TAMPERE UNIVERSITY OF TECHNOLOGY

Department of Information Technology

Antti Eronen

# AUTOMATIC MUSICAL INSTRUMENT RECOGNITION

*Master of Science Thesis*

The subject was approved by the Department of Information Technology on the 11th of April 2001.

Thesis supervisors: Professor Jaakko Astola
MSc Anssi Klapuri

# Preface

This work was carried out at the Institute of Signal Processing, Department of Information Technology, Tampere University of Technology, Finland.

First and foremost, I wish to express my gratitude to Mr Anssi Klapuri, who was the initiator of this research and provided guidance, advice and support of all kinds for this work. I wish to thank Professor Jaakko Astola for his advice and comments.

I am grateful for the staff at the Audio Research Group and Insitute of Signal Processing for providing a stimulating working atmosphere. During the recent years, many persons have become as much friends as colleagues. I want to thank Mr Jari Yli-Hietanen and Professor Pauli Kuosmanen for providing the opportunity to start working at the Audio Research Group even before I had learned the basics of DSP.

I wish to thank my parents for understanding the long periods I have been working and not having time to visit them.

Finally, I wish to thank Katri for her love and support.

Tampere, October 2001

Antti Eronen

# Table of Contents

# Tiivistelmä

TAMPEREEN TEKNILLINEN KORKEAKOULU
Tietotekniikan osasto
Signaalinkäsittelyn laitos

ERONEN, ANTTI: Automaattinen soittimen tunnistus
Diplomityö, 69 s.
Tarkastajat: Prof. Jaakko Astola, DI Anssi Klapuri
Rahoittaja: Tampereen teknillinen korkeakoulu, Signaalinkäsittelyn laitos
Lokakuu 2001
Avainsanat: automaattinen soittimen tunnistus, äänilähteen tunnistus, äänenvärin tunnistus, musiikin sisältöanalyysi, laskennallinen kuulema-analyysi

Tässä työssä käsitellään soitinten automaattista tunnistusta, tavoitteena rakentaa järjestelmä, joka pystyy “kuuntelemaan” musiikkiäänitystä ja tunnistamaan siitä soittimen. Järjestelmän testauksessa käytettävä materiaali koostui 5286:sta orkesterisoitinten tuottamasta yksittäisestä nuotista. Kyseisen soitinjoukon äänenväriä on tutkittu hyvin paljon. Työn kirjallisuustutkimusosassa tarkastellaan tutkimuksia, joissa on analysoitu näiden soitinten ääniä, sekä soitinakustiikan tietämystä. Automaattista äänilähteiden tunnistusta käsittelevän kirjallisuuden kanssa nämä muodostavat pohjan työn tärkeimmälle osuudelle: kuinka akustisista signaaleista voidaan irrottaa ihmisten kuulohavainnoille tärkeitä piirteitä.

Työssä toteutettiin ja kehitettiin useita erilaisia piirteenirrotusalgoritmeja sekä hahmontunnistusjärjestelmä, jonka osana algoritmeja käytettiin. Järjestelmän suorituskyky testattiin useissa kokeissa. Järjestelmä saavutti 35 % suorituskyvyn kokeessa, jossa käytettiin tietokantaa, joka sisälsi useita esimerkkejä 29 eri soittimesta. Soitinperhe tunnistettiin oikein kuuden perheen välillä 77 % testitapauksista. Piirrevektorit tässä kokeessa koostuivat kepstrikertoimista ja äänten herätettä, kirkkautta, modulaatioita, asynkronisuutta ja perustaajuutta kuvaavista piirteistä.

Järjestelmän suorituskykyä ja sen tekemiä sekaannuksia verrattiin ihmisten kykyihin. Vertailun perusteella järjestelmän suorituskyky on huonompi kuin ihmisten vastaavassa tehtävässä (ihmisten tunnistustarkkuudeksi on ilmoitettu 46 % yksittäisillä soittimilla ja 92 % soitinperheillä [Martin99]), mutta se on verrattavissa muiden rakennettujen järjestelmien suorituskykyihin. Järjestelmä tekee samanlaisia sekaannuksia kuin ihmiset, joten piirteenirrotusalgoritmeilla on onnistuttu mittaamaan ihmisten havainnoille oleellista informaatiota akustisista signaaleista.

# Abstract

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology
Institute of Signal Processing

ERONEN, ANTTI: Automatic musical instrument recognition
Master of Science Thesis, 69 pages
Examiners: Prof. Jaakko Astola, MSc Anssi Klapuri
Funding: Tampere University of Technology, Institute of Signal Processing
October 2001
Keywords: Automatic musical instrument recognition, sound source recognition, timbre recognition, audio content analysis, computational auditory scene analysis

This thesis concerns the automatic recognition of musical instruments, where the idea is to build computer systems that “listen” to musical sounds and recognize which instrument is playing. Experimental material consisted of 5286 single notes from Western orchestral instruments, the timbre of which have been studied in great depth. The literature review part of this thesis introduces the studies on the sound of musical instruments, as well as related knowledge on instrument acoustics. Together with the state-of-the-art in automatic sound source recognition systems, these form the foundation for the most important part of this thesis: the extraction of perceptually relevant features from acoustic musical signals.

Several different feature extraction algorithms were implemented and developed, and used as a front-end for a pattern recognition system. The performance of the system was evaluated in several experiments. Using feature vectors that included cepstral coefficients and features relating to the type of excitation, brightness, modulations, asynchronity and fundamental frequency of tones, an accuracy of 35 % was obtained on a database including several examples of 29 instruments. The recognition of the family of the instrument between six possible classes was successful in 77 % of the cases.

The performance of the system and the confusions it made were compared to the results reported for human perception. The comparison shows that the performance of the system is worse than that of humans in a similar task (46 % in individual instrument and 92 % in instrument family recognition [Martin99]), although it is comparable to the performance of other reported systems. Confusions of the system resemble those of human subjects, indicating that the feature extraction algorithms have managed to capture perceptually relevant information from the acoustic signals.

# 1 Introduction

Automatic sound source recognition plays an important role in developing automatic indexing and database retrieval applications. These applications have potential in saving the humans from time taking searches through huge amounts of digital audio material available today. For instance, it would be most useful if we could find sound samples that “sound similar” as a given sound example. Music content analysis in general has many practical applications, including e.g. structured coding, automatic musical signal annotation, and musicians’ tools. Automatic musical instrument recognition is a crucial subtask in solving these difficult problems, and may also provide useful information in other sound source recognition areas, such as speaker recognition. However, musical signal analysis has not been able to attain as much commercial interest as, for instance, speaker and speech recognition. This is because the topics around speech processing are more readily commercially applicable, although both areas are considered as being highly complicated. Through constructing computer systems that “listen”, we may also gain some new insights into human perception. This thesis describes the construction and evaluation of a musical instrument recognition system that is able to recognize single tones played by Western orchestral instruments.

A central concept in our study is the quality of sound, i.e. what something sounds like. A musical sound is said to have four perceptual attributes: *pitch, loudness, duration* and *timbre*. These four attributes make it possible for a listener to distinguish musical sounds from each other. Pitch, loudness and duration are better understood than timbre and they have clear physical counterparts. For musical sounds, pitch is well defined and is almost equal to the fundamental frequency. The physical counterpart of loudness is intensity, which is proportional to the square of the amplitude of the acoustic pressure. The third dimension, perceived duration, corresponds quite closely to the physical duration with tones that are not very short. Timbre is the least understood among the four attributes. Traditionally, timbre is defined by exclusion: the quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar [ANSI73]. We are fortunate in the sense that many psychoacousticians have explored the underlying acoustic properties that cause different sound quality, or timbre sensations. Based on this information, and adding the knowledge about the physical properties of sound producing instruments, we can try to construct algorithms that measure this information from digitally stored acoustic signals.

Systems have been built that try to extract perceptually relevant information from musical instrument sounds and recognize their sources. However, the implemented systems are still far from being applicable to real-world musical signals in general. Most of the systems operate either on isolated notes or monophonic phrases. Brown has shown that it is possible to recognize four woodwind instruments in monophonic recordings with an accuracy that is comparable to human abilities [Brown99]. Marques constructed a system capable of discriminating between eight instruments with 70 % accuracy. Martin's system recognized a wider set of instruments, although it did not perform as well as human subjects in the same task [Martin99]. Generally, when the amount of instruments is increased, humans outperform the machines especially in recognizing musical instrument families, i.e., higher-level instrument categories.

Musical instrument recognition is related to many other fields of research. The methods used in implementing musical instrument recognition systems are drawn from different technical areas. The preprocessing and feature extraction techniques can be taken from speech and speaker recognition. Commonly, classification is performed with statistical pattern recognition techniques. Also neural networks and other soft computing techniques have been applied.

Musical instrument recognition and sound source recognition in general are essential parts of computational auditory scene analysis (CASA). In this field, the goal is to analyze complex acoustic environments, including the recognition of overlapping sound events, and thus their sources. In musical synthesis, the model parameters are often analyzed from an acoustic signal. There might be potential in combining these two fields, using physical model synthesis parameters for musical instrument recognition and bringing new methods for feature extraction from musical instrument recognition to physical modeling.

A recent multimedia description standard MPEG-7, developed by the Moving Pictures Expert Group, has two different objectives relating to instrument recognition [Herrera99, Peeters00]. The first, music segmentation according to the played instrument, requires an operating instrument recognition system. The second, segmentation according to perceptual features, means that no universal labels are assigned to the segments, but the segmentation is accomplished using some distance metrics, such as distances between feature values measuring perceptually relevant information calculated from the sample. In multimedia applications, some higher level information is likely to be available, such as structural and semantic information, temporal data, notes, chords or scales.

## 1.1 Motivation for this work

This research originated from the need to build a functional block into an automatic transcription system being constructed at the Institute of Signal Processing at Tampere University of Technology. The project was initiated by Anssi Klapuri who has described the initial steps in his MSc thesis [Klapuri98], and the current state of the project has been recently presented in [Klapuri00, Klapuri01a, Klapuri01b]. The latest paper also describes the first steps towards integrating automatic musical instrument recognition into the other blocks of the transcriber. When complete, this application should be able to transform an acoustic signal into a symbolic representation consisting of notes, their pitches, timings and the instrument label.

The second motivation relates to the more generic problem of sound source recognition and analysis of auditory scenes. The idea is to compile a toolbox of generic feature extractors and classification methods that can be applied to a variety of audio related analysis and understanding problems. In fact, some of the methods implemented for this study and the knowledge gained have been already used in [Peltonen01b].

## 1.2 Defining the problem and selecting an approach

There exists an enormous variety of musical instruments in the world. In practical applications, we naturally train the system with the classes of instruments that are most likely for that particular application. In this thesis, Western orchestral instruments are considered. This is done for two reasons. First, the timbre of these instruments has been extensively studied, providing insights into the information that makes recognition possible and should therefore be attempted to extract from the sounds. Second, recordings of these instruments are easily available, whereas in the cases of more exotic instruments we would first have to make the databases.

In defining the musical instrument recognition task, several levels of difficulty can be found. Monophonic recognition refers to the recognition of solo music or solo notes, and is the most often studied. This study uses isolated notes as test material mainly because samples with annotations were available with a reasonable effort, and there were published isolated note recognition systems with which the performance could be compared. However, this can be generalized to monophonic phrases by introducing a temporal segmentation stage. We present also an alternative approach using Gaussian mixture models that does not require explicit segmentation into notes.

Polyphonic recognition has received much fewer attempts. It is not even clear how the problem should be approached. One way would be to separate the sounds of individual instruments from the mixture and then classify them individually using algorithms developed for monophonic recognition. In this case, the polyphonic musical instrument recognition problem would culminate into reliable sound separation, and the main task of the recognizer block would be to cope with possibly corrupted separated sounds. This is the approach we will pursue. However, the separation approach has received some criticism, too. It has been argued than humans do not separate a single musical instrument from a mixture but more or less consider a mixture of musical sounds as a whole [Scheirer00]. Since there exists algorithms for polyphonic pitch estimation [Klapuri01], separation of concurrent harmonic sounds [Virtanen01], and recognition of musical instruments from the separated tones [Eronen01], it is natural to try the separation approach.

## 1.3 Organization of this thesis

In Chapter 2, we describe a literature review on automatic musical instrument recognition and related fields of interest. Chapter 3 presents an overview of the implemented system and discusses the selected design philosophy. Chapter 4 is devoted to the description of feature extraction algorithms, which include both common front-ends in different audio content analysis applications, and algorithms developed for this thesis. Following the flow of information in the recognition system, Chapter 5 describes the back-end of the system, which consists of alternative classification algorithms. In Chapter 6, the system is evaluated in different tasks, and its performance is compared to reported systems and human abilities. Finally, Chapter 7 summarizes the observations made in this study and suggests some directions for future work.

# 2 Literature review

A literature review was conducted, studying the automatic musical instrument recognition literature and the relating fields of interest. Quite soon it became apparent that we have to go deeper than just the existing feature extraction algorithms and classification techniques for musical instrument recognition, because the field was, and still is, an immature one. In addition to machine hearing, pattern recognition, and digital signal processing, the foundation of this work relies on studies in psychoacoustics and instrument acoustics.

The reasons why human perception is studied in some depth in this thesis are well motivated. Audition reveals what is relevant and irrelevant, and tells about the subjective importance of certain properties. Considering human perception in musical applications is especially important, since musical sounds are designed merely for human audition. Finally, the auditory system is very successful; thus it operates as a *benchmark* for sound understanding systems. If we could imitate the performance of human audition, we would do extremely well [Ellis01].

We start with a discussion on the current knowledge on how humans recognize sound sources. Then we introduce the first benchmark for our system: studies on the human abilities in recognizing musical instruments. Based on the human abilities, we then present criteria for the evaluation of sound source recognition systems, and then introduce the current state-of-the-art in sound source recognition. This is followed by a short comparison to human abilities. The next topic considers about what is the relevant and irrelevant in sounds; first the dimensions affecting sound quality are discussed. Then, we present a model of sound production in order to find explanations on what causes these properties, and describe the musical instruments as sound producing objects. The literature review is concluded with a discussion on several perceptually salient acoustic features that can be used for musical instrument recognition, drawn from the human perception experiments and known acoustic properties of musical instruments.

## 2.1 Psychoacoustics of sound source recognition

Many events and objects can be recognized based on the produced sound alone. Recognition means that what is currently being heard corresponds in some way to something that has already been heard in the past [McAdams93], as for example, when a voice on the telephone, or the footsteps of someone walking down the hall, or a piece of music on the radio are each recognized. However, little is known about how the human sound source recognition actually works [Martin99]. In the following, we will look at some of the ideas presented on how humans perceive sounds and what makes it possible for us to recognize sound sources.

The basic problem in sound source recognition is contextual variation. Sound waves produced by a certain source are different produced at each event. If they were similar, then the recognition could take place simply by comparing the waves into some characteristic templates stored into memory. In the real world, the waves produced at different times are very different. This is due to the fact that the physical process generating the sound is very seldom exactly the same at different times. In addition, the position of a source with respect to a listener, and the acoustic characteristics of the environment affect the sound waves.

The listener must use information that is characteristic to a source and remains constant from one time to another. We call this information *acoustic invariants* [Handel95, McAdams93, Martin99]. Sound producing objects have acoustic properties, which are the result of the production process. These properties enable us to recognize sound sources by listening. The properties include e.g. the type of excitation, the physical construction, the materials, and the shape and size of the resonance structures. The type of excitation varies from instrument to another, and has significant influence on the sound. The resonance structures affect the spectrum of the resulting sound, the temporal development of spectral partials, and so on. By using features that are affected by the invariants, it is possible to move backwards to the invariants themselves, and to the identity of the sound source [Martin99].

However, the situation is complicated by a few things. The acoustic properties evolve over time, typically quite slowly and continuously. The auditory world is transparent and linear; the sound waves from different sources add together and form larger sound sources. For example, the sound of an orchestra is a mixture of the sounds of all the instruments. [Handel95]

Recognition requires learned experience. An auditory percept is evoked by acoustic waves, which are the result of the physical processes of the source. We humans tend to hear the process that has generated the sound, or "see through the sound" into the sound generating mechanism. But for the sake of coping with environmental variation and changes in the production processes, we need to learn the connection between different acoustic properties and their sources. We learn, for instance, how different environments affect some sound. Then, the final recognition is obtained by matching the information in the sound heard with some representation in the long term memory, i.e. a lexicon of sound forms [McAdams93].

McAdams presents recognition as a multistage process illustrated in Figure 1 [McAdams93]. He hypothesizes that the link between the perceptual qualities of the sound source, its abstract representation in memory, its identity and the meanings and associations with other objects in the environment is a result of a sequential process with some feedback connections. In the following, we will briefly discuss the steps in McAdams's model.

The first stage, sensory transduction involves the transmission of acoustic vibration to the

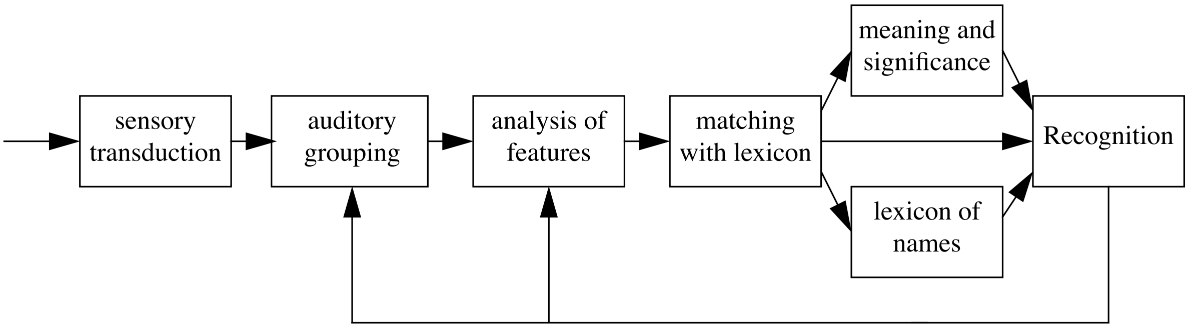

**Figure 1.** Stages of auditory processing in sound source recognition (after [McAdams93]).

cochlea, which is a shell-shaped organ in the inner ear. The cochlea performs initial frequency analysis and dynamic compression. Acoustic vibration is transmitted to a membrane inside the cochlea, namely basilar membrane, of which different frequencies of the input signal set different parts into motion. From the basilar membrane, the movement at different points is transduced into neural impulses that are sent through the auditory nerve to the brain. In the auditory grouping phase, the stream of input information is then processed into separate auditory representations, one for each sound source in the environment [Bregman90]. This means that the components constituting the sound of each source are segregated from the input information (which describes the whole sound mixture), and the components belonging to a certain sound source are integrated into a group. Now we have representations for the sound sources, and analysis of features can begin. It is supposed that in this stage, the brain progressively analyzes the perceptual features relevant to listening at a given moment.

By this point, the initial auditory representation has been changed into a group of abstract properties characterizing the acoustic invariants of each source, such as the spectral, temporal or onset characteristics. In the phase of matching with auditory lexicon, the input representation is matched to classes of similar sound sources and events in memory, and the stimulus is recognized as the class giving the best match. In the next phase, information of the class with respect to the situation, or context, and the listener is available, making it possible to react to an unknown sound, for example. If a verbal description is known for such an event, the final recognition is obtained as a name for the sound source from a verbal lexicon of names.

The feedback loops in Figure 1 are required to explain some phenomena in auditory perception. For example, one's own name is easily recognized even from a very noisy background, or, a much better signal-to-noise ratio is required for understanding foreign languages than one's own native language. One of the best examples is phonemic restoration, meaning that words with corrupted or removed phonemes are heard as if they were not corrupted at all. Bregman refers to these effects as schema-based processing, meaning influence from later stages of processing to auditory grouping and to analysis of features [Bregman90].

## 2.2 Human accuracy in musical instrument recognition

This section reviews some musical instrument recognition experiments made with human subjects. Unfortunately, only a few researchers have used realistic stimuli for the listening experiments; the reported studies have mainly used isolated notes from few instruments and with very limited number of pitches (often from the same pitch).

In [Brown01], Brown summarizes the recognition accuracies in some human perception experiments. The percentage of correct identifications and the number of instruments used are presented in Table 1. The five earliest studies have used isolated tones, the five most recent have used monophonic phrases [Campbell78, Kendall86, Brown99, Martin99, Brown01]. We will now discuss the two most recent studies in more detail. The first includes preliminary results from a listening test conducted by Houix, McAdams and Brown [Brown01]. They conducted a free identification experiment, where fifteen musicians were asked to classify 60 samples into categories, whose number was not told beforehand. The material consisted of solo music excerpts of the oboe, saxophone, clarinet and flute, which were on the average a few seconds in duration. The obtained recognition percentages were 87 for the oboe, 87 for the sax, 71 for the clarinet, and 93 for the flute. The average recognition accuracy was 85 %.

**Table 1: Summary of recognition accuracies in human perception experiments (after [Brown01]).**

| Study | Percentage correct | Number of instruments | |
|---|---|---|---|
| [Eagleson47] | 56 | 9 | |
| [Saldanha64] | 41 | 10 | |
| [Berger64] | 59 | 10 | |
| [Clark64] | 90 | 3 | flute, clarinet and oboe |
| [Strong67] | 85 | 8 | |
| [Campbell78] | 72 | 6 | |
| [Kendall86] | 84 | 3 | trumpet, clar. and violin |
| [Brown99] | 89 | 2 | oboe and sax |
| [Martin99] isolated tones | 46 | 27 | |
| 10-second excerpts | 67 | 27 | |
| [Brown01] | 85 | 4 | oboe, sax, clar. and flute |

With regard to our computer simulations in Chapter 6, the studies reported in [Martin99] are the most relevant. Martin conducted two listening experiments with a wide range of instruments and samples. Fourteen subjects participated in Martin's test, all of whom were either performing musicians or other musical experts. In the first test, 137 isolated notes from the McGill collection [Opolko87] were used, including tones at ten different pitches. The subjects were asked to select the instrument among 27 possibilities. Fourteen instruments were included in the test set: violin, viola, cello, double bass, flute, piccolo, oboe, English horn, bassoon, b-flat clarinet, trumpet, French horn, tenor trombone and tuba. In this test, the overall recognition accuracy was 46 % for individual instruments, and 92 % for instrument families. Martin's instrument families were the strings, brass, double reeds, clarinets and flutes. In the second experiment using 10-second excerpts, the accuracies increased to 67 % and 97 % with individual instruments and families, respectively. In this test, examples of 19 instruments were included in the test set.

In Martin's isolated tone test, the subjects often made confusions within the instrument families. In the string family, there were confusions between the violin and viola, the viola and cello, and the cello and double bass [Martin99, pp. 125]. Within the flute family, the flute was confused as alto flute, and the piccolo as flute. The oboe as b-flat clarinet, English horn as oboe, bassoon as contrabassoon, and b-flat clarinet as oboe were common within the woodwind family (Martin divided these into double reeds and clarinets, though). Within the brass family, the most frequent confusions were as follows: trumpet as cornet, French horn as trumpet or tenor trombone, tenor trombone as bassoon or French horn, and tuba as French horn or tenor trombone. In instrument family classification, the recognition accuracies were best for the strings and flutes [Martin99, pp. 127].

In solo segments, the subjects made very few confusions outside families, as indicated by the high average family recognition accuracy (97 %). The confusions within the families were between the violin and viola, oboe and English horn, and between the saxophones. In addition, the alto and tenor trombone were confused as the French horn. The only instrument family recognized under 90 % accuracy was the double reeds.

## 2.3 Sound source recognition systems

This section first presents criteria for evaluating sound source recognition systems, and then reviews some of the most relevant systems.

### Criteria for system evaluation

Martin has presented several criteria for evaluating sound source recognition systems [Martin99]. First, the systems should be able to generalize, i.e. different instances of the same kind of sound should be recognized as similar. Systems should be robust, they should be able to work with realistic recording conditions, with noise, reverberation and even competing sound sources. Scalability means that the system should be able to learn to recognize additional sound sources without decrement in performance. In addition, when the conditions become worse, the performance of systems should gradually degrade. A system should be able to introduce new categories as necessary, and refine the classification criteria as it gains more "experience". Finally, the simpler out of two equally accurate systems is better.

### Environmental sound recognition

Let us start with the most general case: recognition of environmental sounds and sound effects. It is a very broad field, however, here we will review only a couple examples. Klassner's Sound Understanding Testbed (SUT) was built to recognize specific household and environmental sounds [Klassner96]. It was a trial application for the Integrated Processing and Understanding of Signals (IPUS) architecture [Lesser95], which simultaneously searches for an explanation of a signal and a suitable front-end configuration for analyzing it. SUT had a library of 40 sounds, from which models were derived by hand. The test material was constructed by placing four independent sounds from the library on a five-second recording, and the system's task was to recognize which event happened and when. Depending on whether all models were used as references, or just the ones actually put on the recording, the accuracies were 59 % and 61 %, respectively.

Zhang and Kuo tested a query-by-example scheme for the recognition of sound effects [Zhang00]. The sound effect classes were such as applause, footstep, explosion and raining. With 18 sound effect classes, a performance of 86 % was reported. Dufaux et.al. used hidden Markov models (HMM) for classifying six different classes of impulsive sounds: door slams, glass breaks, human screams, explosions, gun shots, and stationary noises [Dufaux00]. Their front-end consisted of a median-filter based detection stage, and a uniform frequency resolution filterbank whose channel energies were used as features. With a database consisting of 822 sounds, the system was rather robust towards Gaussian noise. It achieved a recognition rate of 98 % at 70 dB signal-to-noise ratio (SNR), and 80 % at 0 dB SNR.

Several systems have tried to recognize vehicles, or other sources of noise. Wu et.al. used features derived from a power spectrum via the principal component analysis (PCA) to cluster car, truck and motor cycle sounds [Wu98]. They did not report any recognition rates, but the system managed to cluster the sounds from different classes into separable clusters using a small database. Jarnicki et.al. used a filterbank front-end as an input to a nearest neighbour classifier [Jarnicki98]. Their system was capable of classifying between military vehicles, transporters and civilian vehicles with over 90 % accuracy. However, little details were given on the amount of testing and training material. A more advanced system was built by Gaunard et.al [Gaunard98]. They used a database of 141 noise events for training, and 43 events for testing. Their classes were car, truck, moped, aircraft and train. Linear prediction (LP) cepstral coefficients, or a 1/3-octave filterbank were used as a front-end for a HMM classifier. The best reported recognition accuracy (95 %) was obtained using ten LP cepstral coefficients and a five-state HMM. The system performed slightly better than six human subjects in a listening test using a subset of the same material.

The problem with environmental sound source recognition systems is that they operate with a very limited set of sounds, while they should be able to handle an enormous variety of different sound sources. Nevertheless, the field is important. Successful computational auditory scene analysis will require the recognition of individual sound sources in the mixture. Recently, Peltonen et.al. presented a human perception experiment concerning a subtask of CASA, where the task was to recognize the acoustic context in which the recording has been made without necessarily interpreting the sounds of single sources [Peltonen01]. However, the results of the study showed that for human subjects, single prominent sound events are the most salient cues for determining the environment.

### Human voice recognition

Speaker recognition is the most studied sound source recognition problem [Martin99]. The human voice recognition task differs from the musical instrument recognition task in various respects. While the number of different instruments is quite limited, there are millions of voices. On the other hand, the fundamental frequency range produced by humans is relatively limited compared to the playing range of most instruments. Nonetheless, a single voice can produce a much greater variety of sounds than a single instrument [Handel95].

In speaker recognition, the idea is to identify the inherent differences in the articulatory organs (the structure of the vocal tract, the size of the nasal cavity, and vocal cord characteristics) and the manner of speaking [Mammone96]. The possible cues for voice identification include the average fundamental frequency as well as the frequency range and contour of the vocal fold vibration [Handel95]. Features relating to the vocal tract include the strengths, frequencies and possibly the bandwidths of the formants, i.e., the resonances of the vocal tract. However, in practice the implemented systems have utilized only features relating to the vocal tract characteristics.

The features used in speaker recognition systems are usually cepstral coefficients based on linear prediction or discrete Fourier transform (DFT), and sometimes include the first and second-order derivatives of these coefficients over time. With the LP coefficients, there is a strong theoretical motivation for modeling the vocal tract as an all-pole filter, as will be seen later in this chapter.

The objective of speaker recognition may be recognition or verification. In the latter, the task is to verify if the person is the one he or she claims to be. The recognition task can be further divided into text-dependent or text-independent with regard of the possible vocabulary. The first approach for speaker recognition used long term average statistics derived from frame-based features, with the motivation that averaging would discard the phonemic variations and retain only the speaker dependent component. In practice, however, speaker dependent information is lost, too. More recent methods aim at comparing the features between similar phonetic sounds within the training and test sets. This is achieved either via explicit segmentation; using a HMM based continuous speech recognizer as a front-end, or through implicit segmentation. The latter method is the most commonly used today, and it involves unsupervised clustering of acoustic features during training and recognition. The most commonly used statistical model is the Gaussian mixture model (GMM). Potential fields of further research in speaker recognition are the use of fundamental frequency information and the speech rhythm.

The problems is speaker recognition include that the performance of systems suffers when acoustic conditions vary from those during testing [Murthy99, Alonso00]. The performance also suffers as interfering sounds are mixed with the speech signal, or when the population size grows [Reynolds95]. As an example, Reynolds reported a system that used 20 mel-frequency cepstral coefficients calculated in 20-ms frames, and used GMMs as the back-end [Reynolds95]. With clean recordings, including only one recording of a particular speaker, the recognition performance was almost perfect for a population of 630 speakers. But under varying acoustic conditions, e.g. using different handsets during training and testing, the performance suffered. With 10 talkers, an accuracy of 94 % was reported, and with 113 talkers, the accuracy was 83 %. However, speaker recognition systems are still the ones closest to practical applicability among the areas of sound source recognition, and the methods are the most developed.

## 2.4 Musical instrument recognition systems

Various attempts have been made to construct automatic musical instrument recognition systems. Researchers have used different approaches and scopes, achieving different performances. Most systems have operated on isolated notes, often taken from the same, single source, and having notes over a very small pitch range. The most recent systems have operated on solo music taken from commercial recordings. Polyphonic recognition has also received some attempts, although the number of instruments has still been very limited. The studies using isolated tones and monophonic phrases are the most relevant in our scope.

### Generation of timbre spaces

A number of experiments has been done in order to generate timbre spaces (i.e. to cluster musical sounds into some space having perceptually relevant dimensions) with techniques attempting to model the human auditory system. These can be considered as relating to instrument recognition. They have hardly shown any performance for actual classification of musical instruments, but give an idea of what could be achieved with this approach. Many of these studies have usually used an auditory model of some kind as an input to a Kohonen self-organizing map (SOM) [Feiten91, DePoli93, Cosi94, Feiten94, Toiviainen95, Toiviainen96, Depoli97]. For instance, De Poli and Prandoni used mel-frequency cepstral coefficients calculated from isolated tones as inputs to a SOM, with their aim to construct timbre spaces [Cosi94, DePoli97]. One tone per instrument was used, all of the same pitch. Six mel-frequency cepstral coefficients (MFCC) from a 27-band filterbank were used as an input to the SOM. In some cases, dimensionality was reduced with principal component analysis (PCA). In [Cosi94], also some other features were used. Unfortunately, the actual performance in classi-

**Table 2: Summary of recognition percentages of isolated note recognition systems using only one example of each instrument.**

| Study | Percentage correct | Number of instruments | |
|---|---|---|---|
| [Kaminskyj95] | 98 | 4 | guitar, piano, marimba and accordion |
| [Kaminskyj00] | 82 | 19 | |
| [Fujinaga98] | 50 | 23 | |
| [Fraser99] | 64 | 23 | |
| [Fujinaga00] | 68 | 23 | |
| [Martin98] | 72 (93) | 14 | |
| [Kostek99] | 97 | 4 | bass trombone, trombone, English horn and contra bassoon |
| | 81 | 20 | |
| [Kostek01] | 93 | 4 | oboe, trumpet, violin, cello |
| | 90 | 18 | |

fying the tones was not reported. Feiten and Guntzel trained a Kohonen SOM with spectral features from 98 tones produced by a Roland Sound Canvas synthesizer. The authors suggest that the system can be used for retrieval applications, but provide no evaluable results [Feiten94].

### Recognition of single tones

These studies have used isolated notes as test material, with varying number of instruments and pitches.

*Studies using one example of each instrument*

Kaminskyj and Materka used features derived from a root-mean-square (RMS) energy envelope via PCA and used a neural network or a k-nearest neighbor (k-NN) classifier to classify guitar, piano, marimba and accordion tones over a one-octave band [Kaminskyj95]. Both classifiers achieved a good performance, approximately 98 %. However, strong conclusions cannot be made since the instruments were very different, there was only one example of each instrument, the note range was small, and the training and test data were from the same recording session. More recently, Kaminskyj ([Kaminskyj00]) has extended the system to recognize 19 instruments over three octave pitch range from the McGill collection [Opolko87]. Using features derived from the RMS-energy envelope and constant-Q transform ([Brown92]), an accuracy of 82 % was reported using a classifier combination scheme. Leave-one-out cross validation was used, and the pitch of the note was provided for the system and utilized in limiting the search set for training examples.

Fujinaga and Fraser trained a k-NN with features extracted from 1338 spectral slices of 23 instruments playing a range of pitches [Fujinaga98]. Using leave-one-out cross validation and a genetic algorithm for finding good feature combinations, a recognition accuracy of 50 % was obtained with 23 instruments. When the authors added features relating to the dynamically changing spectral envelope, and velocity of spectral centroid and its variance, the accuracy increased to 64 % [Fraser99]. Finally, after small refinements and adding spectral irregularity and tristimulus features, an accuracy of 68% was reported [Fujinaga00]. Martin and Kim reported a system operating on full pitch ranges of 14 instruments [Martin98]. The samples were a subset of the isolated notes on the McGill collection [Opolko87]. The best classifier was the k-NN, enhanced with the Fisher discriminant analysis to reduce the dimensions of the data, and a hierarchical classification architecture for first recognizing the instrument families. Using 70 % / 30 % splits between the training and test data, they obtained a recognition rate of 72 % in individual instrument, and after finding a 10-feature set giving the best average performance, an accuracy of 93 % in classification between five instrument families.

Kostek has calculated several different features relating to the spectral shape and onset characteristics of tones taken from chromatic scales with different articulation styles [Kostek99]. A two-layer feed-forward neural network was used as a classifier. The author reports excellent recognition percentages with four instruments: the bass trombone, trombone, English horn and contra bassoon. However, the pitch of the note was provided for the system, and the training and test material were from different channels of the same stereo recording setup. Later, Kostek and Czyzewski also tried using wavelet-analysis based features for musical instrument recognition, but their preliminary results were worse than with the earlier features [Kostek00]. In the most recent paper, the same authors expanded their feature set to include 34 FFT-based features, and 23 wavelet features [Kostek01]. A promising percentage of 90 % with 18 classes is reported, however, a leave-one-out cross-validation scheme probably increases the recognition rate. The results obtained with the wavelet features were almost as good as with the other features.

Table 2 summarizes the recognition percentages reported in isolated note studies. The most severe limitation of all these studies is that they all used only one example of each instrument. This significantly decreases the generalizability of the results, as we will demonstrate with our system in Chapter 6. The study described next is the only study using isolated tones from more than one source and represents the state-of-the-art in isolated tone recognition.

*A study using several examples of each instrument*

Martin used a wide set of features describing the acoustic properties discussed later in this chapter, which were calculated from the outputs of a log-lag correlogram [Martin99]. The classifier used was a Bayesian classifier within a taxonomic hierarchy, enhanced with context dependent feature selection and rule-one-category-out decisions. The computer system was evaluated with the same data as in his listening test which we already reviewed in Section 2.2. In classifying 137 notes from 14 instruments from the McGill collection, and with 27 target classes, the best accuracy reported was 39 % for individual instrument, and 76 % for instrument family classification. Thus, when a more demanding evaluation material is used, the recognition percentages are significantly lower than in the experiments described above.

**Table 3: Summary of recognition accuracies in experiments using monophonic phrases**

| Study | Percentage correct | Number of instruments |
|---|---|---|
| [Dubnov98] | not given | 18 |
| [Marques99] 0.2 seconds | 70 | 8 |
| 2 seconds | 83 | 8 |
| [Brown99] | 94 | 2 oboe, saxophone |
| [Brown01] | 84 | 4 oboe, sax, flute and clarinet |
| [Martin99] | 57 (75) | 27 |

### Recognition of monophonic phrases

The four following systems have operated on solo music taken from commercial recordings.

*A study using one example of each instrument*

Dubnov and Rodet used cepstral and delta cepstral coefficients which were calculated from 18 musical pieces from as many instruments [Dubnov98]. The sequence of features was first vector quantized and then fed to a statistical clustering algorithm. The clustering results were promising since the features from different instruments were clustered into different clusters, however the authors do not report any quantitative recognition rates.

*Studies using several examples of each instrument*

The following three systems represent the state-of-the-art in monophonic musical instrument recognition. They all used material taken from compact disks, and several different examples of each instrument were included. Marques built a system that recognized eight instruments based on short segments of audio taken from two compact disks [Marques99]. The instruments were bagpipes, clarinet, flute, harpsichord, organ, piano, trombone and violin. Using very short, 0.2-second segments she reported an accuracy of 70 % using 16 mel-frequency cepstral coefficients and a support vector machine as a classifier. In classifying 2-second segments, the classification accuracy increased to 83 %.

Brown has used speaker recognition techniques for classifying between oboe, saxophone, flute and clarinet [Brown01]. She used independent test and training data of varying quality taken from commercial recordings. By using a quefrency derivative of constant-Q coefficients she obtained an accuracy of 84 %, which was comparable to the accuracy of human subjects in a listening test conducted with a subset of the samples. Other successful features in her study were cepstral coefficients and autocorrelation coefficients. In an earlier study, her system classified between oboe and saxophone samples with a 94 % accuracy [Brown99].

Martin evaluated his system, which we just described above, with 10-second excerpts from solo recordings. The conditions were the same as in the listening test in Section 2.2: 19 different instruments in the test set, and a forced choice between 27 instruments. The best reported accuracies were 57 % in individual instrument, and 75 % in instrument family classification. The system outperformed three out of the fourteen subjects in the listening tests.

Table 3 summarizes the results from studies using solo phrases. The two remaining tasks in musical instrument recognition are content based retrieval, and polyphonic recognition, which are now briefly introduced.

### Content based retrieval

The MPEG-7 standard presents a scheme for instrument sound description, and it was evaluated in a retrieval task as a collaboration between IRCAM (France) and IUA/UPF (Spain) in [Peeters00]. The evaluated features, or descriptors in MPEG-7 terminology, were calculated from a representation very similar to our sinusoid envelopes, which are later discussed in Chapter 4. The authors performed an experiment, where random notes were selected from a database of sound samples, and then similar samples were searched using the descriptors, or just random selection. The subjects were asked to give a rating for the two sets of samples selected in the alternative ways. A "mean score" of approximately 60 % was obtained using one descriptor, and approximately 80 % when using five descriptors.

### Polyphonic recognition

In the field of CASA, interesting work has been done that can be considered as first polyphonic musical instrument recognition systems. Godsmark and Brown used a "timbre track" representation, in which spectral centroid was presented as a function of amplitude to segregate polyphonic music to its constituent melodic lines [Godsmark99]. In assigning piano and double bass notes to their streams, the recognition rate was over 80 %. With a music piece consisting of four instruments, the piano, guitar, bass and xylophone, the recognition rate of their system decreased to about 40 %. However, the success of this task depends also on other metrics. Nevertheless, the results are interesting.

The work of Kashino et.al. in music transcription involves also instrument recognition. In [Kashino95], a system transcribing random chords of clarinet, flute, piano, trumpet and violin with some success was presented. Later, Kashino and Murase have built a system that transcribes three instrument melodies [Kashino98, Kashino99]. Using adaptive templates and contextual information, the system recognized three instruments, violin, flute and piano with 88.5 % accuracy after the pitch of the note was provided. More recently, the work was continued by Kinoshita et.al. [Kinoshita99]. The authors presented a system that could handle two note chords with overlapping frequency components using weighted template-matching with feature significance evaluation. They reported recognition accuracies from 66 % to 75 % with chords made of notes of five instruments.

Although the number of instruments in these studies has still been quite limited, these systems present the state-of-the-art when multiple simultaneous sounds are allowed. The task is definetely more challenging than monophonic recognition, and using only a few instruments makes the problem addressable with a reasonable effort.

### Discussion

Based on the discussion here, the problem in assessing the performance musical instrument recognition systems is that there is great variation in the methods of evaluation. Only few systems have used material that includes several examples of a particular instrument recorded in different environments. There is no guarantee of the ability to generalize with those systems that use material from a single source for training and testing. Unfortunately, the evaluation of musical instrument recognition systems will be difficult until there exist publicitly available databases, on which tests can be made. The McGill collection [Opolko87] is commonly used, unfortunately it includes only solo tones and one example of each instrument, and therefore poses only a minor challenge for recognition systems if used as the only source for evaluation data.

For these reasons, drawing conclusions of the relative performance of various features and classification methods is also difficult. The only feature reported succesful in all experiments that have used it are the mel-frequency cepstral coefficients. However, good results have been obtained without using cepstral features at all in [Martin99]. In Chapter 6, we experimentally evaluate a wide selection of the features used in these studies in a solo tone recognition task. The effect of classification methods on the recognition performance is probably minor compared to the effect of features, as has been reported for example in the field of speech/ music discrimination [Scheirer97].

## 2.5 Comparison between artificial systems and human abilities

The current state-of-the-art in artificial sound source recognition is still very limited in its practical applicability. Under laboratory conditions, the systems are able to successfully recognize a wider set of sound sources. However, if the conditions become more realistic, i.e. the material is noisy, recorded in different locations with different setups, or there are interfering sounds, the systems are able to successfully handle only a small number of sound sources. The main challenge for the future is to build systems that can recognize wider sets of sound sources with increased generality and in realistic conditions [Martin99].

In general, humans are superior with regard to all the evaluation criteria presented in Section 2.3 [Martin99]. They are able to generalize between different pieces of instruments, and recognize more abstract classes such as bowed string instruments. People are robust recognizers because they are able to focus on a sound of a single instrument in a concert, or a single voice within a babble. In addition, they are able to learn new sound sources easily, and learn to become experts in recognizing, for example, orchestral instruments. The recognition accuracy of human subjects gradually worsens as the level of background noise, and interfering sound sources increases.

Only in limited contexts, such as discriminating between four woodwind instruments, computer systems have performed comparable to human subjects [Brown01]. With more general tasks, a lot of work needs to be done.

## 2.6 Perceptual dimensions of timbre

The most essential question with regard to human perception here is: what are the qualities in musical instrument sounds making the recognition of their sources possible. There are four main dimensions in sounds; pitch, loudness, duration, and timbre.

The fourth dimension, timbre, or sound “colour”, is the most vague and complex of these dimensions. It is defined as something that enables discrimination when the three other dimensions are equal. Based on the current knowledge, timbre is a complex and multidimensional property. It is unlikely that any one property or a fixed combination of properties uniquely determines timbre. A considerable amount of effort has been done in order to find the most salient acoustic attributes affecting the perception of timbre. Some examples include [Saldanha64, Wedin72, Plomp76, Grey77, Grey78, Wessel79, Iverson93, McAdams95]. Often these studies have involved multidimensional scaling (MDS) experiments, where a set of sound stimuli is presented to human subjects, who then give a rating to their similarity or dissimilarity [McAdams93]. The stimuli usually consists of a small number of tones with equal pitch, loudness and duration. On the basis of these judgements, a low-dimensional space which best accommodates the similarity ratings is constructed and a perceptual or acoustic interpretation is searched for these dimensions. Another technique is discrimination experiments, where the sound samples are modified in some way and then the subjects are asked whether they can hear any differences [McAdams93]. If they cannot, the removed property has been irrelevant with respect to sound quality.

The two acoustic properties described in MDS experiments have usually been spectral centroid and rise time [Grey77, McAdams93, Handel95, Lakatos00]. The first measures the spectral energy distribution in the steady state portion of a tone, which corresponds to the perceived brightness. The second is the time between the onset and the instant of maximal amplitude. The psychophysical meaning of the third dimension has varied, but it has often related to temporal variations or irregularity in the spectral envelope. Good reviews over the enormous body of timbre perception literature can be found in [McAdams93, Handel95]. These available results provide a good starting point for the search of features to be used in musical instrument recognition systems.

## 2.7 A model of sound production

Acoustic properties of sound sources are another source of information on relevant features. Let us start by considering a model of sound production presented in [Handel95]. It consists of two, possibly interacting components: the *source* and the *filter*. The source is excited by energy to generate a vibration pattern to the source, which is then imposed on the filter. For instance, in a guitar the vibrating strings are the source, the pluck from the finger is the excitation, and the body is the filter. The filter acts as a resonator, having different vibration modes. Each mode affects the spectrum of the sound going through the resonator, causing peaks in the frequency spectrum at resonant frequencies.

### Effects of the excitation and source

The source vibration determines the frequency contents of the sound. The relative amplitude of frequency partials can be affected by changing the method and strength of excitation. In plucked strings, the excitation is very short. When plucked with a finger, the high frequencies are dampened because the pluck is soft. Contrariwise, when a guitar is strongly plucked with a plectrum, the sound is sharp and bright, with rich spectrum of prominent high frequencies. With some instruments, the player can vary the source by introducing vibrato, which is periodic frequency modulation between 4 Hz and 8 Hz [Fletcher98]. For example, a violin player can push a string downwards onto the fingerboard, and then periodically bend his or her finger along the direction of the strings, causing the length of the string, and thus the wavelength of the vibration to vary at the same rate.

### Effects of the filter

The filter has two effects on the resulting sound: it changes both the strength and the time relations of frequency partials. Each vibration mode of the resonator, or filter, can be characterized by its resonance frequency and its quality factor Q. The resonance frequency of each mode is the frequency at which the amplitude of vibration is at maximum. The value of Q is defined as the ratio of the system's resonance frequency to the -3 dB bandwidth of the frequency response of the mode. Q measures the sharpness of tuning and the temporal response of a system: the frequency response of a mode with high Q is narrow, and flat for a mode with low Q.

In addition to causing a hearable boosting at the resonance frequency into the frequency spectrum of the sound, a resonance with high Q also causes a longer time delay into the signal components passing through the mode. Generally, as the driving vibrator sharply increases or decreases its amplitude, so does the amplitude of the mode. In a mode with high Q and sharp peak in the spectrum, the amplitude of the vibration changes slowlier than the amplitude of the driving vibrator, thus causing a time delay. This is important, since humans are capable of hearing even a few millisecond time differences in the amplitude patterns of signal components [Karjalainen99]. Similarly, the rate of decay as a response to a sudden decrease in the amplitude of vibration of the source with a mode with high Q is also slower than with a mode with low Q.

In natural systems, the filter contains a multitude of vibration modes, which are usually at inharmonic frequencies and each have a different quality factor [Rossing90, Handel95]. In this case, the modes will have asynchronous temporal patterns, as the modes with low Q reach the maximum amplitude earlier than modes with high Q. If the damping of all the modes is roughly equal, then the modes will reach their maximum amplitudes almost simultaneously.

The sound can be changed by altering the filter characteristics. For example, a trumpet player can use different kinds of mutes to change the resonance characteristics of the tube and the radiation characteristics of the bell.

### The resulting sound

The final sound is the result of the effects caused by the excitation, resonators and radiation characteristics. In sound producing mechanisms that can be modeled as linear systems, the transfer function of the resulting signal is the product of the transfer functions of the partial systems (if they are in cascade), mathematically

$$Y(z) = X(z)\prod_{i=1}^{N} H_i(z), \tag{1}$$

where $Y(z)$ and $X(z)$ are the $z$-transforms of the output and excitation signal, respectively, and $H_i(z)$ are the z-transforms of the $N$ subsystems. This model works well for the human sound production, however, most musical instruments are characterized by a highly nonlinear behavior, which will become apparent in the next section. In voice modeling, the excitation is the vibration at glottis, and the following subsystems are the vocal tract and the reflections at lips. Correspondingly, the sound of a guitar is the result of the plucking excitation, string source, bridge resonator, body resonator and the radiation characteristics [Karjalainen99]. The string, bridge and body resonator are in cascade, but the individual strings in parallel. Actual physical modeling of musical instruments is out of the scope of this thesis, and an interested reader is referred to [Välimäki96] for an overview of this field.

## 2.8 Physical properties of musical instruments

Western orchestral instruments have a representative set of sound types and production mechanisms that have certain basic properties. Traditionally, the instruments are divided into three classes: the strings, the brass and the woodwinds. The sound of the instrument members within each family are similar, and often humans make confusions within, but not easily between, these families. Examples include confusing the violin and viola, the oboe and English horn, or the trombone and French horn [Martin99]. In the following, we briefly present the different members of each family and their physical build. However, in order to gain a deeper understanding into the effects of different sound production mechanisms, Appendix A discusses the acoustics of musical instruments in more detail.

### The strings

The members of the string family include the violin, viola, cello and double bass, in the order of increasing size. These four form a tight perceptual family, and human subjects consistently make confusions within this family [Martin99]. The fifth string instrument considered in this thesis is the acoustic guitar, which differs from the four other string instruments.

The string instruments consist of a wooden body with a top and back plate and sides, and an extended neck. The strings are stretched along the neck and over a fingerboard. At the other end, the strings are attached to the bridge and at the other end to the tuning pegs which control the string tension. The strings can be excited by plucking with fingers, drawing a bow over them or hitting them with the bow (*martele* style of playing). The strings itself move very little air, but the sound is produced by the vibration of the body and the air in it [Rossing90]. They are set into motion by the string vibration which transmits to the body via the coupling through the bridge. The motion of the top plate is the source of the most of the sound, and is a result of the interaction between the driving force from the bridge and the resonances of the instrument body [Rossing90].

The acoustic guitar is commonly used in popular music. Its construction is basically the same as that of the Western orchestral string instruments just presented. It is played by plucking with a finger or with a plectrum.

### The brass

The members of the brass family considered in this thesis include the trumpet, trombone, French horn, and tuba. The brass instruments have the simplest acoustic structure among the three families. They consist of a long, hard walled tube with a flaring bell attached at one end. The sound is produced by blowing at the other end of the tube, and the pitch of the instrument can be varied by changing the lip tension. The player can use mutes to alter the sound, or insert his hand into the bell with the French horn.

### The woodwind

The woodwind family is more heterogeneous than the string and brass families, and there exists several acoustically and perceptually distinct subgroups [Martin99]. The subgroups are the single reed clarinets, the double reeds, the flutes with an air reed, and the single reed saxophones. In wind instruments, the single or double reed operates in a similar way as the players lips in brass instruments, allowing puffs of air into a conical tube where standing waves are then created. The effective length of the tube is varied by opening and closing tone holes, changing the pitch of the played note. [Fletcher98]

*Double reeds*

The double reed subfamily consists of the oboe, English horn, bassoon and contrabassoon, in the order of increasing size. These instruments have a double reed which consists of two halves of cane beating against each other [Rossing90]. The reed is attached into a conical tube. These instruments are commonly played with vibrato [Martin99].

*Clarinets*

The clarinets have a single reed mouthpiece attached to a cylindrical tube [Fletcher98]. There exists several different sized clarinets, the E-flat, B-flat, bass and contrabass clarinets are the members considered in this study. The B-flat clarinet is the most common, and the bass and contrabass clarinets are larger than the B-flat clarinet. The E-flat clarinet is a very small, bright sounded instrument.

*Saxophones*

The members of the saxophone family include the soprano, alto, tenor and baritone saxophone. Although nowadays made of brass, the saxophones are single reed instruments with a conical bore.

*Flutes*

The members of the flute or air reed family include the piccolo, flute, alto flute and bass flute in the order of increasing size. They consist of a more or less cylindrical pipe, which has finger holes along its length. The pipe is stopped at one end, and has a blowing hole near the stopped end [Fletcher98].

## 2.9 Features for musical instrument recognition

This section presents various features that can be used for recognizing musical instruments, summarizing the observations from constructed recognition systems, timbre studies and instrument acoustics.

### Spectral shape

A classic feature is spectral shape, the time varying relative amplitude of each frequency partial [Handel95]. Various measures can be used to characterize spectral shape. The spectral energy distribution measured with the spectral centroid has been an explanation for the results of many perception experiments [Grey77, Iverson93, Plomp76, Wedin72, Wessel79, Poli97, Toiv96]. It relates to the perceived brightness of tones [Wessel79], and measures also harmonic richness. Sounds that have few harmonics sound soft but dark, and those with lots of harmonics, especially strong high harmonics, have a bright and sometimes sharp tone.

Formants are characteristic for many instruments. Their frequencies would be one characterization. However, the exact frequencies are hard to measure reliably, therefore the formant information is usually represented with an approximation of the smooth spectral envelope. We will discuss this in more depth in the description of cepstral analysis algorithms in Chapter 4.

Variance of component amplitudes, or spectral irregularity (IRR) corresponds to the standard deviation of the amplitudes from the spectral envelope [Krimphoff94]. This has been also referred as the spectral flux or spectral fine structure [Kostek99], or spectral smoothness [McAdams99]. Irregularity of the spectrum can indicate a complex resonant structure often found in string instruments. A mathematical formulation for IRR will be given in Section 4.10. Another measures are the even and odd harmonic content, which can be indicative of the cylindrical tube closed at one end used in clarinets [Martin99].

Measuring the spectrum over different portions of the tone reveals information on different properties. In the quasi-steady state (or almost steady state), information on formants is conveyed in the spectrum. During the onset, the spectral shape may reveal the frequency contents of the source vibration, and the differences in the rate of rise of different frequency partials.

### Onset and offset transients, and the amplitude envelope

Onset and offset transients can provide a rich source of information for musical instrument recognition. Some instruments have more rapid onsets, i.e. the duration of the onset period (also called rise time) is shorter than with others. Rapid onsets indicate tight coupling between the excitation and resonance structures. For instance, the flute has a very slow onset, while other wind instruments generally have quite rapid onsets. Rise time has often been a perceptually salient cue in human perception experiments [McAdams93, Handel95, Poli97].

The differences in the attack and decay of different partials are important. For string instruments, the differences are due to variations in the method of excitation, whether bowed or plucked, or variations in the damping of different resonance modes. With the wind instruments, nonlinear feedback causes differences in the development of different partials. A property characterizing the onset times of different partials is onset asynchrony, which is usually measured via the deviation in the onset times and durations of different partials. The absolute and relative onset times of partials reveal information of the center frequencies and the Q values of resonance modes of the sound source [Martin99].

However, there are some problems in using onset features for musical instrument recognition. Iverson and Krumhansl investigated timbre perception using entire tones, onsets only, and tones minus the onsets [Iverson93]. Based on subjects' ratings, they argue that subjects seem to be making judgments on the basis of acoustic properties that are found throughout the tone. Most likely the attack transients become less useful with melodic phrases than with isolated notes, since the music can be continuous having no clear onsets or offsets. In addition, the shapes of onsets and offsets of partials vary across sounds, depending on the pitch and playing style [Handel95].

### Pitch features

The pitch period indicates the vibration frequency of the source. Absolute pitch also tells about the size of the instrument. Large instruments commonly produce lower pitches than smaller instruments. Also, if we can reliably measure the pitch, and know the playing ranges of possible instruments, pitch can be used to rule out those instruments that cannot produce the measured pitch [Martin98].

Variations in the quasi-steady state of musical instruments convey lots of information of the sound source. Vibrato playing is characteristic for many instruments, but it also reveals information of the resonance structures [McAdams93, Martin99]. The frequency modulation causes the instrument's harmonic partials to interact with the resonances, which again causes amplitude modulation, and by measuring this, information of the resonance structure is obtained. The stability of source excitation and the strength of the coupling between the excitation and resonance structures is indicated by random variations or fluctuations in pitch [Martin99]. For example, the onset of brass instruments have an unstable period during the onset, until the pitch stabilizes into the target value. The unstable interaction between the bow and string causes the tones of string instruments to have high amounts of pitch jitter.

### Amplitude and loudness features

Besides pitch, amplitude variations in the quasi-steady state of tones convey lots of important information. Differences in amplitude envelopes contributed to similarity judgements in [Iverson93], furthermore, the dynamic attributes were present not only in the onset, but also throughout the tone. Tremolo, i.e. periodic amplitude modulation is characteristic for many instruments. For instance, flutes produce strong tremolo. In addition, playing flutes in flutter style introduces characteristic amplitude variation into the tones.

Information on the dynamics of an instrument could also aid recognition. The dynamic range of an instrument is defined as the relation between the level of a sound measured when played *forte fortissimo*, and the level when played *piano pianissimo*. Of the Western orchestral instruments, the brass instruments produce the largest sound levels, the strings are about 10 dB quieter. The woodwind instruments produce slightly louder sounds than the strings. [Kostek99]

Using information on the dependence of the qualities of tone on the playing dynamics might be used for recognition. Beauchamp has suggested using the ratio of spectral centroid to intensity; as tones become louder, they become brighter in a relationship that is characteristic for a particular instrument [Beauchamp82].

### Noise and inharmonicity

Many instrument sounds are characterized by initial noise when the excitation is first applied to the source. For instance, bowing a violin creates an initial high frequency scratch before the bowing stabilizes [Handel95]. Continuous noise can be found from flute sounds, where blowing across the mouthpiece creates a “breathy” sound. Inharmonicity is an important characteristic for many instruments, i.e. for plucked strings and the piano.

Neither noise nor inharmonicity properties have been applied for musical instrument recognition so far. A possible approach might be to use sinusoidal modeling techniques to separate the harmonic parts from the signal and analyze the remaining noise residual [Serra97, Virtanen01]. Also, with a reliable algorithm for determining the frequencies of partials, measures for inharmonicity could be obtained.

### Transitions between sounds and higher level knowledge

Transitions between sounds may be significant for recognition [Handel95, Brown01]. On one hand, the overlap of decay and attack of successive tones may mask transient properties making it impossible to use this information. On the other hand, the interaction of the patterns of successive notes may create unique acoustic information not heard in isolated tones.

Generally, a short piece of musical phrase leads to a far better recognition than isolated tones [Kendall86, Martin99]. However, this is more likely due to the use of higher level inference, instead of low level acoustic information at the transition points. For instance, if we hear a jazz piece played by some wind instrument, we often anticipate that it is a saxophone.

### Dependence on context

Based on the research on human subjects, we know that a large number of acoustic properties can determine the qualities of a tone and the identification of instruments. Moreover, no single feature is the reason for recognition; some level of recognition performance will be obtained with a single property. The listeners, or machines, should use those cues that lead to best performance in the given context and task. [Handel95]

There will be correlation in the acoustic cues due to the interactive nature of sound production; the physical sound production processes are hardly isolated, linear processes. In addition, the combination and interaction of the properties of single objects in the mixture of sounds will generate new, emergent properties that belong to the larger entity, such as an orchestra. Most importantly, the cues leading to best performance will depend on the context: the duration, loudness and pitch of tones, the set of training sounds, the task, and with humans, on their experience. The features are also redundant; parts of the sound can be masked and still identification is possible. [Handel95]

These aspects in human perception and also the fact that many features have been explored especially in musical timbre studies provide a starting point for the construction of automatic sound source recognition systems. Thus, although there are no sight of the ultimate definition of which are the relevant acoustic features for humans in recognizing sound sources and events, artificial sound source recognition systems may still utilize the several proposed cues for recognition, implementing a vast number of them and selecting the most suitable for the given task at hand.

# 3 Overview of the system

This chapter presents an overview of the implemented system before giving a more detailed description in the coming chapters.

## 3.1 Functional components

Figure 2 presents a block diagram of the main components of the implemented system. Acoustic signals have been recorded into data files for the ease of training and testing the system. The format of most samples is standard CD quality, 44.1 kHz 16-bit fixed point with the exception that some guitar samples are in 48 kHz format (see Section 6.1). If a real-time implementation was made, we would need a microphone and an analog-to-digital (AD) converter.

In the preprocessing stage, the mean is removed from the input signal, and its amplitude is linearly scaled between -1 and 1. The input signal that is fed to discrete Fourier transform (DFT) analysis and LP analysis is also filtered with a high pass filter $1 -0.97z^{-1}$. This flattens the sound spectrum, and is useful to do prior to measuring the overall spectral envelope because the spectrum of natural sounds has high concentration of energy at low frequencies.

The next four components transform the signal into some compact representation that is easier to interpret than the raw waveform. The representations used were the LP coefficients, outputs of a mel-filterbank calculated in successive frames, sinusoid envelopes, and a short-time RMS-energy envelope. The calculation and limitations of these representations is presented in the next chapter.

In the feature extraction stage, various characteristic features are extracted from the different representations. The mid-level representations containing hundreds or thousands of values calculated at discrete time intervals is compressed into around 1-50 characteristic features for each note (or for each time interval if we are using frame-based features). This means that after this stage, each observation from a class is represented as a point in a space with limited number of dimensions. Chapter 4 is devoted to the detailed description of various feature extraction algorithms.

Model training either stores the feature vectors corresponding to the class of the labeled input signal as a finite number of templates, or trains a probabilistic model based on the observations of the class. In the classification step, the feature stream of the input signal is compared to the stored templates, or a likelihood value is calculated based on the probabilistic models of trained classes. The recognition result is given as the class giving the best match. Chapter 5 describes the used classification methods in more detail.

## 3.2 Discussion on the approach

Some studies have put emphasis on using a single auditory model as a mid-level representation that tries to emulate the operation of the sensory transduction stage at McAdams's model in Figure 1 [Cosi96, Toiviainen96, Martin99]. This is well motivated if the purpose is to gain understanding into the operation of human perception [Martin99]. For the purpose of our research, we chose not to try to imitate the operation of human perception by using an auditory model, but instead utilize psychoacoustic knowledge in our feature extraction algorithms.

One of our central objectives was to examine different possible features for instrument recognition. Since we do not know what features should be used, we chose not to limit ourselves into any particular mid-level representation. When integrated into a larger entity, such as a transcription system, we should use those representations that are available, such as the amplitude and frequency tracks of a sinusoidal modeling block. Therefore, we use several different representations, utilize psychoacoustic knowledge in calculating them by using perceptual frequency scales, for example, and in finding the set of features to be extracted.

The third reason for abandoning auditory models is the fact that the proposed models are computationally very intensive and some information is lost in the process.

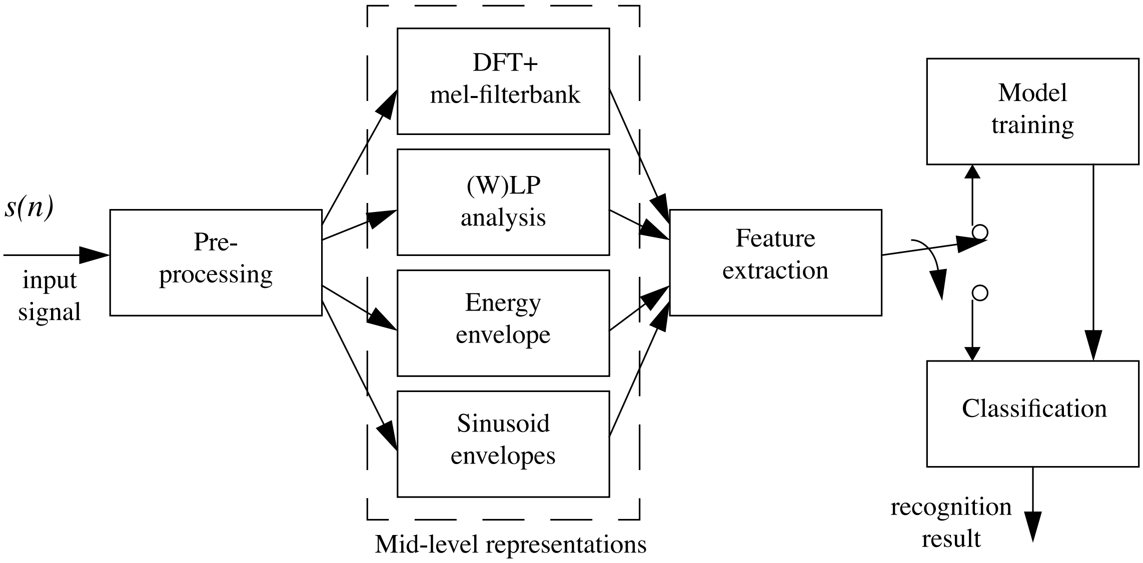

**Figure 2.** Block diagram of the implemented musical instrument recognition system.

# 4 Feature extraction

In this part, a wide selection of perceptually relevant acoustic features for sound source recognition are presented. Along with these, the calculation of the representations used as an input for the feature extraction are described. We intentionally give little detail in describing the feature extraction algorithms, besides the cepstral features. The cepstral feature extraction methods are well developed in the speech and speaker recognition literature, and thus can be presented in great detail. The other algorithms have received only a little development, and each author has presented a different, intuitively motivated way of measuring the different properties known to be important for the perceptions of human subjects. Therefore, mathematical details are given only as much as is necessary with these algorithms. The development of robust feature detection algorithms for these various features is a fertile area of future research, and is likely to be the main source of performance increase for musical instrument recognition systems, since many authors have emphasized the importance of a set of salient features over sophisticated classification schemes.

## 4.1 Cepstral coefficients

Formants are spectral prominences created by one or more resonances in the sound source. They represent essential information for speech and speaker recognition, and also for musical instrument recognition. A robust feature for measuring formant information, or the smooth spectral envelope, are cepstral coefficients. The cepstrum of a signal *y(n)* is defined as

$$c(n) = F^{-1}\{\log|F\{y(n)\}|\}, \tag{2}$$

where *F* stands for the discrete Fourier transform (DFT). Calculating cepstral coefficients from the above equation is not very efficient, since two fast Fourier transforms (FFT) are needed. The coefficients can be more efficiently calculated from a mel-frequency filterbank, or from linear prediction coefficients.

Another reason for not using the above equation is the utilization of psychoacoustic frequency scales. DFT uses linear frequency resolution, so we must use some kind of warping transform to convert the linear frequency scale into a perceptual scale. Also the conventional LP analysis has this limitation, but one way to overcome the problem is to use warped linear prediction (WLP) based features.

## 4.2 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients ([Davis80]) have become one of the most popular techniques for the front-end feature-extraction in automatic speech recognition systems.

Brown has utilized cepstral coefficients calculated from a constant-Q transform for the recognition of woodwind instruments [Brown99, Brown01]. We will use here the conventional FFT-based method utilizing a mel-scaling filterbank. Figure 3 shows a block diagram of the MFCC feature extractor. The input signal is first pre-emphasized to flatten the spectrum. Next, a filterbank consisting of triangular filters spaced uniformly across the mel-frequency scale and their heights scaled to unity, is simulated. The mel-scale is given by

$$Mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right), \tag{3}$$

where $f$ is the linear frequency value. To implement this filterbank, a window of audio data is transformed using the DFT, and its magnitude is taken. By multiplying the magnitude spectrum with each triangular filter and summing the values at each channel, a spectral magnitude value for each channel is obtained. The dynamic range of the spectrum is compressed by taking a logarithm of the magnitude at each filterbank channel. Finally, cepstral coefficients are computed by applying a discrete cosine transform (DCT) to the log filterbank magnitudes $m_j$ as follows:

$$c_{mel}(i) = \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}\left(j - \frac{1}{2}\right)\right). \tag{4}$$

DCT decorrelates the cepstral coefficients, thereby making it possible to use diagonal covariance matrices in the statistical modeling of the feature observations.

In most cases, it is possible to retain only the lower order cepstral coefficients to obtain a more compact representation. The optimal order is examined in our simulations in Chapter 6. The lower coefficients describe the overall spectral shape, whereas pitch and spectral fine structure information is included in higher coefficients. The zeroth cepstral coefficient is normally discarded, as it is a function of the channel gain.

The dynamic, or transitional properties of the overall spectral envelope can be characterized with delta cepstral coefficients [Soong88, Rabiner93]. A first order differential logarithmic spectrum is defined by

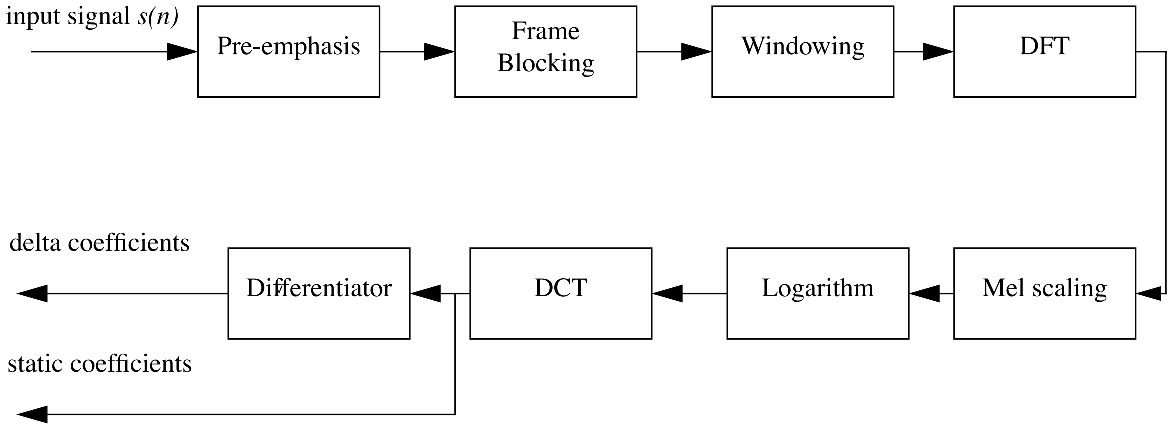

**Figure 3.** Block diagram of the MFCC feature extractor.

$$\frac{\partial \log S(\omega,t)}{\partial t} = \sum_{n=-\infty}^{\infty} \frac{\partial c_n(t)}{\partial t} e^{-jn\omega}, \qquad (5)$$

where $c_n(t)$ is the cepstral coefficient $n$ at time $t$ [Rabiner93, Young00]. Usually the time derivative $\partial c_n(t)/(\partial t)$ is obtained by polynomial approximation over a finite segment of the coefficient trajectory, since the cepstral coefficient sequence $c_n(t)$ does not have any analytical solution. In the case of fitting a first order polynomial $h_1 + h_2 t$ into a segment of the cepstral trajectory $c_n(t), t=-M, -M+1,..., M$, the fitting error to be minimized is expressed as

$$E = \sum_{t=-M}^{M} [c_n(t) - (h_1 - h_2 t)]^2. \qquad (6)$$

The resulting solution with respect to $h_2$ is

$$h_2 = \frac{\sum_{t=-M}^{M} t c_n(t)}{\sum_{t=-M}^{M} t^2}, \qquad (7)$$

and is used as an approximation for the first time derivative of $c_n$ [Rabiner93], which we denote by $\delta_n(t)$. This gives a smoother estimate of the derivative than a direct difference operation. The curve fitting is done individually for each of the cepstral coefficient trajectories $c_n$, $n=1, 2,..., L$.

More features can be obtained by estimating the second order derivative, and the resulting features are referred as acceleration coefficients in the speech recognition literature. The efficiency of MFCCs is due to the mel-based filter spacing and the dynamic range compression in the log filter outputs, which represent the mechanisms present in human hearing in a simplified way [DePoli97].

## 4.3 Linear prediction

Linear prediction analysis is another way to obtain a smooth approximation of the sound spectrum. Here, the spectrum is modeled with an all-pole function, which concentrates on spectral peaks. The human ear is known to be relatively insensitive to zeros. Linear prediction is particularly suitable for speech signals, but can be applied also to musical instrument recognition, although musical instruments rarely can be modeled as linear systems, as became apparent based on the discussion in Chapter 2. Schmid applied LP analysis to musical instrument recognition already in 1977 [Schmid77]. We first describe the conventional linear prediction, and solving the coefficients using the autocorrelation method. Then we discuss a means to modify the LP feature extractor with a cascade of all-pass filters to obtain warped linear prediction based features.

### Conventional forward linear prediction

In classical forward linear prediction, an estimate for the next sample $\hat{y}(n)$ of a linear, discrete-time system, is obtained as a linear combination of $p$ previous output samples:

$$\hat{y}(n) = \sum_{i=1}^{p} a_i y(n-i)\,, \tag{8}$$

where $a_i$ are the predictor coefficients, or linear prediction coefficients. They are fixed coefficients of a predictor all-pole filter, whose transfer function is

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}\,. \tag{9}$$

The goal of linear prediction is to find the set of predictor coefficients $\{a_1, a_2, \ldots, a_p\}$ that minimize the short-time mean-squared prediction error

$$e = E\left\{\left|y(n) - \sum_{i=1}^{p} a_i y(n-i)\right|^2\right\} \approx \sum_{n=-\infty}^{\infty} \left|y(n) - \sum_{i=1}^{p} a_i y(n-i)\right|^2\,, \tag{10}$$

where $E\{\ \}$ denotes expectation. By definition, $e$ is also the prediction error power. Several algorithms exist for minimizing $e$ and solving the predictor coefficients $a_i$, but here we will consider only the most popular approach, the autocorrelation method [Rabiner93].

To solve the above minimization problem, the partial derivatives of $e$ with respect to $a_i$ are set to zero:

$$\frac{\partial e}{\partial a_i} = 0, i = 1, 2, \ldots, p\,. \tag{11}$$

This leads us to a system of normal equations:

$$\sum_n y(n)y(n-k) - \sum_{i=1}^{p} a_i \sum_n y(n-i)y(n-k) = 0\,, k{=}0, 1,\ldots, p{-}1. \tag{12}$$

The autocorrelation function at time $u$ is defined as

$$R_u(k) = \sum_{m=-\infty}^{\infty} y_u(m) y_u(m+k)\,. \tag{13}$$

We can express Equation 12 in terms of the autocorrelation as:

$$\boldsymbol{Ra} = \boldsymbol{r} \Leftrightarrow$$

$$\begin{bmatrix} R_u(0) & R_u(1) & \ldots & R_u(p-1) \\ R_u(1) & R_u(0) & \ldots & R_u(p-2) \\ \ldots & \ldots & \ldots & \ldots \\ R_u(p-1) & R_u(p-2) & \ldots & R_u(0) \end{bmatrix}_{p \times p} \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{bmatrix}_{p \times 1} = \begin{bmatrix} R_u(1) \\ R_u(2) \\ \ldots \\ R_u(p) \end{bmatrix}_{p \times 1}\,. \tag{14}$$

These equations are also called the *Yule-Walker* equations. By applying the structure of the matrix $\boldsymbol{R}$, which is a symmetric Toeplitz matrix, the Yule-Walker equations can be solved in an efficient manner. The most efficient is known as the Durbin's method [Rabiner93, pp. 115], and can be given as follows (the subscript $u$ on $R_u(k)$ is omitted for clarity):

$$E^{(0)} = R(0) \tag{15}$$

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E^{(i-1)}}, \qquad 1 \le i \le p \tag{16}$$

$$a_i^{(i)} = k_i \tag{17}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \qquad 1 \le j \le i-1 \tag{18}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}. \tag{19}$$

Equations 16-19 are solved recursively for $i = 1, 2, \ldots, p$. The final solution is given at the $p^{\text{th}}$ iteration as the LP coefficients $a_j = a_j^{(p)}$, $1 \le j \le p$, and as the reflection coefficients $k_j$.

Now, according to the definition, the linear prediction cepstrum could be calculated directly as the Fourier transform of the filter coefficients. However, the required cepstral coefficients can be more efficiently computed using the recursion [Rabiner93]

$$c_n = -a_n - \frac{1}{n} \sum_{k=1}^{n-1} k c_k a_{n-k} \text{ for } n > 0, \tag{20}$$

where $a_0 = 1$ and $a_k = 0$ for $k > p$.

## 4.4 Warped linear prediction based feature extraction

The conventional LP-analysis suffers from a uniform frequency resolution. Especially in wideband audio applications, poles are wasted to the higher frequencies [Härmä00a]. In wideband audio coding, WLP has proved out to outperform conventional LP based codecs especially with low analysis orders [Härmä00a]. Motivated by this, using cepstral coefficients based on linear prediction on a warped frequency scale was experimented, and the performance is experimentally evaluated in Chapter 6. We begin by reviewing the theory behind the frequency warping transform obtained by replacing the unit delays of a discrete, linear system with first-order all-pass elements. Then we describe a modified LP feature extractor. The discussion here quite slavishly follows the one presented in [Härmä00a]. The technique of warped linear prediction was first proposed by Strube in 1980 [Strube80].

### Obtaining the frequency warping transform

The normalized phase response of a first-order all-pass filter whose transfer function is given by

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \tag{21}$$

is shown in Figure 4a for some real values of $\lambda$. Figure 4b shows the group delays for the same filters. For $\lambda$=0, *D(z)* reduces to a single unit delay having linear phase and constant group delay

If we feed a signal into a cascade of all-pass elements with positive λ, the nonuniform group delay of the elements makes low frequency components proceed slower and high frequency components faster than in a chain of unit delays. Now if we form a new sequence of the values from the outputs of the all-pass chain, we frequency dependently resample the signal.

The resulting mapping from the natural frequency domain to a warped frequency domain is determined by the phase function of the all-pass filter, which is given by

$$\tilde{\omega} = \operatorname{atan}\frac{(1-\lambda^2)\sin(\omega)}{(1+\lambda^2)\cos(\omega)-2\lambda}, \tag{22}$$

where $\omega = 2\pi f / f_s$ and $f_s$ is the sampling rate [Härmä00a].

The temporal structure of the original signal also changes in the warping transformation. The group delay function of the all-pass filter controls the change in length of a sinusoidal signal. The turning point frequency $f_{tp}$ can be expressed as

$$f_{tp} = \pm\frac{f_s}{2\pi}\operatorname{acos}(\lambda), \tag{23}$$

and is equal to the point where the group delay is equal to one sample period. At this point, a warped sinusoid is as long as the original sinusoid and frequency warping does not change its frequency [Härmä00a].

The frequency transformation can be made to approximate the mapping occurring in human ear by selecting the value of λ. The Bark rate scale mapping for a given sampling frequency $f_s$ is given by the expression [Smith99]

$$\lambda_{Bark}(f_s) \approx 1.0674\left[\frac{2}{\pi}\operatorname{atan}(0.06583 f_s)\right]^{\frac{1}{2}} - 0.1916. \tag{24}$$

In our simulations, the value of λ was 0.7564 for 44.1 kHz sampling rate. The mapping occurring in the all-pass chain is a very good approximation of the Bark rate scale mapping.

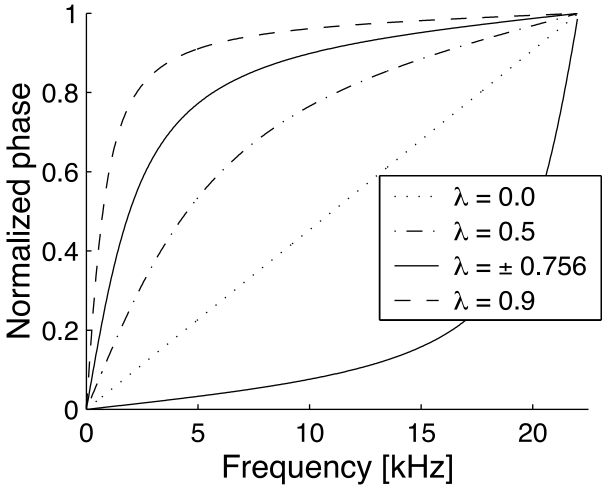

**Figure 4a.** Phase response of a first-order all-pass filter for several values of λ.

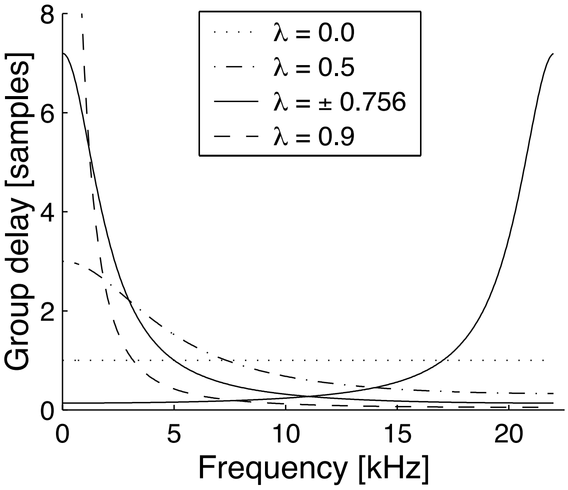

**Figure 4b.** Group delay of a first-order all-pass filter for several values of λ.

Frequency warped signals and systems can be produced by replacing the unit delays of the original system by first order all-pass elements. This can be interpreted in the $z$ domain by the mapping

$$z^{-1} \rightarrow \tilde{z}^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} . \tag{25}$$

## Warped linear prediction

The $z$-transform of Equation 8 is

$$\hat{Y}(z) = \left( \sum_{i=1}^{p} a_i z^{-i} \right) Y(z) . \tag{26}$$

Now, according to the above discussion, the unit delay $z^{-1}$ is replaced by a first-order all-pass filter $D(z)$, given in Equation 21, to warp the system:

$$\hat{Y}(z) = \left[ \sum_{i=1}^{p} a_i D(z)^i \right] Y(z) . \tag{27}$$

In the time domain, $D(z)$ is interpreted as a generalized shift operator which is defined as

$$d_i[y(n)] \equiv \underbrace{\delta(n) \otimes \delta(n) \otimes \ldots \otimes \delta(n)}_{\text{i-fold convolution}} \otimes y(n) , \tag{28}$$

where $\otimes$ denotes convolution and $\delta(n)$ is the impulse response of $D(z)$. Furthermore, $d_0[y(n)] \equiv y(n)$. Now we can write the mean-squared prediction error estimate

$$e = E\left\{ \left| y(n) - \sum_{i=1}^{p} a_i d_i[y(n)] \right|^2 \right\} . \tag{29}$$

The normal equations can be written as

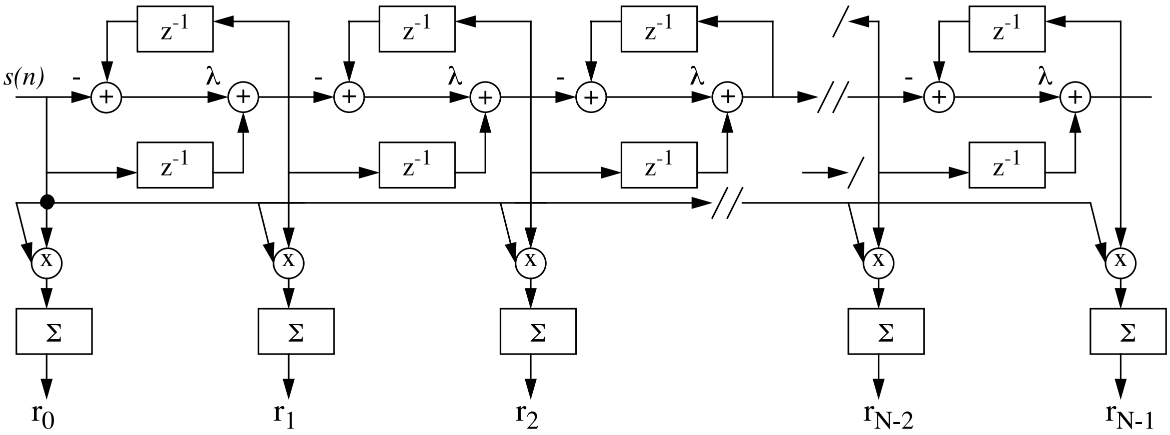

**Figure 5.** Warped autocorrelation network for continuous $N$-tap warped autocorrelation calculation (after [Härmä00a]).

$$E\{d_k[y(n)]d_0[y(n)]\} - \sum_{i=1}^{p} a_i E\{d_i[y(n)]d_k[y(n)]\} = 0\,, k=0,...,p-1. \qquad (30)$$

Since *D(z)* is an all-pass filter,

$$E\{d_k[y(n)]d_l[y(n)]\} = E\{d_{k+m}[y(n)]d_{l+m}[y(n)]\} \quad \forall k, l, m\,, \qquad (31)$$

and the same correlation values appear in both parts of Equation 30 [Härmä00a]. It can therefore be seen as a generalized form of the Yule-Walker equations and can be solved using the autocorrelation method described previously. The autocorrelation values can be computed using the autocorrelation network shown in Figure 5. The network consists of a cascade of all-pass elements (higher part) and blocks for autocorrelation calculation (lower parts).

Now we have means to modify the conventional linear-prediction based feature extraction scheme in order to obtain a more perceptually motivated analysis. A block diagram of the WLP feature extractor is shown in Figure 6. The autocorrelation method of solving the linear prediction coefficients is modified in such a way that the autocorrelation network is replaced with the warped autocorrelation network. The rest of the feature extractor, such as the conversion to cepstral coefficients and delta cepstral coefficient calculation, are kept the same.

The WarpTB toolbox by Härmä and Karjalainen was used for implementing the warped linear prediction calculation [Härmä00b]. It consists of Matlab and C implementations of the basic functions, such as the warped autocorrelation calculation.

## 4.5 Cepstral feature vector formation

For isolated musical tones, it has been found that the onset portion is important for recognition by human subjects. Motivated by this, the cepstral analyses were made separately for the onset and steady state portions of the tone. Based on the RMS-energy level of the signal, it was segmented into onset and steady state parts. This is described in more detail in Section 4.7.

For the onset portion of tones, both LP and mel-cepstral analyses were performed in approximately 20 ms long hamming-windowed frames with 25 % overlap. In the steady state segment, frame length of 40 ms was used. If the onset was shorter than 80 ms, the beginning of steady state was moved forward so that at least 80 ms was analyzed. For the MFCC calculations a discrete Fourier transform was first calculated for the windowed waveform. The length of the

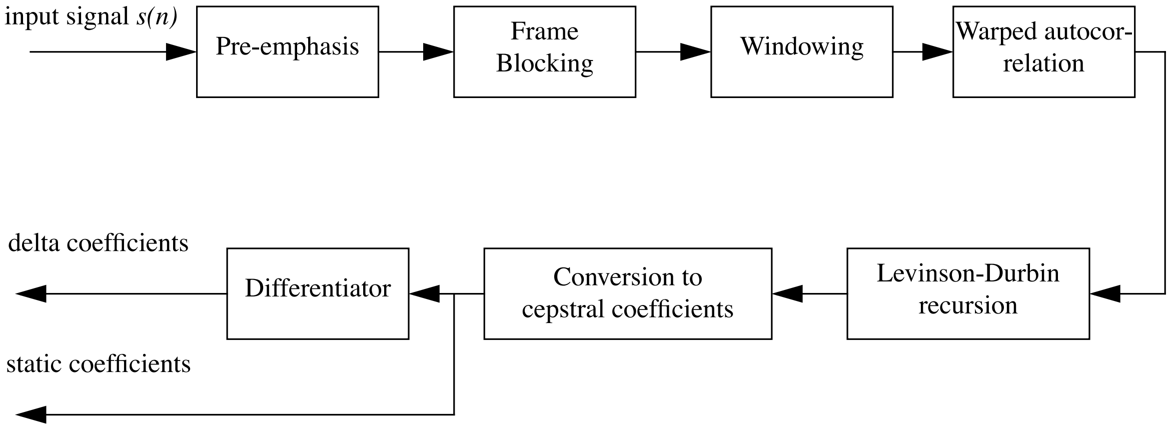

**Figure 6.** Block diagram of the WLPCC feature extractor.

transform was 1024 or 2048 samples for 20 ms and 40 ms frames, respectively. For both LP and mel-cepstral analyses, the median values of cepstral coefficients were stored for the onset and steady state segments. The median was selected instead of the mean for robustness considerations. For instance, if the segmentation scheme failed, few disturbed coefficient values might significantly change the mean of coefficients. For the delta-cepstral coefficients, the median of their absolute value was calculated. We also experimented with coefficient standard deviations in the case of the MFCCs.

For use with Gaussian mixture models, the cepstral and delta-cepstral coefficients were also stored as observation sequences from adjacent frames. Energy thresholding was used here to prevent the silent partitions at the beginning and end of a single note from disturbing the feature values. The frames that had an RMS-energy more than 10 dB below a mean energy were dropped.

## 4.6 Spectral centroid

Spectral centroid (SC) is a simple but very useful feature. Research has demonstrated that the spectral centroid correlates strongly with the subjective qualities of “brightness” or “sharpness”. It can be calculated from different mid-level representations, commonly it is defined as the first moment with respect to frequency in a magnitude spectrum. However, the harmonic spectrum of a musical sound is hard to measure, as we will soon see, therefore more robust feature values are obtained if spectral centroid is calculated from the outputs of a filterbank. We calculated spectral centroid according to the following equation

$$f_{sc} = \frac{\sum_{k=1}^{B} P(k)f(k)}{\sum_{k=1}^{B} P(k)}, \qquad (32)$$

where *k* is the index of a filterbank channel, whose RMS-power is *P(k)*, and center frequency *f(k)*, and *B* is the total number of filterbank channels. We used 1/6-octave bands, meaning that there are six spectral lines per octave. The filterbank was simulated via the FFT. In practise, in the lower frequencies the resolution converged into the linear resolution of the FFT, and at the higher bands the power of the band became the RMS-power of the FFT bins in that channel. The relative spectral centroid was also used, and is defined as

$$f_{scr} = \frac{f_{sc}}{f_0}, \qquad (33)$$

where $f_0$ is the fundamental frequency of a harmonic sound, as given by the algorithm developed by Klapuri [Klapuri99a]. Another way of estimating spectral centroid is to calculate it from the outputs of the mel-frequency filterbank, which slightly simplifies the implementation. However, using 1/6-octave bands gave slightly better results (although not statistically significant), and was used in the final simulations.

The SC of the signal was calculated as a function of time in approximately 20 ms windows with 50 % overlap. Depending on the classifier, the spectral centroid and the relative spectral centroid were stored as a sequence for each note, or the mean and standard deviation were calculated from the observation and used to characterize a note. Standard deviation of spectral centroid and relative SC can be considered as a measure for vibrato. It should be noted that these features depend on the overall colorations of the signal, as well as the pre-processing, or high-pass filtering used.

## 4.7 Amplitude envelope features

Amplitude envelope contains information for instance about the type of excitation; e.g. whether a violin has been bowed or plucked. Tight coupling between the excitation and resonance structure is indicated by short onset durations. The amplitude envelope of a sound can be calculated by half-wave rectification and lowpass filtering of the signal. Another means is the calculation of the short time RMS-energy of the signal, which we found to be more a more straightforward way of obtaining a smooth estimate of the amplitude envelope of a signal. The latter was used in the simulations. We estimated rise-time, decay-time, strength and frequency of amplitude modulation, crest factor and detected exponential decay from a RMS-energy curve calculated in 50 % overlapping 10 ms hanning-windowed frames.

### Onset duration

Rise time, i.e. the duration of onset, is traditionally defined as the time interval between the onset and the instant of maximal amplitude of a sound. However, from natural sounds the maximal amplitude can be achieved at any point, therefore some thresholds have to be used. We implemented a relatively simple algorithm which is based on energy thresholds. First, the starting point of the attack portion is searched, and it is defined as the point where the short-time RMS-energy rises above the -10 dB point below the average RMS-energy of the note.

For onset duration calculation, the 10-base logarithm is taken of the RMS-energy envelope, and multiplied by 10. Then the obtained logarithmic short-time RMS-energy curve is smoothed by convolving it with a 45-ms hanning window. Then a maximum is searched from the smoothed envelope, and the point where the RMS-energy achieves the -3 dB point of the

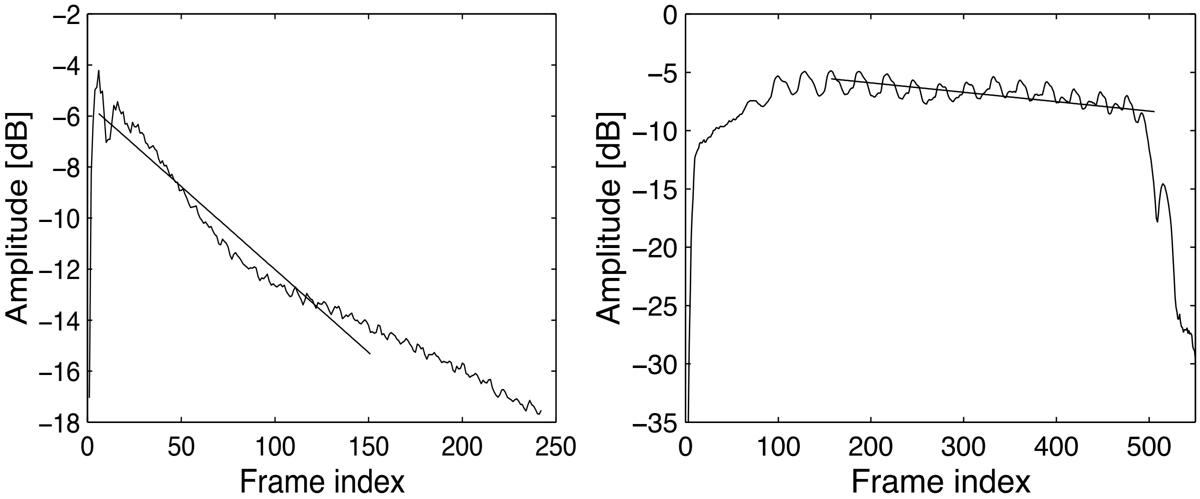

**Figure 7.** Short-time RMS-energy envelopes for piano (left) and violin tones (right). Post-onset decay is measured by fitting a line on dB-scale. The different onset durations, slight beat in the guitar tone, and amplitude modulation in the violin tone are clearly visible.

maximum is taken as the end of attack. Since the point is an index to a frame, we linearly interpolate the end point in samples.

Some ad-hoc rules were added to the algorithm in order that it would cope with different types of notes, pizzicato, sustained and with those where the onset or offset has been removed. For instance, if the onset is very fast, the curve is not convolved at all.

### Other features describing the amplitude envelope

An algorithm quite similar was used to measure the decay time from single note signals. The end of steady state is defined as the point where the short-time RMS-energy goes permanently below the -3 dB point from the maximum. The -10 dB fall after this point is defined as the decay, and its duration as the decay time. For pizzicato tones, this is shorter than for long, sustained tones, however it fails if very short sustained tones are presented for the system, or if the decay portion is removed.

To measure the slope of amplitude decay after the onset, a line is fitted into the amplitude envelope on a logarithmic scale. The fitting was done for the segment of the energy envelope that was between the maximum and the -10 dB point after that. Also, the mean square error of that fit is used as a feature describing exponential decay. Crest factor, i.e. the maximum of amplitude envelope / RMS of amplitude envelope is also used to characterize the shape of the amplitude envelope. These three features aim at discriminating between the pizzicato and sustained tones: the former ones decay exponentially, and have a higher crest factor than sustained tones. The error of line fit may be small for sustained tones with very little fluctuation in the quasi-steady state, however, if there exists amplitude modulation, or some changes, the value becomes larger than with exponentially decaying tones. In addition, the very slow amplitude beating often encountered with plucked strings causes problems. For example, the amplitude of guitar tones often first decays exponentially but then starts to grow again, and a linear fit fails to characterize this decay. Figure 7 shows amplitude envelopes for guitar and violin, and the line fitted after the onset period.

### Amplitude modulation extraction

The RMS-energy envelope, now on a linear scale, is also used to extract features measuring amplitude modulation (AM) properties. Strength, frequency, and heuristic strength (term used by Martin [Martin99]) of amplitude modulation (AM) is measured at two frequency ranges. Rates from 4 to 8 Hz measure tremolo, i.e. AM in conjunction with vibrato, and rates between 10–40 Hz correspond to “graininess” or “roughness” of the tone. The RMS-energy envelope is first windowed with a hanning window. Then, FFT analysis is performed on the windowed envelope, and maxima are searched from the two frequency ranges. The frequency of AM is the frequency of the maximum peak. The amplitude features are calculated as the difference of the peak amplitude and the average amplitude, and the heuristic amplitude is calculated as the difference of the peak amplitude and the average amplitude of the frequency range under consideration. Thus, when performed for these two frequency ranges we end up with a total of six features describing AM characteristics. However, the heuristic strength of AM at the range 10-40 Hz was found irrelevant and was not used in the simulations.

## 4.8 Sinusoid envelopes and onset asynchrony features

Transforms such as FFT or constant-Q transform are designed to give precise frequency information in one frame, but are not effective in measuring the precise time evolution of frequency components. With these transforms, the calculation becomes very inefficient if short analysis steps are needed. Sinusoid envelopes is a representation that was employed to obtain a representation capable of describing the detailed time evolution of harmonic partials. It is quite straightforward to calculate. However, it has some drawbacks which we will soon discuss.

### Calculating the envelope of a single sinusoid

The discrete Fourier transform *X(k)* of a sampled time domain signal *y(n)* is calculated as [Ifeachor93]

$$X(k) = \sum_{n=0}^{N-1} y(n) e^{-jk\frac{2\pi n}{N}}, \tag{34}$$

where *k* is a discrete frequency bin and *N* is the length of the analysis frame. The most common use of this transform is to calculate it over the whole frequency range from *k = 0* to *k = N/2*, i.e. half the sampling rate using the FFT. However, the transform can also be effectively calculated for a certain frequency bin in successive time instants.

When calculating the transform for a certain frequency bin, the length of the time domain analysis frame can be adjusted to be a multiple of the wavelength of that frequency, and we refer it as pitch synchronous analysis. In this case, the frequencies of the harmonic components correspond to the frequencies of the DFT components. In addition, the length of the time frame may be very short, e.g. 3 waves, and the windowing function in the time domain is not needed. Now, a precise sample-by-sample amplitude envelope of that frequency can be extracted through time. First, we calculate the transform for a certain frequency bin as usual, and store all the elements of the sum in Equation 34. Now the transform for the same bin in a time frame one sample later is calculated by subtracting the first element of the sum in the previous frame, and adding a new element calculated by

$$y(n) e^{-jk\frac{2\pi n}{N}}, \tag{35}$$

where *n* now points to the sample right after the previous frame.

### Calculating the representation

The process is then repeated for the different partials in the sound. Thus, this algorithm is asymptotically *O(rm)*, where *r* is the number of sinusoids, and *m* is the length of the analyzed signal, which is not necessarily equal to *N*. For characterizing only the attack portions, 500 ms of the signal would be sufficient. However, in our simulations we calculated the representation for whole signals for the extraction of features relating to the fluctuations in the steady state. The number of sinusoids was limited to *r = 40* in our simulations.

However, it is not perceptually relevant to analyze the high frequency partials independently since the ear’s sensitivity is lower for high frequencies. Therefore, we use a representation having Bark frequency resolution. For each 24 Bark scale bands, we first calculate whether any harmonic frequencies are found on the current band. Then the envelopes are calculated for each harmonic component on that band. If there are more than one components, the resulting band-amplitude is calculated as the mean of the band-amplitudes. Then, an estimate of the intensity in that band is calculated by multiplying the band-amplitude with the center frequency of that band. The intensities are decimated by a factor of about 5 ms to ease the feature computations and smoothed by convolving with a 40 ms half-hanning (raised-cosine) window. This window preserves sudden changes, but masks rapid modulation. Figure 8 displays intensity versus Bark frequency plots for 261 Hz tones produced by flute and clarinet.

The sinusoid envelope representation is relatively compact but still bears high perceptual fidelity to the original sound. However, the problem becomes how to reliably measure the frequencies to be tracked. We can use a fundamental frequency estimation algorithm to find the fundamental frequency of a tone, and then analyze the frequencies equal to the fundamental and its integer multiples. However, musical sounds are only quasi-harmonic, and errors unavoidably occur in the process. There exist methods for following the frequencies of partials, but their description is not relevant in our scope. Another significant source of uncertainty and errors are the failures in estimating the fundamental frequency. Despite these limitations, this representation is a useful first attempt towards measuring the time evolution of partials. The MPEG-7 standard uses a quite similar representation, although the DFT is calculated in frames [Peeters00], which causes limitations with the feasible time resolution. With regard of future developments of our system, using a filterbank instead of sinusoid envelopes would be a simpler and more robust approach.

### Calculating features from the representation

Onset asynchrony refers to the differences in the rate of energy development of different frequency components. The sinusoid envelope representation is used to calculate the intensity envelopes for different harmonics, and the standard deviation of onset durations for different harmonics is used as one feature. For the other feature measuring this property, the intensity envelopes of individual harmonics are fitted into the overall intensity envelope during the onset period, and the average mean square error of those fits was used as feature. A similar measure was calculated for the rest of the waveform. The last feature calculated is the overall variation of intensities at each band.

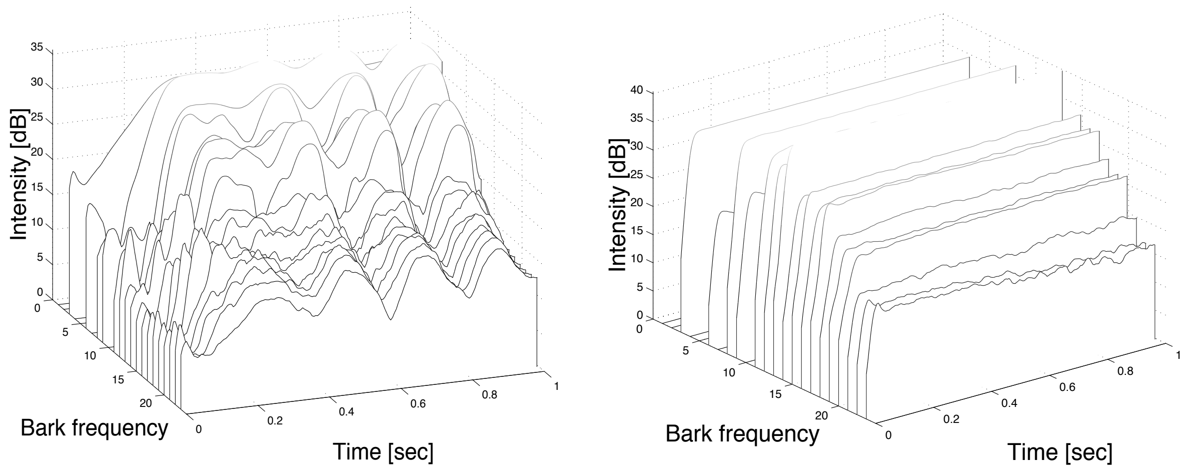

**Figure 8.** Sinusoid envelope representations for flute (left) and clarinet (right), playing the note C4, 261 Hz.

## 4.9 Fundamental frequency and frequency modulation

Different types of frequency modulation are characteristic to some sound sources. The term vibrato refers to periodic modulations, and jitter to random modulations. These features are quite difficult to measure, because many pitch-tracking algorithms require quite a long time frame. This makes especially rapid modulations hard to detect reliably. One interesting approach might be to modify the RAPT algorithm [Talkin95] for modulation tracking. An initial estimate of the fundamental frequency (F0) would be given first, and then a more accurate estimate for the F0 would be searched using cross-correlation in two short windows. Martin estimated frequency modulation from the outputs of a log-lag correlogram [Martin99]

Two of our features were indirect measures of frequency modulation. The standard deviation of spectral centroid, and the standard deviation of F0 estimated in successive frames. However, measuring jitter is not feasible with the current implementation, since the F0-estimation algorithm requires at least a 40 ms window. We used the algorithm presented by Klapuri in [Klapuri99a], whose detailed description is out of the scope of this thesis. A pitch envelope was calculated in 40 ms hanning windowed frames with 50 % overlap, and the mean and standard deviation of F0 estimates were used as features, and also to rule out classes in the first experiment in Section 6.2.

## 4.10 Additional features and discussion

The list of features used in this study are summarized in Table 4. Here we briefly discuss some of the most promising other features presented in the literature which, however, are not evaluated in this study.

Spectral irregularity (IRR) corresponds to the standard deviation of time-averaged harmonic amplitudes from a spectral envelope, and is introduced in [Krimphoff94] as:

$$IRR = 20\log 10\left( \sum_{k=2}^{r-1} \left| A_k - \frac{A_{k-1} + A_k + A_{k+1}}{3} \right| \right), \qquad (36)$$

where $A_k$ is the amplitude of the $k^{\text{th}}$ partial, and $r$ is the number of partials. This has been used in musical instrument recognition by [Kostek99] and [Fujinaga00]. Jensen has presented a modified version

$$IRR = \frac{\sum_{k=1}^{r} (A_k - A_{k+1})^2}{\sum_{k=1}^{r} A_k^2}. \qquad (37)$$

Recently, Brown used a feature relating to these [Brown01]. In her experiments, the bin-to-bin differences in constant-Q coefficients and a quefrency domain derivative gave excellent results in recognizing four woodwind instruments. These features could be calculated from the sinusoid envelope representation, or alternatively from the outputs of a perceptual filterbank implemented either in time or frequency domain.

**Table 4: List of features implemented in this study and the sections in text where they were described.**

| Feature | Feature |
|---|---|
| Onset duration (4.7) | Strength of AM, range 10-40 Hz (4.7) |
| Post onset slope (post onset line fit, 4.7) | Standard deviation of component onset durations (4.8) |
| Exponential decay (goodness of fit, 4.7) | Mean error of the fit between steady state intensities and intensity envelope (4.8) |
| Decay time (4.7) | Mean error of the fit between onset intensities and onset intensity envelope (4.8) |
| Time between the end of attack and the maximum of RMS-energy (4.7) | Overall variation of intensities at each band (4.8) |
| Crest factor (4.7) | Fundamental frequency (4.9) |
| Mean of spectral centroid (SC, 4.6) | Std of fundamental frequency (4.9) |
| Mean of relative SC (4.6) | Linear prediction cepstral coefficients (4.3) |
| Max of relative SC (4.6) | Linear prediction delta cepstral coefficients (4.2) |
| Std of SC (4.6) | Reflection coefficients (based on LP, 4.3) |
| Std of relative SC (4.6) | Warped LP cepstral coefficients (4.4) |
| Frequency of AM, range 4-8 Hz (4.7) | Warped LP delta cepstral coefficients (4.2) |
| Strength of AM, range 4-8 Hz (4.7) | Reflection coefficients (based on WLP, 4.4) |
| Heuristic stregth of AM, range 4-8 Hz (4.7) | Mel-frequency cepstral coefficients (4.2) |
| Frequency of AM, range 10-40 Hz (4.7) | Mel-frequency delta cepstral coefficients (4.2) |

# 5 Classification methods

This chapter presents the different classifiers applied in our experiments. Two main types of classifiers were used: distance-based classifiers and probabilistic classifiers.

## 5.1 Distance-based classifiers

The k-nearest neighbors (k-NN) classifier is a typical example of a distance-based classifier. It stores all the training examples and then calculates a distance between the test observation and all the training observations, thus it employs lazy learning by simply storing all training instances. The class of the most closest training example is given as the classification result (1-NN), or the class appearing most often among the k nearest training observations (k-NN).

A suitable distance metric needs to be chosen with the k-NN. We used the Euclidean distance metric in a normalized space that was obtained with the discrete form of the KL-transform [Parsons87]. The transform is a special case of the principal component analysis if none of the dimensions is dropped. It is also equal to using the Mahalanobis distance with the same covariance matrix for all classes which is estimated from the whole data, and calculating the distance to all the training examples instead of class means. In the normalized space, the features are uncorrelated and the range of variation of each feature is the same. A mathematical formulation of the transform can be found in [Parsons87, pp. 183].

The k-NN classifier is straightforward to implement, and it can form arbitrarily complex decision boundaries. Therefore it was used in many of our simulations. The problem of the k-NN classifier is that it is sensitive to irrelevant features which may dominate the distance metric. In addition, the calculation requires a significant computational load if a large number of training instances is stored.

## 5.2 Probabilistic classifiers

The statistical classifiers used in this thesis assume that the data follows a certain distribution, and try to estimate the parameters of the class distributions from the training observations. Knowing the probability density function of the assumed distribution, the likelihood of each class distribution of generating the test observation can then be calculated.

### Multinormal Gaussian

Let us consider $M$ pattern classes each of which is governed by the multivariate Gaussian distribution:

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)\right\}, \tag{38}$$

with mean vector $\boldsymbol{m}_i$ and covariance matrix $\Sigma_i$. $\mathbf{x}$ is a $D$-dimensional observation vector and $p(\boldsymbol{x}|\omega_i), i = 1, \ldots, M$ are the conditional probabilities of $\boldsymbol{x}$ given the class $\omega_i$. According to the Bayes theorem, the decision function for class $\omega_i$ can be chosen as $d_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)p(\omega_i)$ [Tou74]. However, often it is not desirable to use the *a priori* probabilities $p(\omega_i)$, but they are assumed equal and thus can be discarded. Also the term $(2\pi)^{D/2}$ is the same for all classes and can be discarded.

With small data sets, the remaining problem is in reliably estimating the training data statistics [Tou74]. The amount of training data required grows exponentially with respect to the number of dimensions. Another problem with this classifier is that real feature data often does not follow a Gaussian distribution, however, sometimes the logarithm of the feature follows a Gaussian distribution better.

### Gaussian mixture model (GMM)

A Gaussian mixture model presents each class of data as a linear combination of several Gaussian densities in the feature space. The parameters of the component densities can be iteratively estimated with the well-known expectation maximization (EM) algorithm [Moon96]. Reynolds introduced the use of Gaussian mixture models for speaker recognition [Reynolds95]. His first motivation was that the individual component densities would be able to model some underlying acoustic classes, such as vowels, nasals or fricatives. Second, a linear combination of Gaussian basis functions is capable of forming smooth approximations of arbitrarily shaped densities. Brown has successfully applied Gaussian mixture models for the recognition of woodwind instruments [Brown99, Brown01].

A Gaussian mixture density is a weighted sum of $M$ component densities as given by the equation [Reynolds95]

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x}), \tag{39}$$

where $\boldsymbol{x}$ is a $D$-dimensional feature or observation vector, $b_i(\boldsymbol{x}), i = 1, \ldots, M$, are the component densities and $p_i$ the mixture weights. Each component density is a $D$-variate Gaussian function of the form defined in Equation 38. The mean vectors, covariance matrices and mixture weights of all Gaussian functions together parameterize the complete Gaussian mixture density. These parameters are collectively represented by the notation

$$\lambda = \{p_i, \boldsymbol{m}_i, \Sigma_i\}, i = 1, \ldots, M. \tag{40}$$

The mixture weights satisfy the constraint

$$\sum_{i=1}^{M} p_i = 1. \tag{41}$$

During the training process, the maximum likelihood (ML) estimation is applied to determine the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of $T$ training vectors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the GMM likelihood can be written as

$$p(X|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda). \tag{42}$$

Since this expression is a nonlinear function of the parameters λ, direct optimization is not possible. Therefore, the ML estimates of the GMM parameters are obtained iteratively by using a special case of the EM algorithm. The algorithm begins with an initial model λ and estimates a new model $\tilde{\lambda}$ such that $p(X|\tilde{\lambda}) \geq p(X|\lambda)$. At each iteration, the following reestimation formulas are used which guarantee a monotonic increase in the model's likelihood value [Reynolds95].

- Mixture weight update:

$$\tilde{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda) \tag{43}$$

- Mean vector update:

$$\tilde{\boldsymbol{m}}_i = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)} \tag{44}$$

- Covariance matrix update:

$$\tilde{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)\mathbf{x}_t\mathbf{x}_t^T}{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)} - \tilde{\boldsymbol{m}}_i\tilde{\boldsymbol{m}}_i^T \tag{45}$$

Since we are using diagonal covariance matrices, we need to update only the diagonal elements in the covariance matrices. For an arbitrary diagonal element $s_i^2$ of the covariance matrix of the $i$th mixture, the variance update becomes:

$$\tilde{s}_i^2 = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)\mathrm{x}_t^2}{\sum_{t=1}^{T} p(i|\boldsymbol{x}_t, \lambda)} - \tilde{m}_i^2 \tag{46}$$

where the *a posteriori* probability for the $i$th mixture is given by

$$p(i|\boldsymbol{x}_t, \lambda) = \frac{p_i b_i(\boldsymbol{x})}{\sum_{k=1}^{M} p_k b_k(\boldsymbol{x})} \tag{47}$$

and $s_i^2$, $x_t$, and $m_i$ refer to individual elements of the vectors $\boldsymbol{s}_i^2$, $\boldsymbol{x}_t$, and $\boldsymbol{m}_i$, respectively.

**Initialization of Gaussian mixture models**

Several factors must be considered in training the Gaussian mixture model [Reynolds95].

*Selecting the parameters of the model*. First, the order $M$ of the model must be large enough to represent the feature distributions. However, too large a value will cause problems in the training process, as the amount of data becomes insufficient for a statistical model of many parameters, and the computational cost becomes excessive. In practise, the order needs to be experimentally determined. In Chapter 6, we evaluate the performance of the model using several different orders. Second, the type of covariance matrices for the mixture distributions needs to be selected. In our experiments, we used diagonal covariances since they simplify the implementation and are computationally more feasible than models with full covariances. In addition, a modeling capability of a set of full covariance Gaussians can be equally achieved by using a larger set of diagonal covariance Gaussians [Reynolds95].

*Initialization of the model*. The EM algorithm is guaranteed to find a local maximum likelihood model regardless of the initialization, but different initializations can lead to different local maxima. Since Reynolds found no significant differences in speaker recognition performance among single, random initialization schemes and more elaborate methods, we decided to leave the comparison of initialization methods outside the scope of this thesis. In our experiments, the initial means were randomly selected among the samples from the training data, and then followed by a single iteration of the k-means clustering to initialize the component means, nodal variances and mixture weights.

*Variance limiting*. When there is not enough training data to sufficiently train the variances of the components, or the data is corrupted by noise, the variance values can become very small in magnitude, which causes singularities in the model's likelihood function. To prevent this, a variance limiting constraint was applied to the estimated variances after each EM iteration. Now the variance estimate $\tilde{s}_i^2$ for an arbitrary element of mixture $i^{\text{th}}$ variance vector becomes

$$\tilde{s}_i^2 = \begin{cases} s_i^2, if & s_i^2 > s_{min}^2 \\ s_{min}^2, if & s_i^2 \leq s_{min}^2 \end{cases} \tag{48}$$

where $s_{min}^2$ is the minimum variance value. In our experiments, the value used was $s_{min}^2 = 0.01$.

The H2M Toolbox by Olivier Cappe [Cappe01] was used as an implementation for the Gaussian mixture models. It consists of a combined Matlab and C implementations of the basic structure of the model and the EM-algorithm.

## 5.3 Feature selection and dimensionality reduction

Often when implementing a wide set of features some of them prove out to be irrelevant and may cripple the classification system even if the other features were good. Furthermore, with a small set of high dimensional data we are not able to reliably estimate the parameters of a statistical model. An obvious problem is how to find the relevant features and discard the others. Several techniques have been proposed and applied in the context of musical instrument recognition. Some techniques transform the feature space into a new with reduced number of dimensions which best explain the information in the data, such as the PCA [Kaminskyj95] or the Fisher discriminant analysis [Martin98]. Another technique is to use feature selection or weighting algorithms. In order to find good feature combinations, genetic algorithms have been applied in [Fujinaga98]. We assumed that many of our features were irrelevant, thus a feature selection scheme seemed suitable. Two simple feature selection algorithms were implemented and tested.

The *sequential forward generation* (SFG) starts with an empty set of features and adds features from the set of available features one by one [Liu98]. At each round, the feature the addition of which gives the best performance is selected. This is continued until the performance does not improve, or a desired level of performance is obtained. This algorithm is straightforward to implement, and a ranked list of features can be obtained. However, the algorithm often converges to a suboptimal solution.

Correspondingly, the *sequential backward generation* (SBG) starts removing features one by one from the set of all available features [Liu98]. The least relevant feature is removed at each iteration, i.e. the feature whose removal most improves the classification performance, or does not make it worse. However, in many cases this algorithm does not give the minimal set of features.

These two algorithms complement each other. If the number of relevant features is smaller than $D/2$, where $D$ is the number of available features, SFG is quicker, and if it is greater than $D/2$, SBG performs faster. In our simulations, the SBG gave better results, the SFG often converged to a suboptimal solution. In our second experiment in Chapter 6, we report a subset of features that gave the best performance, and which was obtained with the SBG.

# 6 System evaluation

This chapter describes the experiments which were done to evaluate the system with varying amounts of data, and with different features and classification schemes. We first introduce the evaluation database. Then the computer simulations and results are presented. We present three experiments each with different issues of study. The first is a preliminary experiment with a subset of the evaluation material. Its purpose was to compare the selected approach to earlier reporter experiments using the same material, and to experiment with a hierarchic classification framework. *The second experiment is the most relevant* with regard of the research problem and evaluating the implemented methods. It introduces a realistic performance evaluation with a comprehensive acoustic material. In the third experiment, an alternative approach for the research problem is presented using speaker recognition techniques.

## 6.1 Acoustic material

Samples from five different sources were included in the validation database. The first two sources include the samples from the McGill University Master Samples Collection (MUMS) [Opolko87], as well as recordings of an acoustic guitar made at Tampere University of Technology. The other sources of samples were the University of Iowa website [UIowa00], IRCAM Studio Online [SOL00], and a Roland XP-30 synthesizer. There are different instruments and playing styles included in the instruments from different sources. Table 5 summarizes this information, and along with this the number of independent sources for that particular instruments presented, as well as the pitch range. The pitch range differs slightly from source to source, and is here presented according to the MUMS samples.

The pitch range is presented as MIDI numbers. The conversion from MIDI number $d$ to fundamental frequency $f_0$ in Hz can be made according to the following equation:

$$f_0 = 440 \cdot 2^{\frac{d-69}{12}} . \tag{49}$$

In the following, we shortly summarize the available information on the samples from different sources.

*McGill University Master Samples* (MUMS). Most sounds on the MUMS library were recorded directly to a Sony PCM 3202 DASH recorder. High quality B & K condenser microphones were employed, along with matched B & K microphone pre-amplifiers. Most MUMS samples were recorded in a recording studio. The studio was acoustically neutral, and had a reverberation time of approximately 0.4 seconds. The strings and the piano were recorded in a concert hall with reverberation time varying from 2.5 to 5 seconds.

**Table 5: The evaluation database**

| Instrument | MIDI # | Playing styles | # sources | # notes |
|---|---|---|---|---|
| French Horn | 38-74 | normal, muted | 4 | 373 |
| C Trumpet | 54-87 |  | 3 | 153 |
| Bach Trumpet | 59-91 |  | 1 | 32 |
| Bass Trombone | 29-53 |  | 2 | 38 |
| Tenor Trombone | 40-75 | normal, muted | 3 | 204 |
| Alto Trombone | 65-77 |  | 1 | 13 |
| Tuba | 28-67 |  | 3 | 118 |
| Bass Saxophone | 32-39 |  | 1 | 8 |
| Baritone Saxophone | 36-48 |  | 2 | 39 |
| Tenor Saxophone | 48-61 |  | 3 | 54 |
| Alto Saxophone | 61-74 | vibrato, non vibrato | 4 | 254 |
| Soprano Saxophone | 73-87 | vibrato, non vibrato | 3 | 237 |
| English Horn | 52-81 |  | 2 | 90 |
| Oboe | 58-89 | normal, vibrato | 4 | 233 |
| Contra Bass Clarinet | 30-54 |  | 1 | 25 |
| Bass Clarinet | 37-61 |  | 2 | 38 |
| B-flat Clarinet | 50-86 |  | 4 | 146 |
| E-flat Clarinet | 55-86 |  | 2 | 292 |
| Contra Bassoon | 22-53 |  | 1 | 32 |
| Bassoon | 34-65 | normal, vibrato | 4 | 286 |
| Bass Flute | 48-73 | normal, flutter tongued | 1 | 42 |
| Alto Flute | 55-84 |  | 1 | 30 |
| Flute | 60-96 | vibrato, non vibrato, flutter tongued | 4 | 466 |
| Piccolo | 74-103 | normal, flutter tongued | 2 | 83 |
| Double bass | 24-64 | bowed, martele, muted, plucked, normal mute | 3 | 487 |
| Cello | 36-79 | bowed, martele, muted, plucked, normal mute | 3 | 429 |
| Viola | 48-86 | bowed, martele, muted, plucked, muted (normal, lead) | 3 | 446 |
| Violin | 55-96 | bowed, martele, muted, plucked, muted (normal, lead) | 3 | 441 |
| Acoustic guitar | 40-81 |  | 3 | 197 |
| Piano | 21-108 |  | 1 | 88 |

*University of Iowa Musical Instrument Samples* (UIowa) were recorded in the Anechoic Chamber in the Wendell Johnson Speech and Hearing Center at The University of Iowa on the following equipment: Neumann KM 84 Microphones, Mackie 1402-VLZ Mixer and Panasonic SV-3800 DAT Recorder. The samples were transferred through digital lines to an editing workstation. Three non-normalized dynamic levels are included: *piano pianissimo*, *mezzo forte*, and *forte fortissimo*.

*Roland XP-30 synthesizer* (XP30). These samples were played on the keyboard of a Roland XP 30 synthesizer, and transferred into a Silicon Graphics Octane workstation via analog lines. Samples from several sound banks are included. The dynamic keyboard was switched on, causing clear differences in the dynamics of these samples, although an attempt was made to keep the dynamics as constant as possible. Based on our observations, the quality of the synthesized samples varies, other are very similar to their acoustic counterparts, other of varying quality.

*IRCAM Studio-On-Line* (SOL) samples were recorded in a recording studio, and were originally stored at 24-bits 48 kHz format. When downloaded from the Internet site, a downsampling program was used to convert the samples into 16-bit / 44.1 kHz quality. There were different recording setups available, we downloaded the mono, close microphone channel. The samples from SOL include only the first 2 seconds of the played note, and the end is clipped, thus there is no natural decay.

*The guitar samples recorded at Tampere University of Technology* were recorded in a small room with soft walls, thus having little reverberation. They were recorded with a Sony TCD-D10 DAT recorder and AKG C460B microphones, and then transferred into a Silicon Graphics Octane workstation through digital lines. These samples are in 48 kHz and 16-bit format.

All the samples, except for the SOL samples which were already single tones, were first stored in longer files, each containing a chromatic scale of an instrument. These longer files were then segmented into single notes using a Matlab program which detected the notes using energy thresholds, and stored into separate wave-files. The format of the samples is the standard CD quality of 44.1 kHz / 16-bit, except for our own guitar recordings.

## 6.2 Recognition within a hierarchic framework

After developing several feature extractors, we wanted to evaluate their efficiency. At this point, only the MUMS samples were available for testing. The purpose of this experiment was to:

- Test the set of implemented features
- Propose the simultaneous use of cepstral coefficients and other features
- Analyse a hierarchical classification framework for musical instrument recognition, proposed by Martin in [Martin99].
- Compare the performance to earlier reported experiments employing the same data set [Martin98, Fujinaga98, Fraser99, Fujinaga00]

The results of this study were originally published in [Eronen99].

### Features

All the implemented features were used in this experiment. Eleven cepstral coefficients were calculated separately for the onset and steady state segments based on conventional linear prediction with an analysis order of 9. Thus, the length of the feature vector calculated for each isolated tone included a total of 44 features.

### Hierarchical classification

Musical instruments form a natural hierarchy, where instrument families form an intermediate level. In many applications, classification down to the level of instrument families is sufficient for practical needs. For example, searching for music with string instruments would make sense. In addition to that, a classifier may utilize a hierarchical structure algorithmically while assigning a sound into a lowest level class, individual instrument.

Using a hierarchical classification architecture for musical instrument recognition has been proposed by Martin in [Martin98]. In the following, we give a short review of his principles. At the top level of the taxonomy, instruments are divided into pizzicato and sustained. Second level comprises instrument families, and the bottom level are individual instruments. Classification occurs at each node, applying knowledge of the best features to distinguish between possible subclasses. This way of processing is suggested to have some advantages over direct classification at the lowest end of the taxonomy, because the decision process may be simplified to take into account only a smaller number of possible subclasses.

In our system, at each node a Gaussian or a k-NN classifier was used with a fixed set of features. The Gaussian classifier turned out to yield the best results at the highest level, where the number of classes is two. At the lower levels, the k-NN classifier was used. The features used at a node were selected manually by monitoring feature values of possible subclasses. This was done one feature at a time, and only the features showing clear discrimination ability were included into the feature set of the node.

We implemented a classification hierarchy similar to the one presented by in [Martin98] with the exception that his samples and taxonomy did not include the piano. In our system, the piano was assigned to an own family node because of having a unique set of some features, especially cepstral coefficients. According to Martin, classification performance was better if the reeds and the brass were first processed as one family and separated at the next stage. We wanted to test this with our own feature set and test data and tried the taxonomy with and without the Brass or Reeds node, which is marked with an asterix in Figure 9.

### Results

The validation database consisted of the MUMS samples only. The material included 1498 solo tones covering the entire pitch ranges of 30 orchestral instruments with several articulation styles (e.g. pizzicato, martele, bowed, muted, flutter). All tones were from the McGill Master Samples collection [Opolko87], except the piano and guitar tones which were played

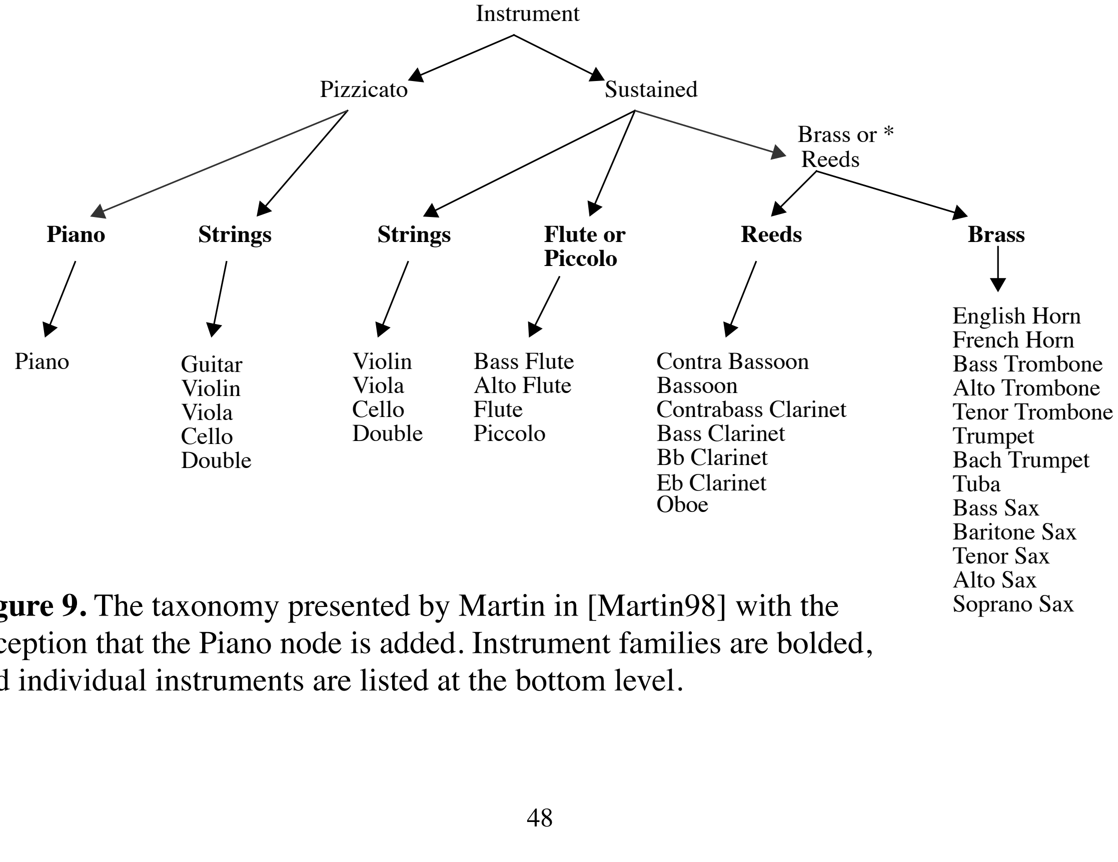

**Figure 9.** The taxonomy presented by Martin in [Martin98] with the exception that the Piano node is added. Instrument families are bolded, and individual instruments are listed at the bottom level.

**Table 6: Results using different classification architectures**

| | Hierarchy 1 | Hierarchy 2 | No hierarchy |
|---|---|---|---|
| Pizzicato / sustained | 99.0 % | 99.0 % | 99.0 % |
| Instrument families | 93.0 % | 94.0 % | 94.7 % |
| Individual instruments | 74.9 % | 75.8 % | 80.6 % |

by amateur musicians and recorded with a DAT recorder. In order to achieve comparable results to those described by Martin in [Martin98], similar way of cross validation with 70 % / 30 % splits of train and test data was used. A difference to the method of Martin was to estimate the fundamental frequency of the test sample before classification, which was then compared to the pitch ranges of different instruments, taking only the possible ones into classification.

In Table 6, the classification results made in the three different ways are presented. Hierarchy 1 is the taxonomy of Figure 9 without the Brass or Reeds node. In the No-hierarchy experiment, classification was made separately for each classification level. The Hierarchy 2 proved out to yield slightly better results, like Martin reported in [Martin98]. But interestingly, in our experiments direct classification with the k-NN in one pass performed best at both tasks, which was not the case in Martin's experiments where Hierarchy 2 yielded the best results. This is probably due to the fact that in this implementation, classification result at the lower level of hierarchy is totally dependent on the results of the higher levels, and the error cumulates as the classification proceeds. In his thesis, Martin obtained the best results with a hierarchical classifier when it was allowed to calculate all the possible paths through the hierarchy [Martin99]. This, however, contradicts the basic idea of the hierarchic that only possible subclasses are taken account as the classification proceeds.

The achieved performance both in instrument family and individual instrument classification was better than reported by Martin in [Martin98]. His system's classification accuracies were approximately 90 % in instrument family and 70 % with individual instruments, while the data set consisted 1023 samples of 15 different instruments, being a subset of our data. Also, when compared to the accuracy of 68 % with 1338 samples of 23 instruments from the MUMS collection reported by Fujinaga and Fraser [Fujinaga98, Fraser99, Fujinaga00], our system performs better.

### Discussion

Great care should be taken when interpreting these results. Only one example of each instruments is included in the MUMS collection, which is a severe limitation of this study and significantly lifts the recognition percentages. With a more realistic data set, the system's performance is significantly worse, as demonstrated in the next experiment. Any study, including this one, using material only from one source has only little value in terms of assessing the performance in realistic conditions. Only a careful conclusion can be made; the achieved performance and comparison to earlier results demonstrates that combining the different types of features succeeded in capturing some extra information about instrument properties.

## 6.3 Comparison of features

A crucial notion in making evaluations is that a system must be able to recognize several instances of an instrument played by different performers in different locations as belonging to the same class. This may be trivial for humans but not for recognition systems. The issues of study in this experiment were

- How does the system perform with a honest task definition and demanding evaluation material that includes several examples of each instrument and both acoustic and synthetic tones
- What are the accuracies obtained with different features and what is the best feature subset
- Is warped linear prediction (WLP) cepstrum a useful feature in musical instrument recognition
- How does the recognition performance of different LP-based features depend on the analysis order
- How is the performance affected by the use of more than one note for recognition
- How does the system perform in comparison to human subjects in a similar task

The results presented here have been accepted for publication in [Eronen01].

### Experimental setup

The whole database described in Section 6.1, except for the piano, was used for testing the system, and cross validation aimed at as realistic conditions as possible with this data set. On each trial, the training data consisted of all the samples except those of the particular performer and instrument being tested. In this way, the training data is maximally utilized, but the system has never heard the samples from that particular instrument in those circumstances before. There were 16 instruments that had at least three independent recordings, so these instruments were used for testing. Table 5 showed the instruments used in the test and train sets. The

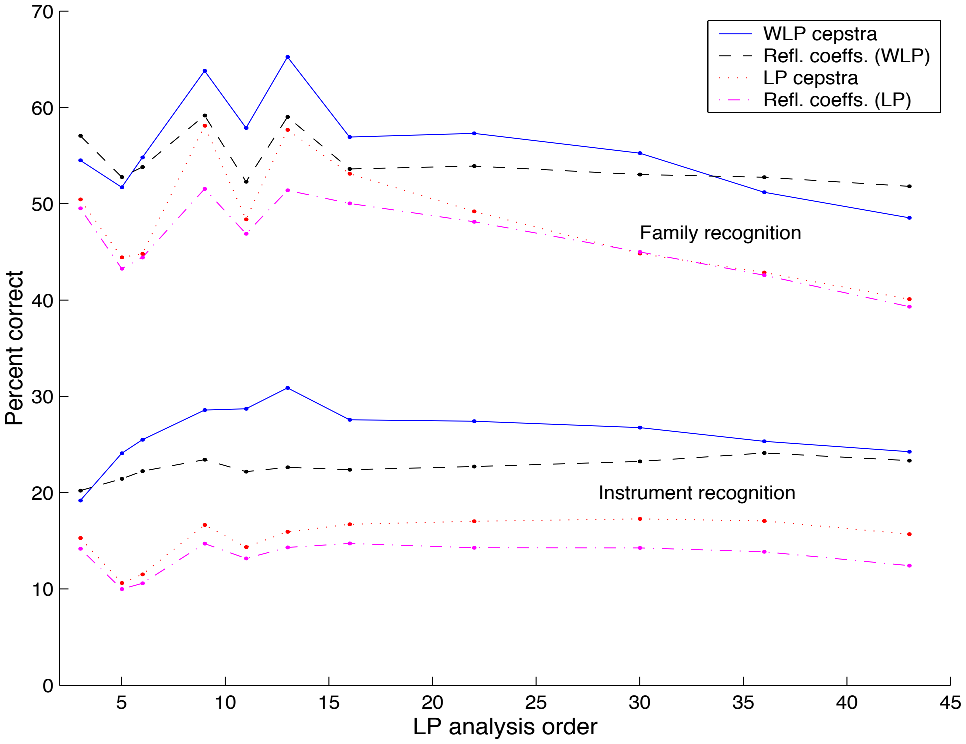

**Figure 10.** Classification performance as a function of analysis order for different LP-based features.

database includes a total of 5286 single tones of 29 instruments, out of which 3337 samples were used for testing. The classifier made its choice among the 29 instruments. In these tests, a random guesser would score 3.5 % in the individual instrument recognition task, and 17 % in family classification. In each test, classifications were performed separately for the instrument family and individual instrument cases. For the sake of simplicity, we did not use the hierarchic classification architecture in this experiment. The k-NN classifier was used, and the values of k were 11 for instrument family and for 5 individual instrument classification.

### Results

Different orders of the linear prediction filter were used to see the effect of that on the performance of LP and WLP-based features. The results for instrument family and individual instrument recognition are shown in Figure 10. The feature vector at all points consisted of two sets of coefficients: medians over the onset period and medians over the steady state. The optimal analysis order was between 9 and 14, above and below which performance degraded. The number of cepstral coefficients was one less than the LP analysis order. WLP cepstral and reflection coefficients outperformed the conventional LP cepstral and reflection coefficients at all analysis orders calculated. The best accuracy among all LP-based features was 33 % for individual instruments (66 % for instrument families), and was obtained with WLP cepstral coefficients (WLPCC) of order 13. There is a peculiar drop in performance at order of 11,

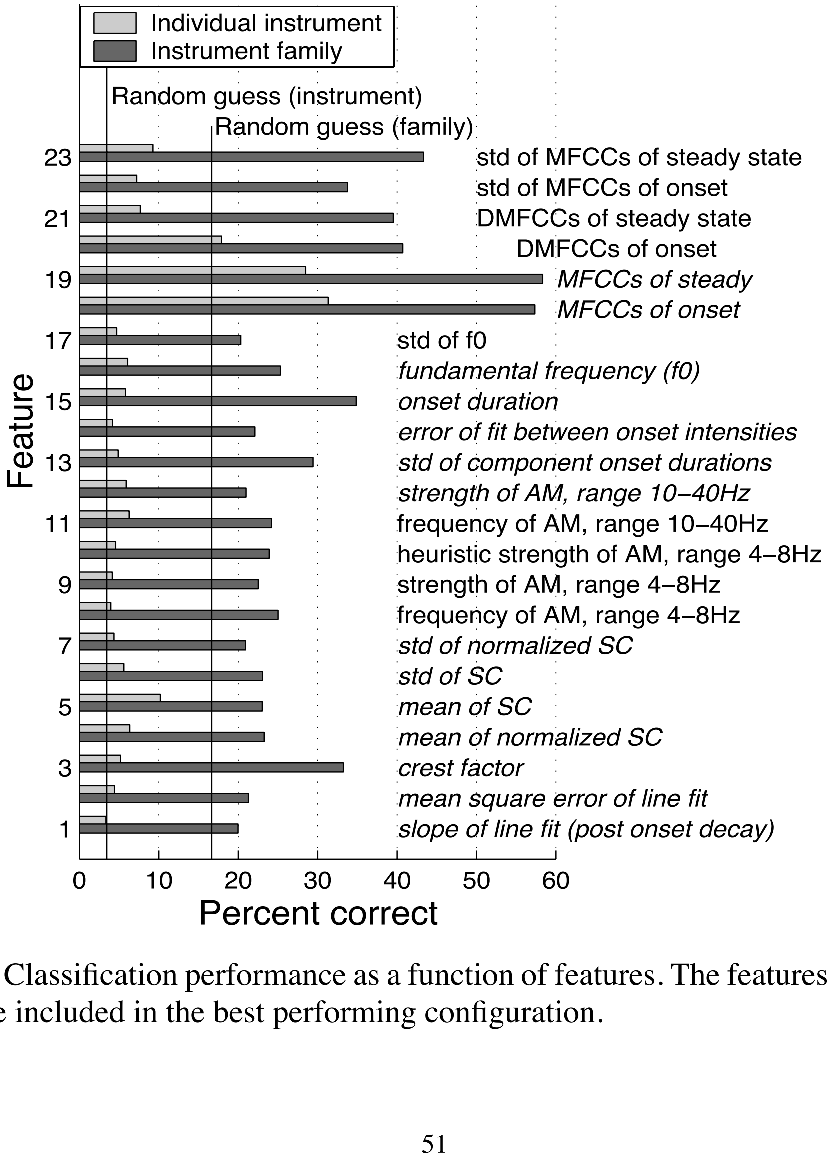

**Figure 11.** Classification performance as a function of features. The features printed in italics were included in the best performing configuration.

where the accuracy in recognizing the strings is worse than at the neighboring orders. We could not figure out the reason for this behavior.

In Figure 11, the classification accuracy of different features is presented. Some features performing below the random guess level are omitted. The cepstral parameters are the mel-frequency cepstral coefficients or their derivatives. The optimal number of MFCCs was 12, above and below which the performance slowly degraded. By using the MFCCs both from the onset and steady state, the accuracies were 32 % (69 %). Because of the computational cost considerations the MFCC were selected as the cepstrum features for the remaining experiments. Adding the mel-frequency delta cepstrum coefficients (DMFCC) slightly improved the performance, using the MFCCs and DMFCCs of the steady state resulted in 34 % (72 %) accuracy.

The other features did not prove out very successful alone. Onset duration was the most successful with 35 % accuracy in instrument family classification. In individual instrument classification, spectral centroid gave the best accuracy, 10 %. Both were clearly inferior to the MFCCs and DMFCCs. It should be noted, however, that the MFCC features are vectors of coefficients, and the other features consist of a single number each.

The best accuracy 35 % (77 %) was obtained by using a feature vector consisting of the features printed in italics in Figure 11. The feature set was found by using a subset of the data and the sequential backward generation algorithm. If the MFCCs were replaced with order 13 WLPCCs, the accuracy was 35 % (72 %).

In practical situations, a recognition system is likely to have more than one note to use for classification. A simulation was made to test the system's behavior in this situation. Random sequences of notes were generated and each note was classified individually. The final classification result was pooled across the sequence by using the majority rule. The recognition accuracies were averaged over 50 runs for each instrument and note sequence length. Figure 12 shows the average accuracies for individual instrument and family classification. With 11 random notes, the average accuracy increased to 51 % (96 %). In instrument family classification, the recognition accuracy for the tenor saxophone was the worst (55 % with 11 notes), whereas the accuracy for the all other instruments was over 90 %. In the case of individual

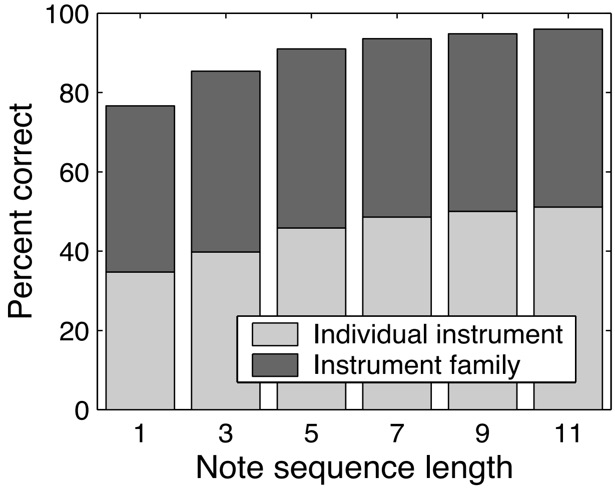

**Figure 12.** Classification performance as a function of note sequence length.

instruments, the accuracy for the tenor trombone, tuba, cello, violin, viola and guitar was poorer than with one note, the accuracy for the other instruments was higher.

The recognition accuracy depends on the recording circumstances, as may be expected. The individual instrument recognition accuracies were 32 %, 87 %, 21 % and 37 % for the samples from MUMS, UIowa, Roland and SOL sources, respectively. The UIowa samples included only the woodwinds and the French horn. For these instruments, the average recognition accuracy for the samples from all sources is 49 %. Thus, the recognition accuracy is clearly better for the UIowa samples recorded in an anechoic chamber. The samples from the other three sources are comparable with the exception that the samples from SOL did not include tenor or soprano saxophone. With synthesized samples the performance is clearly worse, which is probably due to the fact that no synthetic samples of that particular instrument were present in the training set when XP30 samples were tested.

The confusion matrix for the feature set giving the best accuracy is presented in Figure 13. There are large differences in the recognition accuracies of different instruments. The soprano saxophone is recognized correctly in 72 % of the cases, while the classification accuracies for the violin and guitar are only 4 %. French horn is the most common target for misclassifications. Quite interestingly, a similar phenomenon was reported by Martin in [Martin99].

### Comparison to human abilities

It is interesting to compare the behavior of the system to human subjects. As a reference, Martins test described in Section 2.2 is used. In his test, fourteen subjects recognized 137 samples from the McGill collection, a subset of the data used in our evaluations. The differences in the instrument sets are small, Martin's samples did not include any saxophone or guitar samples, but had the piccolo and the English horn, which were not included in our test data. In his test, the subjects recognized the individual instrument correctly in 45.9 % of cases (91.7 % for instrument families). Our system made more outside-family confusions than the subjects in Martin's test. It was not able to generalize into more abstract instrument families as well as humans. In individual instrument classification, the difference is smaller.

The within-family confusions made by the system are quite similar to the confusions made by humans. Examples include the French horn as tenor trombone and vice versa, tuba as French horn, or B-flat clarinet as E-flat clarinet. The confusions between the viola and the violin, and the cello and the double bass were also common to both humans and our system. In the confusions occurring outside the instrument family, confusions of the B-flat clarinet as soprano or alto sax were common to both our system and the subjects.

### Discussion

Warped linear prediction based features proved to be successful in the automatic recognition of musical instrument solo tones, and resulted in a better accuracy than that obtained with corresponding conventional LP based features. The mel-frequency cepstral coefficients gave the best accuracy in instrument family classification, and is the best selection also from the point of view of computational complexity. The best overall accuracy was obtained by augmenting the mel-cepstral coefficients with features describing the type of excitation, brightness, modulations, synchronity and fundamental frequency of tones.

Care should be taken while interpreting the presented results on the accuracy obtained with different features. First, the best set of features for musical instrument recognition depends on

| Presented \ Responded | French horn | Trumpet | Bach trumpet | Bass trombone | Tenor trombone | Alto trombone | Tuba | Bass sax | Baritone sax | Tenor sax | Alto sax | Soprano sax | English horn | Oboe | Contrabass clar. | Bass clarinet | E-flat clarinet | B-flat clarinet | Contrabassoon | Bassoon | Bass flute | Alto flute | Flute | Piccolo | Double bass | Cello | Violin | Viola | Guitar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| French horn | **50** | 3 |  | 2 | 12 |  | 18 |  |  |  |  | 1 |  |  |  |  |  |  |  | 8 |  |  | 1 |  | 5 | 1 |  |  |  |
| Trumpet | 8 | **23** | 7 |  | 24 |  | 2 |  |  | 11 |  | 2 |  | 2 |  |  |  | 3 |  |  |  |  | 5 | 1 | 3 | 1 | 4 | 4 | 1 |
| Tenor tromb. | 31 | 17 |  | 24 | **10** | 6 | 6 |  |  |  |  |  |  |  |  |  |  |  |  | 5 |  |  | 1 |  |  |  |  |  |  |
| Tuba | 76 |  |  | 8 | 4 |  | **7** |  |  |  |  |  |  |  |  |  |  |  |  | 3 |  |  |  |  | 2 |  |  |  |  |
| Tenor sax | 6 | 2 | 2 | 2 | 9 |  |  |  | 15 | **22** | 2 | 6 |  |  |  |  |  |  | 4 | 2 |  |  |  |  |  | 7 | 6 | 17 |  |
| Alto sax |  | 8 |  |  | 1 |  |  |  | 1 |  | **64** | 5 | 2 | 1 |  |  | 3 | 1 |  | 1 |  |  |  |  | 2 |  | 1 | 12 |  |
| Soprano sax | 4 | 3 |  |  | 4 |  |  |  |  |  | 2 | **72** |  | 2 |  |  |  | 5 |  |  |  |  | 10 |  |  |  |  |  |  |
| Oboe | 3 | 7 |  |  | 1 |  |  |  |  |  | 1 | 6 | 3 | **68** |  |  |  | 3 |  |  |  |  | 3 | 2 |  |  |  | 3 |  |
| B-flat clar. | 6 | 4 |  |  | 1 |  | 1 |  |  | 2 | 11 | 16 |  | 4 |  | 1 | 17 | **30** |  | 1 |  |  | 5 |  |  | 1 | 1 | 3 |  |
| Bassoon | 16 | 1 |  |  | 3 |  | 1 |  |  |  | 1 |  |  |  |  |  |  |  | 1 | **70** |  |  | 3 |  | 1 |  |  |  |  |
| Flute | 1 | 1 | 8 |  | 6 | 2 |  |  | 1 |  | 4 | 1 | 1 | 1 |  |  |  | 2 |  | 3 | 1 | 4 | **59** | 2 | 1 | 1 |  | 2 |  |
| Double bass | 2 | 1 |  |  |  |  |  |  |  |  | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  | **56** | 31 | 2 | 5 |  |
| Cello | 1 |  |  |  |  |  |  |  |  | 1 | 4 |  |  |  |  |  |  |  |  |  |  |  | 1 |  | 31 | **30** | 5 | 28 |  |
| Violin | 1 | 1 | 2 |  |  |  |  |  |  |  | 3 | 3 | 1 |  |  |  |  | 2 |  |  |  |  | 4 | 1 | 3 | 8 | **4** | 67 |  |
| Viola |  |  |  | 1 |  |  |  |  |  | 2 | 4 | 1 |  | 1 |  |  | 1 | 1 |  |  |  |  |  |  | 6 | 25 | 45 | **13** |  |
| Guitar | 2 | 8 |  |  | 1 |  | 1 | 1 |  |  | 2 | 1 |  |  |  |  |  |  |  |  |  |  |  |  | 43 | 38 | 1 | 1 | **4** |

**Figure 13.** Confusion matrix for the best performing features set. Entries are expressed as percentages and are rounded to the nearest integer. The boxes indicate instrument families.

the context [Handel95, Martin99]. Second, the extraction algorithms for features other than cepstral coefficients are still in their early stages of development. However, the accuracy clearly improves by adding these features.

Comparison to other systems is difficult because of the wide variety of data sets and cross-validation methods used. The most direct comparison can be made with Martin's system which used the same data as in the listening test above [Martin99]. His best accuracies with this data were 38.7 % and 75.9 %, which are very close to our results (35 % and 77 %). It should be noted, however, that his test set included 137 tones, whereas we tested 3337 tones. In addition, our feature extractors are far simpler than those in Martin's implementation. Nevertheless, the similarity in the performance of these two systems is indicative of the difficulty of the problem of recognizing single tones. Another interesting thing are the feature sets used by our system and that of Martin's: our best features were the MFCCs, whereas Martin did not use cepstral coefficients at all. He does not give any details whether his best features (or most often selected by the context dependent feature selection) were related to the spectral shape or if they were more temporally related. Therefore, it remains unclear whether the information describing the spectral shape has been the most relevant information in these two experiments. Nevertheless, there certainly exist different features giving the same recognition accuracy, and the approach of combining a wide set of features and using the most suitable in the current context seems to be the most promising.

## 6.4 Using Gaussian mixture models for instrument recognition

Our final simulations tested the use of Gaussian mixture models for the data set described in the previous experiment. This approach is different from the one we used for single tone recognition, since here we model the sounds with sequences of observation calculated in adjacent frames, instead of segmenting the music into notes and calculating one feature vector per note. Therefore, the GMM approach is directly applicable to solo music. The earlier described approach could also be applied to musical phrases, however it requires first an integration with an onset detection algorithm, such as the one proposed by Klapuri in [Klapuri99b].

In this experiment, only a subset of our features was used. These were the MFCCs, DMFCCs, and the spectral centroid. The most important criterion in selecting these features was that the feature must be calculated within a single time frame. On the other hand, the MFCCs and the SC were among the best performing features in the previous test. The frame length in this test was 40 ms, which is rather long, and is likely to destroy some information in rapid attack transients.

### Evaluation method

Test and training sets slightly differ from that used in the previous experiment. Using in each test run all the samples other than those of the current instrument instance would have caused an excessive computational load, and therefore four different training and testing set combinations was formed, according to the four main sources for the samples (MUMS, UIowa, XP30 and SOL). For instance, the training set for all test samples from the MUMS collection consisted of the samples from the UIowa, XP30 and SOL sources.

We tested the system with different note sequence length. A longer sequence was formed by catenating the feature vectors from adjacent notes in a chromatic scale using different note sequence lengths $L$. In each evaluation, the chromatic scale was proceeded in steps of $L$ notes, and each subsequence was classified. Using adjacent notes probably gives the classifier less information for recognition than if random notes were used, and thus decreases the recognition accuracy, but this way the computational load is reasonable and all notes are certainly tested. The final classification result is calculated as the percentage of correctly classified *L*-length sequences of the total number of sequences from that chromatic scale.

### Results and discussion

Tables 7 and 8 present the recognition results with varying features, note sequence lengths and model orders for individual instrument and instrument family recognition tasks, respectively. The best results at each note sequence length are bolded. For single notes (*L*=1), the best recognition accuracy in individual instrument recognition was 36 %, obtained with a GMM with 16 component densities and the MFCCs and DMFCCs as features. The best accuracy in instrument family recognition, 58 %, is obtained by using a GMM with 8 components, and adding the spectral centroid to the set of features. The accuracy of the k-NN method using the same test and training sets, and the best set of features in the previous experiment, is 39 % (68 %). Thus, the GMM performs slightly worse keeping in mind that the k-NN would neither achieve as good performance with longer sequences as in the previous experiment due to the worse performance in the single note case.

Brown reported that combining cepstral coefficients neither with delta cepstral coefficients nor the spectral centroid increased the accuracy in recognizing the oboe, saxophone, clarinet and flute [Brown01]. Our results suggest that in some cases using the delta cepstrum increases performance. Also, adding the spectral centroid would seem to slightly increase the accuracy in instrument family recognition, but the differences are indeed too small to make strong conclusions. But these results would again suggest the fact that the relevant features depend on the context, as is apparent based on the discussion of human perception in Chapter 2.

**Table 7: GMM recognition accuracies - individual instrument task.**

| Features used | Model Order | Test sequence length / notes | | | | | |
|---|---|---|---|---|---|---|---|
| | | L=1 | L=3 | L=5 | L=7 | L=11 | L=23 |
| MFCC | M=2 | 31.8 | 35.8 | 38.7 | 41.5 | 47.1 | 54.6 |
| | M=4 | 32.9 | 38.2 | 42.2 | 45.0 | **49.1** | 55.1 |
| | M=8 | 31.7 | 37.2 | 38.8 | 41.3 | 44.7 | 52.5 |
| | M=16 | 34.1 | 38.4 | 42.2 | 43.8 | 46.0 | 49.1 |
| | M=32 | 32.6 | 36.7 | 38.7 | 40.1 | 42.2 | 49.3 |
| MFCC + DMFCC | M=2 | 34.4 | 38.2 | 41.5 | 43.9 | 48.6 | 55.2 |
| | M=4 | 34.8 | 39.0 | 41.3 | 43.3 | 46.8 | 53.1 |
| | M=8 | 34.8 | 39.9 | 41.1 | 42.7 | 46.2 | 54.5 |
| | M=16 | **36.1** | **41.0** | **43.4** | **46.0** | 47.1 | 51.7 |
| | M=32 | 35.1 | 40.1 | 42.5 | 44.4 | 46.8 | 51.3 |
| MFCC + DMFCC + SC | M=2 | 33.6 | 38.7 | 41.5 | 43.0 | 48.4 | **57.3** |
| | M=4 | 34.9 | 40.0 | 41.8 | 43.6 | 48.3 | 55.1 |
| | M=8 | 35.6 | 40.9 | 42.8 | 44.9 | 47.2 | 52.6 |
| | M=16 | 35.9 | 40.9 | 42.2 | 43.7 | 45.7 | 50.1 |
| | M=32 | 35.5 | 40.1 | 42.4 | 43.2 | 45.9 | 52.3 |
| MFCC + SC | M=2 | 29.5 | 34.7 | 37.6 | 39.4 | 44.4 | 55.9 |
| | M=4 | 32.9 | 37.7 | 40.3 | 42.6 | 47.2 | 55.5 |
| | M=8 | 32.8 | 36.2 | 38.1 | 39.6 | 42.5 | 52.1 |
| | M=16 | 31.5 | 35.7 | 37.8 | 39.3 | 42.8 | 48.7 |
| | M=32 | 29.7 | 35.5 | 37.1 | 40.1 | 42.7 | 46.5 |
| SC | M=2 | 6.4 | 7.9 | 8.3 | 9.3 | 11.3 | 15.9 |
| | M=4 | 6.5 | 8.3 | 9.6 | 10.8 | 12.1 | 15.6 |
| | M=8 | 7.8 | 8.7 | 9.6 | 10.7 | 12.3 | 16.6 |
| | M=16 | 7.6 | 9.2 | 10.0 | 10.3 | 12.1 | 17.1 |
| | M=32 | 7.9 | 9.2 | 10.4 | 10.7 | 12.0 | 18.5 |

**Table 8: GMM recognition accuracies - instrument family task.**

| Features used | Model Order | Test sequence length / notes | | | | | |
|---|---|---|---|---|---|---|---|
| | | L=1 | L=3 | L=5 | L=7 | L=11 | L=23 |
| MFCC | M=2 | 51.4 | 55.7 | 58.2 | 61.0 | 66.8 | 73.6 |
| | M=4 | 51.2 | 56.3 | 60.7 | 63.2 | 66.7 | 72.8 |
| | M=8 | 50.8 | 56.9 | 58.3 | 60.3 | 64.2 | 71.5 |
| | M=16 | 52.9 | 56.6 | 59.4 | 60.4 | 62.7 | 65.9 |
| | M=32 | 52.1 | 56.1 | 57.4 | 58.8 | 60.9 | 67.1 |
| MFCC + DMFCC | M=2 | 54.2 | 58.4 | 61.2 | 63.5 | 69.2 | 74.6 |
| | M=4 | 55.7 | 59.5 | 61.9 | 63.9 | 66.7 | 70.2 |
| | M=8 | 55.1 | 60.4 | 62.1 | 63.8 | 67.3 | 73.0 |
| | M=16 | 57.5 | 62.5 | **65.1** | **67.2** | 68.6 | 72.0 |
| | M=32 | 55.5 | 59.7 | 61.9 | 63.7 | 65.9 | 69.0 |
| MFCC + DMFCC + SC | M=2 | 55.2 | 60.6 | 62.8 | 64.2 | 68.9 | 76.7 |
| | M=4 | 56.0 | 61.9 | 64.2 | 65.9 | **69.7** | 73.4 |
| | M=8 | **57.6** | **62.6** | 64.8 | 66.8 | 69.1 | 72.9 |
| | M=16 | 56.0 | 61.2 | 62.3 | 64.0 | 66.2 | 69.2 |
| | M=32 | 55.7 | 59.3 | 61.3 | 61.9 | 64.5 | 70.5 |
| MFCC + SC | M=2 | 51.0 | 56.1 | 59.1 | 60.6 | 66.2 | **77.1** |
| | M=4 | 52.5 | 57.7 | 60.5 | 62.6 | 66.8 | 72.4 |
| | M=8 | 52.5 | 56.1 | 58.3 | 59.7 | 62.4 | 70.6 |
| | M=16 | 51.3 | 55.6 | 57.9 | 59.6 | 63.3 | 68.2 |
| | M=32 | 49.3 | 55.8 | 57.8 | 60.8 | 62.4 | 64.9 |
| SC | M=2 | 22.6 | 27.0 | 29.2 | 31.4 | 35.3 | 40.1 |
| | M=4 | 22.8 | 26.2 | 29.2 | 33.0 | 36.8 | 38.0 |
| | M=8 | 23.5 | 25.8 | 28.4 | 31.6 | 36.4 | 38.3 |
| | M=16 | 23.6 | 25.8 | 28.2 | 31.0 | 35.8 | 39.5 |
| | M=32 | 24.3 | 26.7 | 28.9 | 31.1 | 35.9 | 40.8 |

## 6.5 Future work

The main challenge for the construction of musical instrument recognition systems is increasing their robustness. Many factors influence the features calculated from real sounds. These include the different playing styles and dynamics that vary the sound spectrum. Very few features are constant across the pitch range of an instrument. Instruments radiate sound unevenly at different directions. In addition, the recording environment affects, samples recorded in an anechoic chamber are well recognized, whereas more realistic environments, or synthetic samples pose much extra difficulty for the task. The problem of generalizing is by no means a trivial one: the system must recognize different pieces of violin as belonging to the same class and different members of the string family as a part of the string class.

We are currently collecting a database of solo music, and will continue with some simulations with the GMM approach. It is likely that using the MFCCs and DMFCCs is not enough for this task, and therefore means to effectively combine the various other features with cepstral features should be examined. The approach of combining classifiers is one interesting alternative [Kittler98]. For instance, it would be worth experimenting to combine the GMM or the Hidden Markov Model using cepstral features calculated in frames, and the k-NN using features calculated for each note, via the voting scheme.

The potential applications will of course partly determine the direction into which a system should be developed. A preliminary attempt has been made towards *streaming* together the sounds coming from a single source in the presence of several sources. A musical piece was first transcribed, and a separating algorithm then tried to match the harmonic partials with their sources. A set of separated notes, along with their onset times was then given to our streaming algorithm. The best set of features reported in the second experiment were calculated from the tones, and the feature vectors were then k-means clustered [Klapuri01]. Using no time information at all, the clustering was rather successful with this song which included notes from a flute, bass and chords. However, with more complex music, having more different instruments which may be also more corrupted in the separation process, this straightforward approach will most likely fail. Therefore, Viterbi-type algorithms finding optimal paths through a sequence of observations should be deployed.

In polyphonic music the interfering sounds make the recognition task extremely difficult. In addition to having features that are robust against environment and instrument instance variations, we will have to cope with different kinds of disturbances caused by other sounds in the mixture. As even humans cannot recognize solo instruments based on isolated tones better than with 46 % accuracy, we are sceptical about whether reliable polyphonic recognition of several instruments from note mixtures will be possible based on low level information only. Using longer sequences improves the performance with human subjects and with computer systems, as does limiting the recognition into instrument families. Therefore, recognition of instrument families from longer pieces of polyphonic music would seem a task that could be approached. However, we again face the problem of generalizing: it is difficult to find features and models for a family that would enable generalizing between different members within a single family, or between different instances of a single instrument class. Moreover, if we choose the separation approach, unsuccessful separation will destroy important information. Therefore, recognition of mixtures of notes without separating the notes of single instruments, and integrating top-down knowledge, for instance, in the form of limiting the search space of a musical instrument recognizer, should also be considered.

# 7 Conclusions

We have described a system that can listen to a musical instrument and recognize it. The work started by reviewing human perception: how well humans can recognize different instruments and what are the underlying phenomena taking place in the auditory system. Then we studied the qualities of musical sounds making them distinguishable from each other, as well as acoustics of musical instruments. The knowledge of the perceptually salient acoustic cues possibly used by human subjects in recognition was the basis for the development of feature extraction algorithms. Some alternative approaches were implemented as back-end classifiers: the hierarchic classification architecture, straight classification at the bottom level with a distance based classifier, and the Gaussian mixture model approach.

In the first evaluation, a combined use of cepstral coefficients and various other features was demonstrated. Using the hierarchic classifier architecture could not bring improvement in the recognition accuracy. However, it was concluded that the recognition rates in this experiment were highly optimistic because of insufficient testing material. The next experiment addressed this problem by introducing a wide data set including several examples of a particular instrument. The efficiency of various features was tested, including a feature not used for musical instrument recognition before, the warped linear prediction cepstrum. The best accuracy was comparable to the state-of-the-art systems, and was obtained by combining the mel-frequency cepstral coefficients with features describing the type of excitation, brightness, modulations, synchronity and fundamental frequency of tones. The within-instrument-family confusions made by the system were similar to those made by human subjects, although the system made more both inside and outside-family confusions. In the final experiment, techniques commonly used in speaker recognition were applied for musical instrument recognition. The benefit of the approach was that it is directly applicable to solo phrases.

Using warped linear prediction was more successful than conventional linear-prediction based features. The best selection as cepstral features were the mel-frequency cepstral coefficients. Most of the performance would have been achieved by applying common speaker recognition tools for the problem, however, it was shown that the accuracy of this kind of system using cepstral features can be improved by adding other perceptually relevant features taken from instrument acoustics and psychoacoustics. Nevertheless, their successful implementation requires a substantial amount of work and experimentation.

In order to make truly realistic evaluations, more acoustic data would be needed, including monophonic material. The environment and differences between instrument instances proved out to have a more significant effect on the difficulty of the problem than what was expected at the beginning. In general, the task of reliably recognizing a wide set of instruments from realistic monophonic recordings is not a trivial one; it is difficult for humans and especially for computers. It becomes easier as longer segments of music are used and the recognition is performed at the level of instrument families.

# 8 References

[Alonso00] Alonso-Martinez, Faundez-Zanuy. (2000). “Speaker identification in mismatch training and test conditions“ In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2000.

[ANSI73] American National Standards Institute. (1973). “American national psychoacoustical terminology“. American Standards Association, New York.

[Berger64] Berger. (1964). “Some factors in the recognition of timbre“. *J. Audio. Eng. Soc.* 30, pp. 396-406.

[Bregman90] Bregman. (1990). “Auditory Scene Analysis”. MIT Press.

[Brown92] Brown, Puckette. (1992). “An Efficient Algorithm for the Calculation of a Constant Q Transform”. *J. Acoust. Soc. Am.* 92, pp. 2698-2701.

[Brown99] Brown. (1999). “Computer identification of musical instruments using pattern recognition with cepstral coefficients as features”. *J. Acoust. Soc. Am.* 105(3), March 1999.

[Brown01] Brown. (2001). “Feature dependence in the automatic identification of musical woodwind instruments”. *J. Acoust. Soc. Am.* 109(3), March 2001.

[Campbell78] Campbell, Heller. (1978). “The contribution of the legato transient to instrument identification“. In Proc. of the Research Symposium on the Psychology and Acoustics of Music. University of Kansas, Lawrence, pp. 30-44.

[Cappe01] Cappe. (2001). “H2M : A set of MATLAB/OCTAVE functions for the EM estimation of mixtures and hidden Markov models”. http://www-sig.enst.fr/~cappe/h2m/index.html.

[Clark64] Clark, Milner. (1964). “Dependence of timbre on the tonal loudness produced by musical instruments“. *J. Audio. Eng. Soc.* 12, pp. 28-31.

[Cosi96] Cosi, De Poli, Lauzzana. (1994). “Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification“. *Journal of New Music Research*, Vol. 23, pp. 71-98, 1994.

[Davis80] Davis, Mermelstein. (1980). “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences“. *IEEE Trans. on Acoustics, Speech and Signal Proc.* Vol. 28, No. 4, oo. 357-366.

[DePoli93] DePoli, Prandoni, Tonella. (1993). “Timbre clustering by self-organizing neural networks“ Proc. of X Colloquium on Musical Informatics. University of Milan.

[DePoli97] De Poli, Prandoni. (1997). “Sonological Models for Timbre Characterization“. *Journal of New Music Research*, Vol. 26, pp. 170-197, 1997.

[Dubnov98] Dubnov, Rodet. (1998). “Timbre Recognition with Combined Stationary and Temporal Features”. Proceedings of International Computer Music Conference, 1998.

[Dufaux00] Dufaux, Besacier, Ansorge, Pellandini. (2000). “Automatic sound detection and recognition for noisy environment“. In Proc. of the X European Signal Processing Conference, EUSIPCO 2000, Tampere, Finland.

[Dufournet98] Dufournet, Jouenne, Rozwadowski. (1998). “Automatic Noise Source Recognition“. In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1998.

[Eagleson47] Eagleson, H. W., Eagleson, O. W. (1947) “Identification of musical instruments when heard directly and over a public-address system“. *J. Acoust. Soc. Am.* 19, pp. 338-342.

[Ellis96] Ellis. (1996). “Prediction-driven computational auditory scene analysis”. Ph.D. thesis, MIT.

[Ellis01] Ellis. (2001). Lecture notes on course EE E6820: Speech & Audio Processing & Recognition, Department of Electrical Engineering, Columbia University.

[Eronen00] Eronen, Klapuri. (2000). “Musical instrument recognition using cepstral coefficients and temporal features“ In Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2000.

[Eronen01] Eronen. (2001). “Comparison of features for musical instrument recognition“. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001.

[Feiten91] Feiten, Frank, Ungvary. (1991). “Organization of sounds with neural nets“. In Proc. International Computer Music Conference, 1991.

[Feiten94] Feiten, Guntzel. (1994). “Automatic indexing of a sound database using self-organizing neural nets“. *Computer Music Journal,* Vol. 18, No. 3, pp. 53-65.

[Fletcher98] Fletcher, Rossing. (1998). “The Physics of Musical Instruments“. Springer-Verlag New York, Inc.

[Fraser99] Fraser, Fujinaga. (1999). “Towards real-time recognition of acoustic musical instruments”. Proceedings of the International Computer Music Conference, 1999.

[Fujinaga98] Fujinaga. (1998). “Machine recognition of timbre using steady-state tone of acoustic musical instruments”. Proceedings of the International Computer Music Conference, 1998.

[Fujinaga00] Fujinaga. (2000). “Realtime recognition of orchestral instruments”. Proceedings of the International Computer Music Conference, 2000.

[Gaunard98] Gaunard, Mubikangiey, Couvreur, Fontaine. (1998). “Automatic Classification of Environmental noise events by Hidden Markov Models”. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3609-3612, 1998.

[Godsmark99] Godsmark, Brown. (1999). “A blackboard architecture for computational auditory scene analysis“. *Speech Communication,* Vol. 27, pp. 351-366.

[Goldhor93] Goldhor. (1993). “Recognition of Environmental Sounds”. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993.

[Grey77] Grey. (1977). “Multidimensional perceptual scaling of musical timbres”. *J. Acoust. Soc. Am.*, Vol. 61, No. 5, May 1977.

[Grey78] Grey, Gordon. (1978). “Perceptual effects of spectral modifications of musical timbres“. *J. Acoust. Soc. Am.*, Vol. 63, 1978.

[Handel95] Handel (1995) "Timbre Perception and Auditory Object Identification". In eds. Moore, "Hearing".

[Herrera99] Herrera, Serra. (1999). “A proposal for the description of audio in the context of MPEG-7”. Proceedings of the CBMI’99 European Workshop on Content-Based Multimedia Indexing, 1999.

[Härmä00a] Härmä, Karjalainen, Savioja, Välimäki, Laine, Huopaniemi. (2000). “Frequency-Warped Signal Processing for Audio Applications“. *J. Audio. Eng. Soc.* Vol. 48, No. 11, pp. 1011-1031.

[Härmä00b] Härmä, Karjalainen. (2000). “WarpTB - Matlab Toolbox for Warped DSP (pre-release)”. Available at http://www.acoustics.hut.fi/software/warp/.

[Ifeachor93] Ifeachor, Jervis. (1993). “Digital Signal Processing-A practical approach”. Addison-Wesley Publishing Co.

[Iverson93] Iverson, Krumhansl. (1993). “Isolating the dynamic attributes of musical timbre“. *J. Acoust. Soc. Am.*, Vol. 94, pp. 2595-2603.

[Jarnicki98] Jarnicki, Mazurkiewicz, Maciejewski. (1998). “Mobile Object Recognition Based on Acoustic Information”. In Proceedings of the 24th Ann. Conf. of the IEEE Idustrial Electronics Society, IECON’98, Vol. 3, pp. 1564-1569, 1998.

[Jensen99] Jensen. (1999). “Timbre Models of Musical Sounds“. Ph.D. Thesis, Department of Computer Science, University of Copenhagen, 1999.

[Kaminskyj95] Kaminskyj, Materka. (1995). “Automatic Source Identification of Monophonic Musical Instrument Sounds”. Proceedings of the IEEE Int. Conf. on Neural Networks, 1995.

[Kaminskyj00] Kaminskyj. (2000). “Multi-feature Musical Instrument Sound Classifier“. In Proc. Australasian Computer Music Conference, Queensland University of Technology, July 2000.

[Karjalainen99] Karjalainen. (1999). “Kommunikaatioakustiikka“. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Report 51, 1999. In Finnish.

[Kashino95] Kashino, Nakadai, Kinoshita, Tanaka. (1995). “Application of Bayesian probability network to music scene analysis”. Proceedings of the International Joint Conference on AI, CASA workshop, 1995.

[Kashino98] Kashino, Murase. (1998). “Music Recognition Using Note Transition Context”. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP’98, Vol. 6, pp. 3593-3596, 1998.

[Kashino99] Kashino, Murase. (1999). “A sound source identification system for ensemble music based on template adaptation and music stream extraction”. *Speech Communication,* Vol. 27, pp. 337-349.

[Kinoshita99] Kinoshita, Sakai, Tanaka. (1999). “Musical Sound Source Identification Based on Frequency Component Adaptation“. In Proc. of the IJCAI-99 Workshop on Computational Auditoru Scene Analysis (CASA’99), August 1999, Stockholm.

[Kendall86] Kendall. (1986). “The role of acoustic signal partitions in listener categorization of musical phrases“. *Music Perception* 4, pp. 185-214.

[Kittler98] Kittler, Hatef, Duin, Matas. (1998). “On Combining Classifiers“. *IEEE Transactions on Pattern Analysis and Intelligence*, Vol. 20, No. 3, March 1998.

[Klapuri98] Klapuri. (1998). “Automatic transcription of music”. Master of Science Thesis, Tampere University of Technology, Department of Information Technology, Laboratory of Signal Processing.

[Klapuri99a] Klapuri. (1999). “Pitch estimation using multiple independent time-frequency windows“. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New Paltz, New York, 1999.

[Klapuri99b] Klapuri. (1999). “Sound onset detection by applying psychoacoustic knowledge“. In Proc. ICASSP 1999.

[Klapuri00] Klapuri, Virtanen, Holm. (2000). "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals". In Proc. COST-G6 Conference on Digital Audio Effects, DAFx-00, Verona, Italy, 2000.

[Klapuri01a] Klapuri. " Multipitch estimation and sound separation by the spectral smoothness principle". In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001.

[Klapuri01b] Klapuri, Virtanen, Eronen, Seppänen. (2001). “Automatic Transcription of Musical Recordings“. In Proc. of the Consistent & Reliable Acoustic Cues for sound analysis Workshop, CRAC’01, Aalborg, Denmark, September 2001.

[Klassner96] Klassner. (1996). “Data Reprocessing in Signal Understanding Systems“. Ph.D. thesis, Department of Computer Science, University of Massachusetts Amherst, September 1996.

[Kostek 1999] Kostek. (1999). “Soft Computing in Acoustics: Applications of Neural Networks, Fuzzy Logic and Rough Sets to Musical Acoustics”. Physica-Verlag, 1999.

[Kostek00] “Automatic Classification of Musical Sounds” In Proc. 108th Audio Eng. Soc. Convention.

[Kostek01] Kostek, Czyzewski. (2001). “Automatic Recognition of Musical Instrument Sounds - Further Developments”. In Proc. 110th Audio Eng. Soc. Convention, Amsterdam, Netherlands, May 2001.

[Krimphoff94] Krimphoff, J., McAdams, S. & Winsberg, S. (1994). "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysiques." *Journal de Physique* 4(C5): 625-628.

[Lakatos00] Lakatos, Beauchamp. (2000). “Extended perceptual spaces for pitched and percussive timbres“. *J. Acoust. Soc. Am.*, Vol. 107, No. 5, pp. 2882.

[Lesser95] Lesser, Nawab, Klassner. (1995). “IPUS: An Architecture for the Integrated Processing and Understanding of Signals”, *AI Journal* 77(1), 1995.

[Liu98] Liu, Motoda. (1998), “Feature selection for knowledge discovery and data mining“. Kluwer Academic Publishers.

[Mammone96] Mammone, Zhang, Ramachandran. (1996). “Robust Speaker Recognition”. *IEEE Signal Processing Magazine* 13(5), 58-71, Sep. 1996.

[Marques99] Marques, Moreno. (1999). “A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines“. Compaq Corporation, Cambridge Research laboratory, Technical Report Series CRL 99/4, June 1999.

[Martin98] Martin. (1998). “Musical instrument identification: A pattern-recognition approach“. Presented at the 136th meeting of the Acoustical Society of America, October 13, 1998.

[Martin99] Martin. (1999). “Sound-Source Recognition: A Theory and Computational Model”. Ph.D. thesis, MIT.

[McAdams93] McAdams. (1993). “Recognition of Auditory Sound Sources and Events. Thinking in Sound: The Cognitive Psychology of Human Audition”. Oxford University Press, 1993.

[McAdams95] McAdams, Winsberg, Donnadieu, De Soete, Krimphoff. (1995). “Perceptual scaling of synthesized musical timbres: common dimensions, specifities and latent subject classes“. *Psychological Research*, Vol. 58, pp. 177-192.

[McAdams99] McAdams, Beauchamp, Meneguzzi. (1999). “Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters“. *J. Acoust. Soc. Am.*, Vol. 105, pp. 882-897.

[Moon96] Moon. (1996). “The expectation-maximation algorithm“. *IEEE Signal Processing Magazine,* pp. 47-70, Nov. 1996.

[Moore95] Moore (ed.). (1995). “Hearing. Handbook of Perception and Cognition (2nd edition)”. Academic Press Inc.

[Murthy99] Murthy, Beaufays, Heck, Weintraub. (1999). “Robust Text-Independent Speaker Recognition over Telephone Channels“. *IEEE Trans. on Acoustics, Speech and Signal Proc.,* Vol. 7, No. 5, pp. 554-568.

[Opolko87] Opolko, F. & Wapnick, J. “McGill University Master Samples” (compact disk). McGill University, 1987.

[Parsons87] Parsons. (1987). “Voice and Speech Processing“. McGraw-Hill series in electrical engineering.

[Peeters00] Peeters, McAdams, Herrera. (2000). “Instrument Sound Description in the Context of MPEG-7“. In Proc. of the International Computer Music Conference 2000, Berlin.

[Peltonen01a] Peltonen, Eronen, Parviainen, Klapuri. (2001). “Recognition of Everyday Auditory Scenes: Potentials, Latencies and Cues“. In Proc. 110th Audio Eng. Soc. Convention, Amsterdam, Netherlands, May 2001.

[Peltonen01b] Peltonen. (2000). “Computational Auditory Scene Recognition”. MSc thesis, Tampere University of Technology, Department of Information Technology, August 2001.

[Plomp76] Plomp. (1976). “Aspects of tone sensation“. London, Academic Press.

[Poli97] Poli, Prandoni, "Sonological Models for Timbre Characterization" *Journal of New Music Research,* Vol. 26, pp. 170-197.

[Rabiner93] Rabiner, Juang. (1993). “Fundamentals of speech recognition”. Prentice-Hall 1993.

[Reynolds95] Reynolds, Rose. (1995). “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”. *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, January 1995.

[Rossing90] Rossing. (1990). “The Science of Sound“. Second edition, Addison-Wesley Publishing Co.

[Saldanha64] Saldanha, Corso. (1964). “Timbre cues and the identification of musical instruments“. *J. Acoust. Soc. Am.*, Vol. 36, pp. 2021-2026.

[Scheirer97] Scheirer, Slaney. (1997). “Construction and evaluation of a robust multifeature speech/music discriminator”. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’97, pp. 1331 - 1334.

[Scheirer00] Scheirer. (2000). “Music-Listening Systems“. Ph.D. dissertation, MIT, April 2000.

[Schmid77] Schmid. (1977). “Acoustic Pattern Recognition of Musical Instruments“. Ph.D. thesis, University of Washington.

[Serra97] Serra. (1997). “Musical Sound Modeling With Sinusoids Plus Noise”. Roads, Pope, Poli (eds.). “Musical Signal Processing”. Swets & Zeitlinger Publishers.

[SOL00] Studio-On-Line. (2000). http://www.ircam.fr/studio-online, http://soleil.ircam.fr.

[Soong88] Soong, Rosenberg. (1988). “On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition“. *IEEE Trans. Acoustics, Speech and Signal Proc,* Vol. 36, No. 6, pp. 871-879.

[Strong67] Strong, Clark. (1967). “Perturbations of synthetic orchestral wind instrument tones“. *J. Acoust. Soc. Am.*, Vol. 41, pp. 277-285.

[Strube80] Strube. (1980). “Linear Prediction on a Warped Frequency Scale“. *J. Acoust. Soc. Am.*, Vol. 68, pp. 1071-1076.

[Talkin95] Talkin. (1995). “A Robust Algorithm for Pitch Tracking“. In *Speech Coding and Synthesis*, Kleijn and Paliwal (eds.), Elsevier Science, 1995.

[Toiviainen95] Toiviainen, Kaipainen, Louhivuori. (1995). “Musical timbre: similarity ratings correlate with computational feature space distances“ *Journal of New Music Research*, Vol. 24, No. 3, pp. 282-298.

[Toiviainen96] Toiviainen. (1996). "Optimizing Auditory Images and Distance Metrics for Self-Organizing Timbre Maps" *Journal of New Music Research,* Vol. 25, pp. 1-30.

[Tolonen98] Tolonen. (1998). “Model-Based Analysis and Resynthesis of Acoustic Guitar Tones”. Master’s thesis. Report no. 46, Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland, Jan. 1998.

[Tou74] Tou, Gonzalez. (1974). “Pattern Recognition Principles“. Addison-Wesley Publishing Company, Massachusetts, 1974.

[UIowa00] University of Iowa. (2000). University of Iowa Musical Instrument Samples page. http://theremin.music.uiowa.edu.

[Virtanen01] Virtanen, Klapuri. (2001) "Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation". In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001.

[Välimäki96] Välimäki, Takala. (1996). “Virtual musical instruments - natural sound using physical models“. *Organized Sound,* Vol. 1, No. 2, pp. 75-86.

[Wedin72] Wedin, Goude. (1972). “Dimension analysis of the perception of instrumental timbre“. *Scandinavian Journal of Psychology*, 13, pp. 228-240.

[Wessel79] Wessel. (1979). “Timbre space as a musical control structure”. *Computer Music Journal*, Vol. 3, No. 2, 1979.

[Wu98] Wu, Siegel, Khosla. (1998) “Vehicle Sound Signature Recognition by Frequency Vector Principal Component Analysis”. Proceedings of the IEEE Instrumentation and Measurement Technology Conference, 1998.

[Young00] Young, Kershaw, Odell, Ollason, Valtchev, Woodland. (2000). "The HTK Book (for HTK Version 3.0)". Cambridge University Engineering Department, July 2000.

[Zhang00] Zhang, Kuo. (2000). “Content-based audio classification and retrieval for audiovisual data parsing“. Kluwer Academic Publishers, 2001.

# Appendix A: Acoustics of musical instruments

This appendix discusses the acoustics of musical instrument in some detail, emphasizing the effects of the different sound production mechanisms on the resulting sound.

### The bowed string instruments

When a string instrument is bowed, the interaction between the bow and the strings is a periodic, but a complex phenomenon. During the greater part of vibration, the friction causes the bow to stick into the string, and the string is carried along by the bow. Then the string detaches itself, and moves rapidly back with almost no friction until it is again caught by the moving bow [Rossing90, Karjalainen99]. This movement continues periodically, and is referred as the Helmholtz motion according to its discoverer Herman von Helmholtz.

An ideal, completely flexible string vibrating between two fixed end supports would excert a sideways force with a sawtooth waveform and the harmonics in the spectrum varying in amplitude as $1/n$ on the bridge, where $n$ is the number of the harmonic [Rossing90]. However, the frequency content of the spectrum depends on the pressure of the bow against the string and the position of the bow with respect to the bridge. Bowing close to the bridge with high bowing pressure gives a loud, bright tone, whereas bowing further from the bridge produces a more gentle and darker tone [Rossing90]. Some modes can even be almost completely damped due to the bowing position. In addition, in a real string the vibration is not exactly triangular, and variation in the period of the waveform causes frequency jitter [Fletcher98]. During the attack, the spectrum is not very harmonic [Martin99], and there often is a high frequency scratch before the bowing stabilizes [Handel95].

The body of string instruments has many different modes of vibration, which consist of the coupled motions of the top plate, the back plate and the enclosed air [Rossing90]. The modes cause many narrow resonances with high values of Q. Usually, the low modes are carefully tuned, often according to the open frequencies of some of the strings [Fletcher98], whereas the tuning of higher resonances varies in different instrument pieces. The violin has significant formants at 275, 460 and 700 Hz, and a broad concentration of resonances around 3 kHz, which corresponds to the singer's formant with opera singers [Rossing90]. Detailed description of the tuning of the strings and the resonance modes of the other members of the string family is beyond the scope of this thesis, but an interested reader is referred to [Fletcher98, Rossing90 and Martin99] for a more detailed discussion. The body of string instruments also determines the directional properties of the produced sound since the plates and the air cavity do not radiate in a uniform way to all directions [Fletcher98, Rossing90].

The bridge has a significant effect on the sound of the string instruments, since nearly all of the vibration force of the strings must go through the bridge coupling. Its own resonances also color the sound, for instance, the broad bridge resonances for violin are around 3 and 6 kHz [Rossing90]. These resonances can be lowered by attaching a mute on the bridge, creating a darker tone.

Bowed string instruments are often played with vibrato, which is produced by a periodic rolling motion of a finger on the string, which causes frequency modulation at the range of 5-8 Hz. Because of body resonances, a frequency deviation also generates amplitude modulation in the partials, which may be at the vibrato frequency of twice that. The changes in the sound level due to the amplitude modulation can be of 3 to 15 dB [Fletcher98].

### The guitar

In the following, we briefly discuss the basics of sound production in a plucked guitar, however, these principles also apply to the other strings when plucked.

The spectrum of a plucked string depends on the plucking position and style. If a completely flexible string with no mass attached to rigid supports is plucked at its center, the vibration consists of the fundamental plus the odd numbered harmonics [Rossing90]. If it is plucked from one fifth distance from one end, the fifth harmonic is missing from the spectrum. The player can alter the tone also by changing the plucking style; a loud tone with a sharp attack is obtained with a fast pluck, i.e. with large finger velocity [Tolonen98]. Releasing the string to vibrate with a relatively small initial displacement causes a soft tone.

However, the spectrum of a plucked real string is not completely harmonic. In a stiff string with a mass, waves of different velocities travel at different speeds along the string, which is called dispersion [Fletcher98]. This causes the resulting tone to be slightly inharmonic. Also nonlinearities cause shifts in mode frequencies, see the discussion in [Tolonen98].

The lowest body resonance of an acoustic guitar is typically between 90 and 100 Hz, which corresponds to the first mode of the air cavity or the Helmholtz mode of the instrument body. The second one corresponds to the first mode of the top plate, and is located between 170 and 250 Hz. [Tolonen98] A more detailed discussion on guitar body resonances can be found in [Rossing90].

### The brass

In the brass instruments, the sound is produced by blowing at the other end of the tube. In this process, the player’s tensed lips allow puffs of air into the tube. The pulses travel to the other end of the tube, and partly reflect backwards at the bell because of the impedance mismatch caused by the bell, causing standing waves building up in the tube. This is a similar mechanism to the human sound production system, where the glottis pulses excite the vocal tract. The difference is that the vibration at glottis is relatively independent of the properties of the vocal tract. In wind instruments, there generally exists a tight coupling between the excitation and resonance structures, and they are synchronized. In brass instruments, a positive feedback in the form of air pulses returning from the bell force the player’s lips to vibrate at the natural frequency of the tube [Karjalainen99]. The pitch of the instrument can be varied by changing the lip tension, which changes the vibration mode that is excited. Another way is to change the length of the tube by pressing valves or moving a slide.

The onsets of brass instruments have some unique characteristics. Before the stable oscillation begins, the instrument is not stable. It can take several round trips for the standing waves to build up [Fletcher98]. The duration of attack is typically 50 ± 20 ms and does not change significantly with the pitch of the note [Fletcher98]. Since the bell reflects low frequency energy more effectively than high, the low energy modes build up more quickly than the high frequency modes. This causes the onset partials to be skewed; the low energy partials build up fast and in close synchrony, whereas it takes a longer time for the higher partials [Rossing90]. Another characteristic of the brass instrument onset is the possible wandering of pitch. The pitch of the note may oscillate around a target value before it stabilizes [Martin99].

In wind instruments, the spectrum of the resulted sound depends both on the spectrum of the standing waves within the instrument and the portion of the sound energy that leaks outside [Rossing90]. In brass instruments, the radiation curve of the bell is of high pass type, and it filters the internal mode frequency spectrum. The resulting steady state spectral envelope of brass instruments has a cutoff frequency, above which the amplitude of partials decreases sharply with frequency, and below which all radiated spectral components are approximately equal or slightly increase with frequency. The rate of fall above the cutoff is typically 15-25 dB / octave, and the rate of rise below cutoff 2-4 dB / octave [Fletcher98].

However, the spectrum actually perceived by the listener is more complicated. First, the spectra of brass instruments change substantially with changes in pitch and loudness [Rossing90]. When the instrument is played more loudly, the partials near and above the cutoff become stronger [Fletcher98]. The slope below cutoff increases and the slope above cutoff decreases as the intensity level grows. Second, the bell makes the horn more directional at high frequencies, causing the perceived spectrum to depend on the angle between the bell's axis and the listener's position [Rossing90].

The player may deliberately modify the sound by using mutes. For example, different types of mutes can be used in the trumpet and trombone. In general, they are ment to mute the sound of the instrument, however, the effect is frequency dependent and the sound quality also changes. Special colorations occur at frequencies above 100 Hz, where the mutes have resonances and antiresonances [Fletcher98]. A special technique is used with the French horn, as the player can place his or her hand into the horn. This inhibits the radiation of the bell and increases the reflection of higher frequencies, making it easier to play the higher notes [Rossing90].

### The woodwind

The sound production mechanism in the woodwinds is similar as in brass instruments. Air pulses propagate from the reed, and a positive feedback synchronizes the vibration of the reed to that of the tube. However, the reed has almost no control over the frequency, while the player's lips have considerable mass allowing substantial level of lip control over the frequency in brass instruments [Rossing90]. The excitation of the flute is different, in this instrument a blow of air towards a hole in the tube vibrates at the frequency occurring in the tube [Karjalainen99]. Of these three major families, the woodwinds tend to have the most rapid attack transients, except for the flute [Martin99].

An important characteristics of wind instruments is the acoustic cutoff frequency of the air column, caused by the open tone holes. Below this frequency, sound is reflected back and resonances build up, but above it sound radiates freely to the environment [Fletcher98]. The sound outside the instrument is not as clearly of low pass type, since the high partials are strengthened by the more efficient radiation [Rossing90]. However, this cutoff frequency is essential to the tone of the particular instrument, and is quite independent of the pitch of the note [Martin99]. Fletcher and Rossing present an idealized spectral envelope for reed-woodwind instruments [Fletcher98]. Below the tone-hole lattice cutoff, the radiated power in harmonics falls about -3 dB per octave. Above it, the rolloff is from -12 dB to -18 dB per octave. If the instrument has a cylindrical bore, the power of the even harmonics rises about 3 dB per octave for frequencies below the cutoff.

The directional characteristics of woodwinds are more complicated than with the brasses since the radiation from the open mouth of the bell is supplemented by radiation from the open finger holes. The harmonics below the cutoff frequency radiate mainly from the first one or two tone holes. Higher partials propagate along the open hole part and radiate both from the open tone holes and the open bell [Fletcher98]. The cutoff frequency also limits the highest note readily playable on the instrument. We will now look at the subgroups in a little more detail.

*Double reeds*

The oboe has two broad resonances, the stronger is near 1 kHz and a weaker and more variable is near 3 kHz [Rossing90]. The first is related to the tone hole lattice cutoff, which is within the range 1000-2000 Hz for “musically satisfactory” oboes [Fletcher98]. The second is due to the mechanical properties of the reed [Rossing90]. The radiated spectrum of the oboe rises gradually with increasing frequency until it starts falling about 12 dB per octave above the cutoff. The behavior is almost the same with different playing levels, causing a bright and “reedy” tone [Fletcher98].

The English horn is an alto version of the oboe. It has a prominent resonance near 600 Hz, and a weaker one near 1900 Hz. Above the resonances, the spectrum rolls off with 20 dB per octave. [Martin99] The English horn has a pear-shaped bell that effects distinctively notes near its resonance [Rossing90].

The bassoon is much larger than the oboe and the English horn. The spectrum of bassoon contains a complete range of harmonics. The radiated fundamental is weak in the low notes because of the small tube diameter. Because of the relatively low cutoff between 300-600 Hz, the tone of a bassoon is mellow rather than bright. It has two formants, a strong one at 440-500 Hz and a weaker one at 1220-1280 Hz. As with the two previous instruments, the lower is related to the transition at the tone hole cutoff, and the higher is probably due to the reed [Fletcher98]. The contrabassoon is a larger bassoon, with dimensions about twice that of the bassoon [Martin99].

*Clarinets*

The spectrum of clarinets is limited by the tone-hole cutoff, which varies from 1200-1600 Hz with the B-flat clarinet, depending on the instrument piece [Fletcher98]. The B-flat clarinet, like all reed woodwinds, is rich in harmonics**.** The relative strengths of the odd and even partials depend on their frequencies and on the played note. In the low register below the cutoff, the odd partials are much stronger than the even partials, the second harmonic may be almost completely absent from the spectrum [Fletcher98, Martin99].

*Saxophones*

Saxophones are popular instruments for example in jazz music, but their sound has been studied only a little [Martin99]. The mouthpiece of a saxophone has a significant effect on the tone quality of saxophones. Its resonance is typically comparable to the lattice cutoff frequency, and causes a clear formant in that frequency. For an alto saxophone, the formant is around 850 Hz [Fletcher98].

*Flutes*

In these instruments, the vibrating element is a jet of air blown by the player towards the blowing hole, and is sometimes called an air reed. As in wind instruments, there exists positive feedback, however, now the input flow is controlled by the direction of air flow due to standing waves in the air column, not by pressure pulses [Rossing90].

The onset of a flute has some distinctive characteristics. Generally, the flute has a very slow, smooth attack which can last over 150 ms [Martin99]. The sound starts with noise due to the blow, after which the vibration steadily grows. The spectrum of the sound changes during the onset, as the high frequencies grow slower in the beginning [Karjalainen99].

The steady state spectrum of a flute sound is characterized by strong low harmonics in the low register and a resonance maximum near 600 Hz, with a high frequency rolloff from 10-30 dB per octave. The sound is nearly sinusoidal at frequencies above 800 Hz [Martin99].

Players often use a type of vibrato with the flute in the range of 5-6 Hz by introducing a rhythmic variation into the blowing pressure. The level of the fundamental changes only little, but the variation in the amplitude of higher harmonics can be remarkable. The frequency changes very little [Fletcher98]. A unique style of playing is the flutter style, where the player flutters his or her tongue while blowing air into the hole.