e.g. guitars and singing voice around 500-800 Hz, which increases the complexity of the recognition problem.

Feature	Statistic	Index
Barkbands	mean	0, 1, 24, 25
Barkbands	var	1
Barkbands	dmean	0, 1, 22
Barkbands	dvar	1, 2, 8, 22
MFCC	mean	1, 4-11
MFCC	var	0
MFCC	dvar	0
Spectral contrast	mean	0
Spectral contrast	dmean	0-2
Spectral valleys	mean	0, 4, 5
Spectral valleys	dmean	1, 2
Spectral valleys	dvar	4
Barkbands spread	var	–
Barkbands spread	dvar	–
Pitch confidence	var	–
Pitch salience	mean	–
Spectral flatness	mean	–
Spectral kurtosis	mean	–
Spectral kurtosis	dmean	–
Spectral spread	mean	–
Spectral spread	dmean	–
Odd2even ratio	dmean	–

Table 4.5: Selected features for the percussive model. Legend for the statistics: mean (mean), variance (var), mean of difference (dmean), variance of difference (dvar).

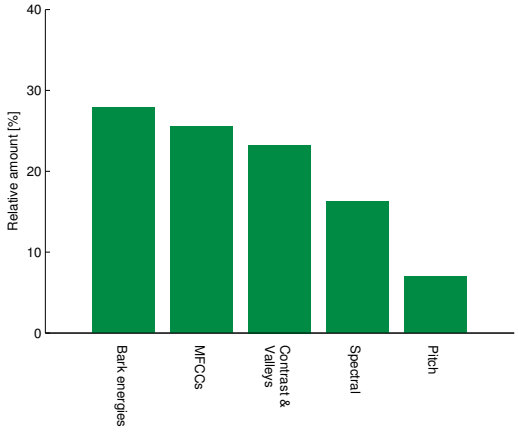


Figure 4.13: Selected features for percussive timbre recognition grouped into categories representing the acoustical facets they describe.

4.2.4.3 General results

Table 4.6 presents the results after 10 independent runs of 10 Fold CV. Again, we compare the performance obtained by the proposed SVM architecture with several classification algorithms from

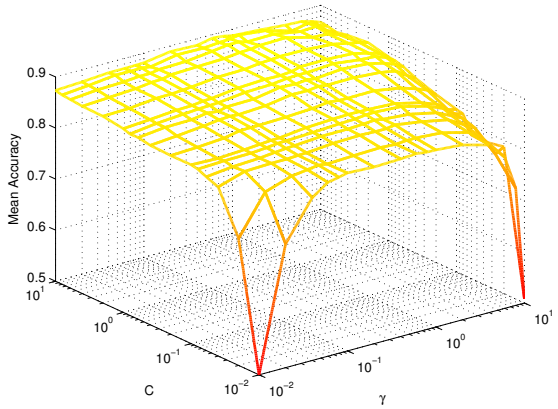


Figure 4.14: Mean accuracy of the percussive timbre model with respect to the SVM parameters. Here, classifier's complexity C and the RBF kernel's γ are depicted.

$\overline{A}_{null}$	C4.5	NB	10NN	MLP	SVM*
50%	83.9%	81.8%	87.7%	87.2%	89%±0.27pp

Table 4.6: Recognition accuracy of the percussive timbre classifier in comparison to various other classification algorithms; a Decision Tree (C4.5), Naïve Bayes (NB), Nearest Neighbour (NN), and Artificial Neural Network (MLP). Due to the simplicity of the problem compared to the pitched instruments, the recognition performance of the shown classifiers lie closer together. Hence, even conceptually simple algorithms such as the C4.5 score good accuracies. The proposed SVM architecture is still superior, although its performance can be regarded as equivalent to the ones of 10NN and MLP, since the proposed SVM is the only approach applying a grid search for parameter optimisation. The asterisk denotes mean accuracy across 10 independent runs of 10-Fold CV.

	Drums	*no-Drums*
Drums	**1026**	128
no-Drums	136	**1018**

Table 4.7: Confusion matrix of the percussive timbre model. The vertical dimension represents the ground truth annotation, while the horizontal dimension represents the predicted labels of the respective instances.

the software package WEKA. As can be seen from the table even simple algorithms such as Decision Trees (C4.5) or Naïve Bayes (NB) score high accuracy values. Since the recognition task at hand is far more simple compared to the one of the pitched instruments, the gap to complex algorithms such as the Artificial Neural Network (MLP) or the SVM is not that big. Since the proposed SVM architecture applies parameter optimisation via grid search, its performance can be regarded as equivalent to the MLP and 10NN algorithms, although it shows the highest value in recognition accuracy. Additionally, Table 4.7 shows the confusion matrix obtained from a single 10-Fold CV.

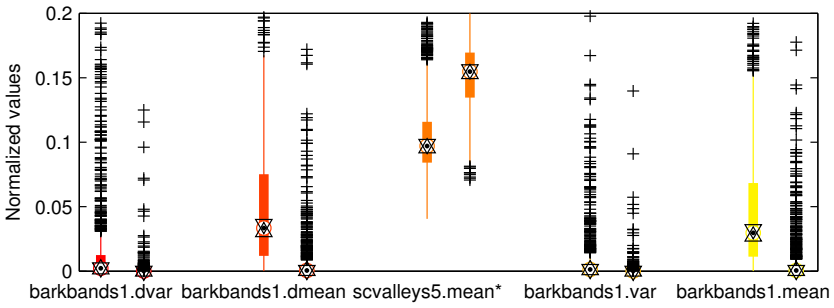


Figure 4.15: Box plots of the 5 top-ranked features for percussive recognition. The asterisk at the 5th spectral valleys coefficient indicates compression of the values in order to fit into the range of the Bark energy features for better visibility.

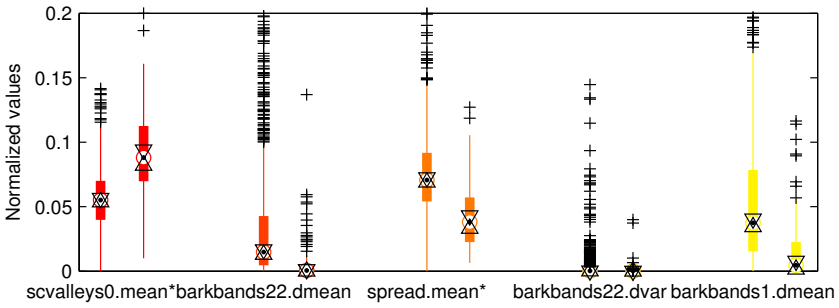
4.2.4.4 Feature analysis

In this section we perform an analysis of the most important audio features involved in the percussive recognition task, as similarly applied for the pitched instruments. We therefore perform a ranking of all selected features based on their χ^2 statistic with respect to the classes (Eq. 4.13). The top-ranked features are assumed to carry the most information for discriminating the target categories. Figure 4.15 shows the 5 top-ranked features resulting from this analysis.

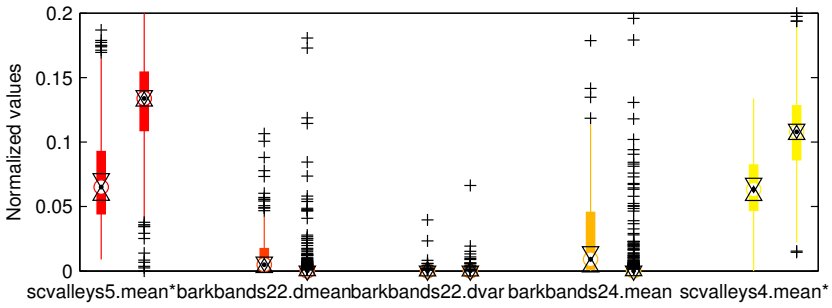
The figure demonstrates the importance of low-energy spectral components between 50 and 100 Hz, as represented by the 2nd Bark energy band (again, note that in our zero-based indexing the 2nd Bark band is denoted as *barkbands1*, see the Appendix for a listing of the Bark energy bands according to the used indexing scheme). All statistical descriptions of the feature's time evolution along the 3 second excerpt exhibit high discriminative power. This indicates the major importance of the Bass Drum, which carries its main energy in the aforementioned frequency region. A perceptually missing Bass Drum therefore often causes wrong predictions, as discussed in Section 4.2.4.5.

Next, we construct binary datasets by grouping the instances of each category into correct and incorrect predictions, performing the same ranking procedure as described above. An analysis of the top-ranked features reveals those audio features primarily involved in the confusions of the respective categories. Moreover, we can deduce the most prominent acoustical analogies involved in this discrimination. Figure 4.16 shows the 5 top-ranked features of the respective dataset.

Interestingly, the figure reveals high frequency components – 22nd and 24th Bark energy bands – to be of major importance in the misclassification of the respective categories. That is, missing high frequency components cause instances labelled as *Drums* to be predicted as *no-Drums*. Vice versa, given a considerable amount of energy in this frequency region, the model predicts instances without drums wrongly as *Drums*. This indicates that apart from low-frequency properties of the signal also high-frequency characteristics are incorporated by the model in the decision process. Remarkably, the former are mostly attributable for correct recognition, while the latter are prominently involved in wrong predictions.



(a) Drums.



(b) No-Drums.

Figure 4.16: Box plots of the 5 top-ranked features for percussive confusions. The asterisk indicates compression of the feature values in order to fit into the range of the Bark energy features for better visibility.

4.2.4.5 Error analysis

In this section we perform a qualitative error analysis by perceptually evaluating the confused instances in Table 4.7. We again group these confusions according to regularities of perceptual, acoustical, or musical kind, and subsequently describe the most prominent of these groups. This gives some insights into the acoustical properties of the data involved in the most prominent confusions.

For both types of confusions, i.e. *Drums* recognised as *no-Drums* and vice versa, instances exhibiting only a sparse amount of drums (e.g. a short drum fill) frequently produce wrong predictions. In the case of confusions of *Drums* as *no-Drums*, the only identified, but prominent regularity is the sound of the Drumset when played with "Brush" sticks, as typically found in jazz performances. Here, the Bass Drum is not audible any more, probably due to the soft playing style and the thereby involved masking effects via other concurrent instruments, while the only remaining perceptual sensation is the noise introduced by the Brushes in the mid and high frequency range. On the other hand, for *no-Drums* wrongly predicted as *Drums*, the following can be observed; first, if the recording contains a constant amount of perceptually important noise, it is predicted as *Drums*. The high frequency components introduced by the noise seem to trigger the *Drums* decision in the model. Second, percussively played pitched instruments, such as the *acoustic Guitar* and the *electric* or *acoustic Bass*, cause confusions, most probably due to the impulsive character of the observed sound events (e.g. "Slap

Bass"). Moreover, different other kinds of percussive instruments are spuriously predicted as *Drums*, among which the most prominent are Shakers or Tambourines, producing high frequency components, and percussion containing only Bongos and Congas. Remarkably, we found almost a double amount of annotation errors in the *no-Drums* (24) category than in the *Drums* category (13). This suggests that identifying the presence of drums is easier for a human than recognising its absence in perceptually ambiguous situations.

4.2.4.6 Discussion

The observed recognition accuracy of almost 90%, presented in Table 4.7, suggests that the timbral characteristics of the Drumset are captured by the developed recognition model. The conclusions drawn from the feature analysis and the perceptual error analysis further indicate that the model uses the corresponding acoustical characteristics for modelling the timbre of the Drumset. Furthermore, the acoustical properties extracted by the applied audio features resemble the properties of the individual instruments of the Drumset; especially frequency regions in the lower and upper range of the spectrum are decisive – and used by the model – for recognising the timbre of the Drumset. Moreover, the prominent presence of descriptions of the spectral envelope (e.g. MFCCs) in the applied audio features can be assigned to the opposite category, i.e. sounds containing only pitched instruments.

4.3 Labelling

4.3.1 Conceptual overview

In this section we describe the approaches taken to infer labels related to the instrumentation of a given music audio signal of any length from the frame-based classifier estimates, described in the previous section. Given the consecutive predictions of the models along time, context analysis is used to translate the probabilistic output by the classifiers into instrumental labels and corresponding confidence values. Due to the stationary character of predominant musical instruments inside a musical context, i.e. when entering in a musical phrase the particular instrument will be active for a certain amount of time and will not stop unexpectedly, labels are derived from longer time scales by exploiting the statistical or evolutionary properties of those instruments therein. In this course we avoid the direct inference of labels from sections containing unreliable classifier decisions, i.e. exhibiting a great variability in the respective probabilistic estimates over time, since the context analysis will rely on portions of the signal with rather unambiguous instrumental information. In other words, sections containing instrumental confusions have less influence on the inferred labels, while those sections with predominant instruments are the main source for label inference, since strong tendencies of the probabilistic estimates towards these instruments are observable there.

Moreover, context analysis with a focus on predominant instruments increases the method's robustness against all kinds of noise. Here, the most apparent noise is represented by musical instruments that are not modelled by the classifiers. Since the categorical space of the system is rather limited – we only model 12 categories from the population of musical instruments – these unknown sources will frequently appear in the input data. Given an unknown instrument at the input of the model, its probabilistic output should ideally not indicate a preference for any modelled instrument. Moreover, the temporal sequence of the classifier's probabilistic estimates should exhibit a great variability, again showing no preference for any category along this dimension. Hence context analysis prevents the method from labelling according to short-term predictions resulting from unknown instruments. However, in case of strong confusions, even context analysis does not provide means for filtering the spurious labels.

The label inference process itself is based on the temporal integration of the instrumental models' probabilistic output. As a first step, a time-varying representation of "instrumental presence" is generated, starting with a frame-wise extraction of the information encoded by the classifiers. That is, a texture window[14] is applied, wherein audio features are both extracted and integrated, and the respective SVM model evaluated. Label inference is then performed on the generated time series by integrating the classifiers' decisions along time. Due to their musically different adoptions, we derive separate labelling approaches for pitched and percussive instruments, which outputs are combined afterwards. Parts of the approaches described in this section have been published by Fuhrmann & Herrera (2010).

4.3.2 Data

For evaluating our labelling approaches we constructed a dataset containing a total number of 235 pieces of Western music, composed of a diversity of musical genres and instrumentations. We asked our lab colleagues – most of them music enthusiasts – to supply us with, at least, five pieces of their favourite music. Additionally, we queried the platform allmusic.com[15] with the modelled instruments and gathered one randomly selected track from each artist of the resulting list. This data gathering process resulted in a diversified set of music pieces, hence guaranteeing for a manifold in musical genres, composition and production styles, and, most importantly, instrumentations. Moreover, we excluded all tracks from the preliminary evaluation collection that were used in the training process of the instrumental models. We applied the fingerprinting algorithm provided by MusicBrainz[16] to unambiguously compare both sets of music pieces. We identified 15 mutually used tracks, resulting in an effective collection size of 220. Additionally, we assigned genre labels to each track in the collection by evaluating the output of 5 human annotators to obtain a consistent description of the musical genres involved in the collection (see Section 6.1.3 for more details and some further remarks on this genre annotation).

[14] The size of this texture window is given by the results of the time scale experiments described in the previous sections, i.e. 3 seconds for both the pitched and percussive labelling.

[15] http://www.allmusic.com

[16] http://musicbrainz.org/doc/PicardDownload

jaz	blu	roc	cla	pop	met	cou	reg	dis
50	7	31	44	64	13	1	1	9

Table 4.8: Genre distribution inside the labelling evaluation dataset. The categories are derived from the genre dataset of Tzanetakis & Cook (2002), see Section 6.1.2 for more details. Legend for the genre labels: Jazz (jaz), Blues (blu), Rock (roc), Classical (cla), Pop (pop), Metal (met), Country (cou), Reggae (reg), and Disco (dis).

Two subjects were paid for annotating the respective half of the collection. After completion, the data was swapped among the subjects in order to double-check the annotation. Moreover, all so-generated annotations were reviewed by a third person to guarantee maximum possible correctness of the data.

Table 4.8 illustrates the distribution of the tracks in the collection with respect to their musical genre according to the human annotations. Note the diversity in musical genres and the atypical *dis* (i.e. Disco) category. Also note that due to the absence of an explicit *Electronic* class in this specific genre taxonomy[17], many electronic pieces are distributed among the *Pop* and *Disco* categories. These tracks mainly exhibit instrumentations involving instruments that are not modelled by the classifiers. Here, mainly synthesiser-based musical instruments are adopted by the composers, exceptionally some pieces feature the modelled instruments *singing Voice* and *Drums*.

In every file the start and end times of nearly all instruments were marked manually, whereas no constraints in the nomenclature were imposed. This implies that in addition to the 11 pitched instruments modelled and the label *Drums*, every instrument is marked with its corresponding name. Hence, the number of categories in the evaluation corpus is greater than the number of categories modelled by the instrumental classifiers. Moreover, if the subject doing the manual annotation could not recognise a given sound source, the label *unknown* was used. To illustrate the distribution of labels inside this music collection, Figure 4.17 shows a cloud of the instrumental tags assigned to the music tracks. As can be seen, all 12 modelled categories exhibit a certain prominence in the cloud, which indicates their importance in Western music. Note especially the weight of the *unknown* category; a statistical analysis of this "category" shows that each music track in the collection contains, on average, 1.61 *unknown* instruments. Moreover, Figure 4.18 depicts the histogram of the number of labels annotated per track, indicating the instrumental complexities covered by this collection.

4.3.3 Approaches

In this section we present the respective algorithms developed for the extraction of labels from the frame-based model decisions. Again, the following is divided into pitched and percussive instruments.

[17] We here adopted the taxonomy proposed by Tzanetakis & Cook (2002). See Section 6.1.2 for the motivations behind this adoption, a detailed analysis of the human genre ratings, and some further taxonomic issues.

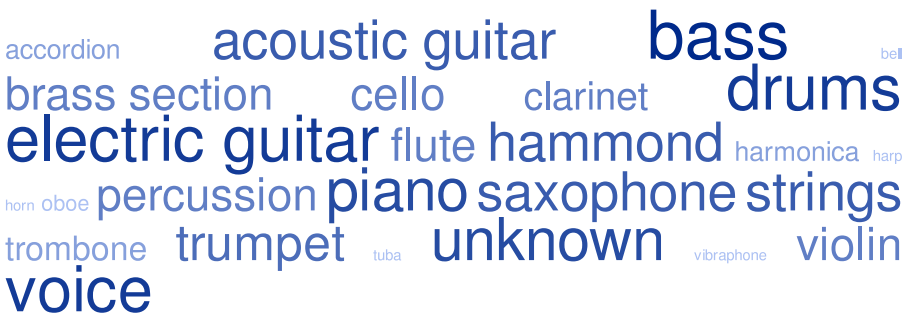


Figure 4.17: Tag cloud of instrumental labels in the evaluation collection. Font size corresponds to frequency. Note the prominence of the 12 modelled categories.

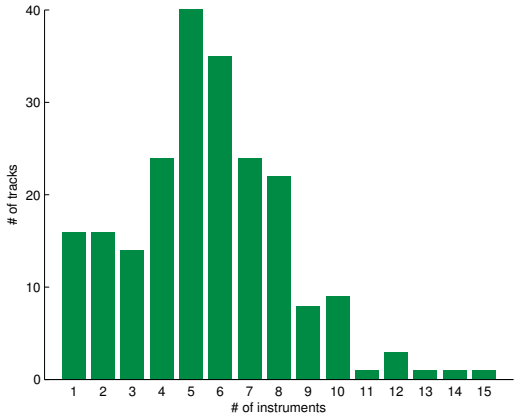


Figure 4.18: Histogram of the number of per-track annotated labels in the evaluation collection.

4.3.3.1 Pitched instruments

The inference of pitched instrumental labels is based on an analysis of the "instrumental presence" representation, which is generated by applying the instrumental model sequentially to the audio signal using a hop size of 0.5 seconds. The resulting multivariate time series is then integrated to obtain the final labels and corresponding confidence values.

The first step consists of estimating the reliability of the segment's label output; given the 11 generated probabilistic output curves, a threshold θ_1 is applied to their mean values along time. This is motivated by experimental evidence that segments with a high number of unknown instruments or heavy inter-instrument occlusion show mean probabilities inside a narrow, low-valued region (note that the instrument probabilities sum to 1 for every frame). If all mean probability values fall below this threshold, the algorithm discards the whole segment and does not assign any pitched label to it. A second threshold θ_2 is then used to eliminate individual instruments showing low activity, which can be regarded as noise. If the mean value of a given probability curve along the analysed signal falls below this threshold, the respective instrument is rejected and not included in the labelling procedure. Figure 4.19 shows an example of the probabilistic representation together with the used

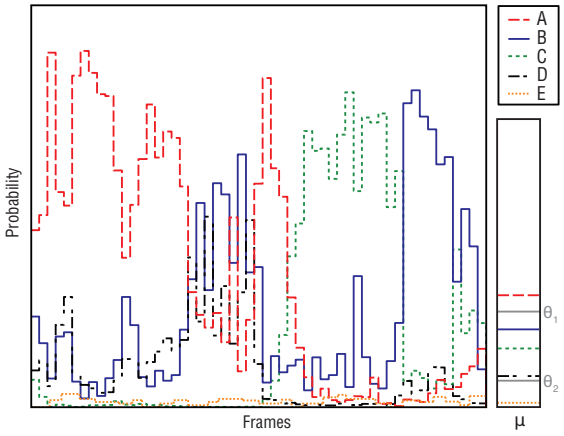


Figure 4.19: An example of the probabilistic representation used for pitched instrument labelling, derived from a 30 second excerpt of music. The main figure shows the probabilistic estimates for sources A-E, the right panel the mean values together with the thresholds used for instrument filtering. The excerpt is used for labelling since A's mean falls above θ_1, and E is rejected as its mean is below θ_2. Note the sequential predominance of the instruments A, C, and B.

threshold parameters. Based on the resulting reduced representation, we derive three approaches for label inference accounting for different granularities of the data's temporal characteristics. Those approaches are:

1. **Mean Value (MV).** The simplest of the considered approaches determines the respective labels by selecting those n_{MV} instruments with the highest mean probabilistic value. The strategy neglects all temporal information provided by the classifier's decisions and derives its output by simply averaging the input. Moreover, it assumes that the predominance of the respective sources is sufficiently reflected in the obtained mean values, e.g in the case of two sequentially predominant sources the highest two mean values should indicate the corresponding instruments. The resulting label confidences are determined by the mean probabilistic value of the respective instruments. Hence, the temporal information and thus the musical context is only incorporated by the predominance of a given instrument with respect to time.

2. **Random Segment (RS).** Segments of length l_{RS} are taken randomly from the reduced probabilistic representation to account for variation in the instrumentation. Within each of these segments, a majority vote among the instruments holding the highest probabilistic values is performed to attach either one or – in the case of a draw – two labels to the signal under analysis. The assigned confidences are derived from the number of the respective instrument's frames divided by both the length l_{RS} and the total number of random segments n_{RS} extracted from the input. All labels are then merged, at which the confidences of identical labels are summed. Here, the temporal dimension of the music is not incorporated, since the information is extracted locally without considering the evolution of the instruments' probabilities along the entire signal.

3. **Curve Tracking (CT).** Probably the most elaborate and plausible approach from the perceptual point-of-view; labels are derived from regions of the excerpt where a dominant instrument can be clearly identified. Decisions in regions where overlapping components hinder confident estimations are inferred from context. The probabilistic curves of the determined instruments are therefore scanned for piece-wise predominant sections. If an instrument is constantly predominating (i.e. it holds the highest of all 11 probabilistic values) within a section with a minimum length l_{CT}, the instrument is added to the excerpt's labels along with a confidence defined by the relative length of the respective section. Moreover, we allow for short discontinuities in these sections of predominance, in order to account for temporary masking by other instruments. This process is repeated until all sections with predominating instruments are identified. Finally, confidence values for multiple labels of the same instrument are summed.

After the respective labelling method we apply a final threshold θ_3 to the estimated confidence values. Labels holding confidences which fall below this threshold are rejected in order to discard unreliable tags.

4.3.3.2 Percussive instruments

In order to determine the presence of the Drumkit, we use a simple voting algorithm working on the classifier's estimates; labelling is performed by accumulating the detected events and deciding on the basis of their frequency. Similarly to the pitched labelling method the developed timbre model is sequentially applied to the audio by using a hop size of 0.5 sec. We then threshold the frame-based probabilistic estimates with a value of 0.5 to obtain a binary representation of classifier decisions. Next, a simple majority vote is performed to determine the presence of the Drumkit. That is, if more than half of the binary decisions are positive, the audio is labelled with *Drums* and the respective confidence is set to the fraction of positive votes.

4.3.4 Evaluation

4.3.4.1 Data

For evaluating the respective labelling methods we use 30-second excerpts, extracted randomly from the music pieces of the evaluation collection described in Section 4.3.2. This strategy to reduce data is common in MIR research, many genre and mood estimation systems use excerpts of 30-second length to represent an entire piece of music [18] (Laurier et al., 2010; Scaringella et al., 2006; Tzanetakis & Cook, 2002). Moreover, this length provides a sufficient amount of data for evaluating the different labelling methods, involving time-varying instrumentations while excluding repetition of

[18] As subsequently discussed in Section 5.2.2.1 an excerpt of 30 seconds is not representative in terms of instrumentation for the entire music piece. Contrary to the concepts of genre and mood, which are rather stable along a music track, instrumentations may vary to a great extent. However, for the purpose targeted here, this length exhibits enough information for evaluating the labelling methods.

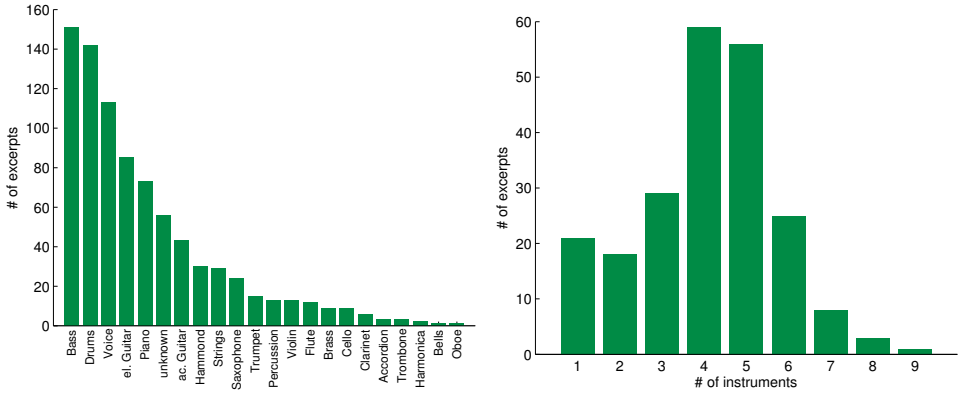


Figure 4.20: Distribution of labels inside the labelling evaluation dataset. Part (a) shows the label frequency, while part (b) depicts the histogram of annotated instruments per excerpt.

Acronym	Value	Description
θ_1	$[0.1, 0.15, 0.2, 0.25, 0.3]$	threshold to filter unreliable input
θ_2	$[0.1, 0.15, 0.2, 0.25, 0.3]$	threshold to filter non-active instruments
θ_3	$[0.1, 0.15, 0.2, 0.25, 0.3]$	threshold to filter low-confidence labels
n_{MV}	$[1, 2, 3]$	number of top-ranked instruments used as labels
l_{RS}	$[4, 5, 6, 7]$	length of the decision window in frames
n_{RS}	$max.4$	number of segments to use for labelling
l_{CT}	$[5, 7, 9, 11]$	minimum length of the section in frames

Table 4.9: Acronyms and respective discrete values of the pitched labelling parameters used in the grid search. The right column shows a short description of the parameter's functionality. See text for more details on the parameters.

the instrumental information due to the musical form of the piece. Figure 4.20 shows the distribution of labels in this dataset together with the histogram of annotated instruments per excerpt.

4.3.4.2 Methodology

Since the three algorithms for the labelling of pitched instruments require a parameter estimation step, evaluation is performed in a 3-Fold CV procedure. That is, in each rotation 2/3 of the data is used for estimating the proper parameter values of the respective algorithms and the remaining 1/3 for performance estimation. A one-stage grid search procedure is carried out during parameter estimation to determine the optimal values from a predefined sampling of the parameter space. Table 4.9 lists all parameters to estimate along with the respective values evaluated during the grid search. In consequence, all forthcoming experiments of this chapter report mean values and corresponding standard deviations.

4.3.4.3 Metrics

To estimate the performance of the label inference approaches we regarded the problem as multi-class, multi-label classification (cf. Turnbull et al., 2008). That is, each instance to evaluate can hold an arbitrary number of unique labels of a given dictionary. By considering $\mathcal{L}$, the closed set of labels to evaluate, $\mathcal{L} = \{l_i\}, i = 1 \dots L$, we first define the individual precision and recall metrics for each label by

$$P_l = \frac{\sum_{i=1}^{N} \tilde{y}_{l,i} \cdot \hat{y}_{l,i}}{\sum_{i=1}^{N} \tilde{y}_{l,i}}, \quad \text{and} \quad R_l = \frac{\sum_{i=1}^{N} \tilde{y}_{l,i} \cdot \hat{y}_{l,i}}{\sum_{i=1}^{N} \hat{y}_{l,i}}, \tag{4.14}$$

where $\hat{\mathcal{Y}} = \{\hat{y}_i\}, i = 1 \dots N$, and $\tilde{\mathcal{Y}} = \{\tilde{y}_i\}, i = 1 \dots N$, with $\tilde{y}_i \subseteq \mathcal{L}$, denote, respectively, the set of ground truth and predicted labels for the elements x_i of a given audio dataset $\mathcal{X} = \{x_i\}, i = 1 \dots N$. Here, $\tilde{y}_{l,i}$ ($\hat{y}_{l,i}$) represents a boolean variable indicating the presence of label l in the prediction (ground truth annotation) of the instance x_i. Furthermore, we derive the individual label F-metric by combining the aforementioned via their harmonic mean, i.e. $F_l = \frac{2P_l R_l}{P_l + R_l}$.

To estimate the cross-label performance of the label inference, we define macro- and micro-averaged F-metrics (Fan & Lin, 2007). First, the macro-averaged F-measure F_{macro} is derived from the individual F-metrics by calculating the arithmetic mean, resulting in

$$F_{\text{macro}} = \frac{1}{L} \sum_{l=1}^{L} F_l = \frac{1}{L} \sum_{l=1}^{L} \frac{2 \sum_{i=1}^{N} \tilde{y}_{l,i} \cdot \hat{y}_{l,i}}{\sum_{i=1}^{N} \tilde{y}_{l,i} + \sum_{i=1}^{N} \hat{y}_{l,i}}. \tag{4.15}$$

Furthermore, we define the micro-averaged F-metric F_{micro}, taking the overall label frequencies into account, hence

$$F_{\text{micro}} = \frac{2 \sum_{i=1}^{N} \sum_{l=1}^{L} \tilde{y}_{l,i} \cdot \hat{y}_{l,i}}{\sum_{i=1}^{N} \sum_{l=1}^{L} \tilde{y}_{l,i} + \sum_{i=1}^{N} \sum_{l=1}^{L} \hat{y}_{l,i}}. \tag{4.16}$$

Moreover, to provide a global estimate of the system's precision and recall, we introduce the micro-averaged, i.e. weighted cross-label average, analogues as defined by

$$P_{\text{micro}} = \frac{1}{\sum_{l=1}^{L} \sum_{i=1}^{N} \tilde{y}_{l,i}} \sum_{l=1}^{L} \sum_{i=1}^{N} \tilde{y}_{l,i} P_l = \frac{\sum_{l=1}^{L} tp_l}{\sum_{l=1}^{L} tp_l + \sum_{l=1}^{L} fp_l}, \tag{4.17}$$

$$R_{\text{micro}} = \frac{1}{\sum_{l=1}^{L} \sum_{i=1}^{N} \hat{y}_{l,i}} \sum_{l=1}^{L} \sum_{i=1}^{N} \hat{y}_{l,i} R_l = \frac{\sum_{l=1}^{L} tp_l}{\sum_{l=1}^{L} tp_l + \sum_{l=1}^{L} fn_l}, \tag{4.18}$$

where tp_l, fp_l, and fn_l denote, respectively, the true positives, false positives, and false negatives of category l. Note that all micro-averaged metrics are weighted according to the instance frequency of

the respective categories, hence more frequent categories have more impact on the respective metric than less frequent ones. On the other hand, macro-averaged metrics apply a simple arithmetic mean across categories, thus all classes contribute to the same extent to the metrics regardless their frequency.

4.3.4.4 Baseline systems

To frame the experimental results we introduce a comparative baseline systems which incorporates the label frequencies in the used evaluation collection. This null model is generated by drawing each label from its respective prior binomial distribution and averaging the resulting performance over 100 independent runs (Ref_{prior}).

4.3.5 General results

Table 4.10 shows the obtained results for all three considered approaches for labelling pitched instruments along with the prior-informed baseline. Note that the pitched and percussive labels are evaluated jointly, thus the methods in the first column of the table only account for the labels of pitched instruments, to which the estimated Drumset label is added. The depicted metrics are related to the amount of erroneous and correct predictions (P_{micro} and R_{micro})[19], as well as the global labelling performance based on the amount of instances (F_{micro}) and categories (F_{macro}).

First, it can be seen that the proposed labelling methods are performing well above the prior-informed label assignment Ref_{prior}. This substantiates the representativeness and validity of the recognition models and confirms the hypothesis that the contextual information is an important cue for label inference, at all levels of granularity exploited here. In total, the algorithms are able to extract almost 60% of all annotated labels correctly, which results in a F score across categories of 0.45. This difference in the two applied F metrics reflects the imbalance of individual instrumental labels in the applied testing collection. However, the models' recognition performance is maintained (see Section 4.2), but the here-evaluated labelling approaches are not limited to predominant sources; all annotated labels are weighted equally in this evaluation.

Second, regarding the three different labelling methods for the pitched instruments, we can observe that none of the proposed methods performs superior than the others. This is even more surprising when considering the conceptual difference of taking just the mean probability of the instruments along the analysed signal (MV) and scanning their output probabilities for piece-wise maxima (CT). On the contrary, we only observe a slightly better performance in terms of both F metrics of MV and CT against RS. We may explain this observation by the fact that if an instrument is predominant it is recognised by all three methods, since all account for the sources' predominance inside the signal. On the other hand, if the algorithm is faced with a too ambiguous scenario, the methods perform similarly bad. The observed small differences between RS on the one side and MV and CT on the

[19] In case that a given instrument i is never predicted for any audio file to evaluate, its respective value of P_i is undefined. We therefore substitute the precision value with the instrument's prior probability, as used for the baseline approach Ref_{prior} (cf. Turnbull et al., 2008).

Method	P_{micro}	R_{micro}	F_{micro}	F_{macro}
Ref_{prior}	0.4 ± 0.02	0.4 ± 0.02	0.4 ± 0.02	0.21 ± 0.02
MV	0.7 ± 0.083	0.58 ± 0.018	0.63 ± 0.03	0.045 ± 0.034
RS	0.61 ± 0.042	0.59 ± 0.087	0.6 ± 0.041	0.43 ± 0.005
CT	0.7 ± 0.083	0.57 ± 0.042	0.62 ± 0.031	0.45 ± 0.035

Table 4.10: General result for the labelling evaluation. Note that the output of the two distinct labelling modules is already merged, even if the compared labelling method only apply for the pitched instruments. The results are, however, proportional.

other side however result from their different analysis "scopes", since the latter two are incorporating the entire instrumental information of the signal, thus are able to better account for the temporal continuity of predominant information inside the signal. The former only applies local information and may thereby extract more likely short-term spurious information, as manifested in the low value of the precision P_{micro} in Table 4.10.

Figure 4.21 furthermore shows the F score for the individual instrumental categories. Again, we cannot observe any significant differences among the three examined pitched labelling approaches, which emphasises the conclusions drawn above. Moreover, the noticeable spread in the standard deviations of particular instruments is related to their annotation frequency inside the used evaluation collection; the higher the depicted standard deviation the less frequently the respective instruments appears in the dataset. What follows is a detailed examination of the performance of individual instruments with respect to the applied evaluation metric. This will further reveal factors influencing the performance of the developed labelling methodology. Moreover, by comparing the here-presented individual performance figures to the metrics obtained in the evaluation of the recognition models (Figure 4.8), we can derive conclusions about the nature of the data and the role of the covered musical instruments therein.

First, we can observe that usually prominent instruments such as the *singing Voice*, the *electric Guitar*, or the *Saxophone* show best performance among the evaluated pitched instruments. This predominance in opposition to the other modelled instruments is neither reflected in the training nor the evaluation process of the recognition models, thus explaining the difference in the respective performance figures. Especially the *singing Voice* improves with respect to the performance in the model evaluation, an indicator for the influence of the context analysis for label inference in case of highly predominant instruments. The same applies to the *Saxophone*, being the worst instrument in the model examination (Figure 4.8). We may explain the here-observable performance with its predominant character in solo phrases, which are typical for this instrument. Moreover, the estimation of the Drumkit performs satisfying, owing again to its predominance in the mixture, indicating that the employed label inference method is appropriate for the task at hand. Besides, the mean value of the resulting confidence values of the *Drums* label for all evaluated instances exceeds 80%, which suggest that the applied methodology for label inference, based on the majority vote, is properly suited.

Next, the worst performance can be observed for the instruments *Clarinet* and *Flute*. Here, the combination of, on the one hand, their sparse appearance in the evaluation data (see Figure 4.20a) and, on the other hand, the usual absence of a predominant character explains the algorithms' low

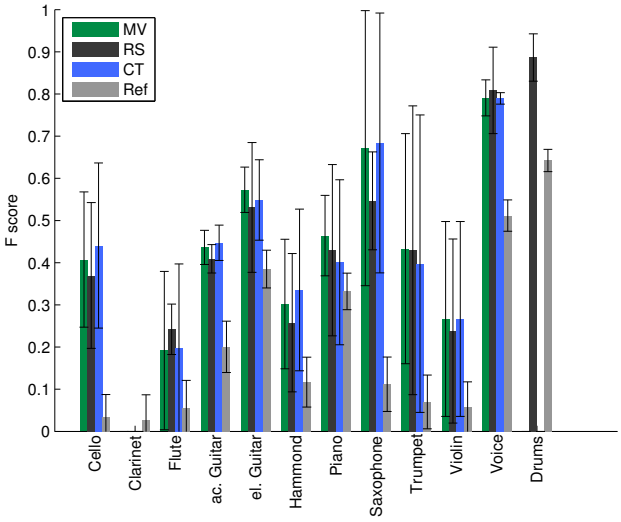


Figure 4.21: Labelling performance of individual instruments. Note that the legend is not applying for the *Drums* category, since the compared labelling methods only account for the pitched instruments. Legend of the labelling approaches: mean value (MV), random segment (RS), and curve tracking (CT). See Section 4.3.3 for details.

labelling performance for these two instruments compared to the other ones. In particular, the *Clarinet* is never predicted correctly on a total of 6 instances containing the respective label (besides, in only 3 of these 6 instances the instrument exhibits a predominant character). In this context, the *Violin*, performing similar to the *Flute*, must by treated slightly different since its performance is underestimated in this analysis. The instrument is often predicted for instances labelled with *Strings* (see Table 4.11), resulting in a lower value of its precision value, hence degrading the corresponding F score. Since we can regard a *Violin* label predicted for a string section as a correct prediction, these "correct" decisions are not reflected by the applied statistical metrics.

At last, the worse performance of *acoustic Guitar* and *Piano* with respect to the corresponding metrics observed in evaluation of the recognition models (Figure 4.8) results from their accompaniment character in most part of the considered musical scope. Thus, in many cases these instruments do not exhibit a predominant character, which is reflected in their poor labelling performance. Note also the close performance of the baseline in the case of the *Piano*; this particular combination of an accompaniment instrument with a high annotation frequency (Figure 4.20a) further decreases the gap to the simple prior-based performance.

4.3.5.1 A note on the parameters

By examining the combinations of parameters resulting in the best performance for the respective pitched labelling methods we observe that, for each method, several parameter combinations lead to the same labelling performance. Noticeable here is the trade-off between the two filters θ_2 and θ_3, which determine, respectively, those instruments to be considered in the labelling process and the threshold for discarding weak labels. In general, these parameters control the algorithm's precision and recall in a given range of values, while keeping the overall labelling performance rep-

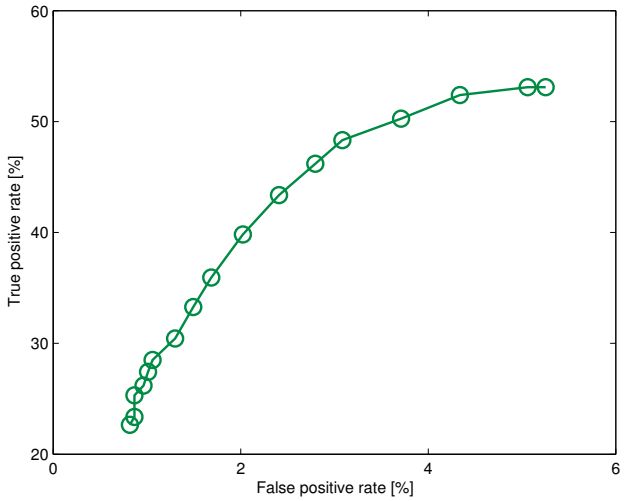


Figure 4.22: ROC curve of labelling performance for variable θ_2. The plot shows the true positive against the false positive rate for decreasing values of the parameter.

resented by the F-metrics the same. Best performance can be particularly observed if one of them takes control of the algorithm's precision while the other controls the recall value; a low value of θ_2, for instance, results in a high recall while the corresponding high value of the label filter θ_3 causes a high precision. Approximately the same performance figures can be accomplished by inverting the two parameters' values; a high value of θ_2 guarantees the high precision and the low value of θ_3 the corresponding high value in recall. Moreover, the importance of a given parameter depends on the labelling method; that is, θ_1 is crucial for the RS method, while the other approaches, i.e. MV and CT, control their performance with θ_2 and θ_3, setting θ_1 to 0. This is reasonable since RS does not incorporate all instrumental information but is rather acting on a limited time support; it is therefore more dependent on a prior elimination of possible spurious information, while MV and CT can filter these locally appearing errors using the musical context. To demonstrate the influence of a single parameter on the algorithm's labelling performance, Figure 4.22 shows a Receiver Operating Characteristics (ROC) curve, depicting the resulting true and false positive rate at a variable θ_2. Usually ROC curves are used to graphically illustrate the trade-off between the hit and false alarm rate of a machine learning algorithm, given a varying system parameter (Witten & Frank, 2005). Hence, we bypass θ_1 as well as θ_3 and vary θ_2 from 0.9 downwards to 0, processing all files in the collection by applying the CT labelling method. Since the conception of the labelling methodology does not allow for the entire range of both ordinate and abscissa in the figure, i.e. in practice it is not possible to reach both 100% true and false positive rate, the range of the curve is limited. However, it still can be seen that the composite of correct and incorrect predictions can be adjusted by different settings of the parameter, at which the optimal trade-off is located around 50% of the true positive rate[20].

Next we look at the approaches' individual parameters, i.e. the number of top-ranked instruments n_{MV} for the MV approach and the minimal length l_{CT} for the CT method. The optimal parameter

[20] The optimal performance is found at the particular point where the tangent to the curve exhibits the same slope as the diagonal of the full-range plot.

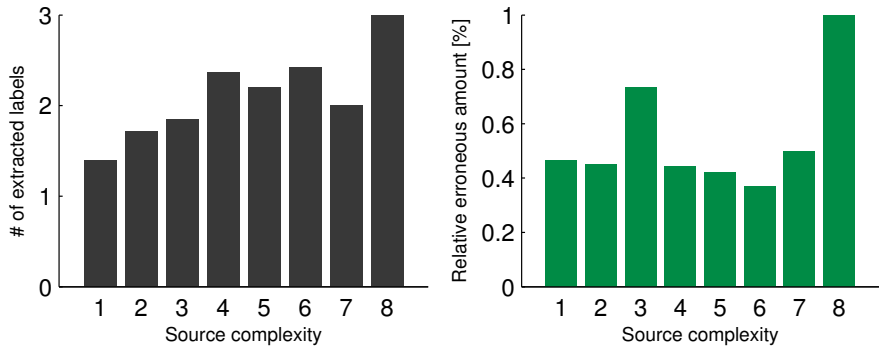
values for each rotation in the 3-Fold CV show a value of 2 for the first parameter. This suggest that also the simple MV method is able to extract multiple predominant instruments from a single music signal. Here we speculate that the method is even able to handle both the case of two sequential predominant instruments and the case of simultaneous predominant sources. A value greater than 2 is however decreasing the labelling performance, indicating that the third mean value already comprises to a large part spurious information. The second individual parameter, the CT method's l_{CT} parameter, leads to best labelling performance for small values, i.e. values of 5 to 7 consecutive classification frames. Since the labelling threshold θ_3 is used to discard labels with low confidence values, i.e. resulting from segments of short duration, it seems that the functionality of this small value for the l_{CT} parameter is primarily to enhance already found labels by increasing their confidence values (recall that the confidence values of multiple identical labels are summed). A label derived from a single predominant occurrence in the probabilistic representation of "instrumental presence" of this short length would probably fall below θ_3 and therefore be eliminated.

4.3.6 Analysis of labelling errors

Similar to the analysis of classification errors we here perform a qualitative analysis of labelling errors. Again, we concentrate on the consistent confusions which show regularities across several instances while trying to disregard noisy artefacts. Moreover, we focus on the wrongly predicted rather than on missed labels; since the data used for evaluation is not providing evidence about the predominance of the instrument inside the mixture, evaluating why a certain instrument has not been predicted is more difficult than estimating why a certain label has been wrongly predicted. The latter can mostly be deduced from a confusion with a perceptually predominant instrument, while the former may simply result from the accompaniment character of the source. In particular, we first evaluate the influence of the music's timbral complexity on the labelling performance and then examine the inter-instrument confusions by means of analysing a cross-confusion matrix. Finally, we concentrate on the impact of not-modelled categories and their respective complexity on the output of the presented method.

For the sake of simplicity, we perform all subsequent experiments with the CT labelling method for the pitched instruments in the 3-Fold CV with the respective best parameter settings. Thus the label output of the instance of all 3 evaluation folds is merged and used to perform the following analyses of errors. Since all three labelling methods presented in Section 4.3.3 perform in the same range of accuracy, we expect the here-derived conclusions to be valid for each of the methods.

To quantify the influence of the data's timbral complexity, i.e. the number of concurrent annotated sound sources, on the labelling performance, Figure 4.23a illustrates the efficiency in terms of extracted labels of the applied method. Hence, the number of extracted labels is plotted against the number of annotated labels, showing a ceiling in the number of extracted labels of around 2.3 for complexities greater than 3. This seems reasonable since in most cases only one or two predominant pitched instrument together with the possible label *Drums* is extracted. Following the conventions of the Western musical system it is very unlikely that within 30 seconds of music more than 2 pitched instruments exhibit a predominant character.



(a) (b)

Figure 4.23: Total and relative-erroneous amount of labels attached with respect to the source complexity. Part (a) refers to number of labels attached to an audio instance, while (b) depicts the relative amount of instances producing wrong predictions. Here, the extreme value of 1 at the 8th bar results from a single audio file containing 8 instruments, which the algorithm labels with a wrong label.

The second part of the figure shows the relative amount of instances producing an erroneous prediction, again with respect to the timbral complexity of the instance. It can be seen that there is hardly any direct dependency between the algorithm's error rate and the number of instruments in the signal. The errors are rather uniformly distributed among all different complexities, when disregarding the outliers at complexities of 3 and 8. Especially the high value at the 8th bar in the figure can be neglected since it results from a single audio file containing 8 instruments[21]. This suggests that predominant information of musical instruments is available across all levels of source density to nearly the same extent (of course except for complexities of 1, where only predominant information is present). Moreover, it indicates that the presented algorithm is able to handle all kinds of source complexities with a reasonable constant error rate which is not directly dependent on the number of concurrent sound sources.

Next, Table 4.11 shows a "confusion matrix" extracted from the error predictions of the algorithm's output in the 3-Fold CV procedure. Here, modelled instruments are plotted against annotated ones, resulting in the noticeable imbalance between the horizontal and vertical dimension. For a given wrongly predicted label (column index) we augment the entries of all respective annotated instruments (row index) in the matrix. As a result all diagonal entries of the modelled categories hold 0. Note that an observed prediction error is affecting all musical instruments annotated in the analysed instance, as the error cannot be attributed to a single acoustic source in the ground truth. Hence, a given prediction error is contributing to multiple rows in the table, depending on the number of annotated instruments of the given instance. Even though some of the instrumental combinations shown in Table 4.11 are not informative – for instance the row containing the instances annotated with the label *Drums* does not give any evidence about the confusions with pitched instruments, a result from the universal adoption of the Drumset in the analysed music – this error representation gives useful insights into the functionalities of the presented labelling method. We can particularly deduce conclusions about the labelling performance on both the modelled and not-modelled categories.

[21]Unfortunately, we could not find a straightforward explanation for the increased value of the 3rd bar.

	Cello	*Clarinet*	*Flute*	*ac. Guitar*	*el. Guitar*	*Hammond*	*Piano*	*Saxophone*	*Trumpet*	*Violin*	*Voice*	*Drums*	Σ
Cello	0	0	0	0	0	0	0	1	0	0	0	0	0.11
Clarinet	1	0	1	0	0	0	0	1	1	0	0	0	0.67
Flute	1	1	0	0	0	3	1	0	0	2	0	2	0.83
ac. Guitar	1	1	3	0	0	2	3	1	0	5	0	9	0.58
el. Guitar	2	3	8	2	0	10	4	1	3	2	2	4	0.48
Hammond	1	1	6	1	0	0	3	0	1	1	0	3	0.57
Piano	4	4	8	3	2	4	0	7	5	3	3	7	0.68
verdammt.*Saxophone*	1	3	3	1	2	1	0	0	4	0	1	0	0.67
Trumpet	1	1	1	0	1	0	0	1	0	1	1	0	0.47
Violin	1	1	1	0	0	1	0	0	1	0	0	1	0.46
Voice	6	1	9	3	3	14	3	7	2	5	0	14	0.59
Drums	1	3	11	5	6	18	4	6	5	5	5	0	0.49
Strings	3	1	1	1	0	1	1	1	2	8	1	3	0.79
Brass	0	0	0	0	0	1	1	4	0	0	1	0	0.78
Bass	1	5	13	4	3	16	4	6	7	4	5	8	0.5
Unknown	2	1	10	5	5	6	3	0	1	4	1	3	0.73
Percussion	0	1	0	0	1	2	1	1	0	0	0	6	0.92
Trombone	0	0	0	0	0	0	0	2	1	0	0	0	1
Harmonica	0	0	1	1	0	0	0	0	0	0	0	0	1
Accordion	0	0	0	0	0	1	0	0	0	0	0	2	1
Bells	0	0	0	0	0	0	1	0	0	0	0	0	1
Oboe	0	0	0	0	0	0	0	0	0	1	0	0	1
Horn	0	0	0	0	0	0	0	0	0	0	0	0	0
Tuba	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.11: Confusion matrix for labelling errors. The vertical dimension represents the ground truth annotation, while the horizontal one denotes the predicted labels. Note that only wrongly predicted labels are considered, i.e. missed labels are not counted. Moreover, a given error is assigned to all instruments in the respective annotation, hence depending on the number of instruments annotated in the respective audio file, appearing multiple times in the matrix. The last column represents the relative weight of the categories' errors.

Analysing the modelled categories, as indicated by the light-grey rectangle in Table 4.11, we can confirm several observations from our previous, "frame-based" error analyses, e.g. see Section 4.2.3.6; for instance, the row containing instances labelled with *acoustic Guitar* shows a significant amount of wrongly predicted labels *Drums*, a fact that has been already observed in Section 4.2.4.5. Similarly, the confusions between *electric Guitar* and *Hammond organ*, attributed to the distortion effect frequently applied by both instruments, or between *Hammond organ* and *Flute* can be found here.

The right-most column in Table 4.11, denoted with Σ, shows the relative amount of erroneous instances of a given category. Thus, we can rank the modelled categories according to this quantity. Similar to the results presented in Figure 4.21, *Flute* performs worst among all modelled instruments

with a total of 83% of wrongly labelled instances, followed by *Piano*, *Saxophone* and *Clarinet*. The latter produces typical confusions with other blown instruments, i.e. *Flute*, *Saxophone* and *Trumpet*. Also the *Saxophone*, albeit performing 2nd best of the pitched instruments in Figure 4.21, shows a fraction of 67% wrongly labelled instances. Here, similar confusion patterns as in the analysis of classification errors (Section 4.2.3.6) can be observed, particularly with the other blown instruments *Clarinet*, *Flute*, and *Trumpet*. At last, the low performance of the *Piano* in this representation can be again explained by both its usual accompaniment character and the fact that it is the only instrument equally employed in all the covered musical genres.

Surprisingly, some of the previously encountered mutual confusions between certain musical instruments are not represented in Table 4.11. We observe a good separation between *acoustic* and *electric Guitars*, which cannot be found in Table 4.3. Correspondingly, *Cello* and *Violin* do not show those strong confusions as illustrated in the confusion matrix of the classification performance. These results may be explained by both the sparsity of some labels in the dataset used for this evaluation and the different adoptions of the instruments depending on the musical context.

An analysis of those categories not modelled by the classifiers in Table 4.11 shows most instances producing "natural" confusions, i.e. confusions expected when considering the acoustical properties of those instruments. In particular, String and Brass sections are labelled in a large part with the respective containing instruments, that is *Cello* and *Violin* labels for *Strings*, and *Saxophone* for the *Brass* category. Also the instances annotated with *Trombone* exhibit these corresponding predictions, i.e. labels *Saxophone* and *Trumpet*, an indication that the acoustical characteristics of those instruments have been encoded properly by the models. Moreover, the *Percussion* category shows strong confusions with the label *Drums*, as similarly observed in Section 4.2.4.5. Finally, and not surprisingly, we identify the *unknown* category as frequent source for labelling errors. Here, conclusions concerning the confusions with the modelled instruments are more than speculative, since the acoustical properties of those unknown sources are not known beforehand.

Finally, we examine the labelling performance with respect to the number of unknown sources present in the evaluation instances. That is, we group the output of the CV related to the amount of not-modelled sources and calculate the evaluation metrics (Section 4.3.4.3) for all resulting groups. Figure 4.24 shows the results in terms of the obtained F-metric F_{micro}. It can be seen that for numbers of 1 to 3 unknown sources the performance of the algorithm degrades gracefully, as stated in the requirements for recognition systems presented in Section 3.3. However, the low value for those instances containing no unknown instrument does not fully agree with this conclusion. We may speculate that, on the one hand, the imbalance in instances between the different groups causes this unexpected value (34, 113, 64, and 9 for numbers of 0, 1, 2, and 3 unknown instruments, respectively). On the other hand, since the musical role of the not-modelled instruments is not known beforehand – it may exhibit accompaniment or solo characteristics – their total amount is only slightly influencing the system's performance on average. Of course, the greater their number the higher the probability a given unknown source exhibits a predominant character, thus causing wrong predictions, which explains the degrading performance of higher-order groups in Figure 4.24. Hence, we can conclude that the number of unknown instruments plays a subordinate role for our recognition system, more important for the labelling performance is the predominance a certain source – both known and unknown – exhibits.

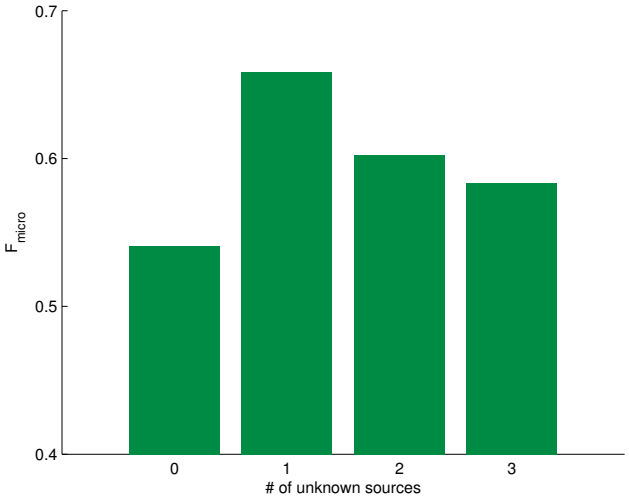


Figure 4.24: Labelling performance with respect to the amount of unknown sources.

4.4 Discussion

4.4.1 Comparison to the state of the art

Here, we shortly relate the presented method on the basis of the obtained results to the corresponding literature in the field of automatic recognition of musical instruments. A reasonably fair comparison is only possible to those studies using real music audio data for evaluation with a similar timbral complexity to ours (in our experiments the maximum is 9 concurrent sources, see Figure 4.20b) and not using any prior information regarding the data in the recognition process. Hence, from the works listed in Table 3.1, Barbedo & Tzanetakis (2011); Eggink & Brown (2004); Essid et al. (2006a); Kobayashi (2009); Leveau et al. (2007); Simmermacher et al. (2006) fulfil the aforementioned criteria. If we then consider the variety in musical styles and genres in the respective studies' evaluation data, we can only keep the works by Kobayashi (2009) and Barbedo & Tzanetakis (2011) for an adequate comparison of recognition performance. Among those three – the two aforementioned and the approach presented in this thesis – Kobayashi (2009), who applies a conceptually similar approach, scores best with 88% of total accuracy for the 50 track evaluation collection. However, this work is incorporating the fewest categories, which moreover include compound categories such as string and brass sections. Here, Barbedo & Tzanetakis (2011), relying on multi-pitch estimation rather than extensive machine learning, is ahead with 25 different categories, which strengthens the impact of the obtained F-score of 0.73. The authors however included neither any "not-known" instruments nor heavy percussive sources in the evaluation data (moreover, the authors note that in the presence of heavy percussion the recognition performance drops to a value of around 0.6 in terms of the F-score). This fact in turn hampers a direct comparison to both the work of Kobayashi (2009) and the here-presented approach. Furthermore, the evaluation data of our method is the most versatile of all three studies, thus incorporating a great amount of not-modelled sources along with the greatest variety in musical styles and genres, including even electronic music.

In conclusion, yet a reduction to the most similar approaches in literature does not guarantee a direct and fair comparison between the respective works. Only in the presence of a general evaluation framework, including a constant taxonomy together with a corresponding evaluation dataset, a comparative analysis becomes possible. In the context of the above-mentioned we regard the performance of our method as state-of-the-art, albeit the existence of a large head-room for improvement of the labelling performance. In this course we want to contribute to the research community with the public availability of the data used in this thesis and thereby hope to improve the comparability of the different approaches in literature. The training data excerpts, the annotations of the evaluation tracks along with an extensive list of the corresponding audio files can be found under http://www.dtic.upf.edu/~ffuhrmann/PhD/data.

4.4.2 General discussion

In this chapter we have presented our approach towards the inference of labels related to musical instruments from polytimbral music audio signals of any length. We combine the frame-level output of pre-trained statistical models of musical instruments (both pitched and percussive) with musical knowledge, i.e. context analysis, to develop a method that robustly extracts information regarding the instrumentation from unknown data. Our focus thereby lies on the the development of a general purpose method, i.e. a method that can be used without additional information[22], thus reflecting an everyday music listening situation. The resulting computational implementation is further thought to be embedded into typical MIR systems performing operations such as music indexing or search and retrieval. We therefore conceptualise the presented method under these constraints, i.e. we adapt the algorithmic design, the taxonomy, and the resulting system's complexity to the envisioned task, i.e. the recognition of musical instruments from Western musical compositions in connection with the integration inside a typical MIR framework.

In the beginning of this chapter we stated 3 hypotheses reflecting our main assumptions prior to the design process of the presented method (Section 4.1). We now are able to validate these 3 theoretical claims by examining the results presented in the respective evaluation sections of this chapter. In particular, we recapitulate the following from our observations and relate it to these hypotheses:

Hypothesis 1 – the ability of extracting instrument specific characteristics from polytimbral music audio signals given a certain amount of predominance of the target – is clearly validated by a reflection on the results presented in Sections 4.2.3.4 and 4.2.4.3, and the corresponding analyses of the involved acoustical features. The performance of both the pitched and percussive recognition model is far in excess of the used null model $\overline{A}_{null}$. Moreover, the presented algorithmic implementation outperforms or is equivalent to all other tested methods in the respective case, i.e. pitched and percussive recognition. Next, the analyses of the most important descriptions in terms of audio features revealed those acoustical dimensions that are widely known to define the different timbres of the employed instrumental categories. In particular, the features selected by our feature selection procedure resemble those features determined to be important in perceptual studies using mono-

[22] We note that the inference process does not need any a priori information, thus the method can be applied to any piece of music regardless of its genre, style, instrumentation, number of concurrent sources, etcetera.

phonic input data. In essence, the information extracted from the audio signal and subsequently applied in the modelling process corresponds to the acoustical properties – or invariants – of the respective instruments.

Hypothesis 2 – the importance of contextual information for label inference – is validated by the results of the labelling algorithms presented in Section 4.3.5. Here, we compare 3 labelling methods, each incorporating a different amount of contextual information, at which all methods clearly outperform the comparative null model Ref_{prior}[23], which is based on the prior distribution of the categories inside the used data collection. Moreover, we observe an advantage of increasing contextual information for labelling performance; those methods which incorporate the full contextual scope score slightly better than the method which uses only local context for label inference. Since the data used to evaluate the labelling methods does not account for predominant instruments, i.e. the ground truth annotations consider all instruments equally (see Section 4.3.2), the importance of the context analysis is also apparent when considering the properties of the labelling approaches. By focussing on those sections with the most confident classifier output while disregarding model decisions on frames where overlapping sources are hindering reliable estimations, a robust label inference is guaranteed. This is also substantiated by the maintenance of performance in comparison to the frame-level evaluations of Sections 4.2.3.4 and 4.2.4.3.

Hypothesis 3 – the validity of the extracted information inside a typical MIR framework – is confirmed by the results obtained from the analysis of labelling errors in Section 4.3.6. Apart from the noise that can be observed in the main confusion matrix of Table 4.11, the most prominent confusions as well as the algorithm's performance on the not-modelled categories can be identified as reasonable. Mutual confusions between modelled categories can mainly be attributed to their similar acoustical properties, while the algorithm mostly predicts acoustically similar instruments on data containing prominent unknown categories, which are present in the evaluation data. Additionally we show that neither the timbral complexity nor the amount of unknown categories is affecting the method's labelling performance to a great extent. This indicates that the method can be used inside a typical MIR framework, since it is able to handle Western music pieces of all kind. Hence, we can conclude that the extracted semantic information enables a meaningful modelling of musical instruments, as assumed in the hypothesis.

Nevertheless, compared to the human ability of recognising sounds from complex mixtures – still the measure of all things – we notice a clearly inferior performance of the developed labelling algorithm, although we are lacking a direct comparative study. This is however evident from the noise that can be observed in all confusion matrices presented in this chapter (Tables 4.3, 4.7, and 4.11), which was never observed in perceptual studies including human subjects (e.g. Martin, 1999; Srinivasan et al., 2002). Humans, in general, tend to confuse particular instruments on the basis of their acoustical properties, a property that is also observable with the presented method.

[23] We want to note the good performance of this baseline system as shown in Table 4.10. Even though the baseline is using the same data for training and testing, which evidently results in an overestimation of its performance, the figures suggest that a lot of the information is already covered in the prior distributions of the respective instruments. Hence, future research in instrument labelling should incorporate this source of information, at least in the evaluation to properly estimate the respective system's performance.

Recapitulating, we believe that the results of this chapter, including both the classification and the labelling steps, not only suggest valuable information for automatic musical instrument recognition research, but for MCP research in general. One of the evident findings in the course of this chapter is that information on sound sources can be obtained directly from the mixture signal; hence a prior separation of the concurrent streams is not implicitly necessary for modelling perceptual mechanisms such as sound source identification! Therefore, our results support the music understanding approach, introduced in Section 2.2, which combines information regarding the music itself with perceptual and cognitive knowledge for music analysis. Analogously, our observations disapprove the transcription model, where a score-like representation is regarded as the universal primary stage for all music analysis systems. Moreover, given the results presented in Figure 4.23b, we can further speculate that not the source complexity itself, but rather the noisy nature of the extracted information is causing the model's confusions. This again favours the music understanding model, since a perceptually inspired modelling of the respective sources together with the provided context should be able to reduce the noise and thereby increase the algorithm's labelling performance. Thus a context-informed enhancement of the source components together with an adequate modelling of the sources – recall that for the human mind learning is a life-long process (see Section 3.3) – seems to be sufficient for a robust recognition of musical instruments in the presence of concurrent sources and noise.

5

Track-level analysis

Methods for an instrumentation analysis of entire music pieces

In the previous chapter we yet concentrated our efforts on processing music audio signals of any length, by presenting a general methodology for automatic musical instrument recognition. The thereby analysed music was not subjected to any convention with regard to formal compositional rules, we particularly evaluated our system on randomly extracted musical excerpts of 30 seconds length. In this chapter we want to exploit the properties that these formal aspects, typically found in Western music, offer to guide the extraction of instrumental labels form entire pieces of music. Like in the previous chapter, we here introduce our main hypotheses that lead to the developments described in the course of this chapter. These assumptions refer to the main criteria we consider prior to the design process of the specific algorithms and will be validated subsequently. They can be stated as follows:

1. The instrumental information that can be extracted from predominant sources represents an essential part of the composition's instrumentation. Therefore, most of the instruments playing in a given music piece appear at any time in a predominant manner.
2. The recurrence of musical instruments, equivalent to the redundancy of instrumental information, within a musical composition can be exploited for reducing the data used for the label inference process, hence alleviating the total computational load of the system.

In particular, we hypothesise that using knowledge derived from the global characteristics of a given music piece is beneficial for instrument recognition in several respects; we may only process those sections where recognition is more reliable or reduce the overall amount of analysed data by exploiting redundancies in the instrumentation. More precisely, the presented methods consider higher-level properties of musical compositions such as structural and instrumental form. In general, this

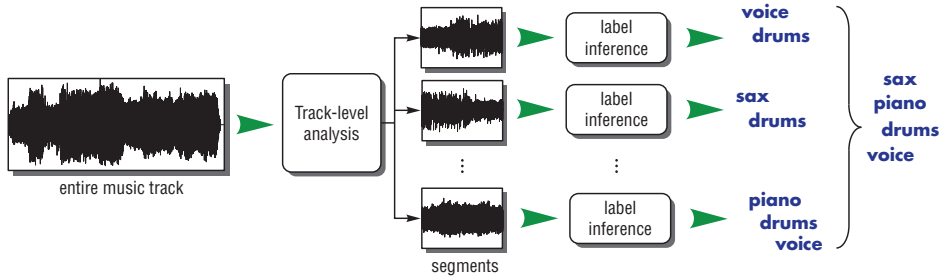


Figure 5.1: The general idea behind the track-level approaches; given an entire piece of music the respective track-level method outputs a set of segments according to its peculiar specifications. We then apply the label inference algorithm to these segments to derive the instrumental labels for the piece under analysis.

so-called *track-level analysis* supplies the subsequent instrument recognition with a list of segments, which indicate where and how often the label inference algorithm has to be applied to extract the most confident or representative instrumental labels. We then evaluate these approaches with respect to the correlation between the obtained labelling performance and the amount of data used for inference. Figure 5.1 illustrates the general idea behind the track-level analysis.

In the following we present two conceptually different approaches towards the recognition of instruments from entire music pieces; first, we describe a knowledge-based approach which identifies sections inside a musical composition exhibiting a certain degree of predominance of one of the involved musical instruments (Section 5.1). Second, we present several agnostic approaches to select the most relevant sections in terms of the analysed track's instrumentation, optimising the problem of both maximising the recognition performance and minimising the computational costs (Section 5.2). These methods are then evaluated in the instrument recognition framework (Section 5.3), considering both the overall labelling accuracy in Section 5.3.4 and their performance with respect to the amount of data used for processing (Section 5.3.5). We finally close this chapter with a discussion of the obtained results and concluding remarks (Section 5.4).

5.1 Solo detection – a knowledge-based approach

The key idea behind this first track-level approach is to locate those sections inside a given piece of music that conform best with the assumptions we have taken in the design process of the label inference method. That is, the existence of a single predominant source, as incorporated in the training data of the recognition models. Furthermore, we already identified the predominance of a single musical instrument being a crucial factor for a successful label extraction. Hence, the developed method explicitly looks for segments in the musical composition, where one single source is pre-

dominating the presumably polytimbral mixture. Due to the relatedness of our definition of the predominance of a source in a musical context (see Section 4.1) and the musical concept of a *Solo*, we derived the name *Solo detection*[1]. In this context, we use the definitions of a *Solo* proposed by the Grove Dictionary of Music and Musicians (Sadie, 1980):

> "[...] a piece played by one performer, or a piece for one melody instrument with accompaniment [...], *and*, [...] a section of the composition in which the soloist dominates and the other parts assume a distinctly subordinate role."

5.1.1 Concept

Our aim is to derive a segmentation algorithm which partitions the music audio signal into *Solo* and *Ensemble* sections. Following the definition from above, we regard all sections of a musical composition a *Solo*, which exhibit a single predominant instrument. In this context, the definition also includes, apart from all possible pitched instruments[2], the *singing Voice*. In Western music the *singing Voice* usually exhibits a strong predominance inside the music, a result from the common mixing and mastering process.

We utilise general acoustical and perceptual properties related to the existence of such a predominant source for segmenting the audio data into blocks of consistent information. The underlying hypothesis is that given a sufficient amount of representative data together with a proper encoding of the relevant information, we can apply a pattern recognition approach to learn the differences that music audio signals with and without a single predominant source exhibit. These learnt models can then be applied to identify, in a given piece of music, those section containing predominant instrumental information.

From this it follows that one key aspect in this analysis involves determining the proper encoding of the information that discriminates best the target categories. The main criterion thereby is to describe the general characteristics of predominant instruments regardless of the instrument's type. Here, we rely on spectral and pitch related characteristics of the signal, described by low-level audio features. Hence, we expect the signal of a predominant sound in general to be different from other sounds not comprising such instruments in terms of these descriptions of the audio signal.

Stated differently, we look for sections in the signal of a given music piece, where instrument recognition is "easier" than for other sections. Typical *Solo* sections exhibit less overlapping components of concurrent musical instruments which simplifies the extraction of the instrument's timbre from the mixture signal. Parts of the here-presented work have previously been published by Fuhrmann et al. (2009b).

[1] We will use the term *SOLO* in the remainder of this chapter.

[2] Here, we are not directly considering percussive instruments since those instruments anyway show a predominant character along the entire piece of music. Thus we assume that if percussive sources are present in the track under analysis, the selected segments contain enough information for their successful recognition.

5.1.2 Background

In this section we summarise the scarce works targeting the problem of detecting predominant instruments in music. The problem itself can be regarded as special variant of the general class of supervised audio segmentation, i.e. partitioning the audio data in homogeneous regions and assigning the corresponding class label to the respective segments.

Peterschmitt et al. (2001) used pitch information to locate solo phrases in pieces of classical music. In this study the mismatch index of a monophonic pitch estimator, derived from the deviation of the observed to the ideal harmonic series, indicates the presence of a predominant instrument. The authors trained the pitch detector using examples of a given instrument and applied the developed detection function to unknown data. Although the initial observations were promising, the overall results did not satisfy the prospects of the research; the derived decision function was far too noisy to discriminate between solo and ensemble parts and resulted in a percentage of 56% correctly assigned frames.

Similarly, Smit & Ellis (2007) applied the error output of a cancellation filter based on periodicity estimation for locating single voice sections in opera music. In particular, the output of an autocorrelation analysis directed a comb filter, which cancelled the harmonic parts of the analysed signal. Then, a simple Bayesian model classified the error output of this filter and a final HMM extracted the best label sequence from the resulting likelihoods. The final segmentation output of the system showed superior performance over the baseline method, namely applying MFCC features in the same Bayesian classification structure.

By adopting a methodology based on pattern recognition Piccina (2009) developed a system for locating mainly guitar solos in contemporary rock and pop music. Similar to the here-presented method a pre-trained model was applied sequentially to the audio data to assign, to each frame, the proper class label. A subsequent post-processing stage refines the raw classifier-based segmentation to obtain homogeneous segments. The author tested the system on 15 music pieces and reported, among other performance measures, a classification accuracy of 88% correctly assigned frames.

In a previous study we applied parts of the here-presented methodology for detecting solo sections in classical music (Fuhrmann et al., 2009b). We analysed a corpus consisting of excerpts taken from recordings of various concerti for solo instrument and orchestra and identified 5 relevant audio features to discriminate between the target categories. We then developed a segmentation and labelling algorithm which combines the output of a local change detection function with the frame-based decisions of a pre-trained SVM model. In this constrained scenario we could report acceptable results for the overall segmentation quality of the system, including a classification accuracy of almost 77% using an evaluation collection of 24 pieces.

Recently, Mauch et al. (2011) proposed a methodology combining timbre features with melodic descriptions of the analysed signal. The authors aimed at detecting both instrumental solo and voice activity sections from popular music tracks by combining 4 audio features. These features were extracted frame-wise and partially derived from a prior estimation of the predominant melody using the technique of Goto (2004), the statistical learning of the respective categories was further accomplished via a SVM-based HMM. The evaluation experiments, which applied a collection of

102 music tracks in a 5-Fold CV procedure, showed that a combination of all tested features is beneficial for the overall recognition performance. Moreover, compared to our results presented in the aforementioned study as well as in the forthcoming section of this chapter, a similar performance in terms of frame accuracy was reported.

5.1.3 Method

To derive a method for segmenting and labelling the input audio signal into the targeted categories we apply a simple model-based approach (e.g. Lu et al., 2003; Scheirer & Slaney, 1997). In particular, we make use of pre-trained classifiers which model the difference between *Solo* and *Ensemble* signals in terms of selected audio features. These models are sequentially applied to the input data and the resulting probabilistic output smoothed along time. We then binarize the resulting representation and further post-process it by applying additional filtering. This final binary sequence indicates the presence of a predominant source for each time frame.

As already mentioned above, the main assumption behind this approach implies that the relevant properties of the data can be encoded in certain descriptions of the audio signal. Hence, we first analyse our previously used large corpus of audio features (Section 4.2.1) to determine those features which best separate the training data in terms of the two categories *Solo* and *Ensemble*. We then use these selected features to train a statistical model using the training data.

Given these features we then construct an SVM classifier to model the decision boundary between the two classes in the audio feature space. First, we extract the features frame-wise from the raw audio signal of all instances in the training collection using a window size of 46 ms and an overlap of 50%. We then integrate the instantaneous and first difference values of these raw features along time using mean and variance statistics to derive a single feature vector for each audio instance. To determine the optimal parameter values for the SVM classifier a two-stage grid search procedure is applied as described in Section 4.2.3.3. Once the parameter values have been identified we train the model using the data from the training collection.

We then use this model to assign the labels *Solo* or *Ensemble* to each classification frame of an unknown music track. That is, we apply the model sequentially to the audio signal by using proper values for the size and the overlap of the consecutive classification frames. This framesize is defined by the results of the time scale experiment outlined below, hence 5 seconds of audio data, while the overlap is set to 20%. We smooth the obtained probabilistic output of the classifiers along time by applying a moving average filter of length l_{ma}, in order to remove short-term fluctuations in the time series of classifier decisions. This time series is subsequently converted into a binary representation by thresholding the values at 0.5, indicating, for each frame, the target categories. For post-processing we finally apply morphological filters of kernel length l_{mo} (Castleman, 1996) to promote longer sections while suppressing shorter ones. These filters have been previously applied for music processing (Lu et al., 2004; Ong et al., 2006). Figure 5.2 shows a schematic illustration of the processes involved in the presented algorithm.

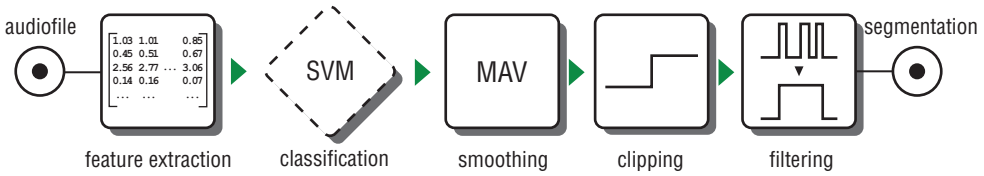


Figure 5.2: Block diagram of the presented track-level process for solo detection

5.1.4 Evaluation

In this section we evaluate the derived solo detection segmentation. We first describe the data used in the design and evaluation process of the presented method, which is followed by a section covering the most important parameters and their respective estimation. We then introduce the metrics applied for estimating the segmentation quality of the algorithm and subsequently asses the performance of the entire system.

5.1.4.1 Data

Here, we outline the data we collected for this research. In particular, we constructed two sets of data, one for training the statistical model, the other for evaluating the segmentation algorithm. It should be noted that no tracks have been used in both training and evaluation collection.

For the training collection we gathered 15-second excerpts from polytimbral music audio signals, containing either a single predominant source or an ensemble section. As already mentioned above we include the *singing Voice* in the corpus of solo sounds due to its common predominant character inside the mixture signal. Furthermore, the data account for various musical genres, hence maximising the generality and representativeness of the developed model. Since the overall goal is to apply the developed algorithm in conjunction with our label inference method, the model has to cope with a maximum variety in musical instruments and styles.

In total we accumulated around 500 excerpt for the *Ensemble* and more than 700 for the *Solo* category, where parts of these excerpts are taken from the training data of the pitched instrument recognition, described in Section 4.2.3. To avoid any bias towards one of the category we again always work with balanced datasets by randomly subsampling the category with the greater amount of instances to the level of the other one. To illustrate the diversity of this dataset, Figure 5.3 shows the distribution of the instances with respect to their musical genre. Moreover, Figure 5.4 depicts a tag cloud of the musical instruments contained in the *Solo* category of the collection.

We evaluate the presented method on entire pieces of music taken from classical, jazz, as well as rock and pop music. In total, we collected, respectively, 24, 20, and 20 musical compositions from the aforementioned musical genres, at which each piece is taken from a different recording. These tracks contain various predominant, i.e. solo instruments, and partially singing voice. We marked the start and end points of all respective sections of *Solo*, *Voice*, and *Ensemble* in these music pieces.

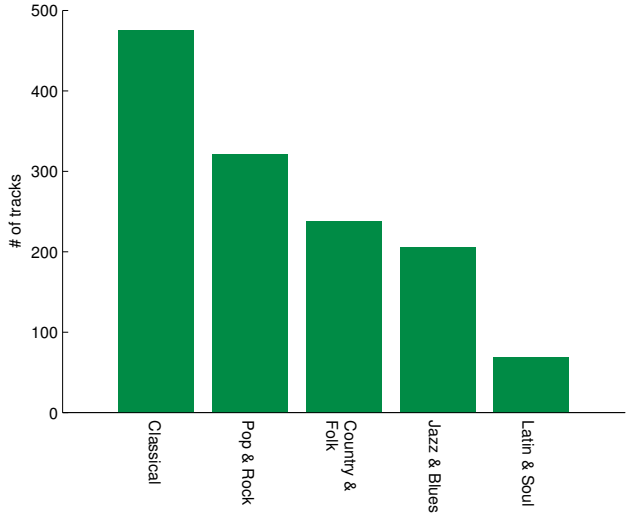


Figure 5.3: Genre distribution of all instances in the solo detection training collection.

Figure 5.4: Frequency of instruments in the *Solo* category of the collection used for training the solo detection model represented as a tag cloud.

5.1.4.2 Parameter estimation

In this section we describe the steps we have taken in the development of the solo detection model. Hence, we apply the typical pattern recognition scheme involving training and testing as described

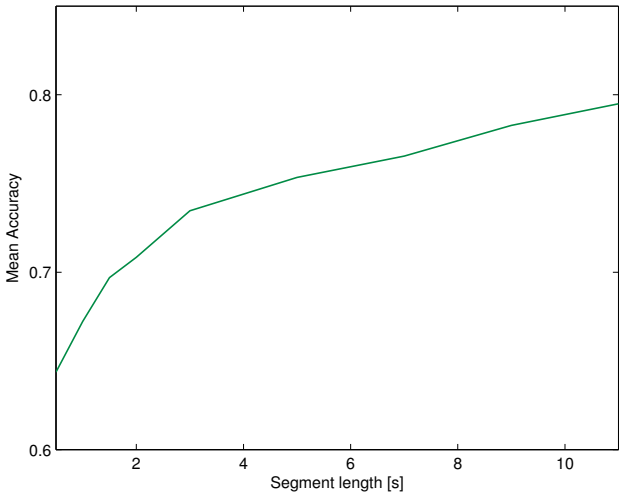


Figure 5.5: Time scale estimation for the solo detection model.

in Section 4.2.1. The here-presented methodology is therefore similar to the process of developing the instrument recognition models in the previous chapter.

Time scale. Similar to the instrument modelling the first step consists of identifying the optimal time scale the model uses to predict the respective labels. That is, we want to determine the optimal amount of data, which corresponds to the audio signal's length, from which a single prediction is performed. We therefore build multiple datasets, each exhibiting audio instances of a different length, taken from all excerpts in the training collection, and compare the average accuracies $\overline{A}$ resulting from a 10×10-Fold CV using standard parameter settings. Figure 5.5 shows the obtained graph, depicting the classification performance against the length of the audio instance. It can be seen that longer time scales are beneficial for the recognition accuracy of the model. However, to assure a reasonable temporal resolution of the final system, we chose the value of 5 seconds; it provides a trade-off between good recognition performance and acceptable temporal resolution of the final segmentation system.

It seems intuitive that the time scale to recognise pitched instruments and to determine solo activity exhibits the same order of magnitude (cf. Section 4.2.3.3). Here, a stronger evidence for the predominance of a given sound source, which increases with longer time scales, enables a more robust recognition. In this regard, longer time scales allow for more accurate sound source recognition.

Feature selection. Here we determine those out of our large set of audio features, which best discriminate the target classes. We therefore employ the same 10-Fold feature selection procedure as described in Section 4.2.1.3; Table 5.1 lists the resulting features. In total, the algorithm selects 30 features for modelling the data in the training collection. Contrastingly, in our previous work we identified 5 features when studying the same problem but focusing exclusively on data taken from classical music (Fuhrmann et al., 2009b). The here-observed excess in number of selected audio features indicates that the problem is far more complex across musical genres. Hence, the distinct recording and production styles employed in different musical genres complicate the extraction of

Feature	Statistic	Index
Pitch confidence	mean	–
Pitch confidence	var	–
Pitch confidence	dvar	–
Spectral crest	mean	–
Spectral spread	dmean	–
Barkbands	var	12
Barkbands	dvar	8
LPC	var	2
LPC	dvar	3
MFCC	mean	5, 9-11
MFCC	var	3-12
Spectral contrast	var	1-3
Spectral valleys	dmean	3
Spectral valleys	dvar	3,4
Tristimulus	var	1

Table 5.1: Selected features for the solo detection model. Legend for the statistics: mean (mean), variance (var), mean of difference (dmean), variance of difference (dvar).

a few significant characteristics that describe the acoustical and perceptual differences between the targeted categories.

As can be seen from Table 5.1 the feature describing the pitch strength takes a prominent role in the list. This seems intuitive since solo sections usually carry stronger pitch sensation than sections without predominant harmonic sources. Hence, the corresponding pitch is easier to extract when applying an estimator designed for monophonic processing. Consequentially, the corresponding confidence scores higher in sections containing predominant instruments. Moreover, the description of the spectral envelope is important due to the relative frequency of MFCC and spectral contrast and valleys features in the table. It seems that ensemble sections exhibit general differences in the spectral envelope than sections containing a soloing instrument that are encoded by these features. Remarkably here is the strong presence of the higher-order MFCC coefficients' variance – in total 10 coefficients – which may describe the existence of a stable spectral envelope in sections containing a predominant source. Furthermore, considering the results from the feature analysis in Section 4.2.3.5, the variance of the first difference of the 9^{th} Bark energy band (630 - 770 Hz, index 8!) seems to be primarily involved in the modelling of the *singing Voice*.

Classification. The statistical modelling part of the presented method is again realised via the SVM implementation provided by the LIBSVM library. For assessing the recognition performance of the solo detection model, we fist determine the optimal combination of classifier and kernel along with their respective parameters. Here, we follow the same 2-stage grid search process as described in Section 4.2.3.3 to estimate the best values for classifier and kernel type together with their relevant parameters. For illustration purpose, Figure 5.6 shows the parameter space spanned by the classifier's cost parameter C and the RBF kernel's parameter γ, and the resulting mean accuracy $\overline{A}$ on the entire dataset.

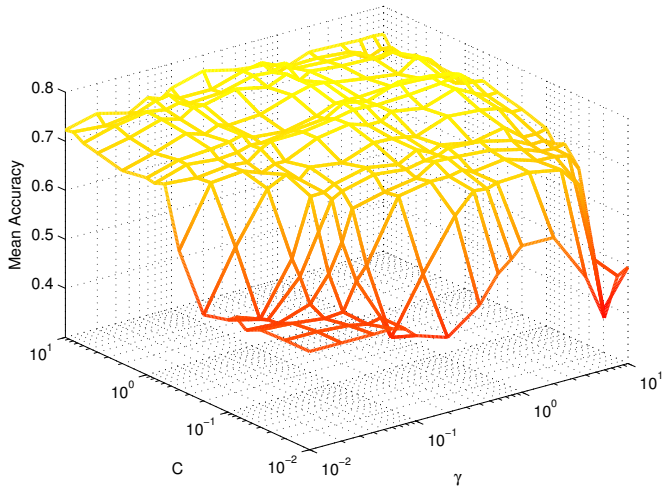


Figure 5.6: Accuracy of the solo detection model with respect to the SVM parameters. Here, the classifier's cost parameter C and the RBF kernel's γ are depicted.

$\overline{A}_{null}$	C4.5	NB	10NN	MLP	SVM*
50%	70.6%	70%	75.4%	74%	75.8%±0.86pp

Table 5.2: Recognition accuracy of the solo detection model in comparison to various other classification algorithms; a Decision Tree (C4.5), Naïve Bayes (NB), Nearest Neighbour (NN), and Artificial Neural Network (MLP). The asterisk denotes mean accuracy across 10 independent runs of 10-Fold CV.

We then estimate the classification performance of the trained model by evaluating the accuracy in a 10×10-Fold CV process. Additionally, we compare the obtained results to the performance of other classifiers typically found in related literature. Table 5.2 shows the results of all tested methods for the solo detection classification problem. As can be seen from the table the recognition accuracy of the presented SVM architecture scores around 75%, hence well above $\overline{A}_{null}$ but far from perfect, leaving a headroom for improvement. The performance of the nearest neighbour (10NN) and the neural network classification (MLP) can be regarded as equivalent to the SVM model, conceptually simpler approaches such as the decision tree and the Naïve Bayes however perform worse. Despite this moderate performance in recognition accuracy we believe that the output of the model, though not perfect, can be used in our instrument recognition framework by providing information regarding the acoustical and perceptual prominence of musical instruments in certain sections of a given composition.

5.1.4.3 Metrics

For a quantitative evaluation of the segmentation we use the notions of true and false positives respectively negatives, thus tp, fp, tn, and fn, on a frame basis. In particular, we apply the true positive rate tpr together with the true negative rate tnr,

$$tpr = \frac{tp}{tp + fn}, \quad \text{and} \quad tnr = \frac{tn}{tn + fp}. \tag{5.1}$$

These metrics account for the percentage of correctly assigned frames in each class, *Solo* and *Ensemble*, respectively. To avoid any bias towards one of the categories due to imbalances in the evaluation collection, we then use the arithmetic mean of the aforementioned to generate an overall measure of classification accuracy, i.e.

$$A_{\text{mean}} = \frac{tpr + tnr}{2}. \tag{5.2}$$

Additionally, we introduce the overall accuracy A_{tot} by considering the total number of correct frame predictions across categories.

For a qualitative assessment of the segmentation we furthermore introduce performance measures originating from image segmentation. In contrast to the aforementioned quantitative metrics these capture the segmentation quality of the system by evaluating the intersections of the output and the reference segments. Following Ortiz & Oliver (2006), we adapt measurement indices taking the correct grouping of frames, under-, and oversegmentation into account. Here, undersegmentation refers to the coverage of several ground truth segments by one single output segment. Accordingly, oversegmentation results from the splitting of a single ground-truth segment into several output segments. For qualitatively capturing these effects we first construct the overlapping area matrix (OAM) (Beauchemin & Thomson, 1997), using, respectively, the output of our algorithm and the ground-truth annotation. Every entry $C_{i,j}$ of this matrix contains the number of frames that the output segment j is contributing to the reference segment i. For perfect segmentation (i.e. same number of segments in reference and output segmentation and no over- and undersegmentation) the OAM contains non-null entries only on its diagonal, each representing the number of frames of the corresponding segment. In the case of segmentation errors non-null off-diagonal entries can be found, characterising the amount of error due to over- and undersegmentation. Then, $\sum_j C_{i,j}$ denotes the number of frames in the ground-truth segment i, and $\sum_i C_{i,j}$ is the number of frames in the output segment j. From this matrix we derive three evaluation indices, according to Ortiz & Oliver (2006):

Percentage of correctly grouped frames.

$$CG(p) = \frac{100}{n_{\text{t}}} \sum_{i=1}^{N_{\text{r}}} \sum_{j=1}^{N_{\text{o}}} cr(S_{\text{ref},i}, S_{\text{out},j}, p)\, C_{i,j} \quad [\%] \tag{5.3}$$

with

$$cr(S_{\text{ref},i}, S_{\text{out},j}, p) = \begin{cases} 1 & \text{if } \frac{C_{i,j}}{n(S_{\text{out},j})} \geq p, \\ 0 & \text{otherwise,} \end{cases} \tag{5.4}$$

and

$$n(S_{\text{out},j}) = \sum_{k=1}^{N_{\text{r}}} C_{k,j}, \tag{5.5}$$

where N_r and N_o denote the number of segments in the ground-truth and output segmentation, respectively, and n_t the total number of frames in the analysed audio. Furthermore, $S_{\mathrm{ref},i}$ refers to the reference segment i, $S_{\mathrm{out},j}$ to the output segment j, while p represents a penalty factor. Hence, CG accounts for those frames in a ground-truth segment $S_{\mathrm{ref},i}$, which are concentrated in a single output segment $S_{\mathrm{out},j}$. For perfect segmentation its value is 100% and any single frame error would reduce it dramatically. We therefore introduce the penalty factor p to relax the constraint of perfect segmentation to *nearly* perfect segmentation, where the term *nearly* depends on the value of p. The parameter thus represents the amount of segmentation error tolerated by the performance measures (a value of 1 indicates the most restrictive scenario).

Percentage of undersegmentation.

$$US(p) = \frac{100}{n_\mathrm{t}} \sum_{j=1}^{N_\mathrm{o}} \left(1 - ur(S_{\mathrm{out},j}, p)\right) n(S_{\mathrm{out},j}) \;\; [\%] \tag{5.6}$$

with

$$ur(S_{\mathrm{out},j}, p) = \begin{cases} 1 & \text{if } \frac{\max_{k=1,\ldots,N_\mathrm{r}}(C_{k,j})}{n(S_{\mathrm{out},j})} \geq p, \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

Thus, US represents the amount of frames, belonging to a single output segment $S_{\mathrm{out},j}$ while covering several segments of the ground truth $S_{\mathrm{ref},i}$. The penalty factor p is similarly introduced to tolerate a certain amount of output errors. Here, the function $ur(S_{\mathrm{out},j}, p)$ works over the columns of the OAM, taking those output segments $S_{\mathrm{out},j}$ into account which overlap with at least one reference region $S_{\mathrm{ref},i}$ is greater or equal than $p \times 100\%$.

Percentage of oversegmentation.

$$OS(p) = \frac{100}{n_\mathrm{t}} \sum_{i=1}^{N_\mathrm{r}} \left(1 - or(S_{\mathrm{ref},i}, p)\right) n(S_{\mathrm{ref},i}) \;\; [\%] \tag{5.8}$$

with

$$or(S_{\mathrm{ref},i}, p) = \begin{cases} 1 & \text{if } \frac{\max_{k=1,\ldots,N_\mathrm{o}}(C_{i,k})}{n(S_{\mathrm{ref},i})} \geq p, \\ 0 & \text{otherwise,} \end{cases} \tag{5.9}$$

and

$$n(S_{\mathrm{ref},i}) = \sum_{k=1}^{N_\mathrm{o}} C_{i,k}. \tag{5.10}$$

Hence, OS accounts for those output segments $S_{\mathrm{out},j}$ splitting a single ground-truth segment $S_{\mathrm{ref},i}$. The function $or(S_{\mathrm{ref},i}, p)$ works over the rows of the OAM, accounting for those rows, represented by the reference segments $S_{\mathrm{ref},i}$, exhibiting more than one non-null entry. These indicate the splits caused by the corresponding output segments $S_{\mathrm{out},j}$. Again, we introduce the penalty factor p, tolerating a certain amount of segmentation error.

Since these evaluation metrics derived from the OAM consider the overlap between output and reference segmentation, they capture the quality of the segmentation to a certain degree. Instead

of working on a frame basis, these metrics – unlike many others – act on a segment basis; first, those output segments meeting the specific criteria (correct grouping, under-, or oversegmentation) are marked as erroneous, and second, all frames of these segments accumulated and related to the total amount of frames in the analysed audio. Thus, the amount of error the segment contributes to the metric depends on its size. Furthermore, the incorporation of the penalty factor p allows to disregard small segmentation errors, which are common with this kind of problems.

In all our subsequent evaluation experiments we use a penalty factor of 0.8. This value was set ad hoc, mostly to relax constraints in the evaluation metrics and maximize its meaningfulness. Here, this specific value refers to the relaxed constraint that 80% of the data of the analysed segment has to meet the measure-specific requirements. Exemplified, a segment of 10 seconds length must agree in 8 of its seconds with the specific condition in order to be regarded as correct. The remaining 2 seconds represent an affordable error for many music audio description systems, and especially for automatic segmentation methods.

5.1.4.4 Results

Here, we assess the performance of the developed solo detection algorithm in segmenting the entire music pieces of the applied music collection with respect to the human-derived annotations. We evaluate the presented segmentation algorithm in a 3-Fold CV procedure, using, in each rotation, 2/3 of the data for testing and the corresponding 1/3 for performance estimation. During testing we perform a grid search in the relevant parameter space to determine the optimal values of the 2 parameters l_{ma} and l_{mo}. We thereby uniformly sample the parameters between 0 and 20 seconds, using a step size of 2 seconds. In this grid search, the performance of the system is estimated with the mean accuracy A_{mean}, hence averaging the performance on *Solo* and *Ensemble* sections[3]. As a result of the CV, all reported performance figures denote mean values across the respective folds.

Table 5.3 lists the evaluation metrics for the presented supervised segmentation algorithm. It can be seen that apart from the expected value for the total accuracy (76.6%), which is in line with the observed classifier accuracy in Table 5.2, the mean accuracy A_{mean} and especially the accuracy on the *Ensemble* sections, i.e. tnr, show a lower performance. Due to the imbalance in the dataset – note that the *Solo* category contains both instrumental solos and sections with *singing Voice* – the respective values of the two system parameters l_{ma} and l_{mo}, and accordingly the overall accuracy is biased towards tpr. This consequentially leads to a low value in the correct grouping of frames CG, since many *Ensemble* segments do not meet the requirement in Eq. (5.4), hence do not contribute their frames to the metric. Analogously, many short annotated *Ensemble* sections are likely to be covered entirely by predicted *Solo* sections, resulting in the relative high value of 57.9% of the US metric. Correspondingly, we observe a low value for the OS figures.

To emphasise the importance of the two system parameters l_{ma} and l_{mo}, representing, respectively, the length of the moving average filter and the length of the kernel of the morphological filter used for post-processing, Figure 5.7 shows the mean accuracy with respect to varying values of the afore-

[3]As a result of the imbalance of categories inside the evaluation collections, the best overall performance A_{tot} would result in an assignment of every frame with the label *Solo*. Thus, we use the average of individual class accuracy to estimate the performance of the system, avoiding any bias towards a particular category.

$\overline{tpr}$	$\overline{tnr}$	$\overline{A}_{\text{mean}}$	$\overline{A}_{\text{tot}}$	$\overline{CG}$	$\overline{US}$	$\overline{OS}$
87.7%	41.4%	70.1%	76.6%	40.4%	57.9%	15.1%

Table 5.3: Evaluation of the solo detection segmentation. The figures represent mean values resulting from the 3-Fold CV procedure.

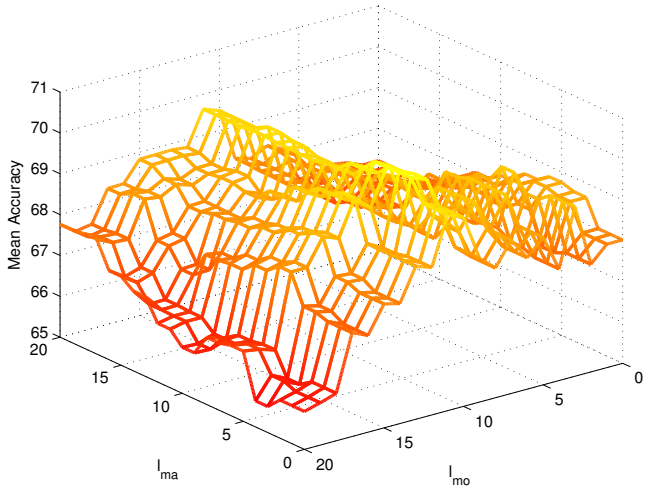


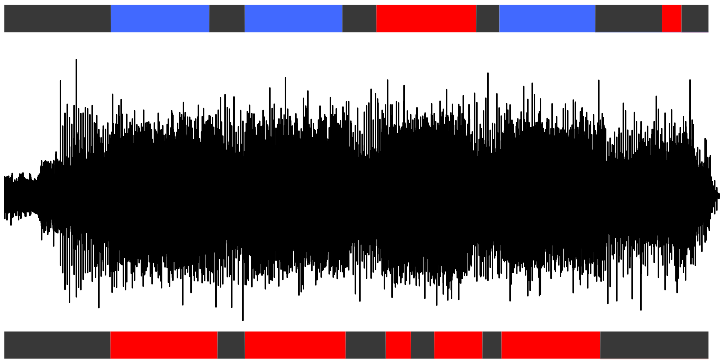
Figure 5.7: Frame recognition accuracy with respect to different parameter values. The y and x axis cover, respectively, the smoothing length l_{ma} and the filter kernel length l_{mo}.

mentioned. It can be seen that while l_{ma} exhibit only minor influence, the value of l_{mo} determines the system's segmentation accuracy, here a kernel length of 10 seconds leads to the best performance. It should be noted that the choice of the metric used to evaluate the system's performance heavily influences the optimal values for the two parameters. Hence, depending on this metric the location of the peak performance in the parameters' value space may vary to a great extent.

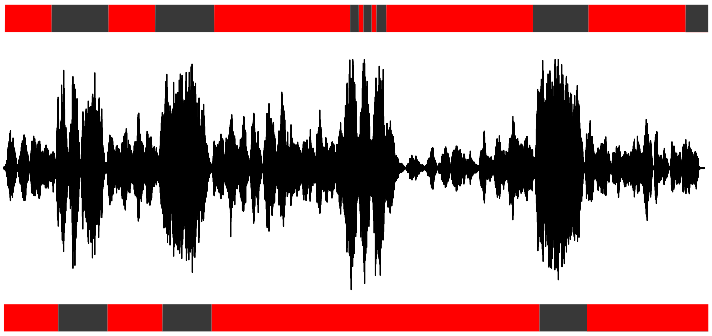
5.1.4.5 Error analysis

Here, we perform a qualitative analysis of the segmentation output by perceptually evaluating the resulting partition of all pieces in the used collection. First, the overall impression of the segmentation output's quality is that the system fulfils its prospects by performing the task with a subjective good performance. However, several regularities can be observed which we shortly outline in detail.

Given the nature of the task – a technological inspired implementation of a musical concept – we observe several ambiguities in both the manual annotations and the output of the segmentation algorithm. For instance, many ground truth ensemble sections exhibit one or several predominant instruments which are therefore labelled with *Solo*. Hence, the mostly subjective decision of classifying a certain musical section into *Solo* or *Ensemble* is not only based on the presence of a single predominant musical instrument; it rather involves higher-level contextual information. Applying only low-level information sources cannot cope with this problem, thus we have to accept a certain upper bound in the segmentation performance of the presented system.



(a) Deep Purple - Smoke on the Water.



(b) W. A. Mozart - Clarinet concerto in A major - II. Adagio.

Figure 5.8: Two examples of the solo detection segmentation. Part (a) depicts a Rock piece while part (b) shows a concert for solo instrument with orchestral accompaniment. The top bar in each figure represents the ground truth annotation, the lower bar the segmentation output of the presented algorithm. The colours red, grey, and blue denote, respectively, the categories of *Solo*, *Ensemble*, and *singing Voice*. Note that the segmentation output only contains the *Solo* and *Ensemble* classes, where a section containing *singing Voice* is regarded as a *Solo* section.

Moreover, certain playing styles often lead to ambiguous values of the selected features, partially producing wrong predictions; the system often recognises unison sections, where the same melody is played by several instruments, as a *Solo* section. Here, the predominance of a single pitch may bias the decision towards the *Solo* class. Next, Brass sections, which often consists of unison lines, show a similar behaviour; it mainly depends on the mixture of the involved instruments whether the section is classified as *Solo* or *Ensemble*. Finally, we observe that the employment of heavily-distorted background instruments hinders the recognition of *singing Voice* or *Solo* sections.

Furthermore, we want to note that many pop, rock, and jazz pieces hardly contain any *Ensemble* sections, in case of regarding sections containing *singing Voice* as *Solo*. This, on the one hand, accounts for the imbalance of the target categories in the evaluation collection. On the other hand, the fact that most of the instrumental information exhibits predominant character partially confirms the 1st hypothesis we stated in the beginning of this chapter, i.e. given a music piece, most of the involved instruments appear at any time in a predominant manner. We will come back to this issue in the second part of this chapter. For illustration purpose, Figure 5.8 shows two examples of the derived segmentation with respect to the annotated ground truth.

5.1.5 Discussion

In this section we presented a knowledge-based algorithm for segmenting a given piece of music into parts containing predominant instrumental information. The method uses a trained model to assign, to each analysis frame, the label *Solo* or *Ensemble*, indicating the presence of a single predominant source. For capturing the intrinsic properties of the aforementioned categories the algorithm applies selected audio features describing spectral and pitch-related properties of the signal. We then evaluated the presented method on a specifically designed dataset in both a quantitative and qualitative manner.

The figures presented in the preceding sections, assessing the performance of the solo detection model itself as well as the overall segmentation system, show acceptable performances with respect to the corresponding null models. The fact that the segmentation output deviates from perfect to a certain extent illustrates the complexity of the addressed task. Anyhow, these figures seem reasonable given the nature of the studied problem; as already mentioned above, the applied definitions of the underlying musical concepts (*Solo*) are quite loose, hence leaving a great margin for (subjective) interpretation, and are only partially represented by the here-employed description in terms of low-level audio features. In addition, genre-related divergences in the target concept of a musical solo complicate the development of a generalising model. Due to the different adoption of soloing instruments in the respective genres and the evident differences in the recording, mixing, and mastering processes, the targeted concepts exhibit obviously different descriptions in terms of audio features across musical genres. Here we speculate that by relaxing the aforementioned generality claims to the model better performance can be achieved. For instance, a genre-dependent parameter selection could already improve the segmentation quality, since the post-processing filter could be adapted to the specific distribution of *Solo* and *Ensemble* sections in the respective genres.

In general, we hypothesise that the employed features are not fully able to describe the targeted concepts, hence representing the main shortcoming of the method. Many short sections exhibiting predominant instrumental combinations are labelled *Solo* by the presented supervised segmentation, which do not fall inside the applied definition of a *Solo*. Hence, the selected spectral and pitch related descriptions of the audio signal partially carry ambiguous information related to the class assignment. Here, descriptions of higher-level melodic aspects of the music can exhibit complementary information and help improving the performance of the system. Since the existence of a consistent melody, played by a single instrument, is a perceptual key property of a *Solo* section, such features should improve the robustness of the system by avoiding both spurious *Solo* and *Ensemble* sections.

Nevertheless, this imperfect output of the segmentation algorithm can be used in the developed instrument recognition framework. The conception behind the presented approach is to locate sections inside a music piece where a single instrument is predominating, in order to improve the robustness of the subsequent instrument recognition. The label inference is then applied to each of the selected segments and the resulting labels merged (see Figure 5.1). Since the aforesaid label inference method should be able to deal with sections not exhibiting predominant character, slight segmentation errors should not affect the performance of the label extraction to a great extent. Moreover, the implemented contextual analysis can compensate for inconsistencies in the segmentation output.

5.2 Sub-track sampling – agnostic approaches

In this section we develop knowledge-free methods to select relevant segments in terms of the instrumentation from an entire music piece. In general, the approaches presented here do not consider the constraints we have introduced in the design process of the instrument recognition models. The resulting output data is rather selected in terms of its representativeness with respect to the overall instrumentation of the music piece. This implies that the subsequent label inference works on any data regardless its complexity. In the algorithms' design process we additionally consider the trade-off between recognition performance in terms of musical instruments and the amount of data processed by the system. The overall aim is to guide the label inference stage with information on where and how often the models have to be applied given the piece under analysis, in order to provide a robust estimate of the piece's instrumentation while keeping the computational costs low. We thereby apply the concepts of local versus global processing of the data; here the ideal combination of localised extraction of the instrumental informations leads to a full description at the global scope, i.e. the instrumentation of the entire track.

In this part of the chapter we consider several approaches which apply the aforementioned concept of extrapolation of locally extracted information to the global scope. We compare their properties in terms of data coverage and musical plausibility, and further evaluate their peculiar functionalities. In the subsequent part of the chapter we then employ these approaches – among others – in the instrument recognition framework and compare the effects of the respective specificities on the recognition performance. Parts of the here-presented have been published by Fuhrmann & Herrera (2011).

5.2.1 Related work

The methods presented along this section partially incorporate information regarding structural aspects of the analysed music piece. Extracting the structure of a musical composition is a research field on its own, hence a review of the related literature goes beyond the here-presented. We therefore refer the interested reader to the recently published comprehensive state of the art overview of Paulus et al. (2010). However, musical structure has been frequently applied in conjunction with several other problems of MIR research. In this context, such works include audio fingerprinting (Levy et al., 2006), music similarity estimation (Aucouturier et al., 2005), cover song detection (Gómez et al., 2006), loop location (Streich & Ong, 2008), or chord analysis (Mauch et al., 2009), to name just a few, all of them using the repetitiveness of the musical structure as a cue for approaching their specific problem.

In general, two distinct methodologies towards the estimation of the musical structure can be identified. The first one evaluates frame-to-frame distances in terms of a pair-wise similarity matrix, from which repeating sequences of events are extracted. Foote (2000) introduced this technique which has been used extensively in music structure research. The second class of approaches towards the extraction of musical structure and its inherent repetitiveness estimates sections inside a given music

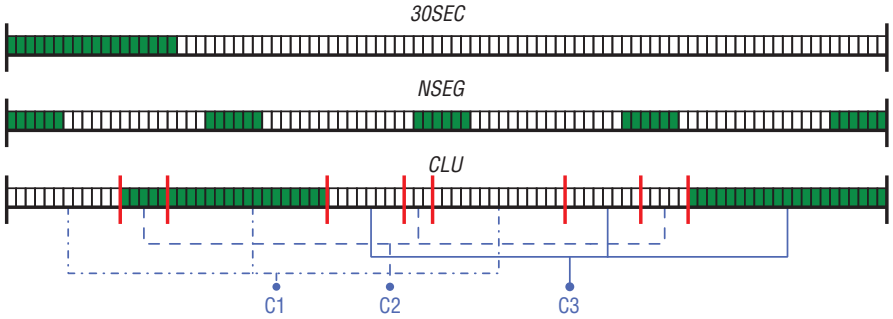


Figure 5.9: Conceptual illustration of the presented agnostic track-level approaches; the green filled frames denote the respective algorithm's output data. Segmentation (red) and clustering (blue) are indicated for the *CLU* method, while *NSEG* applies the values of $l_{\mathrm{N}} = 10$ (sec) and $n_{\mathrm{N}} = 5$. See text for details on the functionalities of the approaches.

piece wherein a certain parameter of interest, e.g. timbre, exhibits rather stable characteristics. This corresponds to the detection of relevant change points in the temporal evolution of the parameter under consideration. Hence, such systems apply change detection algorithms such as the Bayesian Information Criterion (BIC) (Chen & Gopalakrishnan, 1998) or evaluate local frame-to-frame distances (e.g. Tzanetakis & Cook, 1999) to determine structural segments with a stable value of the relevant parameter. To estimate the repetitions inside the musical structure, typical approaches then use clustering techniques such as k-means to group the detected segments with respect to some relevant parameters.

In this regard, our subsequently presented musical structure analysis follows the latter of the aforementioned approaches. Here, we aim at detecting sections of persistent instrumentation and their repetitions inside a given piece of music, hence applying a timbral change analysis in conjunction with a hierarchical clustering analysis.

5.2.2 Approaches

In this section we present 3 conceptually different approaches, which output segments representative for the instrumentation of the analysed track, to pre-process an entire piece of music for label inference. Since the instrumentation and its temporal evolution of a piece of music usually follows a clear structural scheme, we expect, inside a given music track, a certain degree of repetitiveness of its different instrumentations. The described methods exploit this property of music and the resulting redundancy to reduce the amount of data to process. In short, the presented approaches are accounting – some of them more than others – for the time-varying character of instrumentation inside a music piece. Figure 5.9 depicts the underlying ideas.

5.2.2.1 30 seconds (*30SEC*)

This widely used approach assumes that most of the information is already accessible within a time scale of 30 seconds. Many genre, mood, or artist classification systems use an excerpt of this length

to represent an entire music track (e.g. Laurier et al., 2010; Scaringella et al., 2006; Tzanetakis & Cook, 2002). The process can be regarded as an extrapolation of the information obtained from these 30 seconds to the global scope, i.e. the entire piece of music. Since the aforementioned semantic concepts are rather stable across one single piece, the data reduction seems not to affect the significance of the obtained classification results. Instrumentations, however, usually change with time, thus we expect the instrumentation of the entire piece to be poorly represented by the information covered by this approach. In our experiments we extracted the data from 0 to 30 seconds of the track.

5.2.2.2 Segment sampling (*NSEG*)

To extend the previous approach towards an incorporation of the time-varying characteristics of instrumentation, we sample the track uniformly without using knowledge about the actual distribution of musical instruments inside. This enables a distributed local extraction of the information which is combined to a global estimate of the instrumental labels. In particular we extract n_N excerpts of l_N seconds length, at which we take a single segment from the beginning for $n_N = 1$, or one segment from the beginning and another from the end of the music track for a value of 2. For $n_\mathrm{N} > 2$ we always take the segments from the beginning and the end and select the remaining $n_\mathrm{N} - 2$ segments from equal distant locations inside the piece. The parameters n_N and l_N are kept variable for the experiments to be conducted in Section 5.3.

5.2.2.3 Cluster representation (CLU)

Certainly the most elaborated approach from the perceptual point-of-view; we represent a given piece of music with a cluster structure, at which each cluster corresponds to a different instrumentation. In general, composers use timbral recurrences, along with other cues, to create the musical form of the piece (Patel, 2007), serving to guide listeners' expectations by establishing predictability or creating surprise (Huron, 2006). The here-developed structure representation is thought to reflect the overall instrumental form of the piece where sections containing the same instruments group together.

In particular, the presented approach applies unsupervised segmentation and clustering algorithms to locate the different instrumentations and their repetitions. At the end, only one segment per cluster is taken for further analysis. Hence, this approach directly takes advantage of the repetitions in the instrumental structure to reduce the amount of data to process, while the local continuity of the individual instruments is preserved to guarantee a maximum in instrument recognition performance. Moreover, it explicitly uses an estimate of the global distribution of the musical instruments to locally infer the labels from a reduced set of the data by exploiting redundancies among the instrumentations in the piece of music. Finally, the method passes the longest segment of each resulting cluster to the label inference algorithm. Figure 5.10 shows the schematic layout of this approach.

Segmentation. As a first step, the algorithm applies unsupervised change detection to the entire music track to detect changes in the instrumentation. Since the instrumentation is directly linked to timbre, we use MFCCs to represent it in a compact way. The features are extracted frame-wise

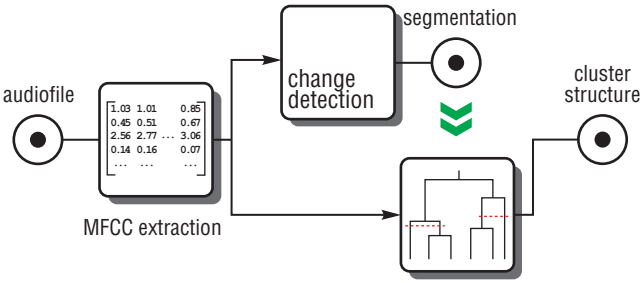


Figure 5.10: Block diagram of the *CLU* approach. The method applies unsupervised segmentation and clustering to represent a given music piece by a cluster structure, where each cluster ideally contains one of the different instrumentations of the piece. In doing so the algorithm exploits the redundancy inherent to the instrumental form of the music.

and analysed to detect local changes in their values (we again use a frame size of 46 ms with a 50% overlap). A segmentation algorithm based on the Bayesian Information Criterion (BIC) processes these data to find local changes in the features' time series. Borrowed from model selection (Schwarz, 1978), a texture window is shifted along the data in order to find the desired changes within the local feature context. Therein the hypothesis is tested whether one model covering the entire window or two models of two sub-parts of it, divided by a corresponding change point, better fit the observed data[4]. If the latter hypothesis is confirmed an optimal change point is estimated. In particular, given N the sample size and Σ the estimated covariance matrix, with indices 0,1, and 2 representing, respectively, the entire, first, and second part of the window, the algorithm uses the likelihood ratio test

$$D(i) = N_0 \log|\Sigma_0| - N_1 \log|\Sigma_1| - N_2 \log|\Sigma_2|. \tag{5.11}$$

to compare the hypotheses at split point i. The BIC value is then estimated as follows,

$$BIC(i) = D(i) - \lambda \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N_0, \tag{5.12}$$

where λ denotes a penalty weight and d the dimensionality of the data. If $BIC(i) > 0$ the data is better explained by two different models of the distribution while $BIC(i) \leq 0$ indicates a better modelling by a single distribution. In the former case, the optimal change point is found by the maximum value of $BIC(i)$ (see the works by Chen & Gopalakrishnan (1998) and Janer (2007) for details on the implementation).

Clustering. In order to group the resulting segments with respect to their instrumentations we employ hierarchical clustering (HC) techniques to find their repetitions inside a given music track. To represent the timbral content of a segment the system again applies the frame-wise extracted MFCCs. We calculate the pair-wise distance matrix between all segments of the music piece by

[4]Here we fit the respective data to a single Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, with μ and Σ representing, respectively, the mean vector and the covariance matrix.

computing the symmetric KL divergence between the Gaussian distributions $\mathcal{N}(\mu, \sigma)$, with μ denoting the mean vector and σ representing the vector containing the standard deviations, of the respective MFCC frame vectors. An agglomerative HC then groups the segments using the generated distance matrix (Xu & Wunsch, 2008). The segments are merged iteratively according to these distances to form a hierarchical cluster tree, a so-called dendrogram, where a specific linkage method further measures proximities between groups of segments at higher levels. In particular, we tested average (UPGMA), complete (CL) and single (SL), i.e. furthest and shortest distance, and weighted average (WPGMA) linkage. The final clusters are then found by pruning the tree according to an inconsistency coefficient, which measures the compactness of each link in the dendrogram. The algorithm thereby applies a threshold c_i to determine the maximum scatter of the data inside the respective branch of the structure. We used the implementation provided by Matlab's statistics toolbox[5].

5.2.3 Evaluation

In this section we evaluate the performance of the presented track-level approaches in terms of their peculiar functionalities. Since only the *CLU* method exhibits algorithmic properties to evaluate, the following covers the experiments for assessing the performance of this particular approach.

5.2.3.1 Data

For evaluating the *CLU* method's segmentation and clustering steps we use the data described in Section 4.3.2. Due to the annotations' character these data contain both ground-truth change points in the instrumentation and the resulting segments labelled with the respective instrumental combination. Accordingly, we use the former for evaluating the segmentation quality of the presented algorithm and the latter for assessing the performance of the HC.

A short analysis of the nature of the different instrumentations inside the data collection shows a mean value of 7.9 different instrumentations along with 14.7 annotated segments per track on average, which indicates that already about 50% of the data explain all observed instrumental combinations. Moreover, instrumentations usually do not change abruptly along a music piece; there rather exists a particular set of musical instruments which determines the piece's main timbral characteristics, hence being active most of the time in the track. That is, we expect, at a given instrumentation change point, a change in only few of the involved instrument. From this follows that most of the aforementioned identified combinations are subsets of others, which implies that there exist an even higher degree of redundancy in terms of individual musical instruments. This confirms the 2$^{\text{nd}}$ hypothesis stated in the beginning of this chapter, concerning the repetitive nature of instrumentations and the resulting redundancy of individual musical instruments in a musical composition.

[5] http://www.mathworks.com/products/statistics/

5.2.3.2 Metrics

We estimate the performance of the segmentation algorithm by using the standard metrics of precision, recall, and F-score (P, R, and F, see Section 4.2.2), usually found in related works. In particular, we regarded a computed change point as correct if its absolute difference from the next annotated one is not greater than one second. Given the nature of the task, i.e. segmenting music into parts of consistent instrumentations, this value seems to be appropriate since segment boundaries are often blurred and cannot be assigned perceptually to a single time instance due to the overlap of the instrumental sounds starting and ending in this particular point. However, even a greater value could be accepted, although we did not want to overestimate the performance of the algorithm. In addition, we use the metrics accounting for the segmentation quality, derived from the OAM matrix, as introduced in Section 5.1.4. We note that the evaluated segmentation is performed on a timbral basis, which is not always reflecting the instrumentation. If the timbre of the same instrumentation changes, the algorithm produces a change point which is not reflected in the ground truth annotation. Consider, for instance, an *electric Guitar* in an accompaniment and solo context, where the timbre of the instrument may exhibit strong differences. The same can apply for the *singing Voice* in verse and chorus sections. Consequently, there exists an upper bound in the performance estimation of the algorithm, which is difficult to assess in view of the aforementioned. However, this bias is reflected in all parameters to evaluate, thus enabling a qualitative comparison.

For evaluating the performance of the HC stage we relate its output to the reference data from the respective annotation. Hence, we input the audio segments taken from the annotation to the algorithm and compare the resulting grouping to the ground truth segment labels. We thereby avoid a propagation of the errors introduced by the segmentation algorithm into the evaluation of the clustering quality. All reference segment boundaries with a mutual distance in time of less than one second are merged to a single time instance in order to ensure the representativeness of the distance estimation. We then assess the clustering quality by computing the normalised Hubert's statistic $\hat{\Gamma}$ (Xu & Wunsch, 2008), which generally measures the correlation of two independently drawn matrices. Given $\mathcal{C}$ and $\mathcal{G}$, denoting, respectively, the generated cluster structure and ground-truth derived grouping of a given track $\mathcal{X} = \{x_i\}, i = 1 \dots N$, consisting of N segments, we accumulate, for all pairs of segments (x_i, x_j), the number of pairs falling into the same cluster for both $\mathcal{C}$ and $\mathcal{G}$ (a), the number of pairs clustered into the same cluster but belonging to different reference groups (b), the number of pairs falling into different clusters in $\mathcal{C}$ but having the same reference group (c), and finally the number of pairs which neither belong to the same cluster in $\mathcal{C}$ nor $\mathcal{G}$ (d). We can then write $\hat{\Gamma}$ as

$$\hat{\Gamma} = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}}, \tag{5.13}$$

with $M = a + b + c + d$, $m_1 = a + b$, and $m_2 = a + c$, resulting in a correlation of $\mathcal{C}$ and $\mathcal{G}$ with a value between 0 and 1.

The metric considers all pairs of instances in both the reference and algorithmically derived tree and relates their respective distributions. This leads to an objective assessment of clustering quality by directly comparing the generated data representation to the ground truth. However, as already

stated above, timbral changes do not always correspond to changes in the instrumentation. Hence, the clustering algorithm generates a data representation solely relying on timbral qualities, while the reference clusters are built upon the respective annotated instances of the musical instruments. Thus, we have to reckon a similar upper bound in performance as discussed above.

5.2.3.3 Methodology

We perform all evaluation experiments in the 3-Fold CV process as similarly applied in the previous evaluation sections. That is, for each CV rotation, we use 2/3 of the data for estimating the proper parameter values by performing a grid search while reserving the remaining 1/3 for performance estimation. In particular, we estimate the optimal values for the BIC segmentation's parameters ws_{BIC}, ss_{BIC}, and λ_{BIC} denoting, respectively, the size of the analysis window, the increment of the change point hypothesis inside this window, and the penalty term in Eq. (5.12) along with the linkage method and the inconsistency threshold c_i for the HC[6].

5.2.3.4 Results

This section covers the results related with the performance evaluation of the *CLU* method's particular algorithms. We perform the quantitative and qualitative evaluation of segmentation and clustering performance separately, hence dividing the following into two subsections.

Segmentation. Table 5.4 lists the evaluation metrics for the BIC segmentation algorithm, depicted as mean values across the folds of the CV. As can be seen from the table, the algorithm is working far from perfect, but is performing comparably to state-of-the-art approaches on related problems (e.g. Fuhrmann et al., 2009b; Goto, 2006; Ong et al., 2006; Turnbull et al., 2007). The problem at hand is even more complex compared to the aforementioned references in terms of the variety of the input data which requires the algorithm to operate on all kinds of musical genres.

Evaluation of the optimal parameter values indicates that a size between 5 and 10 seconds for the BIC analysis window ws_{BIC} and a step size ss_{BIC} of the change point hypothesis of approximately 1 second effects in the best performance across folds. Regarding the penalty term λ_{BIC}, the highest tested value of 5 shows best performance in all 3 folds, suggesting that a high value of precision, i.e. few false positives, leads to the best performance evaluation of the segmentation output in terms of the F-score F. Hence, true change points show in general higher values in the log-likelihood function D (Eq. (5.11)) than spurious ones.

By performing a subjective, i.e. perceptual, analysis of the segmentation output we observe that if the timbre of the music track under analysis is compact, i.e. not subject to fluctuations in dynamics on a short time scale, a good segmentation result is obtained. Moreover, clear changes in timbre (e.g. starting of *singing Voice*) are perfectly hit, while small changes in instrumentation (e.g. starting of a background *String* section) are obviously more problematic. In terms of musical genres good performance is achieved for rock, pop, jazz, and electronic music, especially with tracks that show

[6]During the grid search, we evaluate the segmentation performance with the F-score F and the clustering quality using $\hat{\Gamma}$.

$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{CG}$	$\overline{US}$	$\overline{OS}$
0.4	0.54	0.43	77.6%	20.5%	55.7%

Table 5.4: Evaluation metrics for the *CLU*'s segmentation algorithm.

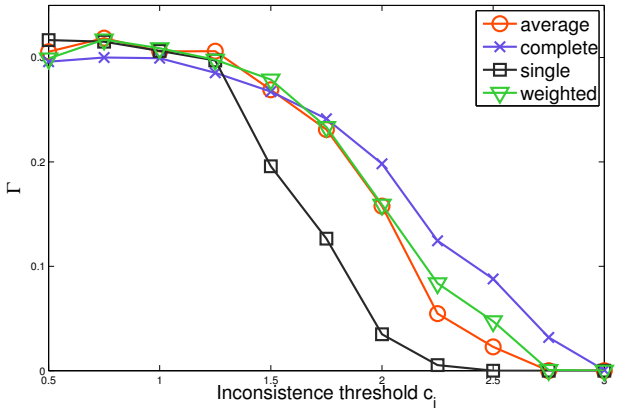


Figure 5.11: Performance of different linkage methods used in the hierarchical clustering algorithm. The figure shows the resulting $\hat{\Gamma}$ values against the inconsistency threshold c_i for the whole dataset.

clear timbral changes in its structure. On the other hand, the algorithm fails on segmenting classical music properly. This may results from the aforementioned fluctuations in dynamics and the thereby caused small changes in timbre.

Additionally, we observed problems with fade and crescendo sections, and sound textures (e.g. *String* sections), which often behave similar, as the algorithm either misses the instrumental change at all or creates a false positive based on the change in volume of the respective sound source. This points towards a more general evaluation problem of this segmentation task; the overlapping instrumental sounds in the transition between different instrumentations can last up to several seconds, which poses difficulties in reflecting the instrumental change in the respective ground truth annotation. In this case, deviations of the generated change point from the annotated one of several seconds may be advisable, whereas other instrumental changes only require several hundreds of milliseconds as parameter value. Nevertheless, the algorithm generally provides a useful segmentation output since it always produces a couple of segments consistent in instrumentation with acceptable length, i.e. greater then 10 seconds, which can be used in the subsequent labelling stage.

Clustering. Figure 5.11 shows the obtained performance estimation of the considered clustering algorithms with respect to the used inconsistent threshold c_i over the entire evaluation collection. From theses results we can identify an optimal parameter range for the threshold; between values of 0.5 and 1.25 the specific linkage method seems not to be decisive, i.e. all methods produce very similar cluster structures. In general, small values of c_i generate a greater number of clusters which better reflects the ground truth grouping. The same behaviour can be observed when assessing the best parameter values estimated for each training set in the rotations of the 3-Fold CV.

Qualitatively, we conclude in a similar manner as in the evaluation of the BIC segmentation output; tracks exhibiting a consistent timbre along time tend to group better with respect to the annotated ground truth. On the other hand, confusions in the cluster assignments often arise for music pieces with heavy fluctuations in the dynamics of the music. Hence, similar to the segmentation evaluation, tracks from the Pop, Rock, Jazz, and Electronic genres show better figures in the clustering performance estimation metrics presented above. This all let us speculate that, different than being the respective algorithms responsible for the observed imperfect performance, it is rather that the applied representation, used to encode the general acoustical properties, is not capturing the desired information. Hence, the MFCC features seem to encode the timbre of the analysed music signal properly for music exhibiting a consistent short-term timbre (e.g. rock music with heavy percussive elements exhibits these characteristics since the predominant sound components introduced by these instruments show very consistent timbral evolutions), while the features fail at describing the overall timbral sensation for pieces containing inconstancies in the dynamics such as classical music. However, a much deeper analysis of the correlations between timbral encodings and the different musical genres' acoustical properties would be needed to derive stronger evidence for the aforesaid hypothesis.

5.2.4 Discussion

This section covered the basics of three unsupervised approaches for partitioning a given music piece into relevant segments in terms of the piece's global instrumentation. Two of them use a sampling heuristic to derive these segments, hence they do not incorporate any information regarding the underlying structure of the track. The third approach performs a timbral analysis to group parts of the given musical composition with respect to the different instrumentations therein. Since the former two do not exhibit any algorithmically parameters to evaluate, only the latter is evaluated in terms of its performance in segmenting a piece according to timbral changes therein as well as its ability to group segments of the same instrumentation into corresponding clusters.

The performed quantitative and qualitative evaluation suggests that the algorithm fulfils the requirements and groups the different instrumentations of a given music piece consistently into a musically reasonable structure. However, like in many other automatic music analysis approaches, the method is upper-bounded; it seems that the segmentation and clustering approach is limited by the underlying encoding of timbre. Subjective evaluation suggests that for certain types of music, e.g. classical music, the applied MFCC features seem to fail in modelling the desired timbral properties. However, tracks from other musical genres exhibiting more consistent timbral sensations show a very good performance in terms of both segmentation and clustering quality.

In the following section we now apply all presented track-level analysis methods as front-end processing for automatic instrument recognition. We will then be able to estimate the influence of the respective conceptual and musical characteristics of the approaches on the recognition performance and the amount of data needed to maximise it.

5.3 Application to automatic musical instrument recognition

In this section we apply the track-level approaches described above to the task of automatic musical instrument recognition. That is, a particular track-level approach acts as pre-processing stage for the actual label extraction in our recognition framework; it outputs a set of segments from which the label inference method described in Chapter 4 determines labels related to the instrumentation of the respective segment. We then combine the label output of all segments of a particular approach to form the final labels for the track under analysis (see Figure 5.1). By using this kind of pre-processing we are either able to select specific excerpts of the analysed piece which enable a more reliable label inference, or, depending on the respective approach, exploit the inherent redundancies in the structure of the track to reduce the amount of data used for extracting the labels.

In the evaluation of the different systems, we perform a quantitative estimation of the system's performance in terms of recognition accuracy as well as its efficiency. This will lead to an assessment of the minimal amount of data needed to maximise the recognition accuracy of our label inference method.

5.3.1 Data

In the experiments we evaluate all approaches using the music collection and corresponding annotations described in Section 4.3.2, as already applied for evaluating the *CLU* method in Section 5.2.3. Here, we merge all annotated musical instruments of a particular track to represent the ground truth for its overall instrumentation. More details about this collection can be found in Section 4.3.2.

5.3.2 Methodology

In order to provide a robust estimate of the methods' performance with respect to the parameters to evaluate, we again perform all our experiments in the 3-Fold CV framework. Hence, for each rotation we use the data of 2 folds for estimating the optimal parameter settings and subsequently test on the remaining fold. We then obtain mean values and corresponding standard deviations by averaging the evaluation results of the respective predictions of all three runs. Parameter estimation itself is performed in a grid search procedure over the relevant parameter space. For each of the studied approaches described in this chapter the parameters are evaluated separately to guarantee maximal comparativeness of the respective results. In all conducted experiments we apply the CT labelling method as described in Section 4.3.3, at which the method's specific parameters are determined via the aforementioned grid search.

metric	Ref_{prior}	*30SEC*	$3SEG_{10}$	$3SEG_{20}$	$6SEG_{10}$	$6SEG_{20}$	*CLU*	*SOLO*	Ref_{all}
P_{micro}	0.4	0.61	0.64	0.63	0.61	0.6	0.63	0.71	0.65
R_{micro}	0.4	0.49	0.59	0.66	0.72	0.78	0.73	0.64	0.73
F_{micro}	0.4	0.55	0.61	0.64	0.66	0.67	0.68	0.67	0.69
F_{macro}	0.26	0.43	0.47	0.49	0.51	0.53	0.54	0.53	0.54
data	–	0.12	0.12	0.25	0.25	0.5	0.66	0.62	1

Table 5.5: Labelling performance estimation applying the different track-level approaches.

5.3.3 Metrics and baselines

We evaluate the labelling performance of the presented approaches using the same metrics as introduced in Section 4.3.4.3. That is, we apply the precision and recall metrics P_{micro} and R_{micro}, as well as the F-scores F_{micro} and F_{macro}, working, respectively, on the instance and category level.

To establish an upper performance bound for the track-level approaches we introduce the Ref_{all} system; by processing all frames with the presented label inference method we perform a global analysis of the instrumentation of the track. However, no data reduction is obtained with this approach. Since the method uses all data available it acts as an upper baseline both in terms of recognition performance and amount of data processed, which all other methods using less data compete with. Furthermore, we generate a lower bound by drawing each label from its respective prior binomial distribution, inferred from all tracks of the collection, averaging the resulting performance over 100 independent runs (Ref_{prior}).

5.3.4 Labelling results

The upper part of Table 5.5 contains the results (mean values) of the applied metrics in the CV obtained for all studied algorithms. In particular, we generate 4 different systems from the *NSEG* concept additionally to the *30SEC*, *CLU*, and *SOLO* approaches, and the two baselines Ref_{prior} and Ref_{all}; by setting n_N and l_N, respectively, to 3 and 6, and accordingly 10 and 20 seconds we synthesise the systems $3SEG_{10}$, $3SEG_{20}$, $6SEG_{10}$, and $6SEG_{20}$. Additionally, figures regarding the relative, with respect to the all-frame processing algorithm Ref_{all}, amount of data used for label inference are shown in the lower panel.

The figures presented in Table 5.5 show that all considered approaches outperform the lower baseline Ref_{prior}, operating well above a knowledge-informed chance level. Moreover, we can observe several apparent regularities in these results; first, the overall amount of data used for label inference is correlated with the recognition performance to a certain extent, e.g. $3SEG_{10} \rightarrow 6SEG_{10} \rightarrow Ref_{\text{all}}$. Here, the recognition performance steadily increases with a growing amount of data used for label inference, at which at a given point a ceiling is reached; adding more data does not affect the overall labelling accuracy. Second, we remark that the location of the data where the labels are extracted from positively affects the recognition accuracy; both the local continuity of the instru-

mentation and its global structure affect the extracted labels when keeping the amount of data fixed. Here, even a uniform sampling introduces a greater variety in the instrumentation, which leads to a higher recognition rate, e.g. $30SEC \rightarrow 3SEG_{10}$. Remarkable is also the high value in precision P_{micro} the *SOLO* approach exhibits in comparison to the *CLU* method. Due to its explicit focus on sections containing predominant instruments wrong predictions are less likely than in the other approach. However, the amount of correctly predicted labels is correspondingly low, indicating that the utilised parts of the signal contain only one single predominant source.

Furthermore, the similar performance figures of the *CLU*, *SOLO* and Ref_{all} approaches suggest that there exists a minimal amount of data from which all the extractable information can be derived. Hence more data will then not result in an improvement of the labelling performance. The next section will examine this phenomenon in more detail, in particular by determining the minimum amount of audio data required to maximise labelling performance.

5.3.5 Scaling aspects

The observations in the previous section suggest a strong amount of repetitiveness inside a music piece. Additionally, many excerpts – even though differing in instrumentation – produce the same label output when processed with our label inference method. To quantify those effects we use the *SOLO*, *CLU* and *NSEG* methods to process the entire piece under analysis, as all three offer a straightforward way to vary the amount of data used by the label inference algorithm. In particular, we study the effect of an increasing amount of segments to process on the labelling performance. In case of the *NSEG* method we constantly increase the amount of segments used by the label inference, thus augmenting the method's parameter n_N. Additionally, we perform the subsequent experiment for two values of l_N, namely 10 and 20 seconds. In case of the *CLU* method we sort the clusters downwards by the accumulated length of their respective segments, start processing just the first one, and iteratively add the next longest cluster. We then similarly rank the output of the *SOLO* algorithm according to the length of the respective segments labelled with *Solo*, and apply the label inference to an increasing amount of segments. For all methods we track the performance figures as well as the amount of data used for inference. Figure 5.12 depicts both performance and amount of data for the first 20 steps on the evaluation data (mean values of CV outputs).

As can be seen from the figure the performance of all tested systems stagnates at a certain amount of segments processed. Due to the different conceptions behind the algorithms those values vary to a great extent, ranging from 2 for the *SOLO* to 7 for the SEG_{20} approach. Accordingly, the "data-blind" sampling approaches reach the stagnation point later than the *CLU* and *SOLO* systems. In general, the latter two perform slightly better compared to the former, we speculate that the local continuity of the instrumentation causes this minor superiority. However, the performance figures of all approaches seem to be too close to identify one outstanding or discard any of them.

Regarding the recognition accuracy, incorporating global timbral structure, as implemented by *CLU*, most benefits labelling performance at the expense of algorithmic pre-processing. Here, the timbral variety has a greater positive impact on the recognition performance than, for instance, the presence

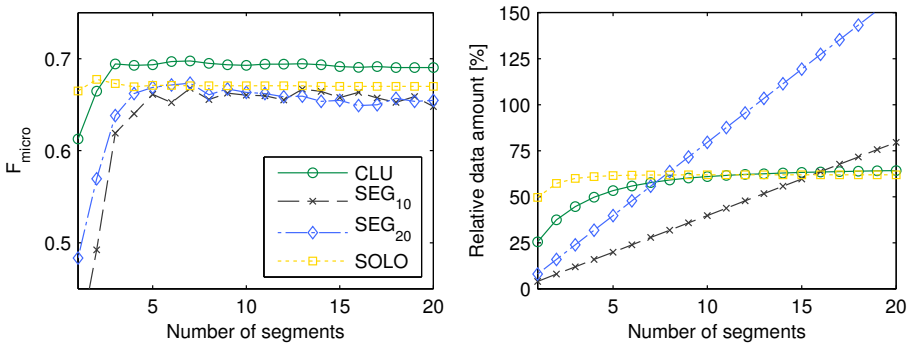


Figure 5.12: Scaling properties of the studied track-level algorithms. Part (a) shows the obtained instrument recognition performance in terms of the F-score F_{micro}, while part (b) depicts the relative amount of data, in relation to the total number of frames, applied by the respective algorithms. Both graphs show the number of processed segments on the abscissa. Mean values across CV-Folds are shown.

of a single predominant source, as implemented by the *SOLO* method. Analogously, contextual-unaware methods such as the sampling approaches show the worst of all studied performances. Moreover, with these sampling methods, an increment in segment size is only constructive for a small number of processed segments, since no difference between SEG_{10} and SEG_{20} can be observed for greater values. In general, the results suggest that, on average, both timbre-informed clustering and knowledge-based segmentation as performed by the *Solo detection* does not result in a significant increase in performance, though they might be of advantage in specialised applications (e.g. working on a single genre which exhibits clear recurrent structural sections).

In terms of the applied data amount, SEG_{10} is superior, reaching its ceiling at around 20% of the data processed. This is followed by the *CLU* method, which already needs around 45% to show its maximum value at 3 processed segments. Then, *SOLO* yet processes around 60% of the data in the first 2 segments, where the peak in recognition accuracy is observed. Finally, SEG_{20} shows a similar performance in terms of the amount of data applied, since it processes around 55% of all frames at the maximum in performance. It should be noted that both *SOLO* and *CLU* only exhibit a maximum of 5-10 relevant segments on average, since no changes in the performance figures can be observed for values greater than these.

5.4 Discussion and conclusions

In this chapter we introduced several track-level approaches which pre-process an entire piece of music for automatic musical instrument recognition. We presented both knowledge-based and agnostic methods to perform a prior segmentation of the audio signal; all approaches output a set of segments from which label output we form the final set of instrumental labels. By applying this kind

of pre-processing we can either restrict the automatic instrument recognition on portions of the analysed music piece, where a more robust extraction of the instrumental information seems possible, or exploit the structural repetitions of the track to minimise the data amount used for processing. Here, we first focused on the location of sections inside a given piece of music, which conform best with the assumptions we have taken in the design process of the recognition models. More precisely, we identify segments in the track which exhibit a single predominant source, in order to guide the instrument recognition to reach a maximum in performance. Second, we aim at exploiting the redundancy in terms of instrumentation – a result from the formal structures typically observed in Western musical compositions – for instrumental label inference (we may extract just a few labels from a thousands of audio frames). Here, we identify the most representative, in terms of the global instrumentation of the piece, portions of the signal to reduce the amount of data processed while maintaining the recognition accuracy. Lastly, we combine the two aforementioned aims and estimate the minimum amount of data needed to reach the maximum in labelling performance

The obtained results suggest that all presented approaches perform comparably, hence we are not able to identify a superior approach or discard any of them. Both the supervised and unsupervised methods show similar qualities in the label output, at which the knowledge-based approach outputs less wrong predictions whereas the knowledge-free algorithms produce a higher hit rate, i.e. correctly extracted labels. This similar performance indicates that most of the extracted labels are originating from predominant sources, which are highly redundant. Therefore a given label may be extracted from multiple locations in the signal, no matter which pre-processing has been applied. Nevertheless, the presented approaches show an average recognition performance in terms of the instance-based F metric of 0.7.

In a further experiment we analysed the dependencies between the labelling performance and the amount of data processed by the different track-level approaches; here, we could observe a strong correlation between the data amount and recognition accuracy for all tested systems. Up to a certain point in data size the labelling performance improves with an increasing amount of data. However, we also observe a subsequent stagnation for all methods. Remarkably, an additional dependency on the location of the data from which we extract the information can be observed.

Furthermore, we can use the here-presented results to validate the two hypotheses stated in the beginning of this chapter. First, from the performance of the *SOLO* approach, which is comparable to all other approaches, we conclude that most of the instrumental information appears in a predominant manner in Western music pieces, hence approving hypothesis 1. Moreover, we can observe a great redundancy in the instrumentations of Western musical compositions, since our best performing approach only needs 45% of the input data to reach its peak performance in recognition accuracy[7]. This confirms hypothesis 2, stating that the redundancy of instrumentations can be exploited for automatic musical instrument recognition.

Recapitulating, a timbre-based analysis of the musical structure, as implemented by the *CLU* method, seems to cope best the dilemma of maximising the recognition performance against minimising the amount of data to process. Furthermore, the stagnation in labelling performance, observable for all

[7]Remarkably, the same factor of about $1/2$ can also be observed when comparing the number of different instrumentations to the overall number of segments in the ground truth annotations of all files in the used music collection, see Section 5.2.3.

studied approaches, indicates a kind-of "glass ceiling" that has been reached. It seems that with the presented classification and labelling methodology we are not able to extract more information on the instrumentation from a given piece of music. Nevertheless, we can observe that predominant instrumental information is highly redundant inside a given Western piece of music from which around 2/3 of the labels can be correctly predicted along with a small proportion of spurious labels. Moreover, this fact allows for a great reduction of the effective amount of data used for label inference.

6

Interaction of musical facets

Instrumentation in the context of musical genres

In this chapter we explore the influence of other musical facets on the automatic recognition of musical instruments. Since the choice of a given music piece's instrumentation is by no means independent from other musical factors – musical genre or mood play an evident role in the composer's decision of adopting particular instruments – we aim at investigating these interdependencies inside our instrument recognition framework. More precisely, we study the role of musical genre in detail, since it is probably the most influential of all musical facets on the instrumentation of a given music piece. Moreover, McKay & Fujinaga (2005) particularly argue that many music classification tasks are subsets of automatic genre classification, hence, due to the difficulty of the problem, features found to be important in the genre context are likely to be robust for general music classification, i.e. being probably decisive in several other tasks involving music classification. A related analysis of the interrelations between musical genres and moods has been recently presented by Laurier (2011).

There is a high consensus among researchers that instrumentation is a primary cue for the recognition of musical genre in both humans and machines (see e.g. Alluri & Toiviainen, 2009; Aucouturier & Pachet, 2003, 2007; Cook, 1999; Guaus, 2009; Tzanetakis & Cook, 2002). Furthermore, McKay & Fujinaga (2005; 2010) showed in two modelling experiments the importance of instrumentation in a genre classification task. Moreover, Jensen et al. (2009) hypothesise that the information regarding the two most salient musical instruments is enough to achieve an acceptable genre recognition performance. Our main hypothesis here, however, relates to the reverse; to say that genre information is an important cue for musical instrument recognition. Moreover, we hypothesise that by integrating the information on musical genre in the automatic instrument recognition process of our developed method we can improve its overall labelling performance. In the remainder of this chapter we thus present experiments which evaluate the influence of musical genres on the instrument recognition performance. Before that, in Section 6.1, we first analyse the mutual associations among different categories of musical genre on the one hand, and musical instruments on the other

hand. By using statistical tests we quantify the degree of relatedness between the aforementioned. In the second part of this chapter we then use the information on the musical genre of a given music piece to guide the extraction of labels related to its instrumentation (Section 6.2). We present several combinatorial approaches, at which we apply both the initially developed instrument recognition models and further develop new models based on the genre information provided by the musical instruments' training data.

6.1 Analysis of mutual association

In this analysis we aim at estimating the degree of relativeness between particular musical genres and corresponding instrumentations. Hence, we quantify the association between the respective categories of musical genres and instruments using statistical measures. In other words, given the musical genre, we evaluate to which degree certain musical instruments are statistically likely to form the instrumentation of a given music piece.

According to literature, Aucouturier (2009) improved the similarity ratings between musical pieces by learning the associations among musical facets, including instruments, in the employed dataset. Besides evident relations such as *Rock Band* and *Guitar*, the author found several musical facets associated with instruments, e.g. musical genre (*Hip-Hop* and *Spoken Words*) or mood (*Transversal Flute* and *Warm*). Apart from that, we are not aware, to the best of our knowledge, of any other works studying the dependencies between musical genres and instruments.

In the following we present a two-stage experiment analysing the aforementioned associations. In the first part we estimate the degree of co-occurrence between musical instruments and genres from entirely human-assigned data. We employ the dataset described in Section 4.3.2, applied for evaluating the label inference, and attach a genre label to each track by evaluating the genre assignments of 5 human experts. By means of statistical tools we then quantify the degree of association between particular genres and instruments involved in these data. Hence, the analysis in the experiment's first part adopts information derived from human expert listeners for both musical genre and instrumentation. We then estimate, in the second part, the associations between human-assigned musical genres and automatically predicted instrumentations. Here, we utilise the same genre information and predict the instrumental labels for each track by applying the methods developed in Chapters 4 and 5. By comparing the outcomes of the two analyses we can assess the representativeness of the predicted instrumental information in terms of its associations with musical genres. We then partially apply the results from these association studies for the automatic instrument recognition experiments in the second part of this chapter, where we combine the information on musical genre and instruments.

	Jazz	*no Jazz*	Σ
Piano	a	b	(a+b)
no Piano	c	d	(c+d)
Σ	(a+c)	(b+d)	n

Table 6.1: Contingency table for an exemplary genre-instrument dependency. Here, the musical genre *Jazz* is depicted over instrument *Piano*.

6.1.1 Method

To determine the hypothesised associations we relate the genre labels to the instrumentations of the tracks in the applied music collection. In essence, we want study if we can observe a higher occurrence of particular instrumental labels given a particular musical genre, compared to others. In order to quantify these associations, we apply the odds ratio (Cornfield, 1951), a statistical measure working on contingency tables. An illustration example of such a contingency table is shown in Table 6.1 for the instrument *Piano* and the musical genre *Jazz*.

In particular, the odds ratio describes the magnitude of coherence between two variables. From the contingency table it can be calculated as follows:

$$OR = \frac{ad}{bc}. \tag{6.1}$$

A value close to 1 indicates independence between the data of the two variables, while increasing deviations from 1 denote stronger associations. Since this value is bounded between $[0\ \infty]$, we introduce the signed odds ratio by mapping the values for a negative associations, i.e. values between $[0\ 1[$, to $]-\infty\ -1[$. Hence, the signed odds ratio is given by

$$SOR = \begin{cases} OR, & \text{if } OR \geq 1 \\ -\frac{1}{OR}, & \text{if } OR < 1. \end{cases} \tag{6.2}$$

6.1.2 Data

As already mentioned above, we applied the music collection used for evaluating the label inference (Section 4.3.2) in this analysis. For deriving the genre annotations we asked 5 expert listeners to assign a single genre label to each of the 220 tracks in the collection in a forced-choice task. We applied the genre taxonomy of the dataset collected by Tzanetakis & Cook (2002), in order to maintain consistency with all following experiments. Hence, the particular genre categories were *Hip-Hop*, *Jazz*, *Blues*, *Rock*, *Classical*, *Pop*, *Metal*, *Country*, *Reggae*, and *Disco*. We then simply applied a majority vote among the annotators' ratings to derive a genre label for each piece of music (in case of a draw

we randomly assigned one of the genres in question). The distribution of the 10 annotated genres inside the collection has already been shown in Table 4.8.

Analysis of inter-rater reliability showed a Krippendorff α of 0.56 (Krippendorff, 2004). This generally low agreement among the human annotators indicates the partial ambiguity of the used data in terms of musical genre as well as some limitations of the applied, too-narrow taxonomy. Moreover, personal communication with the annotators revealed that ambient and electronic music tracks posed the most difficulties in the annotation process. In this respect, an additional category for mostly electronically generated music was identified as missing. A further analysis of the prominent ambiguities in the ratings of musical genres showed the pairs *Rock – Pop*, *Pop – Disco*, and *Rock – Metal* as top-ranked, containing, respectively, 29, 9, and 8 ambiguous pieces. Here, we consider a track as being ambiguous when observing a 3-2 or 2-2-1 constellation among rated genres, indicating a strong disagreement among judges. To exclude unreliable tracks from the subsequent analysis, we furthermore rejected all pieces exhibiting no agreement on a single genre by at least three judges. This let us discard 19 tracks, resulting in a collection of 201 music pieces for the following association experiments.

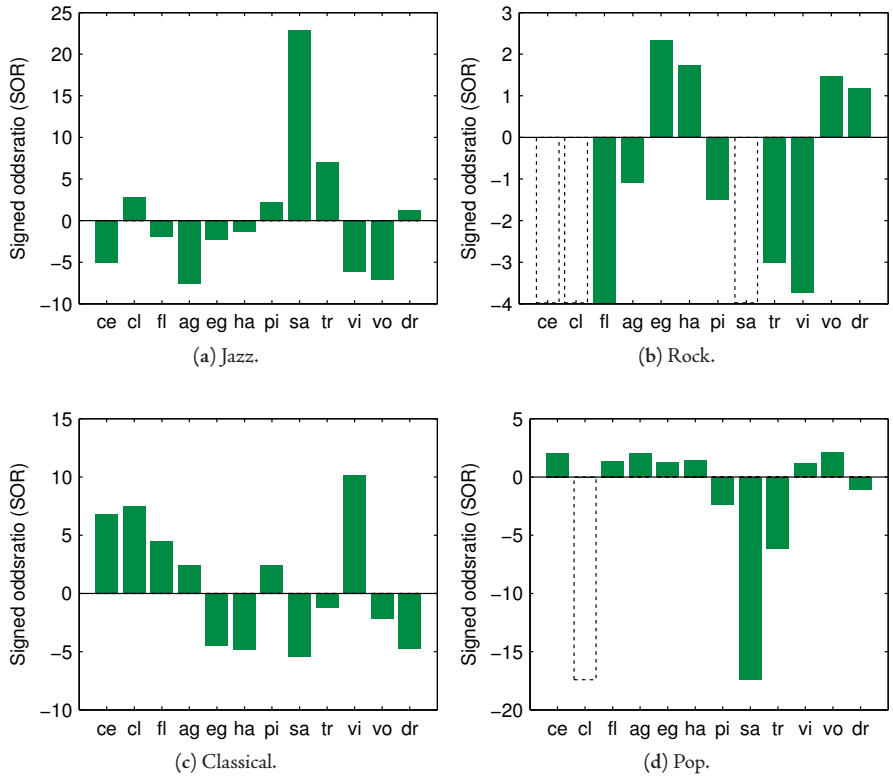
These results indicate the conceptual difficulties which arise when working with some musical genres. These problems, resulting from the mostly socially and culturally grounded definitions of musical genres, together with its adoption for computational modelling, have fuelled much debate in related literature. See, for instance, Aucouturier & Pachet (2003) for the ill-defined nature of the concept of musical genre, not founded on any intrinsic property of the music.

6.1.3 Experiment I – human-assigned instrumentation

Here, we analyse the co-occurrence of musical instruments and genres from entirely human-assigned data. Hence, we avoid all kinds of errors originating from imperfectly working computational systems in this experiment. We rather rely on the knowledge from expert listeners for both musical genres and instruments.

Since an analysis of mutual association via the signed odds ratio is only meaningful for categories with a certain minimum number of observations – those categories containing only few instances do not form a representative sample of the target population – we limit the here-presented results to the four prominent genres *Jazz*, *Rock*, *Classical*, and *Pop*, observable from Table 4.8. Hence, we discard all remaining musical genres due to the lacking number of assigned tracks. Figures 6.1 (a) - (d) show, for each considered musical genre, the signed odds ratio for all musical instrument. Here, a particular plot can be regarded as an "instrumentation profile" of the respective genre. Instruments exhibiting large positive or negative magnitudes for the signed odds ratio indicate, respectively, frequently or rarely observable musical instruments in the particular genre. Values close to ± 1 accordingly suggest that the given instruments does not occur statistically more, or less frequently than in other genres.

It can be seen from Figures 6.1 (a) - (d) that the depicted charts match the instrumentations expected for the considered genres very well. Apart from some few atypical associations, which result from the



(a) Jazz. (b) Rock.

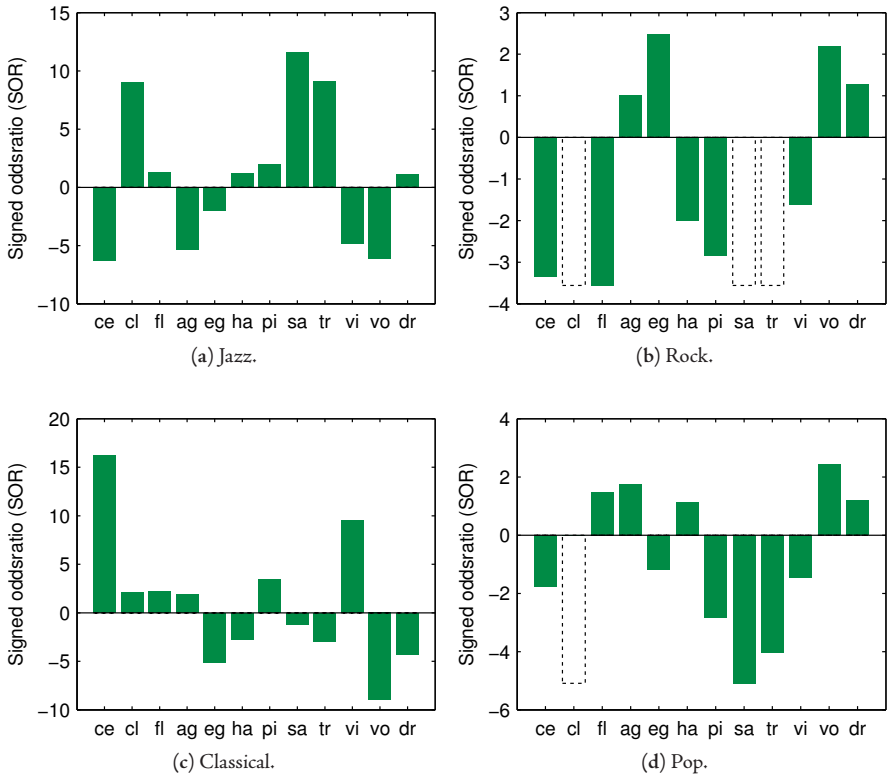
(c) Classical. (d) Pop.

Figure 6.1: Signed odds ratios for human-assigned instrumentation. Note the differences in scaling of the respective ordinates. Also note that uncoloured bars with dashed outlines represent those instruments absent in the particular genre (i.e. strong negative association). The respective values are set, for illustration purpose, to the negative maximum absolute value of the remaining categories for a given musical genre. Legend for the abscissae: *Cello* (ce), *Clarinet* (cl), *Flute* (fl), *acoustic Guitar* (ag), *electric Guitar* (eg), *Hammond organ* (ha), *Piano* (pi), *Saxophone* (sa), *Trumpet* (tr), *Violin* (vi), *singing Voice* (vo), and *Drums* (dr).

peculiarities of the applied dataset, we can observe many intuitive genre-instrument combinations. The *Jazz* category, for instance, exhibits prominent positive associations with the musical instruments *Clarinet*, *Saxophone*, and *Trumpet*, while strong negative co-occurrences can be observed for *Cello* and *Violin*. Similarly, we can see the typical positive associations with *Cello*, *Clarinet*, *Flute*, and *Violin* for classical music, while *electric Guitar*, *Hammond organ*, and *Drums* exhibit the expected negative scores. However, we also observe the surprising negative association of the *singing Voice* with the *Classical* and *Jazz* genres, resulting from the absence of Opera and jazz pieces containing singing voice. Similar considerations apply for the *Saxophone* in the Pop figure (Figure 6.1 (d)).

6.1.4 Experiment II – predicted instrumentation

In this experiment we apply the output of the instrumentation analysis method developed in Chapters 4 and 5 for the association analysis. In particular, we use the label inference approach



(a) Jazz. (b) Rock.

(c) Classical. (d) Pop.

Figure 6.2: Signed odds ratios for predicted instrumentation. Note the differences in scaling of the respective ordinates. Also note that uncoloured bars with dashed outlines represent those instruments absent in the particular genre (i.e. strong negative association). The respective values are set, for illustration purpose, to the negative maximum absolute value of the remaining categories for a given musical genre. Legend for the abscissae: *Cello* (ce), *Clarinet* (cl), *Flute* (fl), *acoustic Guitar* (ag), *electric Guitar* (eg), *Hammond organ* (ha), *Piano* (pi), *Saxophone* (sa), *Trumpet* (tr), *Violin* (vi), *singing Voice* (vo), and *Drums* (dr).

employing the Curve Tracking (CT) algorithm from Section 4.3.3, and the *CLU* track-level approach described in Section 5.2.2.3. Moreover, we process the first three clusters output by the *CLU* method for label inference, following the results from Section 5.3.5.

The applied methodology corresponds to the one described in the first part of the experiment. Consequently, we again use the 4 prominent genres from Table 4.8 for the association analysis. Figure 6.2 shows the resulting odds ratios, plotted for each of the analysed musical genres, where prominent positive and negative values indicate, respectively, frequently and rarely adopted musical instrument given a particular genre. Values close to ± 1 indicate no association between the musical instrument and the genre.

Again, from Figures 6.2 (a) - (d) we can observe that the depicted information conforms with the expected co-occurrences of musical instruments and genres. A comparison with the figures obtained for human-assigned information regarding the instrumentation of the tracks shows that most of the information is overlapping. Although we can identify some deviations from the exact values in

Figure 6.1, the basic "shape" of all genre-typical instrumentation profiles is preserved. This substantiates the findings of the previous chapters related to the validity and informativeness of the extracted instrumental information. Moreover, this result suggests that most of the errors are not resulting from any characteristics of a given musical genre; the errors are rather equally distributed among the different genres. To conclude, the predicted instrumentations using our developed instrument recognition method generally reflects the typical instrument-genre associations.

6.1.5 Summary

In essence, the association analyses led to the expected results, namely that musical instruments and genres are highly dependent, at which most of these dependencies are quite intuitive. Moreover, the comparison of the experiments on human-derived and automatically predicted information regarding the instrumentation of the analysed music pieces suggests that the extracted instrumental labels reflect those associations, hence providing meaningful information with respect to the musical genres inside the analysed music collection. However, the automatic instrument recognition regularly propagates its errors into the association measure, which influences the co-occurrences of musical instruments and genres to a certain extent. We will quantify this influence in the following part of the chapter by partially applying the here-derived findings for the automatic recognition of musical instruments.

6.2 Combined systems: Genre-based instrumentation analysis

The results of the previous section indicate that information regarding the musical genre of a given music piece already contains a lot of information regarding the likelihood of its involved musical instruments. Hence, in this section we examine to what extent we can exploit the associations between musical instruments and genres for automatic musical instrument recognition. We therefore construct several systems, all of which incorporate the genre information in a different manner, and comparatively evaluate them to asses their respective pros and cons with respect to the overall instrument recognition performance.

From an engineering point-of-view, we are aiming at improving instrument recognition by incorporating genre information. Our goal is to construct a system which uses, for a given piece of music, its musical genre to re-evaluate either the intermediate stages or the entire output of the instrument recognition algorithms presented in Chapters 4 and 5. On the one hand, this procedure allows us to eliminate or attenuate spurious genre-atypical instrumental information by incorporating the output provided by a pre-trained genre classifier. On the other hand, it may happen that correct, but genre-atypical, instrumental information is neglected or re-weighted, depending on the respective approach. Moreover, errors introduced by the, presumably, imperfectly working genre classification

are propagated to the instrument recognition stages. In a nutshell, by using information regarding musical genre, we have to accept dropping correctly-assigned labels not typical for the given genre; in return, we can eliminate spurious instrumental labels and thereby increase the overall labelling performance of the system. In the remainder of this chapter we study the influence of all these factors identified by the aforesaid.

Given these considerations it is however to question if such a strategy for performance improvement of an information retrieval system always benefits its user-oriented needs. Here, we may argue that a user has only minor interest in retrieving the genre-typical musical instruments, e.g. querying a given music collection for rock pieces containing *electric Guitar*. (S)he may rather be interested in those instrumental information that is atypical for a given musical genre, e.g. finding classical pieces adopting *Drums* or the aforesaid *electric Guitar*. These considerations are in line with general information theory, which always regards the most infrequent data as being the most informative (Cover & Thomas, 2006). On the other hand, we may also consider the situation where the user is looking for music pieces with absent genre-typical instruments, e.g. querying a database for rock pieces without *electric Guitar*. In this case, the above-presented strategy for labelling performance improvement will not exhibit the aforementioned negative effects on the user's needs.

In all following automatic instrument recognition experiments conducted in Sections 6.2.2 and 6.2.3 we apply the label inference algorithm using the Curve Tracking (CT) labelling approach (see Section 4.3.3). Moreover, since we are analysing only entire pieces of music, we apply the *CLU* approach from Section 5.2.2 to pre-process the respective tracks in order to determine the most relevant instrumentations therein. We then use those segments originating from the three "longest" clusters for extracting the instrumental labels (see Section 5.3.5).

6.2.1 Genre recognition

In this section we describe our approach towards the modelling of musical genre. We apply a standard statistical modelling approach as utilised in many related works (e.g. Aucouturier, 2006; Meng et al., 2007; Pampalk et al., 2005; Tzanetakis & Cook, 2002). First, we describe the adapted data used for building the recognition model, which is discussed in the following part of this section. Finally, we shortly evaluate the constructed classifier on unseen data to assess its prediction abilities.

6.2.1.1 Data

We use the genre classification data collected by Tzanetakis & Cook (2002) for training our genre recognition model. Originally, it covers the above-mentioned categories *Hip-Hop*, *Jazz*, *Blues*, *Rock*, *Classical*, *Pop*, *Metal*, *Country*, *Reggae*, and *Disco*, each of them represented by 100 music audio excerpts of 30-second length. The categories *Classical* and *Jazz* can be further divided, respectively, into the subclasses *Choir*, *Orchestra*, *Piano*, and *String quartet*, and *Bigband*, *Cool*, *Fusion*, *Piano*, *Quartet*, and *Swing*. For a previous application of this collection in MIR research see, for instance, the works by Li & Ogihara (2005), Holzapfel & Stylianou (2008), or Panagakis & Kotropoulos (2009).

For our specific needs we re-distribute the 10 categories into 3 super-classes, namely *Classical*, *Jazz/Blues*, and *Pop/Rock*. That is, we directly adopt the original *Classical* category and merge the original categories *Jazz* and *Blues* into the *Jazz/Blues* class. Finally, we unite all remaining original categories to build the *Pop/Rock* class. This is motivated by the generally small differences in the instrumentations that the sub-categories of a given super-class exhibit (e.g. one can find very similar instrumentations in both jazz and blues music). A further reason for the choice of this rather coarse genre taxonomy is that we can directly map it to the labels of the musical instruments' training collection, as used in Section 6.2.3 (see also Figure 4.4).

For evaluation we use the tracks of our instrument recognition evaluation collection (Section 4.3.2), merging the human-assigned labels similarly to the aforesaid into the super-classes *Classical*, *Jazz/Blues*, and *Pop/Rock*. We then extract excerpts of 30-second length from the audio signal to construct the genre evaluation collection. These data therefore serve as independent test collection for the genre recognition model.

6.2.1.2 Genre classifier

We computationally model the musical genres using a SVM classifier trained with pre-selected low-level audio features. Here, we apply the same modelling methodology as described for the musical instruments in Section 4.2. First, we extract all audio features, presented in Section 4.2.1, frame-wise from each audio instance in the collection by using a window size of approximately 46 ms and an overlap of 50%. The resulting time series of raw feature values are then integrated along time using mean and variance statistics of both the instantaneous and first-difference values. Then, a 10-Fold feature selection procedure selects the most relevant of these audio features for the given task. Next, we estimate the optimal values for the model's parameters by conducting a two-stage grid search in the relevant parameter space. Finally, we train the model using the determined parameters with the pre-selected features extracted from the training collection. It should be noted that we use a flat distribution among the respective musical genres in the dataset in all reported experiments, hence the 3 target categories contain 100 audio instances each.

6.2.1.3 Evaluation

We report an average accuracy $\overline{A}$ following a 10 $\times$ 10-Fold CV of 88.4% $\pm$ 1.31 pp on the training dataset, along with average F values for individual classes of, respectively, 0.96, 0.84, and 0.85 for the *Classical*, *Jazz/Blues* and *Pop/Rock* categories. These figures indicate that the *Classical* category is better modelled than the remaining two classes, hence most errors originate from confusions between *Jazz/Blues*, and *Pop/Rock*. Moreover, the evaluation on the external test set – the 220 tracks from the instrument recognition evaluation collection – result in an accuracy A of 71%. This drop in accuracy following the cross-database evaluation shows the high variability in the modelled musical concepts, a fact that has been previously pointed out by Livshin & Rodet (2003). Furthermore, Guaus (2009) shows the limited generalisation capacities of this particular collection by performing cross-database testing for musical genre classification. Moreover, these results also support the fact that musical genre is by no means a well defined concept from the taxonomic point-of-view, since its perception is highly subjectively and partially influenced by the cultural and social context (Aucou-

turier & Pachet, 2003; Guaus, 2009). Many of the prediction errors may therefore result from these categorical ambiguities. We note that this observed genre classification error is directly translated to the label inference stage of the systems incorporating automatically inferred genre information presented subsequently in this section, since all of these systems use the same data to evaluate their recognition abilities.

6.2.2 Method I - Genre-based labelling

In this section we present combined systems which use the original instrument recognition models as developed in Chapter 4. Thus, the genre information is affecting the output of these models. More precisely, the first here-presented system uses the categorical genre information as a filter on the output of the label inference algorithm. The second combinatorial approach exploits the probabilistic estimates of the genre classifier as a prior to weight the output of the instrument recognition models.

6.2.2.1 Label filtering

As already mentioned above, this approach (S_{LF}) applies the categorical genre information to filter the instrumental labels provided by the label inference algorithm developed in Section 4.3. Hence, given the genre label of the analysed track, we simply reject those predicted labels which are atypical for the musical genre assigned to the piece. In particular, for pitched instruments, we regard *electric Guitar* and *Hammond organ* atypical for the *Classical* genre, *Cello*, *Flute*, and *Violin* atypical for the *Jazz/Blues* category, and finally *Cello*, *Clarinet*, and *Violin* atypical for the *Pop/Rock* genre. We furthermore disregard a predicted label *Drums* for a piece from *Classical* music. This selection of atypical instruments given a particular genre results from general considerations regarding the expected instrumentations of the considered genre, together with the results from the association experiments in the previous part of the chapter. We want to note that we purposely used only those genre-instrument combinations for filtering which take part in both aforementioned sources of consideration. In this regard, we want to avoid any biasing or overfitting of the developed methods towards the applied music collection.

Figure 6.3 depicts the basic concept of this approach in the instrument recognition framework. Note that the illustration is simplified by unifying the pitched and percussive recognition into a single, general, recognition path.

6.2.2.2 Probability weighting

This method, denoted S_{PW}, uses the original models for pitched and percussive timbre recognition and works on the respective output of the classifiers. Here, we directly apply the output of the genre classifier as a prior for the labelling algorithm. In particular, we use the genre probabilities to re-weight the probabilistic estimates of the instrumental classifiers. Figure 6.4 shows a graphical illustration of this approach. Note again that the depiction unites the pitched and percussive

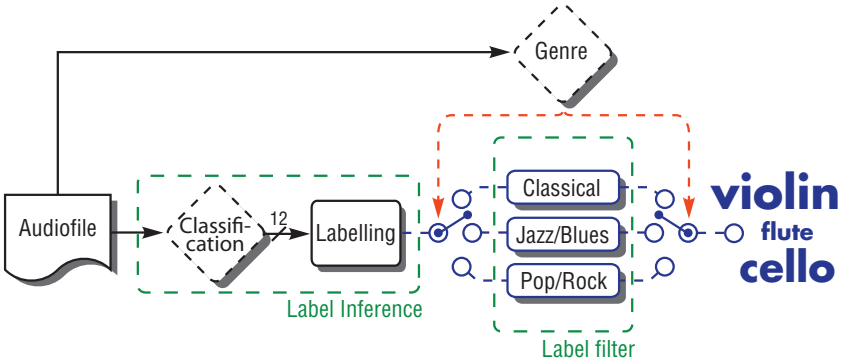


Figure 6.3: Block diagram of combinatorial system S_{LF}. The output of the label inference is simply filtered to suppress genre-atypical instrumental labels. Note that the illustration is unifying the pitched and percussive recognition into a single, general, recognition path.

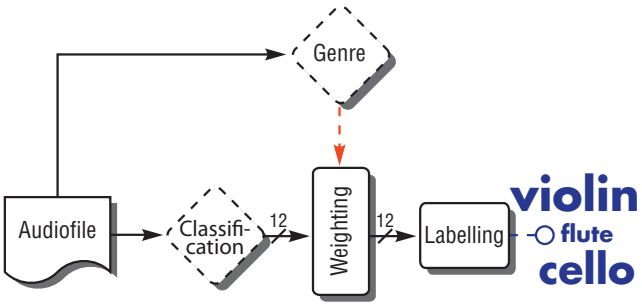


Figure 6.4: Block diagram of combinatorial system S_{PW}. The probabilistic output of the instrument recognition models are re-weighted according to the genre estimate. Note the unified recognition path for pitched and percussive instruments for illustration purpose.

recognition into a single, general, recognition path.

In case of the pitched instrument recognition we first adapt the 3-valued genre probability vector to the instruments' 11 probabilistic estimates. Hence, we assign, to each instrument, its respective genre probability according to the genre-instrument relations defined in the previous approach, and re-normalise the resulting 11-valued genre probability vector so that the probabilities sum to one. We then weight, i.e. multiply, the instrumental probabilities by the respective genre estimates and again re-normalise the resulting vector, doing this for each classification frame. The resulting re-weighted representation of instrumental presence is then passed to the labelling module without any further changes.

We apply a similar procedure for the percussive timbre recognition. Here, we weight the corresponding category with its genre probability, i.e. *no-drums* with *Classical* and *Drums* with the sum of *Jazz/Blues* and *Pop/Rock*, and re-normalise the resulting representation for each classification frame. Drumset detection is then performed via the method described in Section 4.3.3.

	Cello	*Clarinet*	*Flute*	*ac. Guitar*	*el. Guitar*	*Hammond*	*Piano*	*Saxophone*	*Trumpet*	*Violin*	*Voice*	*Drums*
Classical	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×
Jazz/Blues	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓
Pop/Rock	×	×	✓	✓	✓	✓	✓	✓	✓	×	✓	✓

Table 6.2: Categories modelled by the 3 genre-specific instrument recognition models. Note that pitched and percussive categories are represented by two different classifiers.

6.2.3 Method II - Genre-based classification

For the approaches presented in this section we develop new instrument recognition models considering the genre information of the respective training instances. In particular, we exploit the genre labels of the instances in the training collections for the musical instruments (see Figure 4.4) to construct genre-specific statistical models of the 11 pitched instruments[1]. We then use the genre information of the track under analysis to either choose one of the recognition models for prediction (*Classifier selection*) or combine the information provided by the individual models with respect to the genre estimate (*Decision fusion*).

By constructing genre-specific instrument recognition models we consider that certain musical instruments are adopted in a similar manner given a particular musical genre, but exhibit a rather different contextual use across genres (e.g. violins play a predominant role in almost all classical music while their adoption in pop and rock music is more of an accompaniment kind; moreover, in jazz and blues music they appear very rarely). Furthermore, we can take advantage of the different descriptions in terms of audio features a given instrument exhibits in different musical contexts (e.g. an acoustic guitar may be described differently in classical and rock music.). However, the genre-dependent training of the recognition models may add complexity to the overall label inference task, since the information from 3 models may be considered. This can lead to additional spurious information which negatively influences the overall labelling performance.

We develop these new recognition models following the procedure described in Section 4.2.1. First, we construct the 3 datasets according to the genre labels provided by the respective audio instances. Following the distribution of labels and the results from the association analyses in Section 6.1, we use, for constructing the *Classical* model, all available data except the categories *electric Guitar* and *Hammond organ*. The *Jazz/Blues* model comprises all pitched categories except *Cello*, *Flute*, and *Violin*, while the *Pop/Rock* classifier is built using all data aside the categories *Cello*, *Clarinet*, and *Violin*. Table 6.2 summarises the class assignments of the 3 developed genre-specific recognition models for pitched instruments. We note that we mainly use data from the corresponding musical genres in the training data of the respective recognition models (for some rare combinations, e.g. *singing Voice* and the *Classical* model, we had to use the training data from the other musical genres due to lack of relevant instances assigned to the *Classical* genre). Moreover, we again use only flat class distributions in the datasets for all upcoming experiments.

[1] The drumset detection is modified similarly to the approaches in the previous section.

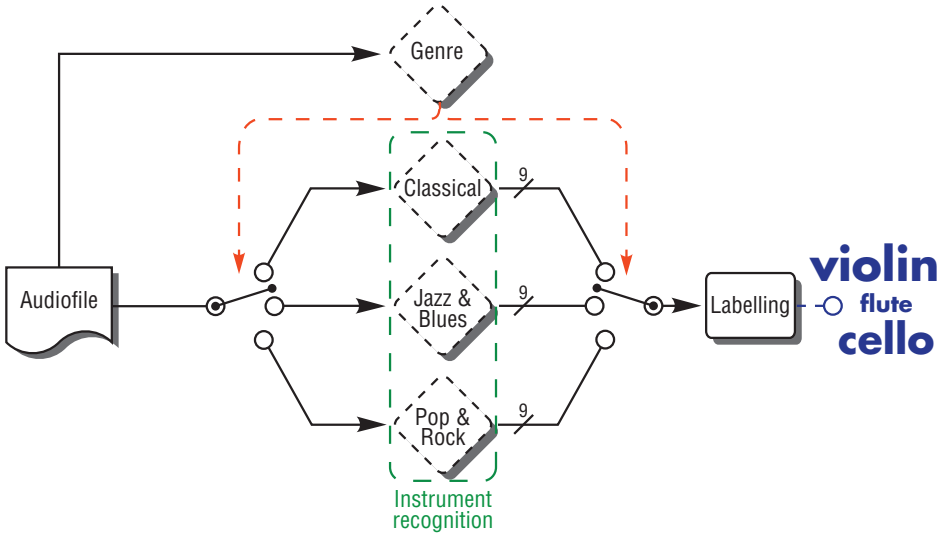


Figure 6.5: Block diagram of combinatorial system S_{CS}. The categorical output of the genre recognition model selects one of the 3 instrumental classifiers for label inference. Note the combined recognition path for pitched and percussive instruments.

Next, we apply the 10-Fold feature selection to each of the 3 generated datasets, identifying the genre-dependent optimal feature subsets for the respective musical instrument recognition task. We then perform the same 2-stage grid search procedure as described in Section 4.2.3.3 to estimate the optimal parameter values of the SVMs for the three classifiers. Finally, we train the 3 models using the respective estimated best parameter settings with the corresponding set of selected audio features extracted from the particular training collection.

6.2.3.1 Classifier selection

This first approach (S_{CS}) explicitly chooses the recognition model to apply considering the genre estimate of the analysed music piece. Hence, it can be regarded as supervised classifier selection (Kuncheva, 2004), where an oracle decides which of the 3 models to use given the data at hand. Label inference for the pitched instruments is then performed by using the predictions of the selected classifier. In case of the percussive labelling we simply disregard the classifier decisions given the label *Classical* for the musical genre. Figure 6.5 shows an illustration of the basic processes involved in the presented approach. Note that the label inference is simplified by showing a combined recognition path for pitched and percussive instruments.

6.2.3.2 Decision fusion

This last approach (S_{DF}) uses the probabilistic genre information to combine the decisions of the 3 independent recognition models. In particular, we apply a weighted sum for decision fusion (Kuncheva, 2004), where the genre probabilities represents the weights. The probabilities of the

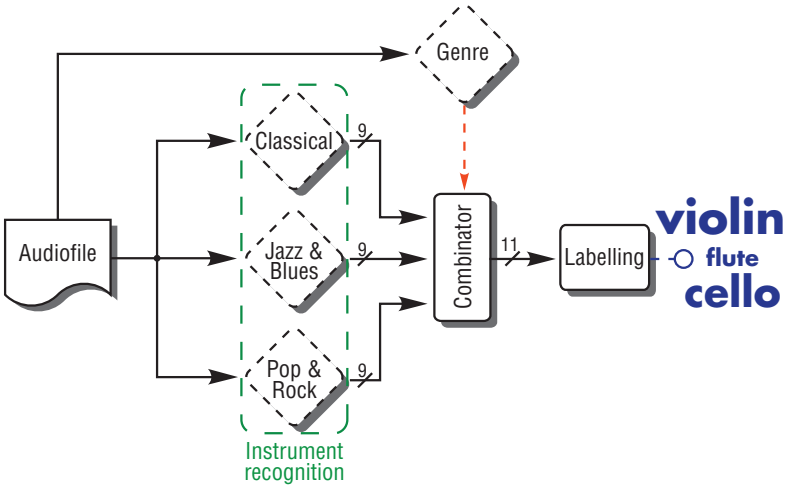


Figure 6.6: Block diagram of combinatorial system S_{DF}. The system uses the probabilistic output of the genre model to combine the probabilistic estimates of the 3 instrumental classifiers. Note that the illustration unites the pitched and percussive recognition paths for simplification.

3 pitched instrument models are simply weighted and summed using the genre information[2], while the weighting process for the percussive model output is implemented as described for the probability weighting S_{PW} in the previous section. Figure 6.6 depicts a block diagram of the decision fusion approach S_{DF}. Note that the pitched and percussive recognition paths are merged to simplify the illustration.

6.2.4 Experiments and results

In the subsequent evaluation we perform all experiments in the 3-Fold CV procedure, as applied in the previous chapters. That is, we use, for each rotation of the CV, 2/3 of the data for parameter tuning and the remaining 1/3 for assessing the labelling performance. To guarantee maximal comparativeness we estimate the parameter values for each system separately. Hence, each combinatorial approach determines its best labelling parameter values using a grid search over the training folds in each CV rotation. We use the metrics presented in Section 4.3.4.3 to evaluate the different aspects of the labelling performance for the respective systems.

Furthermore, we establish three comparative baseline systems. The first, Ref_{Ch5}, is the direct adoption of the *CLU* instrumentation analysis system presented in Chapter 5; identical to all combinatorial approaches, we use the CT track-level analysis to pre-process the entire piece of music, and use the segments of the 3 "longest" determined clusters for label inference. The second baseline, Ref_{prior}, uses the prior probabilities of the modelled musical instruments together with the genre information for label inference. More precisely, this null model is generated by drawing each label from its respective prior binomial distribution and applying the label filtering according to the

[2]We artificially set the probabilistic estimates of the not-modelled categories in the respective genre-specific models to zero to enable a combination of their values.

Metric	Ref_{Ch5}	S_{LF}	S_{PW}	S_{CS}	S_{DF}	$Ref_{\mathrm{prior}}{}^{*}$	Ref_{up}
P_{micro}	0.75	0.78	0.72	0.67	0.63	0.49	1.00
R_{micro}	0.65	0.63	0.68	0.7	0.68	0.42	0.97
F_{micro}	0.69	0.7	0.7	0.69	0.66	0.46	0.98
F_{macro}	0.54	0.53	0.53	0.54	0.51	0.27	0.92

(a) Annotated musical genre.

Metric	Ref_{Ch5}	S_{LF}	S_{PW}	S_{CS}	S_{DF}	$Ref_{\mathrm{prior}}{}^{*}$	Ref_{up}
P_{micro}	0.75	0.76	0.73	0.64	0.66	0.49	1.00
R_{micro}	0.65	0.6	0.65	0.64	0.65	0.41	0.91
F_{micro}	0.69	0.67	0.68	0.64	0.65	0.44	0.95
F_{macro}	0.54	0.53	0.53	0.51	0.5	0.26	0.9

(b) Predicted musical genre.

Table 6.3: Comparative results for all combinatorial approaches. Part (a) of the table shows the evaluation results using the expert-based, i.e. annotated, genre information, while the systems in part (b) use the statistical model to predict the musical genre of the analysed track. The table header includes a reference system from Chapter 5 (Ref_{Ch5}), the label filtering (S_{LF}), probability weighting (S_{PW}), classifier selection (S_{CS}), and decision fusion (S_{DF}) combinatorial approaches, as well as a second and third reference system using, respectively, the prior distribution of the musical instruments and the expert-based instrument annotations, along with the genre information of the respective track for label inference (Ref_{prior}, Ref_{up}). The asterisk denotes average values over 100 independent runs.

musical genre, as described for the label filtering approach S_{LF} (see Section 6.2.2). We estimate its labelling performance by averaging 100 independent runs. Finally, we establish an upper bound (Ref_{up}) by filtering the expert-based annotations with the same label filtering approach.

6.2.4.1 General results

Table 6.3 shows the evaluation's results for all considered systems. To assess the influence of the genre recognition error, we perform the evaluation with both human-assigned and automatically estimated genre information, hence splitting the table into two parts. In case of the expert-based genre information, we use a 3-valued binary vector to represent the probabilistic estimates of the respective genres for the S_{PW} and S_{DF} systems. Moreover, Figures 6.7 (a) and (b) show the methods' performance on individual instrumental categories in terms of the class-wise F-score F. Again, we split the figure into 2 parts, representing, respectively, the results for expert-based and computationally estimated genre. Finally, Figures 6.8 (a) to (h) depict the amount of added and rejected labels in comparison to the output of the baseline Ref_{Ch5}, which uses no genre information for the label inference (see above). In each figure the presented bars are grouped according to the ground truth of the label, i.e. if the respective label does or does not appear in the annotation of the analysed piece[3].

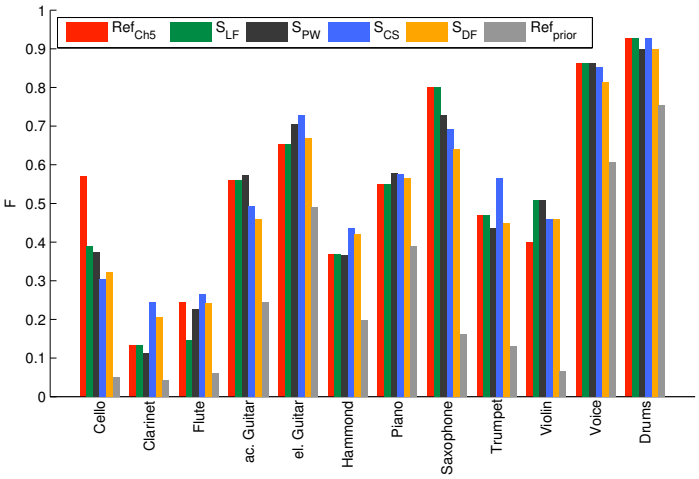
To assess the influence of the genre information on our label predictions we first analyse the results obtained with the expert-based genre information (Table 6 (a)). Here, the upper bound Ref_{up}

[3]We omit the respective performance of the upper bound Ref_{up} in Figures 6.7 and 6.8, since it does not provide relevant information for assessing the performance of the presented combinatorial systems.

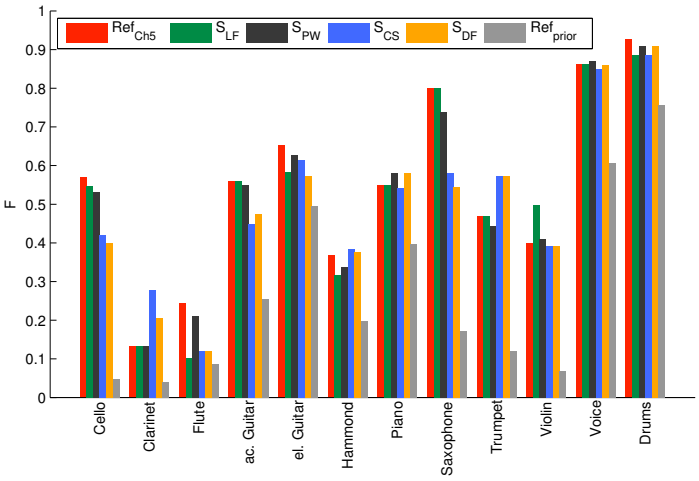
hardly shows deviations from perfect performance, indicating a quite limited number of here-considered genre-atypical labels in the annotations of the music pieces. Moreover, we see a clear improvement in labelling performance for the prior-based reference system Ref_{prior} compared to the figures obtained for the approach using the same prior information but without the genre filtering applied as presented in Section 5.3.4 (see Table 5.5). Since the filtering of genre-atypical labels leads to an increase in the precision P_{micro} by more than 20%, the F-score F_{micro} is analogously improved, namely by 15%. These, however, are quite logical and intuitive results considering the fact that this system's label extraction mechanism is completely genre-blind.

Regarding the four presented combinatorial approaches, we observe no improvement in terms of the overall labelling performance, represented by the two F-scores F_{micro} and F_{macro}, over the null model Ref_{Ch5}, which does not use genre information for label inference. Hence we have to draw the general conclusion that we cannot improve our instrument recognition algorithm using the here-applied genre context, even if we could apply a 100% accurate genre recognition model. A detailed analysis of the performance of the respective combinatorial approaches further shows that the S_{DF} approach clearly performs worse compared to the other three. This suggests that the increase in complexity – the approach combines 25 pitched instrumental probabilities in comparison to a maximum of 11 for all other approaches – degrades the overall labelling performance. All other three systems perform equally in terms of both applied F-scores, being close to the figures of the reference Ref_{Ch5}. We do, however, observe large divergences in the precision and recall metrics P_{micro} and R_{micro} for the different combinatorial approaches from this reference. In particular, we notice a decreasing precision and an increasing recall for the systems S_{LF}, S_{PW}, and S_{CS}, respectively. Since S_{LF} just removes labels, it can only improve its precision but lowers, at the same time, its recall. The S_{PW} approach is less restrictive than the aforementioned and moreover able to predict additional labels compared to the baseline Ref_{Ch5}. Here, the re-weighting of the instrumental probabilistic estimates seems to reveal masked genre-typical instruments, which is reflected in the increased value of the recall. The same process, however, lowers the precision, since also spurious labels are added due to the genre weighting. Finally, the highest value for the recall can be observed for the S_{CS} approach, since it only predicts genre-typical instruments. This, in turn, confirms the findings of the first part of this chapter, i.e. the strong associations between musical genres and instruments. On the other hand, additional confusions are added from the similar acoustical context of the respective model's training data – we trained each classifier with data mostly originating from the genre it represents – resulting in the low value for the precision.

The same trends can be basically observed from the lower part of Table 6.3, which featured approaches apply the estimated musical genre resulting from the genre recognition model described in Section 6.2.1. The figures for the combinatorial systems are proportional lower than the ones from the upper part of the table, which is a result of the propagated genre recognition error. Consequently, these figures are lower than the reference baseline Ref_{Ch5}, indicating that the imperfectly working genre recognition model is degrading the recognition performance in the instrument recognition system. Here, S_{CS} is affected most, since the wrong selection of the classifiers may lead to a series of spurious labels, whereas the genre error's effect for weighting or filtering approaches is more limited.



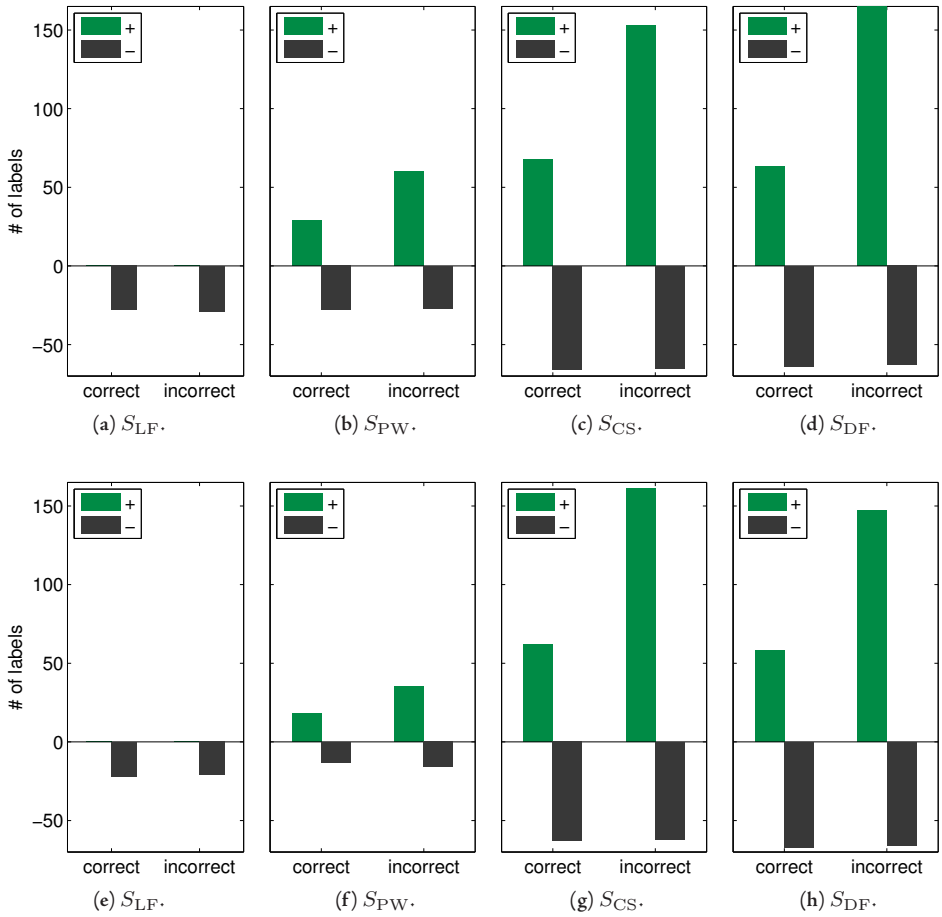
(a) Annotated musical genre.



(b) Predicted musical genre.

Figure 6.7: Performance on individual instruments of all combinatorial approaches. Part (a) shows the labelling performance in terms of the categorical F-score for ground truth genre, while part (b) depicts the same metric for automatically predicted genre labels. Legend for the different approaches: Label filtering (LF), Probability Weighting (PW), Classifier Selection (CS), and Decision Fusion (DF). See text for more details on the compared combined systems and the baseline methods.

Furthermore, Figure 6.7 indicate that the modelled instruments are affected very differently by the incorporation of the genre information. For instance, *singing Voice* and *Drums* show hardly any variations when considering the depicted approaches. On the other hand, instruments such as the *Clarinet*, *Flute*, or *Saxophone* exhibit strong variability with respect to the output of the different systems. This observation may result from the highly skewed frequency of the instruments inside the evaluation collection; more frequent categories are less likely to be affected to a large extent by the genre information, while on less frequent instruments the additional information may have great impact. Moreover, a similar behaviour of the individual F-scores can be observed for the application of annotated and estimated genre information. *Drums*, *singing Voice*, *Piano*, or *Saxophone* show



(a) S_{LF}. (b) S_{PW}. (c) S_{CS}. (d) S_{DF}.

(e) S_{LF}. (f) S_{PW}. (g) S_{CS}. (h) S_{DF}.

Figure 6.8: Quantitative label differences between the respective combinatorial approaches and the reference baseline. The upper part of the figure ((a) - (d)) shows the results for expert-based annotated genre information, while the lower row depicts the results for estimated musical genre ((e) - (h)). The "+" in the legend refers to labels added in comparison to the baseline, while the "−" stands for rejected labels with respect to the output of the reference system. Moreover, the two groups assigned to the abscissa represent whether or not the particular label is featured in the ground truth annotation of the respective track.

very similar patterns for the respective approaches in Figures 6.7 (a) and (b). Some instruments, however, exhibit a contrary behaviour, in case of the *Cello* even benefiting from the error introduced by the genre recognition model. This has to result from the genre-atypical adoption of the particular instrument in several tracks of the used evaluation collection.

Finally, we analyse the quantitative differences in label predictions between the combinatorial approaches and the null model Ref_{Ch5}, as shown in Figure 6.8. Here, an additional correctly predicted label ("+") increases the recall, while a lost correct label ("−") decreases the recall of the respective method. Analogously, an incorrectly added label ("+") decreases the precision, while a removed incorrect instrument ("−") increases it. Interestingly, the amount of removed correct and incorrect labels is approximately the same for all presented approaches, hence affecting the precision the same way positively as the recall negatively. This indicates that the number of wrongly and correctly pre-

dicted genre-atypical instruments is mostly the same. Hence, by only removing labels according to the imposed genre-dependent rules – as implemented by the S_{LF} approach – we cannot improve the labelling performance of the presented approach towards the automatic recognition of musical instruments. Moreover, those approaches applying genre-specific recognition models (S_{CS} and S_{DF}) exhibit the double amount of additional labels, both correctly and incorrectly predicted, which is in accordance to the aforementioned findings. The genre-adapted classifiers encode different acoustical facets of the input audio, hence resulting in greater values for added as well as removed labels in comparison to those approaches using the original recognition models as developed in Section 4.2. Furthermore, the observable greater value in the incorrectly added compared to the correctly added labels corresponds to the greater reduction of the precision in relation to the increase in the corresponding recall.

6.3 Discussion

In this chapter we studied the interrelation between the musical facets of genre and instrumentation. In particular, we analysed the statistical dependencies between particular musical instruments and genres and estimated its influence on the output of our developed instrument recognition approach. In the statistical analysis presented in the first part of the chapter we found strong associations between musical instruments and genres. More precisely, the applied test revealed that each of the modelled instruments is strongly related to at least one of the analysed musical genres. Many instruments, moreover, exhibit associations to several genres. Hence, we can confirm our first hypothesis, stated in the beginning of the chapter, concerning the expected co-occurrences between musical instruments and genres.

By reviewing the results obtained in the second part of the chapter we have to, however, reject the second stated hypothesis – improving the labelling performance of the developed automatic musical instrument recognition method by integrating information on the musical genre. None of our presented approaches combining musical instrument and genre recognition could score over the performance of the null model, which is not using genre information for label inference. Nevertheless, we can identify several reasons for this negative outcome; first, most of the aforementioned null model's prediction error results from genre-typical instruments[4]. Hence, eliminating the spurious genre-atypical labels is not increasing the labelling performance to a great extent; moreover, an additional error is introduced which compensates for this improvement. From a different viewpoint, the information provided by the instrument recognition models and the genre recognition source is not complementary but mostly entirely overlapping; they basically encode the same piece of information. Second, even in the light of the observed associations, many musical instruments – and especially those modelled in this thesis – are adopted in several musical genres, which narrows the prospects of controlling the label inference with genre-related information. Table 6.2 contains 5 out of 12 instruments which are present in all 3 modelled musical genres. This fact may also be rooted

[4]From the initial 1039 labels predicted for all 220 tracks in the used evaluation collection, the categorical filtering approach S_{LF} only removes around 50, indicating this amount of wrongly-predicted genre-atypical instrumental labels.

in the generalising taxonomy chosen for modelling the musical instruments, which can be regarded as a mid-level in the hierarchical taxonomic representation of musical instruments (see Section 4.1). Many instruments further down the hierarchy would exhibit more genre-specific properties (e.g. a *concert Guitar* is mainly adopted in classical music in contrast to the general class of *acoustic Guitar*, which spans multiple musical genres), but at the expense of a higher confusion rate with other instruments of the same family. And third, the error introduced by the imperfectly working genre recognition directly translate to an additional instrument recognition error. Using automatically inferred genre information actually degrades the performance of the labelling, when compared to the null model which is not applying this information source.

These results are rather disappointing when considering the results obtained by McKay & Fujinaga (2005), where the instrumentation was found to be the most important cue for genre recognition[5]. Since the descriptors in the aforesaid work are extracted from symbolic data, the authors could apply fine-grained details about the instrumentation of the analysed pieces. A vector representing all 128 general MIDI musical instruments contained the total time in seconds of the particular instruments in the processed track. Hence, in order to apply genre information for automatic instrument recognition two requirements must be fulfilled; first, a detailed taxonomy of musical instruments must be modelled. This has been already identified above. Second, the information regarding the detected instruments has to be accurate. The relative importance of the instrument in the analysed piece seems to be of importance, given the results obtained by McKay & Fujinaga (2005). Both requirements, however, are only partially met by the presented approach towards the automatic recognition of musical instruments, which explains the negative results in the second part of this chapter.

In this regard it seems plausible that the hypothesis stated in the beginning of the chapter – musical genre information acts as an important cue for musical instrument recognition – only applies for specific genres adopting peculiar instruments. Hence, instruments such as particular percussion instruments (e.g. *Bongos*, *Congas*), genre-typical electronic devices exhibiting distinctive sounds (e.g. Roland's *TR-808* or *TB-303*), or other genre-idiosyncratic instruments such as the *Mellotron* or the *Steel guitar* should be modelled. This consequently would lead to different kinds of model conceptions and thus model architectures. Instead of the here-adopted multi-class models, simple presence/absence models, i.e. one-vs-all, for both a particular musical genre and a particular instrument would then better meet the requirements (e.g. modelling the *Mellotron* for the genre of 1970's *Progressive Rock* as recently developed by Román (2011)).

[5]Here, we hypothesised, without loss of generality, the reverse, namely that musical genre is an important cue for instrument recognition. The results from the association analyses in the first part of this chapter further substantiate this hypothesis.

7

Conclusions

A recap, known problems, and an outlook into the future

After having reviewed the various approaches, implementations, and experimental results of Chapters 4 - 6, we take a step back and reflect on the general outcomes of this work together with their implications for the relevant research field. We motivated our work in Chapter 1 by stating the importance of instrumentation in general music listening; as the musical representation of timbre sensation it strongly influences our mental inference of higher-level musical concepts (Alluri & Toiviainen, 2009). Moreover, instrumentation represents one of the most important semantic concepts humans use to communicate musical meaning. From this viewpoint we identified two main directives of possible research lines; the first one, purely engineering motivated, uses the information regarding the instrumentation of a music piece to provide accurate, i.e. musically meaningful, search and retrieval facilities in large catalogues of multimedia items, as well as personalised music recommendations. The second direction explores the areas of human auditory perception and cognition, where hearing research still knows little about how our mind makes sense of complex acoustical scenes (Yost, 2008). Here, the aim is to contribute to a deeper understanding of sound in general and its processing inside the human brain.

Driven by these bifocal research perspectives, we asked questions such as "what kind of instrumental information do we need for a meaningful description of the music from a user's point-of-view?", or "which sound components of a given musical instrument affect its identifiability among other instruments?". Some of these questions which arose in the course of this thesis could be answered, while others still remain unanswered and subject to future research. In what follows we first summarise the content covered in this thesis (Section 7.1), reflect on the insights we gained from the various experimental results (Section 7.2), point towards directions for future research (Section 7.3), and close this thesis with some concluding remarks (Section 7.4).

7.1 Thesis summary

To the authors' knowledge, this dissertation presents the first thesis work designing approaches for the automatic recognition of musical instruments specifically targeted at the processing of polyphonic, multi-source music audio data. We developed a modular, hierarchically constructed method which incorporates, at each level, psycho-acoustical and musicological knowledge bits. We thereby designed and evaluated the respective components in its corresponding musical context. Moreover, this thesis offers the most extensive evaluation framework, compared to related works in the field, up to now, assessing the method's accuracy, generality, scalability, robustness, and efficiency.

In particular, in Chapter 4 we started at the level of a musical phrase (typically in the range of several seconds), which is known to be the fundamental building block in the human source recognition process (Kendall, 1986; Martin, 1999). Here we developed statistical recognition models which are able to predict the presence of a single predominant musical instrument in a musical mixture (Section 4.2). An in-depth analysis of low-level audio features involved in the decision process showed how the specific acoustical characteristics of the modelled instruments are inherent in the identification process, hence bridging the gap to both perceptual and psycho-acoustic research. In the subsequent thorough error analysis we furthermore identified many prediction errors to be similar to those found in recognition studies using human subjects.

In the second part of Chapter 4 we used an analysis of musical context on top of the models' predictions to infer knowledge regarding the instrumentation of a given music audio signal (Section 4.3). We thereby showed that the applied context analysis allows for a reliable extraction of the instrumental information together with a robust handling of unknown sources. Moreover, we proved the usefulness of the information resulting from predominant sources in the instrument recognition paradigm and showed how to incorporate this information into a multiple instrument recognition system.

Chapter 5 covered the next level in the hierarchy, namely the processing of entire music pieces. Here, we described and compared several approaches, both knowledge-based and agnostic ones, for a complete instrumentation analysis of music tracks. We identified the capacities as well as the limitations of the presented methods and showed how the redundancy in the instrumentation of a given music piece can be exploited to reduce the amount of data used for processing. In short, the approaches were able to correctly extract around 2/3 of the instrumental information along with a low amount of spurious labels by using only a fraction of the available input data. We however identified a ceiling in the recognition performance that could be explained by the constrains applied in the design process of the recognition models.

Finally, in Chapter 6 we even entered a global contextual level by linking the instrumentation of a music piece with its musical genre. We first quantified the statistical dependencies between musical instruments and genres by applying proper measures. In the second part of the chapter we then developed automatic musical instrument recognition systems which integrate the information on the musical genre in the decision process. We could generally conclude that a context-adaptive taxonomy of musical instruments is needed to fully exploit the information provided by the musical genre.

Recapitulating, in this thesis we have taken several fundamentally different paths compared to related works in the field. From a perceptual viewpoint, we directly translated – yet imposing proper constraints to the modelling process – the underlying problem from its very general monotimbral nature into a polytimbral context. Moreover, many "transcriptive" approaches view the problem as inseparable from automatic music transcription, hence performing instrument recognition either simultaneous or subsequent to an estimation of multiple pitches or onsets. Most related studies on automatic instrument recognition further rely on a strict frame-by-frame processing. Our method, on the contrary, infers the information regarding the instrumentation from portions of the signal exhibiting the most confident model predictions. Next, the observed redundancy of the information led us to discard more than half of the available input data, with no reduction of the recognition accuracy. The results presented here – along with various findings from psycho-acoustic and machine listening research – suggest that both, the "transcriptive" viewpoint as well as the strict frame-wise processing, are by no means required for a successful and detailed description of the instrumentation of a musical composition. Furthermore, we strictly evaluated our approaches only against real music data of any timbral and musical complexity in order to assess its recognition performance in a general context, a procedure which is still not standardised in related works. At last, we contextualised the problem by exploiting apparent associations between high-level semantic concepts that are inherent to the analysed music, which, to the best of our knowledge, has not been done before inside the instrument recognition paradigm.

In the light of the aforementioned, we can now draw the connection from the presented approach to the general evaluation criteria for recognition systems presented in Section 3.3. First, the developed method meets criteria 2, 4, and 1 by exhibiting, respectively, good performance in the handling of data complexity and noise, as well as acceptable generalisation capabilities. In corresponding experiments we showed that the recognition error is neither dependent on the complexity nor the amount of unknown sources in the data. Moreover, the generalisation capabilities were revealed by the method's performance on the independent, constraint-free dataset used in Section 4.3 and thereafter. Furthermore, the presented method meets criterion 3 insofar that the applied architecture of the statistical models – we use SVM classifiers – allows for a flexible management of the modelled categories, thus new classes can be added easily provided the necessary labelled data. Finally, the presented algorithm also confirms with criterion 6 since the basic label inference presented in Section 4.3 is based on a sequential analysis of time-series, confirming with the *content understanding* notion of any music processing systems (Martin, 1999). Hence, only criterion 5 – the adaptivity of the employed learning strategy – is not met, but the need for such a flexible, semi-supervised architecture is apparent. However, we leave this issue open for future research directions.

7.2 Gained insights

In this thesis, we developed and evaluated an algorithm for the automatic recognition of musical instruments from music audio signals. Even if our method is working imperfectly, the various evaluation results provide valuable insights into the problem. They can be stated as follows:

1. We do not see a need for complex signal processing, especially source separation, in order to extract high-level cues from music signals. Admittedly, the results provided by this thesis along with several examples from literature (see e.g. Barbedo & Tzanetakis, 2011; Haro & Herrera, 2009) suggest that a certain amount of adaptive pre-processing benefits machine perception. Nevertheless, research has shown that even very untrained human listeners can accurately fulfil tasks such as musical genre or style, emotive character, timbre, or rhythm perception without effort (Martin et al., 1998). In this regard, we may further speculate that those musical instruments, which can only be recognised using perfect source separation as pre-processing, may by no means be important for the description of the musical composition; the given source cannot be perceived by the listener in a way that it would bear relevant descriptive information. We therefore believe – and will state it more explicit in the subsequent section – that studying human auditory processing and its extensive inferential character provides enough information for an accurate modelling of the acoustical scene, including source recognition. In this context, we may cite Hawley (1993), who wrote, referring to an experiment teaching pigeons to differentiate between music composed by J. S. Bach and I. Stravinsky (Neuringer & Porter, 1984) – that

   > "…the pigeon reminds us that even without much general intelligence a machine can glean enough from an audio signal to draw conclusions about its content."

   In this context, we can assume that the discrimination inside the pigeons' brains relied on timbral cues, and not on more musical aspects such as structure or tonality.

2. In our developed framework the predominance of a source is the most important cue for recognition. This is not surprising since we constrained the whole approach to the modelling of predominant sources. However, the presented results further suggest that a certain amount of predominance enables the robust extraction of the source's invariants. Remarkable here is also the amount of information we can explain by concentrating only on predominant sources. Besides, this makes sense from an evolutionary viewpoint since stronger acoustical signals always imply a stronger possible threat. Now let us think think further, if we are able, by means of signal processing, to "predominatise" non-predominant sources, we may boost the accuracy of recognition systems to a great extent (in this context, see also the provided link to the auditory attention mechanisms in the next section).

3. The applied acoustical description of the input signal in terms of low-level audio features and the approach towards the statistical modelling work reasonably well within their respective limitations. We indeed identified the need for a better description of various acoustical aspects of the musical instruments and a more flexible learning environment. However, the observed recognition performance together with the results of both feature and error analyses indicate that not the applied techniques of pattern recognition are the primal source of error, but the data representation itself. That is to say, given the perfect representation, we should be able to increase the performance of the current system to a great extent.

4. Context, in general, plays a pervasive role for recognition systems. Even if the presented approach incorporates musical context only in a very rough manner, we could show very promising directions (see also the work of Barbedo & Tzanetakis (2011), where the contextual

analysis is simplistically incorporated via successive majority votes). Moreover, the results of Chapter 6 suggest that yet a much broader context is needed for an in-depth description of music in terms of musical instruments.

5. The evident recurrence of musical instruments inside the musical compositions requires much more attraction of interest in the algorithmic processing (see again the results presented in Chapter 5 and by Barbedo & Tzanetakis (2011)). Given the conventions of Western music, it is far more likely that an already identified instrumentation continues playing than the occurrence of a sudden major change therein. Hence, knowing where the instrumentation is changing is much more important than the knowledge of the entire instrumentation in each analysis frame. A subsequent label extraction can than be mainly guided by probabilistic inference inside regions of persistent timbre.

6. There is no universal approach towards an instrumentation analysis for Western music pieces. Our results suggest that, although the phrase-level instrument recognition itself has shown to be robust across different genres, different types of music require specialised algorithms to analyse their timbral properties. This is apparent from the outcomes of Section 5.2.3, where the proposed timbre analysis by means of segmentation and clustering of MFCCs showed good performance on structured rock, pop, or jazz music, but failed on pieces from classical music. This further indicates that we have not yet fully understood the underlying processes of music, here especially timbre, in order to describe it in a way for a reliable exploitation of its characteristics to infer higher-level musical concepts (McAdams, 1999).

7. A meaningful description of music in terms of instrumentation, with an envisioned application in MIR systems, goes far beyond the here-presented. One key aspect still remains the identification of the user's need – maybe the most important aspect in our understanding of music. A successful recognition system then fully adapts to this need to return valuable information.

7.3 Pending problems and future perspectives

It goes without saying that the approaches presented in this thesis only represent the beginning in an exhaustive research line towards automatic source recognition from complex auditory scenes. Moreover, many initial goals of this work have only been partially met and the amount of research questions regarding the topic has merely increased than declined. Remarkably, many of the subsequently listed yet appeared in the respective section in Martin's thesis (1999), more than 10 years ago. However, we here identify several (still-) open issues and point towards possible answers for their handling in forthcoming studies.

From our viewpoint, the main effort of future approaches has to be taken to understand, from a signal processing point-of-view, the complex auditory scene. We have presented evidence – along with numerous studies from related literature (Essid et al., 2006a; Fuhrmann et al., 2009a; Haro

& Herrera, 2009; Little & Pardo, 2008; Martin, 1999) – that the recognition process itself can yet be performed in an accurate, reliable, generalising, scalable, and efficient manner, even from complex, i.e. non-monophonic and polytimbral, data. Hence, most unsolved issues originate from the front-end processing of recognition systems for multi-source audio signals. We therefore see a strong need to develop a deeper understanding of complex auditory scenes and its perception, and its incorporation into the algorithmic architecture. More precisely, source recognition from polytimbral data includes – per definition – auditory scene analysis (ASA). Thus, except in some very rare cases, which can mostly be simulated under laboratory conditions, these two areas are inextricable. Therefore, one has to approach both in order to achieve an accurate solution for the problem, e.g. a human-comparable recognition performance. The here-presented approach – mainly applying techniques originating from MIR-related research – only represents a single building block of a complete recognition system, and has to be complemented by algorithms that analyse the auditory scene more in detail.

From a perceptual point-of-view, much of the recognition process is assumed to be based on inference from prior knowledge (Martin, 1999; von Helmholtz, 1954), a process which is only partially understood in general hearing research. Here, we again want to emphasise the importance of top-down control and specifically musical expectations, as shown by Ellis (1996) and, more recently, Cont (2008) and Hazan (2010), which are essential parts of human auditory processing. Hence, modelling these musical expectations can be accomplished in a fully probabilistic architecture and should play a key role in future recognition system. In a much broader CASA sense, the derived representations then serve as additional, high-level timbral cues (i.e. the identity of the specific acoustic sources) in the general hypotheses management system for auditory scene analysis.

The importance of a given instrument's predominance inside a mixture for its successful recognition is one key finding of the presented work. Since the auditory system similarly extracts high-level information from reliable portions of the incoming signal while it infers missing parts from contextual or prior knowledge (cf. Warren, 1970), an explicit location of short-term predominant signal parts seems to be essential for improving recognition performance from mixtures. Hence, automatic musical instrument recognition from polytimbral music signals should be based on the analysis of a single instrument in both the spectral and temporal dimension, at which multiple instruments can be identified sequentially. Consequentially, dissolving a single source from the mixture for recognition seems to be more appropriate than a separation of all containing sources (see, e.g. (Durrieu et al., 2009; Lagrange et al., 2008) for some work on this topic). The resulting signal can then be recognised using standard pattern recognition. Here, we can draw the connection to the attention mechanisms of the human auditory system which enable the listener to focus on a specific source in the incoming sensory data. Hence, the aim is to attenuate concurrent sounds while preserving the characteristics of the target source for a reliable feature extraction (cf. a typical foreground-background modelling paradigm). Moreover, information of already detected sources can then be incorporated in the scene analysis process. This further improves the representation of the target inside the mixture while disregarding potentially ambiguous portions of the signal. In this regard, and to connect the last three paragraphs, speech processing research can provide a good starting point for constructing such flexible recognition systems, incorporating both bottom-up and top-down schemes together with the aforesaid auditory attention mechanisms (e.g. Barker et al., 2010).

From an engineering viewpoint, we identify the possibilities of a signal-adaptive estimation of the acoustical units instead of the here-applied fixed-length paradigm, e.g. the time span of several tacti could be used to comply with the phrase-level paradigm for source recognition. An analysis of changes in the instrumentation, based on the overall timbre and an estimation of the number of concurrent sources, inside the entire signal can further help to improve the recognition performance.

In a more general sense, we see the need for constructing flexible, multi-hierarchical recognition systems, which adapt to the context at hand, in order to develop descriptive algorithms[1]. In particular, multiple overlapping taxonomies covering different levels in the hierarchical representation of musical instruments would be needed to extract a detailed description of the instrumentation from a given music piece (see also the results and conclusions of Chapter 6). Here, a successful recognition system would both require general broad taxonomies at the upper level of the instrumental hierarchy to perform general categorisation tasks and very specialised, fine grained taxonomies to adapt to the musical context at hand for a detailed description of the music in terms of the involved musical instruments. Hence, also more general contextual information has to be involved in the recognition process; detailed genre information, a particular playing style, or even the name of the analysed musical composition can serve as the cue for the selection of the proper taxonomy. The involved recognition models can then be specialised by the incorporation of context-aware feature selection and parameter tuning.

Finally, future recognition systems have to adopt more general flexible learning mechanisms. The aforementioned taxonomies are by no means of a static kind; they usually evolve in time since new categories arise from the data while models of already existing ones keep continuously updated in terms of the underlying training data and the respective model parameters. Automatic identification of musical instruments therefore calls for semi-supervised learning concepts with an active involvement of expert, i.e. human, teachers to prevent incorrect or inaccurate machine knowledge. This can also be viewed from a perceptual viewpoint as for the human mind learning represents a live-long, context-adaptive process.

7.4 Concluding remarks

Not for no reason does this dissertation start, in Chapter 2, with an overview of human auditory perception and cognition. One main conclusion of this work is that these principles are indispensable for successful automatic recognition systems, whether they model them explicitly or just borrow key components. Since human auditory perception and cognition is, after all, our touchstone for this domain, future approaches should incorporate the principles the human brain uses to recognise sound sources in complex acoustic scenes. Hence, from our perspective, psychoacoustics and auditory scene analysis is the right starting point for forthcoming studies. As already stated in the

[1] It is by way not informative and only somewhat descriptive to recognise an electric guitar from a rock piece.

previous section the main effort has to be taken on the processing of multiple simultaneous sound sources up to the level where the actual categorisation is performed, e.g. in a CASA framework.

This thesis has also shown that the automatic recognition of musical instruments is still a very active field of research in MIR, here deductible from the amount of works reviewed in Section 3.5 – and, of course, by the amount of works that has been discarded due to prior exclusion. Recent approaches pay more and more effort to process complex audio data, in our literature survey we could spot several approaches which work solely on real-world music signals (in fact a logical practice that should be requisite for all future studies). This is even more remarkable since at the beginning of the here-presented research, the author was not aware of any study dealing with complex input data of this kind. More precisely, this thesis started in 2007, hence all of the comparative approaches presented in the discussion of Chapter 4 originate after this date. This documents both the steady improvement of the algorithms towards the recognition of sound sources from complex mixtures (approaching the cocktail party!) and the merit and contributions of the insights gained from the previous research works. We hope that the work presented in this thesis is in line with these considerations and thus play its role in the steady improvement of musical instrument recognition approaches, providing the cornerstones for the next generations of approaches towards the problem. Finally we hope to also contribute to the overall scientific goal of a thorough understanding of human abilities to process and resolve complex auditory scenes.

In this light we encourage further comparative research in the field by publishing the data used to construct and evaluate the different modules of the presented approaches. In particular, in the course of this thesis we designed two complete datasets for research on automatic recognition of musical instruments from music audio signals. The complete package along with a list of the corresponding audio tracks can be found under `http://www.dtic.upf.edu/~ffuhrmann/PhD/data`.

Ferdinand Fuhrmann, Barcelona, 30th January 2012

Bibliography

Abdallah, S. & Plumbley, M. (2004). Polyphonic music transcription by non-negative sparse coding of power spectra. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 318–325.

Agostini, G., Longari, M., & Pollastri, E. (2003). Musical instrument timbres classification with spectral features. *EURASIP Journal on Applied Signal Processing*, *2003*(1), 5–14.

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Lecture Notes in Computer Science*, *3201*, 39–50.

Akkermans, V., Serrà, J., & Herrera, P. (2009). Shape-based spectral contrast descriptor. In *Proceedings of the Sound and Music Conference (SMC)*, pp. 143–148.

Alluri, V. & Toiviainen, P. (2009). Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre. *Music Perception*, *27*(3), 223–241.

Alluri, V. & Toiviainen, P. (in Press). Cross-cultural regularities in polyphonic timbre perception. *Music Perception*.

Aucouturier, J. (2006). *Ten experiments on the modelling of polyphonic timbre*. Ph.D. thesis, University of Paris VI.

Aucouturier, J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In J. Minett & W. Wang (Eds.) *Language, evoluation and the brain*, pp. 35–64. City University of Hong Kong Press.

Aucouturier, J. & Pachet, F. (2003). Representing Musical Genre: A State of the Art. *Journal of New Music Research*, *32*(1), 83–93.

Aucouturier, J. & Pachet, F. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, *1*(1).

Aucouturier, J. & Pachet, F. (2007). The influence of polyphony on the dynamical modelling of musical timbre. *Pattern Recognition Letters*, *28*(5), 654–661.

Aucouturier, J., Pachet, F., & Sandler, M. (2005). The way it Sounds: Timbre models for analysis and retrieval of music signals. *IEEE transactions on multimedia*, *7*(6), 1028–1035.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

Barbedo, J. & Tzanetakis, G. (2011). Musical Instrument Classification using individual partials. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 111–122.

Barker, J., Ma, N., Coy, A., & Cooke, M. (2010). Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Computer Speech and Language*, *24*(1), 94–111.

Beauchemin, M. & Thomson, K. (1997). The evaluation of segmentation results and the overlapping area matrix. *International Journal of Remote Sensing*, *18*(18), 3895–3899.

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., & D'Adamo, D. A. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 159–171.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, pp. 97–110.

Bonada, J. (2008). *Voice processing and synthesis by performance sampling and spectral models*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Bregman, A. (1990). *Auditory scene analysis*. The perceptual organization of sound. Cambridge, USA: MIT Press.

Broadbent, D. (1958). *Perception and communication*. Oxford University Press.

Brossier, P. (2006). *Automatic annotation of musical audio for interactive applications*. Ph.D. thesis, Queen Mary University, London.

Brown, G. & Cooke, M. (1994). Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research*, *23*(2), 107–132.

Brown, J. (1991). Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America (JASA)*, *89*(1), 425–434.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *1998*(2), 121–167.

Burred, J. (2009). *From sparse models to timbre learning: New methods for musical source separation*. Ph.D. thesis, Berlin University of Technology.

Burred, J., Robel, A., & Sikora, T. (2010). Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 663–674.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America (JASA)*, *118*(1), 471–482.

Cambouropoulos, E. (2009). How similar is similar? *Musicæ Scientiæ, Discussion Forum 4B*, pp. 7–24.

Carlyon, R. (2004). How the brain separates sounds. *Trends in Cognitive Sciences*, *8*(10), 465–471.

Casey, M. (1998). *Auditory group theory with applications to statistical basis methods for structured audio*. Ph.D. thesis, Massachusetts Institute of Technology (MIT), MA, USA.

Casey, M. & Slaney, M. (2006). The importance of sequences in musical similarity. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5–8.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696.

Castleman, K. (1996). *Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 2nd edn.

Celma, O. & Serra, X. (2008). Foafing the music: Bridging the semantic gap in music recommendation. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(4), 250–256.

Cemgil, A. & Gürgen, F. (1997). Classification of musical instrument sounds using neural networks. In *Proceedings of the IEEE Signal Processing and Communication Applications Conference (SIU)*.

Charbonneau, G. R. (1981). Timbre and the perceptual effects of three types of data reduction. *Computer Music Journal*, *5*(2), 10–19.

Chen, S., Donoho, D., & Saunders, M. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on scientific Computing*, *20*(1), 33–61.

Chen, S. & Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription & Understanding Workshop*, pp. 127–132.

Cherry, C. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America (JASA)*, *25*(5), 975–979.

Cont, A. (2008). *Modeling musical anticipation: From the time of music to the music of time*. Ph.D. thesis, University of California and University of Pierre et Marie Curie, San Diego and Paris.

Cont, A., Dubnov, S., & Wessel, D. (2007). Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 85–92.

Cook, P. (1999). Music, cognition, and computerized sound. MIT Press.

Cooke, M. (1993). Modelling auditory processing and organisation. Cambridge University Press.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, *11*(6), 1269–1275.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cover, T. M. & Thomas, J. A. (2006). *Elements of information theory*. Hoboken, NJ, USA: John Wiley and Sons, 2nd edn.

Crawley, E., Acker-Mills, B., Pastore, R., & Weil, S. (2002). Change detection in multi-voice music: The role of musical structure, musical training, and task demands. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 367–378.

Crummer, G., Walton, J., Wayman, J., Hantz, E., & Frisina, R. (1994). Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *The Journal of the Acoustical Society of America (JASA)*, *95*(5), 2720–2727.

Deliege, I. (2001). Similarity Perception - Categorization - Cue Abstraction. *Music Perception*, *18*(3), 233–243.

Downie, S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(4), 247–255.

Downie, S., Byrd, D., & Crawford, T. (2009). Ten years of ISMIR: Reflections on challenges and opportunities. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 13–18.

Drexler, E. (2009). The antiparallel structures of science and engineering. http://metamodern.com/2009/06/22/the-antiparallel-structures-of-science-and-engineering/, accessed Nov. 2011.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Wiley-Interscience, 2nd edn.

Durrieu, J., Ozerov, A., & Févotte, C. (2009). Main instrument separation from stereophonic audio signals using a source/filter model. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 15–19.

Eck, D., Lamere, P., Bertin-Mahieux, T., & Green, S. (2008). Automatic generation of social tags for music recommendation. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.) *Advances in neural information processing systems*, pp. 1–8. MIT Press.

Eggink, J. & Brown, G. (2003). A missing feature approach to instrument identification in polyphonic music. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 553–556.

Eggink, J. & Brown, G. (2004). Instrument recognition in accompanied sonatas and concertos. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 217–220.

Ellis, D. (1996). *Prediction-driven computational auditory scene analysis*. Ph.D. thesis, Massachusetts Institute of Technology (MIT), MA, USA.

Ellis, D. (2010). A history and overview of machine listening. http://www.ee.columbia.edu/~dpwe/talks/gatsby-2010-05.pdf, accessed Oct. 2011.

Eronen, A. (2001). *Automatic musical instrument recognition*. Master's thesis, Tampere University of Technology.

Eronen, A. (2003). Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proceedings of the IEEE International Symposium on Signal Processing and its Applications*, pp. 133–136.

Essid, S., Leveau, P., Richard, G., Daudet, L., & David, B. (2005). On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments. *Proceedings of the Audio Engineering Society (AES) Convention*.

Essid, S., Richard, G., & David, B. (2006a). Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 68–80.

Essid, S., Richard, G., & David, B. (2006b). Musical instrument recognition by pairwise classification strategies. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1401–1412.

Fan, R. & Lin, C. (2007). A study on threshold selection for multi-label classification. Tech. rep.

Fant, G. (1974). *Speech sounds and features*. Cambridge, USA: MIT Press.

Fayyad, U. & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Articial Intelligence*, pp. 1022–1027.

Fletcher, N. & Rossing, T. (1998). *The physics of musical instruments*. New York: Springer, 2nd edn.

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 452–455.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Data mining, inference, and prediction. New York: Springer.

Fuhrmann, F., Haro, M., & Herrera, P. (2009a). Scalability, generality and temporal aspects in the automatic recognition of predominant musical instruments in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 321–326.

Fuhrmann, F. & Herrera, P. (2010). Polyphonic instrument recognition for exploring semantic similarities in music. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 281–288.

Fuhrmann, F. & Herrera, P. (2011). Quantifying the relevance of locally extracted information for musical instrument recognition from entire pieces of music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 239–244.

Fuhrmann, F., Herrera, P., & Serra, X. (2009b). Detecting Solo Phrases in Music using spectral and pitch-related descriptors. *Journal of New Music Research*, *38*(4), 343–356.

Gibson, J. (1950). The perception of the visual world. Houghton Mifflin.

Gillet, O. & Richard, G. (2006). Enst-drums: an extensive audio-visual database for drum signals processing. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 156–159.

Gillet, O. & Richard, G. (2008). Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(3), 529–540.

Godsmark, D. & Brown, G. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication*, *27*(3-4), 351–366.

Gómez, E., Ong, B., & Herrera, P. (2006). Automatic tonal analysis from music summaries for version identification. In *Audio Engineering Society (AES) Convention*.

Goodwin, M. (1997). *Adaptive signal models: Theory, algorithms, and audio applications*. Ph.D. thesis, University of California, Berkeley.

Goto, M. (2004). A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, *43*(4), 311–329.

Goto, M. (2006). A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1783–1794.

Goto, M., Hashiguchi, H., & Nishimura, T. (2003). RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 229–230.

Gouyon, F. (2005). *A computational approach to rhythm description: Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Gouyon, F. & Herrera, P. (2001). Exploration of techniques for automatic labeling of audio drum tracks instruments. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*.

Gouyon, F., Herrera, P., Gómez, E., Cano, P., Bonanda, J., Loscos, A., Amatrian, X., & Serra, X. (2008). Content processing of music audio signals. In P. Polotti & D. Rocchesso (Eds.) *Sound to sense, sense to sound: A state of the art in sound and music computing*, pp. 83–160. Logos.

Grey, J. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America (JASA)*, *61*(5), 1270–1277.

Grey, J. (1978). Timbre discrimination in musical patterns. *Journal of the Acoustical Society of America (JASA)*, *64*(2), 467–472.

Guaus, E. (2009). *Audio content processing for automatic music genre classification*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Hajda, J., Kendall, R., & Carterette, E. (1997). Methodological issues in timbre research. In I. Deliege & J. Sloboda (Eds.) *Perception and cognition of music*. Psychology Press.

Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations*, *11*(1), 10–18.

Handel, S. (1995). Timbre perception and auditory object identification. In B. Moore (Ed.) *Hearing*, pp. 425–461. New York: Academic Press.

Handel, S. & Erickson, M. (2004). Sound source identification: The possible role of timbre transformations. *Music Perception*, *21*(4), 587–610.

Hargreaves, D. & North, A. (1999). The functions of music in everyday life: Redefining the social in music psychology. *Psychology of Music*, *27*, 71–83.

Haro, M. (2008). *Detecting and describing percussive event in polyphonic music*. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Haro, M. & Herrera, P. (2009). From low-level to song-level percussion descriptors of polyphonic music. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 243–248.

Hastie, T. & Tibshirani, R. (1998). Classification by pairwise coupling. *Annals of Statistics*, *26*(2).

Hawley, M. (1993). *Structure out of sound*. Ph.D. thesis, Massachusetts Institute of Technology (MIT), MA, USA.

Hazan, A. (2010). *Musical expectation modelling from audio: A causal mid-level approach to predictive representation and learning of spectro-temporal events*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Heittola, T., Klapuri, A., & Virtanen, T. (2009). Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 327–332.

Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2000). Towards instrument segmentation for music content description: A critical review of instrument classification techniques. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

Herrera, P., Dehamel, A., & Gouyon, F. (2003). Automatic labeling of unpitched percussion sounds. In *Audio Engineering Society (AES) Convention*.

Herrera, P., Klapuri, A., & Davy, M. (2006). Automatic classification of pitched musical instrument sounds. In A. Klapuri (Ed.) *Signal processing methods for automatic music transcription*, pp. 163–200. Springer.

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

Hsu, C., Chang, C., & Lin, C. (2003). A practical guide to support vector classification. Tech. rep.

Huron, D. (1989). Voice denumerability in polyphonic music of homogeneous timbres. *Music Perception*, *6*(4), 361–382.

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, *19*(1), 1–64.

Huron, D. (2006). *Sweet anticipation*. Music and the psychology of expectation. Cambridge, USA: MIT Press.

Iverson, P. & Krumhansl, C. (1991). Measuring similarity of musical timbres. *Journal of the Acoustical Society of America (JASA)*, *89*(4B), 1988.

Jain, A., Duin, R., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Janer, X. (2007). *A BIC-based approach to singer identification*. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Jensen, J., Christensen, M., Ellis, D., & Jensen, S. (2009). Quantitative analysis of a common audio similarity measure. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 693–703.

Joder, C., Essid, S., & Richard, G. (2009). Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(1), 174–186.

Jordan, A. (2007). *Akustische Instrumentenerkennung unter Berücksichtigung des Einschwingvorganges, der Tonlage und der Dynamik*. Master's thesis, University of Music and Performing Arts Vienna, Austria.

Kaminsky, I. & Materka, A. (1995). Automatic source identification of monophonic musical instrument sounds. In *Proceedings of the IEEE International Conference on Neural Networks*, pp. 189–194.

Kassler, M. (1966). Toward musical information retrieval. *Perspectives of New Music*, *4*(2), 59–67.

Kendall, R. (1986). The role of acoustic signal partitions in listener categorization of musical phrases. *Music Perception*, *4*(2), 185–213.

Kendall, R. & Carterette, E. (1993). Identification and blend of timbres as a basis for orchestration. *Contemporary Music Review*, *9*(1), 51–67.

Kitahara, T., Goto, M., Komatani, K., Ogata, T., & Okuno, H. (2006). Instrogram: A new musical instrument recognition technique without using onset detection nor f0 estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 229–232.

Kitahara, T., Goto, M., Komatani, K., Ogata, T., & Okuno, H. (2007). Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps. *EURASIP Journal on Advances in Signal Processing*, *2007*, 1–16.

Klapuri, A. (2003). *Signal processing methods for the automatic transcription of music*. Ph.D. thesis, Tampere University of Technology.

Kobayashi, Y. (2009). Automatic generation of musical instrument detector by using evolutionary learning method. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 93–98.

Kölsch, S. & Siebel, W. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, *9*(12), 578–584.

Krippendorff, K. (2004). *Content analysis*. An introduction to its methodology. London, UK: Sage Publications, 2nd edn.

Krumhansl, C. (1991). Music Psychology: Tonal Structures in Perception and Memory. *Annual Review of Psychology*, *42*, 227–303.

Kuncheva, L. (2004). *Combining pattern classifiers*. Methods and algorithms. Hoboken, NJ, USA: Wiley-Interscience.

Lagrange, M., Martins, L. G., Murdoch, J., & Tzanetakis, G. (2008). Normalized cuts for predominant melodic source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 278–290.

Lagrange, M., Raspaud, M., Badeau, R., & Richard, G. (2010). Explicit modeling of temporal dynamics within musical signals for acoustical unit similarity. *Pattern Recognition Letters*, *31*, 1498–1506.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, *62*(7), 1426–1439.

Langley, P. (1996). *Elements of machine learning*. San Francisco, CA: Morgan Kaufmann.

Laurier, C. (2011). *Automatic classification of musical mood by content-based analysis*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Laurier, C., Meyers, O., Serrà, J., Blech, M., Herrera, P., & Serra, X. (2010). Indexing music by mood: design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, *48*(1), 161–184.

Leman, M. (2003). Foundations of musicology as content processing science. *Journal of Music and Meaning*, *1*.

Leveau, P., Sodoyer, D., & Daudet, L. (2007). Automatic instrument recognition in a polyphonic mixture using sparse representations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 233–236.

Leveau, P., Vincent, E., Richard, G., & Daudet, L. (2008). Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 116–128.

Levitin, D. (2008). *This Is Your Brain on Music*. The science of a human obsession. London, UK: Atlantic Books.

Levy, M., Sandier, M., & Casey, M. (2006). Extraction of high-level musical structure from audio data and its application to thumbnail generation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1316–1319.

Li, T. & Ogihara, M. (2005). Music genre classification with taxonomy. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 197–200.

Licklider, J. C. R. (1951). Basic correlates of the auditory stimulus. In S. Stevens (Ed.) *Handbook of experimental psychology*, pp. 985–1035. New York: Wiley.

Lincoln, H. (1967). Some criteria and techniques for developing computerized thematic indices. In H. Heckman (Ed.) *Elektronische Datenverarbeitung in der Musikwissenschaft*. Bosse.

Little, D. & Pardo, B. (2008). Learning musical instruments from mixtures of audio with weak labels. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 127–132.

Liu, R. & Li, S. (2009). A review on music source separation. In *IEEE Youth Conference on Information, Computing and Telecommunication (YC-ICT)*, pp. 343–346.

Livshin, A. & Rodet, X. (2003). The importance of cross database evaluation in sound classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

Livshin, A. & Rodet, X. (2004). Musical instrument identification in continuous recordings. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 222–226.

Livshin, A. & Rodet, X. (2006). The significance of the non-harmonic "noise" versus the harmonic series for musical instrument recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 95–100.

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

Lohr, S. (2009). *Sampling*. Design and analysis. Boston, MA: Brooks/Cole, 2nd edn.

Loui, P. & Wessel, D. (2006). Acquiring new musical grammars: a statistical learning approach. In *Proceedings of the International Conference on Music Perception and Cognition*, pp. 1009–1017.

Lu, L., Wang, M., & Zhang, H. (2004). Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 275–282. New York, New York, USA: ACM Press.

Lu, L., Zhang, H., & Li, S. (2003). Content-based audio classification and segmentation by using support vector machines. *Multimedia systems*, *8*(6), 482–492.

Lufti, R. (2008). Human sound source identification. In W. Yost, A. Popper, & R. Fay (Eds.) *Auditory perception of sound sources*, pp. 13–42. New York: Springer.

Mallat, S. (1999). A wavelet tour of signal processing. Academic Press.

Mallat, S. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, *41*(12), 3397–3415.

Mandel, M. & Ellis, D. (2005). Song-level features and support vector machines for music classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 594–599.

Manning, C., Raghavan, P., & Schütze, H. (2009). Introduction to information retrieval. `http://nlp.stanford.edu/IR-book/`.

Marozeau, J., de Cheveigne, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America (JASA)*, *114*(5), 2946–2957.

Marr, D. (1982). *Vision*. A computational investigation into the human representation and processing of visual information. W. H. Freeman & Co.

Martin, K. (1999). *Sound-source recognition: A theory and computational model*. Ph.D. thesis, Massachusetts Institute of Technology (MIT), MA, USA.

Martin, K., Scheirer, E., & Vercoe, B. (1998). Music content analysis through models of audition. In *Proceedings of the ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*.

Mauch, M., Fujihara, H., Yoshii, K., & Goto, M. (2011). Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

Mauch, M., Noland, K., & Dixon, S. (2009). Using musical structure to enhance automatic chord transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 231–236.

McAdams, S. (1993). Recognition of auditory sound sources and events. In S. McAdams & E. Bigand (Eds.) *Thinking in sound: The cognitive psychology of human audition*. Oxford University Press.

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23*(3), 85–102.

McAdams, S. & Cunible, J. (1992). Perception of timbral analogies. *Philosophical Transactions: Biological Sciences*, *336*, 383–389.

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*(3), 177–192.

McAulay, R. & Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34*(4), 744–754.

McKay, C. & Fujinaga, I. (2005). Automatic music classification and the importance of instrument identification. In *Proceedings of the Conference on Interdisciplinary Musicology*.

McKay, C. & Fujinaga, I. (2010). Improving automatic music classification performance by extracting features from different types of data. In *Proceedings of the international Conference on Multimedia Information Retrieval (MIR)*, pp. 257–266. ACM Press.

Meng, A., Ahrendt, P., Larsen, J., & Hansen, L. K. (2007). Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1654–1664.

Meyer, J. (2009). *Acoustics and the Performance of Music*. Manual for acousticians, audio engineers, musicians, architects and musical instrument makers. New York: Springer, 5th edn.

Meyer, L. (1956). *Emotion and meaning in music*. University of Chicago Press.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*, 81–97.

Minsky, M. (1988). *The society of mind*. New York: Simon & Schuster.

Moore, B. (1989). An introduction to the psychology of hearing. Academic Press, London.

Moore, B. (1995). Frequency analysis and masking. In B. Moore (Ed.) *Hearing*, pp. 161–200. New York: Academic Press.

Moore, B. (2005a). Basic auditory processes. In B. Goldstein, G. Humphreys, M. Shiffrar, & W. Yost (Eds.) *Blackwell handbook of sensation and perception*, pp. 379–407. Malden, USA: Wiley-Blackwell.

Moore, B. (2005b). Loudness, pitch and timbre. In B. Goldstein, G. Humphreys, M. Shiffrar, & W. Yost (Eds.) *Blackwell handbook of sensation and perception*, pp. 408–436. Malden, USA: Wiley-Blackwell.

Moorer, J. (1975). *On the segmentation and analysis of continuous musical sound by digital computer*. Ph.D. thesis, Stanford University.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures*. The implication-realization model. University of Chicago Press.

Neuringer, A. & Porter, D. (1984). Music discrimination by pigeons. *Journal of Experimental Psychology: Animal Behavioral Processes*, *10*, 138–148.

Nielsen, A., Sigurdsson, S., Hansen, L., & Arenas-García, J. (2007). On the relevance of spectral features for instrument classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 485–488.

Ohm, G. (1873). Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. *Annalen der Physik und Chemie*, *59*, 513–565.

Olson, H. (1967). *Music, physics and engineering*. Courier Dover Publications.

Ong, B., Gómez, E., & Streich, S. (2006). Automatic extraction of musical structure using pitch class distribution features. In *Workshop on Learning the Semantics of Audio Signals*, pp. 53–65.

Orio, N. (2006). Music retrieval: a tutorial and review. *Foundations and Trends in Information Retrieval*, *1*(1), 1–90.

Ortiz, A. & Oliver, G. (2006). On the use of the overlapping area matrix for image segmentation evaluation: A survey and new performance measures. *Pattern Recognition Letters*, *27*(2006), 1916–1926.

Pampalk, E., Flexer, A., & Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 628–633.

Panagakis, Y. & Kotropoulos, C. (2009). Music genre classification using locality preserving non-negative tensor factorization and sparse representations. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 249–254.

Patel, A. (2007). *Music, language, and the brain*. Oxford University Press.

Paulus, J. & Klapuri, A. (2009). Drum sound detection in polyphonic music with hidden markov models. *EURASIP Journal on Audio, Speech, and Music Processing, 2009*, 1–9.

Paulus, J., Müller, M., & Klapuri, A. (2010). Audio-based music structure analysis. *Proceedings of the 11th International Society for Music Information Retrieval Conference*.

Peeters, G. (2003). Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Audio Engineering Society (AES) Convention*.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. rep.

Pei, S. & Hsu, N. (2009). Instrumentation analysis and identification of polyphonic music using beat-synchronous feature integration and fuzzy clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 169–172.

Peterschmitt, G., Gómez, E., & Herrera, P. (2001). Pitch-based solo location. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, pp. 239–243.

Piccina, L. (2009). *An algorithm for solo detection using multifeature statistics*. Master's thesis, Politecnico di Milano.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, MIT Press*.

Plomp, R. (1964). The Ear as a Frequency Analyzer. *Journal of the Acoustical Society of America (JASA)*, *36*(9), 1628–1936.

Plomp, R. & Levelt, W. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America (JASA)*, *38*(4), 548–560.

Plomp, R. & Mimpen, A. (1968). The Ear as a Frequency Analyzer II. *Journal of the Acoustical Society of America (JASA)*, *43*(4), 764–767.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1992). *Numerical recipes in C*. The art of scientific computing. Cambridge University Press, 2nd edn.

Rabiner, L. & Juang, B. (1993). *Fundamentals of speech recognition*. New York: Prentice Hall.

Reber, A. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863.

Reuter, C. (1997). Karl Erich Schumann's principles of timbre as a helpful tool in stream segregation research. In M. Leman (Ed.) *Music, Gestalt, and Computing - Studies in Cognitive and Systematic Musicology*, pp. 362–374. Springer.

Reuter, C. (2003). Stream segregation and formant areas. In *Proceedings of the European Society for the Cognitive Sciences of Music Conference (ESCOM)*, pp. 329–331.

Reuter, C. (2009). The role of formant positions and micro-modulations in blending and partial masking of musical instruments. *The Journal of the Acoustical Society of America*, *126*(4), 2237.

Robinson, D. & Dadson, R. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, *7*(5), 166–181.

Román, C. (2011). *Detection of genre-specific musical instruments: The case of the Mellotron*. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.) *Cognition and categorization*. Lawrence Erlbaum Associates.

Sadie, S. (1980). *The new Grove dictionary of music and musicians*. New York: Macmillan Press, 6th edn.

Saffran, J., Johnson, E., Aslin, R., & Newport, E. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Sandell, G. (1995). Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. *Music Perception*, *13*(2), 209–246.

Sandell, G. (1996). Identifying musical instruments from multiple versus single notes. *The Journal of the Acoustical Society of America (JASA)*, *100*(4), 2752.

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: A survey. *IEEE Signal Processing Magazine*, *23*(2), 133–141.

Scheirer, E. (1996). Bregman's chimerae: Music perception as auditory scene analysis. In *Proceedings of the International Conference on Music Perception and Cognition*.

Scheirer, E. (1999). Towards music understanding without separation: Segmenting music with correlogram comodulation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 99–102.

Scheirer, E. (2000). *Music-listening systems*. Ph.D. thesis, Massachusetts Institute of Technology (MIT), MA, USA.

Scheirer, E. & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1331–1334.

Schouten, J. (1970). The residue revisited. In R. Plomp & G. Smoorenburg (Eds.) *Frequency analysis and periodicity detection in hearing*. Sijthoff.

Schumann, E. (1929). *Die Physik der Klangfarben*. Berlin: Humboldt University.

Schwarz, D. (1998). *Spectral envelopes in sound analysis and synthesis*. Master's thesis, Universität Stuttgart.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Serrà, J., Gómez, E., Herrera, P., & Serra, X. (2008). Statistical analysis of chroma features in western music predicts human judgments of tonality. *Journal of New Music Research*, *37*(4), 299–309.

Serra, X. (1989). *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. Ph.D. thesis, Stanford University.

Shepard, R. (1964). Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America (JASA)*, *36*(12), 2346–2353.

Shepard, R. & Jordan, D. (1984). Auditory illusions demonstrating that tones are assimilated to an internalized musical scale. *Science*, *226*(4680), 1333–1334.

Simmermacher, C., Deng, D., & Cranefield, S. (2006). Feature analysis and classification of classical musical instruments: an empirical study. *Lecture Notes in Computer Science*, *4065*, 444–458.

Singh, P. G. (1987). Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre? *Journal of the Acoustical Society of America (JASA)*, *82*(3), 886–899.

Slaney, M. (1995). A critique of pure audition. In D. Rosenthal & H. Okuno (Eds.) *Proceedings of the Computational Auditory Scene Analysis Workshop*, pp. 13–18. Erlbaum Associates Inc.

Sloboda, J. & Edworthy, J. (1981). Attending to two melodies at once: The of key relatedness. *Psychology of Music*, *9*(1), 39–43.

Smaragdis, P. & Brown, J. (2003). Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177–180.

Smaragdis, P., Shashanka, M., & Raj, B. (2009). A sparse non-parametric approach for single channel separation of known sounds. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*.

Smit, C. & Ellis, D. (2007). Solo voice detection via optimal cancellation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 207–210.

Southard, K. (2010). The paradox of choice, the myth of growth, and the future of music. `http://www.instantencore.com/buzz/item.aspx?FeedEntryId=137783`, accessed Nov. 2011.

Srinivasan, A., Sullivan, D., & Fujinaga, I. (2002). Recognition of isolated instrument tones by conservatory students. In *Proceedings of the International Conference on Music Perception and Cognition*, pp. 17–21.

Stevens, S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153–181.

Stevens, S. & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329–353.

Streich, S. (2006). *Music complexity: a multi-faceted description of audio content*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Streich, S. & Ong, B. (2008). A music loop explorer system. In *Proceedings of the International Computer Music Conference (ICMC)*.

Sundberg, J. (1987). *The science of the singing voice*. Northern Illinois University Press.

Temperley, D. (2004). *The cognition of basic musical structures*. Cambridge, USA: MIT Press.

Temperley, D. (2007). *Music and probability*. Cambridge, USA: MIT Press.

Tillmann, B. & McAdams, S. (2004). Implicit learning of musical timbre sequences: Statistical regularities confronted with acoustical (dis)similarities. *Journal of experimental psychology, learning, memory and cognition*, *30*(5), 1131–1142.

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 467–476.

Turnbull, D., Lanckriet, G., Pampalk, E., & Goto, M. (2007). A supervised approach for detecting boundaries in music using difference features and boosting. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 51–54.

Tzanetakis, G. (2002). *Manipulation, analysis and retrieval systems for audio signals*. Ph.D. thesis, Princeton University.

Tzanetakis, G. & Cook, P. (1999). Multifeature audio segmentation for browsing and annotation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 103–106.

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, pp. 1–10.

Vapnik, V. (1999). *The nature of statistical learning theory*. New York: Springer, 2nd edn.

Vincent, E., Jafari, M., Abdallah, S., Plumbley, M., & Davies, M. (2010). Probabilistic modeling paradigms for audio source separation. In W. Wang (Ed.) *Machine Audition: Principles, Algorithsm and Systems*. IGI Global.

Virtanen, T. (2006). *Sound source separation in monaural music signals*. Ph.D. thesis, Tampere University of Technology.

von Helmholtz, H. (1954). On the sensations of tone as a physiological basis for the theory of music. Dover, New York.

von Hornbostel, E. & Sachs, C. (1961). Classification of musical instruments (Translated from the original german by Anthony Baines and Klaus P. Wachsmann). *The Galpin Society Journal*, *14*, 3–29.

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.

Weintraub, M. (1986). *A theory and computational model of monaural sound separation*. Ph.D. thesis, Stanford University.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung*, *4*, 301–350.

Wessel, D. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3*(2), 45–52.

Wiggins, G. A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. *Proceedings of the IEEE International Symposium on Multimedia*, pp. 477–482.

Winkler, I., Kushnerenko, E., Horváth, J., Ceponiene, R., Fellman, V., Huotilainen, M., Näätänen, R., & Sussman, E. (2003). Newborn infants can organize the auditory world. In *Proceedings of the National Academy of Science (PNAS)*, pp. 11812–11815.

Witten, I. & Frank, E. (2005). *Data mining*. Practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann, 2nd edn.

Wu, M., Wang, D., & Brown, G. (2003). A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, *11*(3), 229–241.

Wu, T., Lin, C., & Weng, R. (2004). Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, pp. 975–1005.

Xu, R. & Wunsch, D. (2008). *Clustering*. Wiley - IEEE Press.

Yost, W. (2008). Perceiving sound sources. In W. Yost, A. Popper, & R. Fay (Eds.) *Auditory perception of sound sources*, pp. 1–12. New York: Springer.

Yu, G. & Slotine, J. (2009). Audio classification from time-frequency texture. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1677–1680.

Zwicker, E. & Terhardt, E. (1980). Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America (JASA)*, *68*(5), 1523–1525.

Appendices

A

Audio features

Here we provide the complete formulations of all audio features used in the modelling process. All mathematical considerations are derived from the respective references provided in Section 4.2.1. If not stated differently, X_i denotes the magnitude of the FFT bin i and N the total amount of bins resulting from a $2N + 1$-point FFT.

Bark energy bands. First, to map the frequency values in Hertz to the psycho-acoustic Bark scale, we use

$$bark = 13 \arctan(\frac{0.76}{1000} f) + 3.5 \arctan((\frac{1}{7500} f)^2).$$

For calculating the final energy values the magnitudes inside each band are squared and summed,

$$E_i = \sum_{j=f_{\text{start}}}^{f_{\text{end}}} X_j^2,$$

where E_i denotes the energy value of the i^{th} Bark band while f_{start} and f_{end} refer to the its start and end index in terms of FFT bins. In our specific implementation we use 26 bands ranging from 20 to 15 500 Hz. For convenience, Table A.1 lists these bands numbered by the applied indexing schema together with their corresponding frequency ranges.

Spectral energy. The spectral energy is given by the sum over all values of the power spectrum, i.e.

Index	Low	High	Index	Low	High
0	20	50	13	1 480	1 720
1	50	100	14	1 720	2 000
2	100	150	15	2 000	2 320
3	150	200	16	2 320	2 700
4	200	300	17	2 700	3 150
5	300	400	18	3 150	3 700
6	400	510	19	3 700	4 400
7	510	630	20	4 400	5 300
8	630	770	21	5 300	6 400
9	770	920	22	6 400	7 700
10	920	1 080	23	7 700	9 500
11	1 080	1 270	24	9 500	12 000
12	1 270	1 480	25	12 000	15 500

Table A.1: Indexing and frequency range [Hz] of Bark energy bands. In order to improve the feature's resolution at low frequencies, the first 4 bands are created by dividing the original 2 lowest bands. Note that all in-text references to individual Bark bands apply the here-presented indexing schema.

$$E = \sum_{i=1}^{N} X_i^2.$$

Mel Frequency Cepstral Coefficients (MFCCs). The computation of the MFCCs first involves a mapping of the frequency values from Hertz to Mel and a subsequent energy calculation inside each generated band (see above). To convert the frequencies we use

$$mel = \frac{1000}{\log_{10}(2)} \log_{10}[1 + \frac{f}{1000}].$$

After a logarithmic compression of the energy values the resulting signal is transformed into the cepstral domain via the Discrete Cosine Transform (DCT), which is defined as follows:

$$c[n] = 2 \sum_{k=0}^{N-1} X_k \cos(\frac{\pi n(2k+1)}{2N}), \qquad 0 \leq n \leq N-1,$$

where $c[n]$ denotes the n^{th} cepstral coefficient.

Spectral contrast and valleys. First, the raw spectral contrast and valleys features are computed for each considered frequency band separately,

$$C_k = (\frac{P_k}{V_k})^{1/\log(\mu_k)}, \quad \text{with} \quad \mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{k,i},$$

where P_k and V_k represent the description of the peaks and valleys in band k, while N_k connotes the number of FFT bins in the respective frequency band. It can be seen that the ratio of peaks and valleys is weighted by the shape of the band k, implemented by the mean peak value μ_k. The corresponding values for the peaks and valleys description are given by

$$P_k = \frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N_k} X_{k,i}, \quad \text{and} \quad V_k = \frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N_k} X_{k,N_k-i+1},$$

where α denotes the fraction of FFT bins from the ranked list of magnitude values to be used ($0 < \alpha \leq 1$). We use a value of 0.4 in our implementation. Finally, the respective C_k and V_k values are decorrelated by applying a local PCA, using the covariance matrix estimated from the all present frame observations. The resulting features represent the spectral contrast and valleys coefficients.

Linear Prediction Coefficients (LPC). Here, we concentrate on showing how the time-domain prediction coefficients can be regarded as a description of the signal's spectral envelope. First, in a LPC analysis, the signal's sample value $x[n]$ is extrapolated by using a weighted sum of the previous values of the signal,

$$x[n] = \sum_{i=1}^{p} a_i x[n-i],$$

where a_i represent the p prediction coefficients. The coefficients are estimated by minimising the respective error between the actual signal value and its extrapolation,

$$e[n] = x[n] - \sum_{i=1}^{p} a_i x[n-i]$$

By transforming this relation into the z-domain the process can be regarded as a filtering of the input signal x,

$$E(z) = (1 - \sum_{i=1}^{p} a_i z^{-i}) X(z),$$

where the term in brackets denotes the filter's transfer function. Furthermore, this transfer function can be used to minimise the error but also to synthesise the signal from the error given the coefficients, hence the resulting filters are given by

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \quad \rightarrow \quad S(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}.$$

It can be seen that the resulting synthesis filter $S(z)$ takes the form of an all-pole filter, since the function does not exhibit zeros in its numerator, but p zeros in the denominator, which come in complex-conjugate pairs since the a_is only take real values. The transfer function of this filter thus exhibits $p/2$ peaks which describe the spectral envelope of the signal. The actual computation of the coefficients a_i is implements by using the autocorrelation method, as described in the relevant literature.

Spectral centroid. The magnitude spectrum is regarded as a distribution, where the frequencies denote the distribution's values and its magnitudes its observation probabilities. The centroid describes the distribution's barycentre, hence

$$centroid = \frac{\sum_{i=1}^{N} X_i f_i}{\sum_{i=1}^{N} X_i},$$

where f_i represents the centre frequency value of FFT bin i.

Spectral spread. It describes the deviation of the spectral distribution from its mean, thus

$$spread = \frac{\sum_{i=1}^{N} X_i (f_i - \mu)^2}{\sum_{i=1}^{N} X_i},$$

where μ denotes the observed spectral distribution's mean, i.e. the centroid.

Spectral skewness. The spectral skewness is computed from the 3rd order moment, describing the global shape of the spectral distribution,

$$skewness = \frac{\sum_{i=1}^{N} X_i (f_i - \mu)^3}{(\sum_{i=1}^{N} X_i (f_i - \mu)^2)^{3/2}}.$$

Spectral kurtosis. The spectral kurtosis represents the 4th order moment, again describing global shape properties of the spectral distribution,

$$kurtosis = \frac{\sum_{i=1}^{N} X_i (f_i - \mu)^4}{(\sum_{i=1}^{N} X_i (f_i - \mu)^2)^2}.$$

Spectral decrease. It defines the decrease of magnitude values in the spectrum,

$$decrease = \frac{1}{\sum_{i=2}^{N} X_i} \sum_{i=2}^{N} \frac{X_i - X_1}{i - 1}.$$

Spectral flatness. It is defined by the ratio of the geometric and the arithmetic mean of the spectrum, here transformed into decibels,

$$\mathit{flatness}_{\mathrm{dB}} = 10\, log_{10}(\frac{(\prod_{i=1}^{N} X_i)^{1/N}}{\frac{1}{N}\sum_{i=1}^{N} X_i}).$$

Spectral crest. This feature describes the shape of the spectrum by relating the maximum to the mean magnitude. It is calculated by

$$\mathit{crest} = \frac{\max\limits_{i}(X_i)}{\frac{1}{N}\sum_{i=1}^{N} X_i)}.$$

Spectral flux. It is derived by comparing the spectra of the actual and the previous frame,

$$\mathit{flux} = ||X[n] - X[n-1]||_2,$$

where $X[n]$ denotes the magnitude spectrum at frame n and $||\cdot||_2$ the Euclidean norm.

Spectral roll-off. The 85 percentile of the power spectral distribution, that is, it is defined by the frequency below which 85% of the spectral energy lies,

$$\mathit{rolloff} = f_i, \quad \max_i \sum_{j=1}^{i} X_j^2 = 0.85 \sum_{j=1}^{N} X_j^2, i = 1:N$$

where i denotes the FFT bin index where the accumulated spectral energy reaches 85% of the total spectral energy.

High frequency content. A weighted energy calculation, as defined by

$$\mathit{hfc} = \sum_{i=1}^{N} iX_i.$$

Spectral strongpeak. The spectral strongpeak is calculated by dividing the spectrum's maximum by the width of this particular peak, i.e.

$$\mathit{strongpeak} = \frac{\max\limits_{i} X_i}{\log_{10}(k_{\mathrm{hi}} - k_{\mathrm{lo}})},$$

where k_{hi} and k_{lo} represent, respectively, the upper and lower FFT-bin index around the maximum peak of the spectrum where the respective magnitude reaches half the value of the maximum peak.

Spectral dissonance. Given the FFT bin indices of the peaks of a given spectrum, the dissonance is calculated as followed,

$$dissonance = \sum_{i=1}^{P} \sum_{j=1}^{P} (1 - c_{i,j}) X_j,$$

where P denotes the total number of peaks while the function $c_{i,j}$ represents a polynomial implementation of the obtained consonance curves in the original publication.

Spectral complexity. In our implementation the spectral complexity is defined by the number of spectral peaks present in an audio frame. Hence, we apply a peak detection algorithm to the input spectrum to derive the value of this audio feature.

Pitch confidence. Its value is derived from the depth of the deepest valley of the Yin_{FFT}-lag function. The descriptor is computed as followed,

$$pitch_{\mathrm{conf}} = 1 - \min_{\tau}(yin_{\mathrm{FFT}}(\tau)), \quad \text{with}$$

$$yin_{\mathrm{FFT}}(\tau) = \frac{4}{N} \sum_{i=1}^{N} X_i^2 - \frac{2}{N} \sum_{i=1}^{N} X_i^2 \cos(\frac{2\pi i \tau}{N}),$$

with τ denoting the time domain lag in samples.

Pitch salience. It is defined as the local maximum of the normalised autocorrelation function, i.e.

$$pitch_{\mathrm{sal}} = \max_{\tau} \frac{r_x(\tau)}{r_x(0)}, \quad \text{with} \quad r_x(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x[n]x[n+\tau], \quad 0 \leq \tau \leq M,$$

where M corresponds to the maximum time-shift in samples.

Inharmonicity. Given the harmonic components $h_i, i = 1 \ldots H$ of an estimated fundamental frequency f_0 in the signal, the inharmonicity is given by

$$inharmonicity = \frac{2}{f_0} \frac{\sum_{i=1}^{H} |f_{h_i} - i f_0| X_{h_i}^2}{\sum_{i=1}^{H} X_{h_i}^2},$$

where f_{h_i} and X_{h_i} denote, respectively, the value in frequency and magnitude of the FFT bin associated with the i^{th} harmonic.

Odd-to-even harmonic energy ratio. Similarly, the magnitudes of the harmonic components $h_i, i = 1 \dots H$ of an estimated f_0 are used to calculate the ratio between the odd and even harmonics, hence

$$odd2even = \frac{\sum_i X_{h_i}^2}{\sum_j X_{h_j}^2}, \quad i = 1, 3, 5, \dots, H, \quad j = 2, 4, 6, \dots, H.$$

Tristimuli. We derive 3 values for the tristimulus, which account for different energy ratios in the series of harmonics $h_i, i = 1 \dots H$ of an estimated f_0. Hence,

$$T_1 = \frac{X_{h_1}}{\sum_{i=1}^{H} X_{h_i}},$$

$$T_2 = \frac{X_{h_2} + X_{h_3} + X_{h_4}}{\sum_{i=1}^{H} X_{h_i}},$$

$$T_3 = \frac{\sum_{i=5}^{H} X_{h_i}}{\sum_{i=1}^{H} X_{h_i}}.$$

Zero crossing rate. The temporal zero crossing rate is given by

$$zcr = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|,$$

where $x[n]$ represents the time domain signal and N its length in samples.

B

Evaluation collection

Here we provide the complete list of all music pieces used in the evaluation collection described in Section 4.3.2. Table B.1 shows the metadata (artist, album artist, album title, track number, and track title) of each track along with the musical genre and the annotated instrumentation. The corresponding annotation files can be obtained from http://www.dtic.upf.edu/~ffuhrmann/PhD/data.

Artist	Album Artist	Album	Track	Title	Genre	Annotation
10cc	10cc	The original soundtrack	02	I'm not in love	pop	bas / dru / gac / org / pia / unk / voi
AC/DC	AC/DC	Back In Black	01	Hells Bells	met	bas / bel / dru / gel / voi
Acker Bilk	Various Artists	The Heart of Rock 'n' Roll: 1962	20	Stranger on the Shore	j/b	bas / cla / str
Agustin Barrios Mangore	Norbert Kraft	Guitar Favourites	07	Julia Florida	cla	gac
Air France	Air France	No Way Down	02	June Evenings	pop	bas / dru / flu / gel / str / tru / unk / voi
Alexandre Lagoya	Various Artists	Masters of the Guitar	03	Canarios (Sanz)	cla	gac
Alice in Chains	Alice in Chains	Alice in Chains	05	Head Creeps	met	bas / dru / gel / voi
Anathema	Anathema	A Fine Day to Exit	02	Release	roc	bas / dru / gac / gel / unk / voi
Andrew Hill	Various Artists	Blue Note: The Ultimate Jazz Collection	04	Mira	j/b	bas / dru / pia / sax / tru
Angelo Badalamenti	Various Artists	Lost Highway	04	Red Bats With Teeth	j/b	bas / dru / pia / sax
Antonio Orozco	Antonio Orozco	CadizFornia	05	Hoy todo va al reves (feat. Toteking)	roc	bas / dru / gac / gel / org / voi
Aphex Twin	Aphex Twin	Drukqs	01	Jynweythek Ylow	d/e	unk
Arnold Schoenberg	Fritz Widmer	Schoenberg: Pierrot Lunaire	14	Nacht	cla	cel / cla / pia / unk / voi
Art Pepper	Art Pepper	Living Legend	05	Lost Life	j/b	bas / dru / pia / sax
Art Pepper	Art Pepper	Roadgame	01	Roadgame	j/b	bas / dru / pia / sax / voi
Arvo Part	Arvo Part	Alina	01	Spiegel im Spiegel	cla	pia / vio
At the Drive-In	At the Drive-In	Relationship of Command	03	One Armed Scissor	roc	bas / dru / gel / pia / unk / voi
Autechre	Various Artists	Artificial Intelligence	08	The Egg	d/e	bas / dru / unk / voi
B.B. King	B.B. King	The Ultimate Collection	43	The Thrill Is Gone	j/b	bas / dru / gel / org / str / voi
Barney Bigard	Barney Bigard	The Ultimate Jazz Archive (Classic Jazz, Ragtime & Dixieland, disc 4)	01	Sugar	j/b	bas / cla / dru / pia / sax / tru
Barney Kessel	Barney Kessel	Let's Cook!	02	Time Remembered	j/b	bas / dru / gel / pia / unk
Bebo & Cigala	Bebo & Cigala	Lagrimas negras	01	Inolvidable	fol	bas / per / pia / voi
Bert Jansch	Bert Jansch	Avocet	01	Avocet	fol	bas / cel / flu / gac / gel / vio
Bjork	Bjork	Family Tree	22	Bachelorette	pop	str / voi
Black Sabbath	Black Sabbath	Sabotage	03	Symptom of the Universe	met	bas / dru / gac / gel / voi
Blackmail	Blackmail	Friend or Foe	01	Airdrop	roc	bas / cel / dru / gel / str / unk / voi

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Blind Guardian	Blind Guardian	Nightfall in Middle-Earth	09	Mirror Mirror	met	bas / dru / gel / unk / voi
Blind Guardian	Blind Guardian	Somewhere Far Beyond	07	The Bard's Song: In the Forest	met	cel / gac / voi
Bobbi Humphrey	Various Artists	Smooth Jazz Gold	04	You Are the Sunshine of My Life	j/b	bas / dru / flu / gel / org / pia / unk
Booker T. Jones	Booker T. Jones	Potato Hole	01	Pound It Out	roc	bas / dru / gel / org
Boots Randolph	Various Artists	Jukebox Hits of 1963, Volume 1	14	Yakety Sax	fol	bas / dru / gac / sax
Boston	Boston	Boston	01	More Than A Feeling	roc	bas / dru / gac / gel / per / voi
Branford Marsalis	Branford Marsalis	Contemporary Jazz	01	In the Crease	j/b	bas / dru / pia / sax
Branford Marsalis	Branford Marsalis	Renaissance	01	Just One of Those Things	j/b	bas / dru / pia / sax / voi
Brian Eno	Brian Eno	Another Green World	03	St. Elmo's Fire	pop	bas / gac / gel / org / per / pia / unk / voi
Brian Eno	Brian Eno	Apollo: Atmospheres & Soundtracks	01	Under Stars	d/e	bas / gel / org / unk
Bud Shank	Various Artists	Blue Berlin	12	What'll I Do (feat. Bob Cooper)	j/b	bas / dru / flu / gel / sax
Buddy DeFranco & Tommy Gumina	Various Artists	Mercury Records Jazz Story (disc 2)	07	Scrapple from the Apple	j/b	acc / bas / cla / dru
Camaron de la Isla	Camaron de la Isla	La leyenda del tiempo	01	La leyenda del tiempo	fol	bas / dru / gac / org / unk / voi
Caravan Palace	Various Artists	Electro Swing	01	Jolie Coquine	pop	bas / dru / gac / gel / unk / vio / voi
Carlo Gesualdo	Carlo Gesualdo	Quarto Libro di Madrigali (La Venexiana)	03	Il Quarto Libro di Madrigali: Io tacero, W. 4 No. 21	cla	voi
Carlos Santana	Carlos Santana	Oneness - Silver Dreams, Golden Reality	01	The Chosen Hour	roc	per / voi
Carole King	Carole King	Tapestry	01	I Feel the Earth Move	pop	bas / dru / gel / pia / voi
Cassia Eller	Cassia Eller	Acustico	05	Partido alto	fol	bas / dru / gac / unk / voi
Chet Atkins & Jerry Reed	Chet Atkins & Jerry Reed	Sneakin' Around	09	Sneakin' Around	j/b	bas / dru / gac / gel / har / per / unk / voi
Chet Baker	Chet Baker	'Round Midnight (Live at Salt Peanuts Club)	01	Prayer for the Newborn	j/b	bas / dru / pia / sax / tru
Chuck Mangione	Chuck Mangione	Feels So Good	01	Feels So Good	j/b	bas / dru / gac / gel / sax / tru
Clara Nunes	Clara Nunes	Canto das tres racas	01	Canto das tres racas	fol	bas / gac / per / unk / voi
Claude Debussy	Various Artists	Franck/Debussy/Ravel (The Melos Ensemble)	05	Sonata for Violin and Piano, L. 140: I. Allegro vivo	cla	pia / vio

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Claude Debussy	Various Artists	La musique de Chambre	06	Sonata for Cello & Piano: 1 Prologue	cla	cel / pia
Claude Debussy	Various Artists	La musique de Chambre	11	Sonata for Violin & Piano: 1 Allegro vivo	cla	pia / vio
Claude Debussy	Various Artists	The Classical Collection: Debussy-Poetic Impressions (disc 11)	14	Deux Arabesques in E major	cla	pia
Clifford Brown & Max Roach	Clifford Brown & Max Roach	Verve Jazz Masters 44	02	Star Dust	j/b	bas / dru / pia / str / tru
Coleman Hawkins	Coleman Hawkins	Ultimate Coleman Hawkins	11	Beyond the Blue Horizon	j/b	bas / cel / dru / pia / sax / tru
Dave Douglas	Dave Douglas	Strange Liberation	10	The Jones	j/b	bas / dru / org / sax / tru
Deftones	Deftones	Deftones	01	Hexagram	met	bas / dru / gel / voi
Depeche Mode	Depeche Mode	The Singles 86-98	10	World in My Eyes	pop	bas / dru / gel / unk / voi
Depeche Mode	Depeche Mode	Violator	03	Personal Jesus	pop	bas / dru / gel / unk / voi
Doc Watson	Various Artists	Generations of Bluegrass, Volume 3	02	Deep River Blues	fol	gac / voi
Donald Byrd	Donald Byrd	Black Byrd	01	Flight Time	j/b	bas / bra / dru / flu / gel / pia / tru / unk
Dredg	Dredg	Catch Without Arms	01	Ode to the Sun	roc	bas / dru / gel / unk / voi
Eddie Cleanhead Vinson	Eddie Cleanhead Vinson	The Blues Collection 57: Cleanhead Blues	01	Wee Baby Blues	j/b	bas / dru / gel / pia / sax / voi
Edgar Froese	Edgar Froese	Epsilon in Malaysian Pale	01	Epsilon in Malaysian Pale	d/e	bas / flu / str / unk
Edmond Hall	Edmond Hall	Steamin' and Beamin'	21	Continental Blues	j/b	bas / bra / cla / dru / pia / sax / tro / tru
Elton John	Elton John	Honky Chateau	05	Honky Cat	pop	bas / bra / dru / pia / unk / voi
Elvis Presley	Elvis Presley	Elvis Is Back!	01	Stuck on You	pop	bas / dru / pia / voi
Ennio Morricone	Mireille Mathieu	Mirelle Mathieu chante Ennio Morricone	01	Un jour tu reviendras	pop	bas / cel / flu / gac / pia / str / vio / voi
Eric Clapton	Eric Clapton	Complete Clapton	01	I Feel Free	roc	bas / dru / gel / per / pia / voi
Erik Satie	Klara Kormendi	Piano Works (Selection)	26	Trois Gymnopedies: I. Lent et douloureux	cla	pia
Faith No More	Faith No More	Album of the Year	02	Stripsearch	roc	bas / dru / gel / str / unk / voi
Fats Navarro	Fats Navarro	The Fabulous Fats Navarro	04	Boperation	j/b	bas / bra / dru / pia / sax / tro / tru / vib
Felix Mendelssohn	Various Artists	Piano Concertos Nos. 1 & 2	01	II. Adagio (molto sostenuto)	cla	flu / pia / str
Four Tet	Four Tet	Love Cry	01	Love Cry (original version)	d/e	bas / dru / gac / org / unk / voi

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Francisco Tarrega	Andres Segovia	The Art of Segovia	01	Recuerdos de la Alhambra (for Guitar)	cla	gac
Franz Schubert	Franz Schubert	Piano Works (disc 1: Sonatas Nos. 14 & 17) (Alfred Brendel)	05	Piano Sonata No. 14 in A minor, Op. posth. 143: D. 784 I. Allegro giusto	cla	pia
Frederic Chopin	Various Artists	unkown album	01	Sonata for cello and piano in G minor Op.65 1. Allegro moderato	cla	cel / pia
Fripp & Eno	Various Artists	A Brief History of Ambient, Volume 1	03	Evening Star	pop	gac / gel / pia / unk
Genesis	Genesis	A Trick of the Tail	02	Entangled	roc	gac / gel / unk / voi
Gheorghe Zamfir	Gheorghe Zamfir	The Lonely Shepherd	01	The Lonely Shepherd	fol	bas / dru / flu / gac / per / str / tru
Gil Scott-Heron & Brian Jackson	Gil Scott-Heron & Brian Jackson	The First Minute of a New Day	01	Offering	j/b	bas / dru / org / voi
Giovanni Pierluigi da Palestrina	Various Artists	Palestrina/Lassus: Masses (Oxford Schola Cantorum)	07	Stabat mater	cla	voi
Gnawa Diffusion	Gnawa Diffusion	Algeria	08	Bleu Blanc Gyrophare	fol	bas / dru / flu / gac / org / tro / unk / voi
God Is an Astronaut	God Is an Astronaut	All Is Violent, All Is Bright	02	All Is Violent, All Is Bright	roc	bas / dru / gel / org / str / unk
Gyorgy Ligeti	Various Artists	Journey to the Stars	09	Atmospheres	cla	cla / flu / per / str / tru / unk / vio
HammerFall	HammerFall	Legacy of Kings	01	Heeding the Call	met	bas / dru / gel / voi
Hank Mobley	Hank Mobley	A Slice of the Top	01	Hank's Other Bag	j/b	bas / dru / pia / sax / tru / tub
Harold Budd	Harold Budd	Avalon Sutra	01	Arabesque 3	cla	pia / sax / unk
Harry James	Harry James	Verve Jazz Masters 55	02	Walkin'	j/b	bas / bra / dru / pia / sax / tro
Helmet	Helmet	Aftertaste	01	Pure	met	bas / dru / gel / unk / voi
Herbie Mann & Bobby Jaspar	Herbie Mann & Bobby Jaspar	Flute Flight	02	Bodo	j/b	bas / dru / flu / gel / pia
Iron Maiden	Iron Maiden	Seventh Son of a Seventh Son	04	The Evil That Men Do	met	bas / dru / gel / unk / voi
Isaac Albeniz	Julian Bream	Julian Bream Plays Granados & Albeniz	04	Mallorca, Op. 202	cla	gac
Jackie McLean	Jackie McLean	New Wine In Old Bottles	10	Confirmation	j/b	bas / dru / pia / sax
Jaga Jazzist	Jaga Jazzist	What We Must	06	Mikado	j/b	bas / dru / gac / gel / org / per / tru / unk / voi
James Galway	James Galway	Greatest Hits	19	Pennywhistle Jig	fol	bas / bra / dru / flu / gac / hor / str

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Jeff Beck	Jeff Beck	The Late 60's With Rod Stewart	01	Hi-Ho Silver Lining	pop	bas / dru / gel / vio / voi
Jelly Roll Morton	Jelly Roll Morton	The Complete Library of Congress Recordings	12	Honkey Tonk Blues	j/b	pia / voi
Jethro Tull	Jethro Tull	Stand Up	11	Living in the Past	roc	bas / dru / flu / gel / per / vio / voi
Jimi Hendrix	Jimi Hendrix	Experience Hendrix: The Best of Jimi Hendrix	01	Purple Haze	roc	bas / dru / gel / unk / voi
Jimmy Guiffre	Jimmy Guiffre	The Life of a Trio	01	Sensing	cla	bas / pia / sax
Jimmy McGriff	Jimmy McGriff	Best Of The Blues	08	Jumpin' the Blues	j/b	bas / dru / gel / org / sax
Jimmy Smith	Various Artists	Verve Jazz Masters 60: The Collection	07	Organ Grinder's Swing	j/b	bas / dru / gel / org / voi
Joanna Newsom	Joanna Newsom	Ys	01	Emily	pop	cel / har / unk / vio / voi
Joe Jackson	Joe Jackson	Night and Day	01	Another World	pop	bas / dru / gel / org / per / pia / unk / voi
Johann Sebastian Bach	Andras Schiff	English Suites, BWV 806-811	08	Suite No.3 in G Minor, BWV 808-Prelude	cla	pia
Johann Sebastian Bach	Hilary Hahn	Bach Concertos (Los Angeles Chamber Orchestra)	01	Concerto for Violin, Strings and Continuo	cla	str / vio
Johann Sebastian Bach	Various Artists	Essential Bach	02	Orchestral Suite No. 3: Air	cla	cel / str / vio
John Coltrane	John Coltrane	Blue Train	01	Blue Train	j/b	bas / bra / dru / pia / sax / tro / tru
John Coltrane	John Coltrane	The Very Best of John Coltrane	01	A Love Supreme, Part 1: Acknowledgement	j/b	bas / dru / pia / sax / voi
John McLaughlin	John McLaughlin	Thieves and Poets	07	My Romance	cla	bas / gac
Johnny Dodds	Johnny Dodds	The Ultimate Jazz Archive (Classic Jazz, Ragtime & Dixieland, disc 1)	08	I Can't Say	j/b	bra / gac / obo / pia / tru
Josep Archon	Clara Rockmore	The Art of the Theremin	05	Hebrew Melody	cla	pia / unk
Kate Bush	Kate Bush	Aerial (A Sky Of Honey)	08	Nocturn	pop	bas / dru / gac / gel / org / pia / str / unk / voi
Keith Jarrett	Keith Jarrett	Arbour Zena	01	Runes (Dedicated to the Unknown)	j/b	bas / cel / pia / sax / str
Keith Richards	Keith Richards	Main Offender	01	999	roc	bas / dru / gel / org / voi
Kenny Larkin	Kenny Larkin	Azimuth	10	Q	d/e	dru / unk
Kid Koala	Kid Koala	Basin Street Blues	02	Vacation Island	fol	dru / gac / str / unk / voi

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Killing Joke	Killing Joke	Night Time	03	Love Like Blood	roc	bas / dru / gel / org / pia / voi
King Curtis	King Curtis	Soul Meeting	01	Da-Duh-Dah	j/b	bas / dru / pia / sax / tru
Kraftwerk	Kraftwerk	Computerwelt	03	Nummern	d/e	dru / pia / unk / voi
Ladytron	Ladytron	Destroy Everything You Touch	01	Destroy Everything You Touch (radio edit)	d/e	bas / dru / unk / voi
Lali Puna	Lali Puna	Faking the Books	01	Faking the Books	d/e	dru / gac / gel / unk / voi
Laurie Anderson	Laurie Anderson	Mister Heartbreak	04	Kokoku	fol	bas / org / per / unk / voi
Laurindo Almeida & Charlie Byrd	Laurindo Almeida & Charlie Byrd	Tango	01	Orchids in the Moonlight	fol	bas / dru / gac
Led Zeppelin	Led Zeppelin	Physical Graffiti (1975) Remaster (2009)	01	Custard Pie	roc	bas / dru / gel / voi
Lee Morgan	Various Artists	Blue Note Blend, Volume 2	13	The Sidewinder	j/b	bas / bra / dru / pia / sax / tru
Leo Kottke	Leo Kottke	6- and 12-String Guitar	03	Ojo	fol	gac
Leonard Cohen	Leonard Cohen	Greatest Hits	04	The Partisan	pop	bas / gac / har / voi
Liona Boyd	Liona Boyd	Camino Latino (Latin Journey)	01	Carretera Libertad (Freedom Highway)	fol	bas / flu / gac / per / unk
Lisa Germano	Various Artists	Underworld	16	From a Shell	pop	bas / gel / pia / str / vio / voi
Love Spirals Downwards	Love Spirals Downwards	Ardor	01	Will You Fade	roc	bas / dru / gel / org / unk / voi
Mariza	Mariza	Fado em Mim	01	Oica La O Senhor Vinho	fol	bas / dru / gac / voi
Mark O'Connor	Mark O'Connor	Heroes	08	Sadness, Darlin' Waltz	cla	str
Massimiliano Morabito	Massimiliano Morabito	Sende na rionette suna	02	Pizzica Pizzica di Ostuni	fol	acc / gac / voi
Massive Attack	Clint Mansell	Pi Original Soundtrack	06	Angel	pop	bas / dru / gel / str / unk / voi
Mastodon	Mastodon	Crack the Skye	04	The Czar	met	bas / dru / gel / org / pia / voi
Mecanica Popular	Mecanica Popular	Baku: 1922	07	La edad del bronce	fol	dru / unk
Meshuggah	Meshuggah	Catch 33	13	Sum	met	bas / dru / gel / str / unk / voi
Metallica	Metallica	Metallica	08	Nothing Else Matters	roc	bas / dru / gac / gel / org / str / voi
Mike Oldfield	Mike Oldfield	Boxed	04	Argiers	fol	flu / gac
Miles Davis	Miles Davis	Kind of Blue	01	So What	j/b	bas / bra / dru / pia / sax / tru
Mogwai	Mogwai	The Hawk Is Howling	01	I'm Jim Morrison, I'm Dead	roc	bas / dru / gel / pia / str / unk
Mose Allison	Mose Allison	I Don't Worry About a Thing	01	I Don't Worry About a Thing	j/b	bas / dru / pia / voi
New Order	New Order	International	02	Blue Monday	d/e	bas / dru / gel / org / unk / voi
Niccolo Paganini	Eliot Fisk	24 Caprices, Arranged for Guitar	01	Capriccio No. 1 in E major	cla	gac

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Nine Inch Nails	Nine Inch Nails	The Fragile	12	The Great Below	d/e	bas / cel / dru / gac / pia / unk / voi
Norah Jones	Norah Jones	Come Away With Me	05	Come Away With Me	pop	bas / dru / gel / pia / voi
Nusrat Fateh Ali Khan	Various Artists	10 out of 10	02	Mustt Mustt	pop	acc / bas / gel / voi
Onetwo	Onetwo	Instead	02	The Theory of Everything, Part 2	pop	bas / dru / pia / str / unk / voi
Pantera	Pantera	The Great Southern Trendkill	09	Floods	met	bas / dru / gel / unk / voi
Papa John Creach	Papa John Creach	Papa Blues	01	Sweet Life Blues	j/b	bas / dru / gel / pia / sax / vio
Pet Shop Boys	Pet Shop Boys	Minimal (The Remixes)	01	Minimal (radio edit)	d/e	bas / dru / gel / str / unk / voi
Philip Glass	Philip Glass	Glass Reflections	03	Mishima	cla	cel
Pink Floyd	Pink Floyd	A Saucerful of Secrets	03	Set the Controls for the Heart of the Sun	roc	bas / dru / org / unk / voi
Portishead	Portishead	Roseland NYC Live	10	Roads	pop	bas / dru / gel / str / unk / voi
Portugal. The Man	Portugal. The Man	Waiter: You Vultures!	02	Gold Fronts	roc	bas / dru / gac / gel / pia / unk / voi
Propaganda	Various Artists	Electric 80's (disc 2)	08	Dr Mabuse (13th Life mix)	d/e	bas / dru / unk / voi
Queens of the Stone Age	Queens of the Stone Age	Songs for the Deaf	08	Go With the Flow	met	bas / dru / gel / pia / unk / voi
Radiohead	Radiohead	OK Computer	01	Airbag	roc	bas / cel / dru / gel / per / str / unk / voi
Rage Against the Machine	Rage Against the Machine	Rage Against the Machine	05	Bullet in the Head	met	bas / dru / gel / voi
Rahsaan Roland Kirk	Rahsaan Roland Kirk	Volunteered Slavery	01	Volunteered Slavery	j/b	bas / dru / per / pia / sax / tru / voi
Ray Charles	Ray Charles	The Very Best Of Ray Charles	14	Georgia on My Mind	pop	bas / dru / pia / str / voi
Refused	Refused	The Shape of Punk to Come	06	New Noise	met	bas / dru / gel / unk / voi
Reuben Wilson	Various Artists	Blue Note Trip, Volume 3	07	Inner City Blues (Makes Me Wanna Holler)	j/b	bas / dru / gel / org / sax
Richard Groove Holmes	Richard Groove Holmes	Good Vibrations	01	Good Vibrations	j/b	bas / dru / gel / org / sax
Richard Thompson	Richard Thompson	The Life and Music of Richard Thompson (Walking the Long Miles Home-Muswell Hill to LA)	01	Now That I Am Dead	pop	bas / gac / voi
Robert Rich & B. Lustmord	Robert Rich & B. Lustmord	Stalker	02	Synergistic Perceptions	d/e	bas / unk / voi
Rufus Wainwright	Rufus Wainwright	Poses	03	Poses	pop	cel / pia / str / vio / voi
Saratoga	Saratoga	El clan de la lucha	08	Si amaneciera	fol	bas / dru / gac / gel / str / voi
Sergej Rachmaninoff	Various Artists	unknown album	05	Sonata for cello and piano in G minor Op.19 1. Lento. Allegro moderato	cla	cel / pia

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Sigur Ros	Sigur Ros	Hvarf/Heim	02	Hljomalind	pop	bas / dru / gel / unk / voi
Sondre Lerche	Sondre Lerche	Dan in Real Life	03	I'll Be OK	pop	bas / dru / gel / pia / unk / voi
Sonny Rollins	Sonny Rollins	Sonny Rollins, Vol. 1	04	Sonnysphere	j/b	bas / dru / pia / sax / tru
Stanley Turrentine	Stanley Turrentine	Hustlin'	01	Trouble (No. 2)	j/b	bas / dru / gel / org / sax
Steve Miller Band	Steve Miller Band	Fly Like An Eagle	03	Fly Like An Eagle	pop	bas / dru / gel / org / unk / voi
Steve Winwood	Steve Winwood	Arc Of A Diver	06	Night Train	pop	bas / dru / gel / org / voi
Sufjan Stevens	Sufjan Stevens	Michigan	02	All Good Naysayers, Speak Up! Or Forever Hold Your Peace!	pop	bas / dru / gac / gel / org / pia / unk / voi
Super Poti Poti	Super Poti Poti	De Les Millors Concons Populars Catalenes	01	El Patufet	fol	bra / dru / flu / unk / voi
Super Poti Poti	Super Poti Poti	De Les Millors Concons Populars Catalenes	21	La Lluna La Pruna	fol	dru / gel / unk / voi
T-Bone Walker Quintet	Various Artists	Blues Classics 1945-1949	16	Bobby Sox Blues	j/b	bas / dru / gel / pia / sax / tru / voi
Talk Talk	Talk Talk	Spirit of Eden	05	I Believe in You	pop	bas / dru / gel / har / org / pia / str / unk / voi
Tangerine Dream	Tangerine Dream	Exit	01	Kiew Mission	d/e	dru / unk / voi
Telefon Tel Aviv	Telefon Tel Aviv	Immolate Yourself	04	Helen of Troy	d/e	dru / unk / voi
The Beatles	The Beatles	1	16	Eleanor Rigby	pop	str / voi
The Beatles	The Beatles	Revolver	14	Tomorrow Never Knows	pop	bas / dru / gel / pia / unk / voi
The Beatles	The Beatles	Sgt. Pepper's Lonely Hearts Club Band	03	Lucy in the Sky With Diamonds	pop	bas / dru / gel / voi
The Cardigans	The Cardigans	Gran turismo	08	My Favourite Game	pop	bas / dru / gel / org / unk / voi
The Cure	The Cure	Faith	07	The Drowning Man	pop	bas / dru / gel / unk / voi
The Dice Man	Various Artists	Artificial Intelligence	01	Polygon Window	d/e	bas / dru / unk
The Dillinger Escape Plan	The Dillinger Escape Plan	Ire Works	11	Dead as History	met	bas / dru / flu / gac / gel / org / pia / unk / voi
The Haunted	The Haunted	The Dead Eye	11	The Failure	met	bas / dru / gac / gel / voi
The Human League	The Human League	Reproduction	06	Empire State Human	d/e	dru / unk / voi
The J.B.'s	The J.B.'s	Funky Good Time: The Anthology	16	Gimme Some More (very live)	j/b	bas / bra / dru / gel / per / tro / voi
The Mars Volta	The Mars Volta	Frances the Mute	02	The Widow	roc	bas / dru / gac / gel / org / str / tru / unk / voi

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
The Police	The Police	Synchronicity	01	Synchronicity I	pop	bas / dru / gel / org / unk / voi
The Raconteurs	The Raconteurs	Broken Boy Soldiers	02	Store Bought Bones	roc	bas / dru / gel / org / voi
The Shamen	Various Artists	Turn Up the Bass, Volume 15	10	Move Any Mountain	pop	bas / dru / unk / voi
Thelonious Monk	Thelonious Monk	Thelonious Monk Plays Duke Ellington	01	It Don't Mean a Thing (If It Ain't Got That Swing)	j/b	bas / dru / pia
Thursday	Thursday	A City by the Light Divided	01	The Other Side of the Crash/Over and Out (Of Control)	met	bas / dru / gel / pia / unk / voi
Tierra Santa	Tierra Santa	Apocalipsis	01	Neron	met	bas / dru / gel / unk / voi
Tom Waits	Tom Waits	Blue Valentine	01	Somewhere	pop	str / tru / voi
Tomahawk	Tomahawk	Mit Gas	02	Rape This Day	met	bas / dru / gel / org / unk / voi
Tony Scott & Horst Jankowski Trio & Tony Scott	Horst Jankowski Trio	In Concert	02	Yesterdays	j/b	bas / cla / dru / pia
Tool	Tool	Lateralus	09	Lateralus	met	bas / dru / gel / voi
Transistor Transistor	Transistor Transistor	Erase All Name and Likeness	02	And the Body Will Die	met	bas / dru / gel / voi
Trouble Over Tokyo	Trouble Over Tokyo	Pyramides	01	Start Making Noise	pop	dru / gac / unk / voi
U2	Johnny Cash	American III: Solitary Man	04	One	pop	gac / org / pia / voi
Wayne Shorter	Wayne Shorter	Speak No Evil	04	Speak No Evil	j/b	bas / bra / dru / pia / sax / tru
Wes Montgomery	Wes Montgomery	Movin' Wes	01	Caravan	j/b	bas / bra / dru / gel
Wolfgang Amadeus Mozart	Andrew Manze	3 Violin Concertos (The English Concert)	08	Violin Concerto No. 5 in A major, KV. 219, Turkish: II. Adagio	cla	cla / str / vio
Wolfgang Amadeus Mozart	Arthur Grumiaux	Violin Concertos (Complete)	01	Concerto for Violin and Orchestra No. 1 in B-flat major, K. 207: I. Allegro moderato	cla	str / vio
Wolfgang Amadeus Mozart	Gidon Kremer	Mozart: The Complete Violin Concertos	14	Violin Concerto No. 5 In A Major, KV. 219: Adagio	cla	str / vio
Woody Herman	Woody Herman	Verve Jazz Masters 54	01	Don't Get Around Much Any More	j/b	bas / bra / dru / obo / pia / sax / tro / tru
Wynton Marsalis	Wynton Marsalis	He and She	02	School Boy	j/b	bas / dru / pia / sax / tru
Yann Tiersen	Yann Tiersen	Le Fabuleux Destin d'Amelie Poulain	17	Sur le fil	cla	pia

continued on next page . . .

Artist	Album Artist	Album	Track	Title	Genre	Annotation
Yazoo	Yazoo	Upstairs at Eric's	01	Don't Go	d/e	bas / dru / unk / voi
Yes	Yes	Tormato	02	Don't Kill the Whale	roc	bas / dru / gel / pia / str / unk / voi
Yngwie J. Malmsteen	Yngwie J. Malmsteen	The Best of 1990-1999	01	Gimme, Gimme, Gimme	met	bas / dru / gel / voi
Zeca Baleiro	Zeca Baleiro	Lado Z	04	Na Subida do Morro	fol	flu / gac / voi
mum	mum	Summer Make Good	02	Weeping Rock, Rock	roc	dru / flu / gac / gel / pia / tru / unk / vio / voi

Table B.1: Music tracks used in the evaluation collection. Legend for genres: Rock (roc), Pop (pop), Metal (met), Classical (cla), Jazz & Blues (j/b), Disco & Electronic (d/e), Folk (fol). Legend for instruments: Cello (cel), Clarinet (cla), Flute (flu), acoustic Guitar (gac), electric Guitar (gel), Hammond organ (org), Piano (pia), Saxophone (sax), Trumpet (tru), Violin (vio), singing Voice (voi), Drums (dru), Bass (bas), String section (str), Brass section (bra), Bells (bel), Percussion (per), Trombone (tro), Tuba (tub), Oboe (obo), Harmonica (har), Accordion (acc), Horn (hor), Vibraphone (vib), Unknown (unk).

C

Author's publications

Peer-reviewed journals and conference proceedings

Fuhrmann, F., & Herrera, P. (2011). Quantifying the relevance of locally extracted information for musical instrument recognition from entire pieces of music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 239-244*.

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., & Herrera, P. (2011). A Content-based System for Music Recommendation and Visualization of User Preferences Working on Semantic Notions. In *International Workshop on Content-based Multimedia Indexing*.

Bogdanov, D., Haro, M., Fuhrmann, F., Gómez, E., & Herrera, P. (2010). Content-based music recommendation based on user preference examples. In *Proceedings of the ACM Conference on Recommender Systems*.

Fuhrmann, F., & Herrera, P. (2010). Polyphonic instrument recognition for exploring semantic similarities in music. In *Proceedings of the International Conference on Digital Audio Effects (DAFx), pp. 281–288*.

Haro, M., Xambó, A., Fuhrmann, F., Bogdanov, D., Gómez, E., & Herrera, P. (2010). The Musical Avatar - A visualization of musical preferences by means of audio content description. In *Proceedings of Audio Mostly, pp. 103-110*.

Fuhrmann, F., Haro, M., & Herrera, P. (2009). Scalability, generality and temporal aspects in the automatic recognition of predominant musical instruments in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 321–326*.

Fuhrmann, F., Herrera, P., & Serra, X. (2009). Detecting Solo Phrases in Music using spectral and pitch-related descriptors. *Journal of New Music Research, 38(4), 343–356*.

Submitted

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., Herrera, P., & Serra, X. Semantic content-based music recommendation and visualization based on user preference examples. *User Modeling and User-Adapted Interaction*.

One never really finishes his
work, he merely abandons it.

Paul Valéry (1871-1945)

This thesis was written using the XeLaTeX typesetting system.