Computer Models for Musical Instrument Identification

Nicolas D. Chétry

Centre for Digital Music
Department of Electronic Engineering
Queen Mary, University of London

A thesis submitted for the degree of
Doctor of Philosophy
of the University of London

April, 2006

"I certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline."

Nicolas D. Chétry

Abstract

A particular aspect in the perception of sound is concerned with what is commonly termed as texture or timbre. From a perceptual perspective, timbre is what allows us to distinguish sounds that have similar pitch and loudness. Indeed most people are able to discern a piano tone from a violin tone or able to distinguish different voices or singers.

This thesis deals with timbre modelling. Specifically, the formant theory of timbre is the main theme throughout. This theory states that acoustic musical instrument sounds can be characterised by their formant structures. Following this principle, the central point of our approach is to propose a computer implementation for building musical instrument identification and classification systems.

Although the main thrust of this thesis is to propose a coherent and unified approach to the musical instrument identification problem, it is oriented towards the development of algorithms that can be used in Music Information Retrieval (MIR) frameworks. Drawing on research in speech processing, a complete supervised system taking into account both physical and perceptual aspects of timbre is described.

The approach is composed of three distinct processing layers. Parametric models that allow us to represent signals through mid-level physical and perceptual representations are considered. Next, the use of the Line Spectrum Frequencies as spectral envelope and formant descriptors is emphasised. Finally, the use of generative and discriminative techniques for building instrument and database models is investigated. Our system is evaluated under realistic recording conditions using databases of isolated notes and melodic phrases.

Acknowledgements

I always believed that writing my Ph.D. thesis would mark in some ways the end of my education. Looking back on it at the time of writing, I have realised that it is just the beginning of it. The task is certainly not easy but the personal achievements that come out of it are decidedly priceless.

I would like to thank here the colleagues and persons who I have been in contact with, who directly or indirectly accompanied the development of my work, and without whom these three years and a half spent in the Centre for Digital Music wouldn't have been so beneficial.

First and foremost, I would like to thank my supervisor, Mark Sandler, for his expertise and for the confidence he placed in my research. I also greatly appreciated the freedom in research and investigations that the Centre for Digital Music offers to his members. I also would like to thank Mike Davies, my second supervisor, with whom I collaborated on several aspects of my work, for having let me bother him with my various questions which, after all, were not so existential.

My special thanks go to Miki Mond for his friendship and for the advise he gave me before and during this time spent in London. I would like also to thank Gang Feng, a truly fantastic teacher, who is the person at the origins of my interest in digital audio signal processing.

My deep gratitude is due to Nikolaos Mitianoudis, who took some of his precious time for proof-reading this manuscript, and to Juan-Pablo Bello, for his comments regarding the organisation of the thesis.

Finally, I would like to thank the colleagues, researchers and past members of the Centre, for the various collaborations, discussions and for the good times we have spent together. In no particular order, I owe special thanks to Emmanuel, Paul, Dawn, Katy, Chris L., Chris D., Chris H., Ranting, Maria, Yves, Andrew, Samer, Matthew, Xue, Thomas, Peyman and Massimiliano.

"So don't fear, if you hear, a foreign sound to your ear"
It's Alright, Ma (I'm Only Bleeding)
Bob Dylan

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASA	Auditory Scene Analysis
AMDF	Average Magnitude Distance Function
ATH	Absolute Threshold of Hearing
CASA	Computational Auditory Scene Analysis
CDA	Canonical Discriminant Analysis
EM	Expectation–Maximisation
FFT	Fast Fourier Transform
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
HMP	Harmonic Matching Pursuit
IFFT	Inverse Fast Fourier Transform
k-NN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
LP	Linear Predictive
LPC	Linear Predictive Coefficients
LSF	Line Spectrum Frequencies
LSP	Line Spectrum Pairs
MAP	Maximum *a-posteriori*
MDS	Multidimensional Scaling
MFCC	Mel-Frequency Cepstrum Coefficients
MIR	Music Information Retrieval
OLA	Overlap and Add
PARCOR	Partial Correlation (coefficients)
PLSF	Perceptual LSF
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
RBF	Radial Basis Function
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
STFT	Short-term Fourier Transform
WLSF	Warped LSF

Introduction

A particular aspect in the perception of sound is concerned with what is commonly termed as texture or timbre. Timbre is also referred to as tone colour or quality of sound. From a perceptual perspective, timbre is what allows us to distinguish sounds that have similar pitches and loudnesses. Beyond this rather vague definition, it is that property of sound that allows us to understand most our surrounding environment. Without seeing, we can picture a car that is passing from its engine noise, we can imagine someone walking when we hear footsteps in a corridor or we can recognise familiar voices over the telephone. Indeed most people are able to discern a piano tone from a violin tone or able to distinguish different voices or singers.

The understanding of timbre is of great interest for psychologists, musicologists, physicists and scientists. Probably the first research work embodying all these aspects was by of Helmholtz [Hel54], published in 1877. Through a radically experimental approach, Helmholtz, physicist, musician, sets the fundamental basis of what will be later described as *musical timbre perception*.

Helmholtz considered that the majority of the timbral information was contained within the quasi-stationary part of a sound. This view is partially acceptable, especially if one considers the class of pseudo or quasi-stationary signals. However, simple experimental observations revealed the importance of onsets, attacks and other time-related cues in the perception of timbre.

So the facets of timbre are multiple. This led Plomb to describe it as a *multi-dimensional attribute of sound* [Plo70]. At the signal level, timbre depends on a multitude of acoustic properties whose individual contributions to the whole are not well-defined. This thesis proposes to focus on one of these aspects.

Specifically, this research starts from the *formant theory of timbre*. This theory states that acoustic musical instrument sounds have characteristic and salient formant structures that can be used to uniquely characterise them. Following this principle, the

central point of our approach is to propose a computer implementation for building musical instrument identification and classification systems.

The choice of each technique composing our system is carefully explained, justified and interpreted. Drawing on research in speech coding and speaker identification, we propose to tackle the problem from both physical and perceptual points of view. Although this research is oriented towards the development of algorithms that can be used in Music Information Retrieval (MIR) frameworks, its main thrust is to propose a coherent and unified approach to the musical instrument identification problem.

The search for invariance and constancy in timbre is central to the building of instrument models. It is a fact that the perception of the timbre of an instrument or the identification of a particular sound object by humans can be achieved in a wide variety of acoustic circumstances. Simple experiments such as resampling a saxophone melody from 44100 Hz (thus roughly containing all the frequency details that the human ear can perceive) down to 4000 Hz will certainly affect the overall quality of the sound but the saxophone can still be recognised. The resampling operation introduces non-negligible alterations at the signal level: roughly a tenth of the time samples and frequency information are remaining, only the frequencies up to 2000 Hz being preserved. But humans have this extraordinary ability to adapt, map and transpose the knowledge learned from previous experiences to various novel situations that were unknown to them until then.

Building a timbre model consists to a certain extent of mathematically characterising this invariance. However, it is not guaranteed that a computer system that performs well under given conditions will be able to reproduce the human ability of adaptation. In a computer system, this problem is directly transposed to the feature and classifier levels since any change in the signal will be decidedly carried by the acoustic descriptors. From there arise concerns about the importance of using acoustic pre-processing or methods of feature normalisation in order to get independence from particular recording conditions, instrument brands or playing styles. One approach to address this problem is to consider the largest possible amount of representative sounds for each instrument class. Models will therefore have the knowledge of various realisations of the same timbre.

Another approach consists of tuning a system for a particular application, at some expense on the models generalisation properties. For instance, it is understandable that if one wishes to identify a monophonic instrument in a melodic context, models built from melodic phrases will perform better than models trained using isolated notes, as will be shown in chapter 6.

Thesis overview

We propose to recognise acoustic musical instruments based on their sound spectral structure. We describe and evaluate a supervised system composed of three processing layers: parametric modelling, acoustic timbral descriptor extraction and machine learning algorithms. Supervised musical instrument identification systems encountered in the literature are generally composed of these three layers. But depending on the considered approach, emphasis can be placed on any of them. For instance, multi-feature systems such as the ones described in section 2.3.1 pay attention on the nature and number of features that are considered for building the instrument models. Conversely, data-mining approaches attempt to optimise the performances at the classifier level by selecting for each instrument the feature sets that maximise the correct identification rates.

The approach presented in this thesis gives equal contribution to each layer by presenting a unified methodology. Specifically, the focus is on preserving consistency between known perceptual and experimental facts in the perception of timbre, the acoustic descriptors computed from the waveforms and the algorithms used for building the models.

The use of parametric models allows us to describe sounds through mid-level physical and perceptual representations. Another advantage of using parametric models is that they can represent different sounds, having different physical structures, in common and universal frameworks whereby intrinsic differences and similarities can be more easily characterised. In particular, we investigate the use of linear predictive models and sinusoidal analysis/synthesis models in a musical instrument identification context.

We argue that the source-filter linear predictive model, widely used to represent the mechanisms of speech production, can be transposed to the case of acoustic musical instrument sounds. The emphasis is on the consideration of the Line Spectrum Frequencies (LSF) as spectral envelope and formant structure descriptors. We study the influence of pitch on the LSF and propose to use the pitch as a prior for the learning and identification phases.

It is a fact that the perception and interpretation of sound relies on the low-level processing that takes place in the ear. Hence, it can be argued that the consideration of psycho-acoustic knowledge would better fit the mechanisms of timbre perception by humans. For this reason, we propose to include perceptual principles in a sinusoidal analysis/synthesis framework prior to the feature extraction stage.

From a set of multi-dimensional features, machine learning algorithms are used to determine a mathematical invariant that can describe the timbre of an instrument. For this purpose, *generative* and *discriminative* methods are compared. All throughout the thesis, the distinction between these two types of techniques, which yield two different interpretations and implementations of the pattern recognition problem, is clearly made.

A complete evaluation of the approach is presented. It is shown how well it performs under realistic conditions, using databases of mono-timbral isolated notes and melodic phrases.

Thesis outline

This text is organised as follows: chapters 1 and 2 present the background of this thesis, introduce the problem and limit the scope in which this research has been carried out. The approach that is proposed to tackle the problem of identifying and recognising musical instruments is described. In chapters 3 and 4, technical background about acoustic timbral descriptors and machine learning algorithms are reviewed. Experimental evaluations are reported in chapters 5 and 6, while conclusions about this research and extensions of the proposed method close this thesis.

Chapter 1 describes important acoustic and cognitive aspects in the perception of sound and music by humans. Starting with the acoustic representation of audio signals in the auditory peripheral system, this chapter introduces fundamental notions in acoustics such as frequency and temporal masking, pitch and loudness. Principles of two perception theories are outlined. Next, the concepts of texture and timbre of sound are defined. In particular, the principles governing the *formant theory of timbre* are described.

Chapter 2 introduces fundamental principles involved in the design of automated musical instrument identification systems. Experimental results assessing human ability to identify and recognise musical instrument tones are summarised. An overview of available techniques and existing systems encountered in the literature is given. Next, a computer implementation of the *formant theory of timbre* is proposed.

In chapter 3, various acoustic timbral descriptors that can be computed from the sound waveforms are described. The use of parametric models to represent the mechanisms of sound production by musical instruments is investigated. In particular, the theoretical principles of linear predictive models and sinusoidal models are recalled. It is further justified why the linear predictive model can be transposed to represent

the mechanisms of sound production by musical instruments. The emphasis is on the Line Spectrum Frequencies and on their use as formant structure descriptors. Finally, we introduce a perceptually motivated sinusoidal analysis/synthesis technique that will serve as pre-processing layer to the feature extraction stage.

In chapter 4, three machine learning algorithms are presented. The distinction between *generative* and *discriminative* methods is highlighted. The theoretical principles of K-means, Gaussian Mixture Models and Support Vector Machines are analysed. The interpretations of learning characteristic spectral shapes and classifying spectral envelopes when building *timbre* and *database* models are proposed.

Chapters 5 and 6 are concerned with the experimental evaluation of our approach. In chapter 5, a database of isolated notes, containing 3292 tones classified into 10 classes of instrument is used to evaluate the system. The use of a global sound attribute such as the pitch is then proposed for both the modelling and identification phases. We finally show that our approach advantageously takes into account specific acoustic information carried by the onset of musical tones.

Drawing on the conclusions reached in chapter 5, our approach is extended to deal with sounds taken from realistic musical contexts. In chapter 6, our system is evaluated using a database of mono-timbral melodic phrases extracted from commercial recordings.

A summary and a conclusion about this research close the thesis. Perspectives towards further related works in MIR are discussed. In particular, a direct application of our system for the evaluation of spectral and texture similarities between songs is proposed.

1. Acoustic and cognitive aspects of music

The perception and interpretation of sound play important roles in human evolution. From birth to death, humans learn and interact by its means. By listening to sounds, we can gather a multitude of information from our environment, complementing the visual clues. By speaking and listening to others, we are able to exchange ideas and socialise. By listening to music we can relax and experience emotions. These multiple aspects of the importance of sound for humans share the same low-level mechanisms of perception. Attempting to map and quantify what can be described as the *response* of the subject to a stimulus is of particular interest in research in audio and musical signal processing.

At the acoustic level, no distinction is made between the origin and nature of sound, its meaning and the eventual musicality that it conveys. The ear acts as a transducer, receiving and transforming a signal into a suitable form to be processed at the upper levels of the perceptual chain.

This chapter introduces the fundamental acoustic and perceptual principles governing the perception of sound by humans. In section 1.1, details about the acoustic processing that takes place in the ear are described. Next, an overview of two theories of perception is given in section 1.2. The concepts of music semantics and categorisation are outlined in sections 1.2.3 and 1.2.4 respectively.

Section 1.3 focuses on the timbre or texture of sounds. The perceptual attributes that characterise the quality of musical instrument tones are described. In particular, principles of the *formant theory of timbre* are exposed in section 1.3.1.3. In section 1.3.1.4, the classical theory which states that steady-state and harmonic portions of sounds entirely define the timbre is complemented with important experimental facts revealing the importance of onsets, transients and other time-related cues.

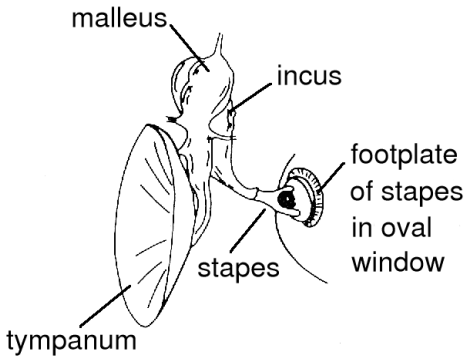


Figure 1.1: Schematic representation of the middle ear. The tympanum delimits the boundary with the external ear, while the oval window marks the beginning of the inner ear.

1.1 Acoustic aspects

The overall physiological mechanisms of hearing are quite well known and a considerable literature exists on this topic ([ZF99], [Moo97]). The collaboration between psycho-acousticians, cognitians and audio engineers has greatly helped to consolidate the foundations of the existing techniques and, at the same time, to broaden the areas of research.

1.1.1 Anatomy of the ear

The auditory system can be divided into two main parts: the peripheral auditory system, commonly called the *ear* (responsible for the low-level signal processing), and the central auditory system beginning with the first neurons and ending with the cortex where the signals are interpreted (responsible for the high-level signal processing).

The mechanisms driving the auditory process are complex, compounded by acoustic, mechanic, hydrodynamic and electro-chemical subsystems. The main function of the ear is to encode the acoustic signal into nervous influx by successively transforming it into different forms. It acts as a transducer, similar to a microphone. In particular, it is composed of:

- **The outer ear**: the purpose of the *pinna* is to collect sound waves. It performs resonant gain, direction-dependent filtering and impedance matching. Waves are then transmitted through the auditory canal (*meatus*) to the eardrum (*tympanic membrane*). This constitutes the external part of the ear and waves are still in their acoustic form.

- **The middle ear**: vibrations of the tympanum are mechanically communicated through a linked triplet of small bones, the ossicles (called *malleus, incus* and *stapes* respectively) which are elastically connected one to the other. The footplate of the stapes is fastened to an elastic membrane (the *oval window*), marking the boundary with the inner ear. A schematic representation of the middle ear is given in figure 1.1.

- **The inner ear**: the inner ear is composed of one main central organ, the *cochlea*, entirely filled with a liquid called the *endolymph*. Vibrations from the oval window are transmitted to the cochlea where thousands of tiny hair cells are stimulated. The hydrodynamic waves are transformed into nerve pulses which are then transmitted to the brain through the auditory nerve.

The first two stages mainly act as an impedance-matching transformer. The analysis and transcription of signals take place in the inner ear. Although the cochlea is dedicated to hearing, the inner ear contains also organs responsible for the balance.

The cochlea is composed of two and a-half turns of a spiral cavity filled with endolymph. Vibrations of the oval window induce waves in this liquid. One particular structure in the cochlea, the basilar membrane, is mapped with approximately 30000 hair cells, which selectively respond to frequencies close to their characteristic frequencies and induce an electric impulse in the auditory nerve. In some way, they behave like bandpass filters whose selectivity is directly related to their position along the basilar membrane. Such a decomposition is termed as *tonotopic*.

In figure 1.2 is depicted a cross-sectional diagram of the inner ear. It can be noticed that the high frequencies are decomposed at the beginning of the first spiral cavity turn (the frequencies 1000 Hz and 2000 Hz are treated at distances 20 mm and 17 mm from the vestibular end respectively) whereas low frequencies (below 200 Hz) are decomposed in the apical turn.

A particular vocabulary is used when describing how sounds are perceived. Musicians, acousticians and scientists share a common language in order to express their sensations: loudness is a measure of the perceived intensity and is a subjective concept; the notion of pitch allows to represent complex musical stimuli on a perceptual frequency scale. Timbre, or quality of a sound helps to differentiate between two sounds having similar pitches and loudnesses.

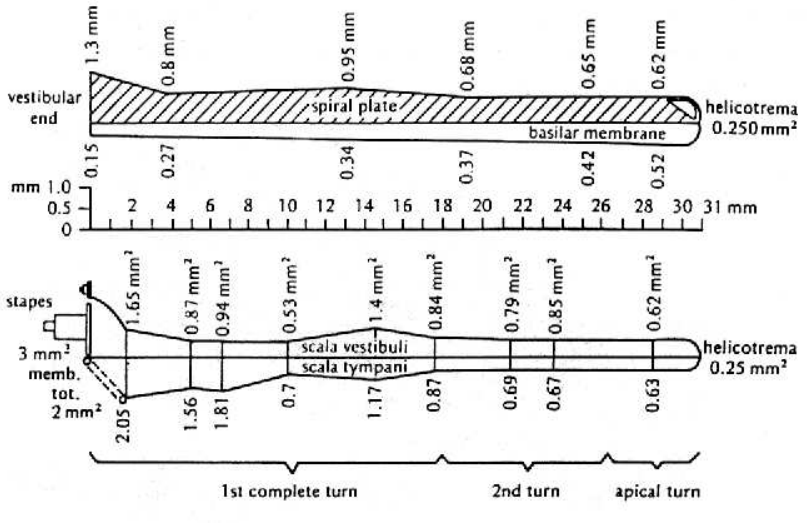


Figure 1.2: Schematic diagrams showing the basilar membrane dimensions (top) and of the scalae of the human cochlea (bottom). Adapted from Fletcher, 1953, by Yost and Nielson, 1977.

1.1.2 Intensity and loudness

Intensity, or the acoustic pressure of a sound, is an absolute physical quantity that can be measured with instruments, while loudness is a subjective aural response depending on each individual's hearing acuity. Intensity is expressed as an energy per time unit and per surface unit. The human ear is sensitive in the range 1 $watt.m^{-2}$ (the pain threshold) to 1 trillionth of a $watt.m^{-2}$ (corresponding to the softest perceptible sound). It is common to adopt a logarithmic scale to deal with such a wide range of variation. According to the Weber-Fechner law, the response of any sense organ is approximately proportional to the logarithm of the magnitude of the stimulus. The Bel (B) is defined as the logarithmic ratio of the considered sound intensity I over the slightest perceptible intensity I_0, corresponding to the faintest pure tone (at frequency $f_0 = 1$ kHz) that can be heard by the human ear. Mathematically, it can be written

$$L_b = \log_{10}(\frac{I}{I_0})$$

1 Bel corresponds to a ratio of 10 in sound intensity. However, as the smallest detectable variation in intensity is 0.1 B, the decibel (dB), defined as

$$L_{db} = 10\log_{10}(\frac{I}{I_0}),$$

represents a standard and psycho-acoustically relevant unit to measure sound intensity

level.

Another useful measure of loudness is the Sound Pressure Level (SPL) and is defined for an acoustic pressure p as

$$L_p = 20 \log_{10}(\frac{p}{p_0})$$

It is also expressed in decibels and measures the relation to a reference pressure p_0. By convention, $p_0 = 20$ μPa, which is approximately the threshold of human hearing in its most sensitive range (1 to 3 kHz, see figure 1.3).

The ear can perceive sounds between 16 Hz and 20 kHz, corresponding to approximately 10 octaves. Its most sensitive range spans frequencies from 1000 to 3000 Hz. The Absolute Threshold of Hearing (ATH), corresponding to the ear's internal noise level, is the intensity level below which audio stimuli are not perceived at all (represented by the curve (a) in figure 1.3). It is often taken as the 0 dB reference. Similarly, the threshold of pain (curve (b) in figure 1.3) is the minimum level for which sounds having intensities located above irreversibly destroy the hair cells (approximately 120 dB). The auditory field is defined as the area delimited by the thresholds of hearing and pain.

The loudness of a pure tone is not only determined by its acoustic pressure but also depends on its frequency. Isotonic curves mark equal perceived loudness for a given sound intensity as a function of the frequency. They correspond to the curves between (a) and (b) in figure 1.3.

1.1.3 Frequency masking and critical bands

The hair cells responsible for the tonotopic decomposition of the signal in the cochlea respond to frequencies close to their characteristic frequencies. Another important property of the transcription mechanism is that a stimulated region of the basilar membrane inhibits the neighbouring hair cells' responses. As a result, when two sounds with different but adjacent frequencies are simultaneously presented, it might happen that one can be masked by the other. This phenomenon is called *frequency masking* and can be total or partial, depending on the two stimuli loudnesses and relative frequencies.

In the presence of a masking sound, the auditory threshold is deformed. Figure 1.4 illustrates the frequency masking phenomenon with a masker with frequency $f = 1$ kHz. The absolute threshold of hearing is deformed and sounds having their intensities below the masking threshold are not audible.

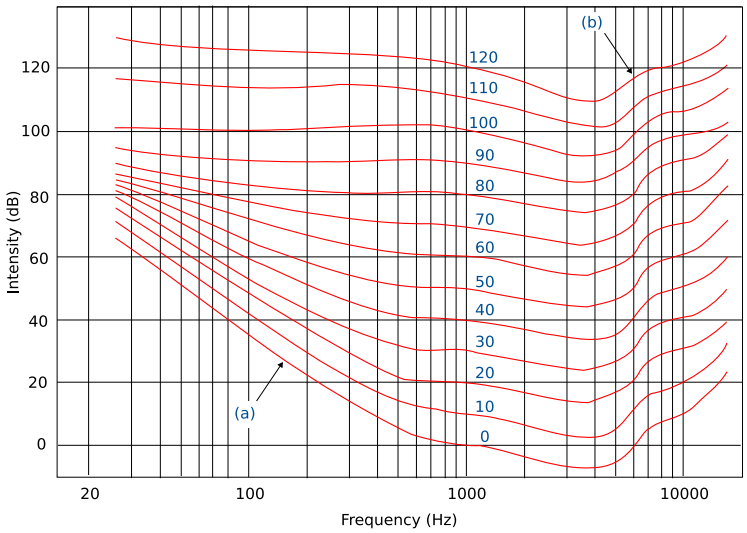


Figure 1.3: Fletcher-Munson lines of equal subjective loudness. Frequencies (Hz) are reported in logarithmic scale in abscissa. The curve (a) is the absolute threshold of hearing, and the curve (b) the threshold of pain. Curves between (a) and (b) are isotonic curves (labelled in phons). Adapted from [FM33].

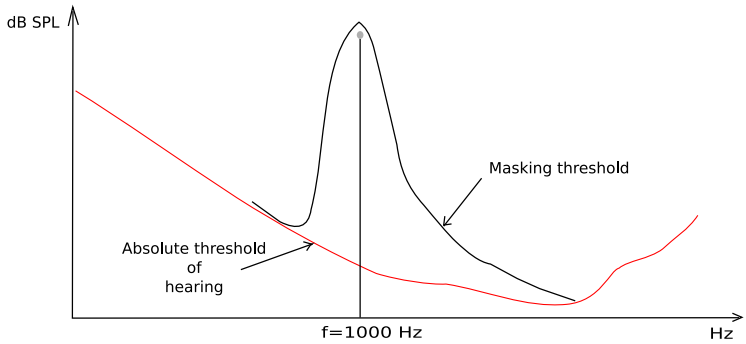


Figure 1.4: Deformation of the absolute threshold of hearing in presence of a masker with frequency $f = 1$ kHz. Sounds having their intensities below the deformed masking threshold are not audible.

The spread in frequency of the masking effect shows that a pure tone excites the auditory system on a scale much broader than its physical spectra. The concept of the auditory filter, introduced by Fletcher [Fle40], suggests that the basilar membrane behaves as a filter bank with interleaved boundaries. A representation of such an auditory filterbank is depicted in figure 1.5. In practice, this model is approximated by a series of rectangular filters whose widths are experimentally determined. Note that only a frequency band, the *critical band*, of the masker participates in the masking.

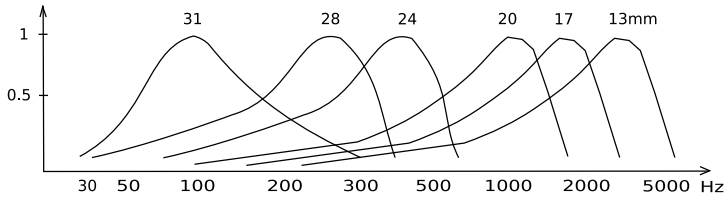


Figure 1.5: Auditory filter bank model showing the filters frequency responses and the distances in millimetres from the vestibular end where the decomposition takes place in the basilar membrane (see figure 1.2).

The *critical band* corresponds to the minimum frequency space for which two harmonics in a complex periodic sound can be discriminated. By definition, one Bark is the width of a critical band, whatever its central frequency is. Its value is 100 Hz up to 500 Hz and approximately 20% of the central frequency above. Critical band boundaries can be approximated using the analytical formula [ZF99]:

$$z = 13 \arctan(0.76\frac{f}{1000}) + 3.5 \arctan((\frac{f}{7500})^2)$$

where f is expressed in Hz and z in Bark.

The perception of frequency is not linear over the entire frequency range. Above 1000 Hz, the frequency of a pure tone has to be more than doubled to be perceived as twice in height by the ear. A constant variation in the Mel scale corresponds to a constant variation in the perceived frequency. The Mel scale can be calculated as [SV40]:

$$M = 1127.01048 \log(\frac{f}{700} + 1)$$

where f is expressed in Hz and M in Mels.

1.1.4 Temporal masking

Temporal masking is also called non-simultaneous masking. Due to temporal inhibition of the nerve fibres' activity following a stimulation, and to a certain extent to the phenomenon of time integration in the ear, two sounds separated by a time gap can mask each other.

Two types of temporal maskings can be distinguished: the *forward masking* (the maskee is before the masker) and the *backward masking* (the masker is before the maskee). These phenomena only occurs when the two sounds are separated by a maximum of a few tens of milliseconds.

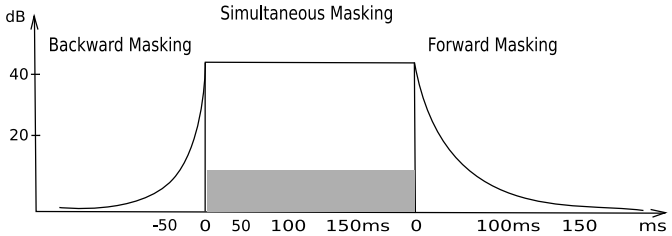


Figure 1.6: Temporal masking phenomenon created by a 200 ms duration pure tone stimulus (grey area). In ordinate is reported the level of just masked pure tone [ZF99].

Important conclusions may be drawn from figure 1.6. Firstly, simultaneous masking is more effective when the frequency of the masked signal is equal to, or higher than that of the masker. Secondly, while forward masking is effective for a significant time after the masker has stopped, backward masking may only be effective for few tens of milliseconds before the masker onset.

1.1.5 Pitch and fundamental frequency

Most of the complex tones that the ear perceives can be classified according to their position on the frequency scale. When these sounds are periodic, a unified pitch is perceived rather than separated frequency components corresponding to the overtones. According to Helmholtz [Hel54], the pitch is mainly determined by the fundamental frequency value f_0, i.e. the value corresponding to the repetition rate of the temporal periodicity of the sound, and by its relative strength compared to the upper partials. In this classical theory, the higher harmonics were thought only to influence timbre.

However, for certain sounds, low pitch perception also occurs even if the *corresponding* fundamental frequency is not physically present in the signal[1], thus refuting Helmholtz theory. Schouten [Sch40] introduced the formulation of the *periodicity pitch theory* in which pitch is mainly derived from the waveform periodicity of the unresolved higher harmonic of the stimulus. In this case, periodicity does not change if a component, such as the fundamental frequency, is missing.

In [Rit67], a frequency dominant region has been associated as mostly influencing the pitch sensation. It has been experimentally shown that partials falling between 500 and 2000 Hz (the *dominant region*) were contributing more to the pitch determination that the other ones. In this theory, low pitch of tones with low fundamental

[1] this phenomenon is known as the *missing fundamental*

frequencies depends on the higher partials while pitch of tones with high fundamental frequencies is determined by the fundamental as it lies in the dominant region.

Pitch is one of the three perceptual attributes of tones, along with the loudness and timbre that is used to completely describe sound. In music, pitch systems like the diatonic-chromatic or the 12-tone in Western music allow us to compare and organise notes of different sounds in similar frequency scales.

The various theories about the perception of pitch raise many questions about the nature of the processing that takes place in the brain. In particular, why are all components of the complex tones perceived as a perceptual unit so that all the partials fuse into one percept? Perception theories attempt to explain this phenomenon from a general cognitive perspective.

1.2 Perception

The most influential theories that have been proposed to explain the mechanisms of perception are outlined in this section. Commonly called cognitive or psychology theories, they intend to relate the *sensory* information to the interpretation of the message at a higher cognitive level. Their use is not restricted to audio and musical stimuli but also serves to describe, among others, the mechanisms driving visual perception.

1.2.1 The Gestalt school

The *structuralist* stream, represented by Wundt [Blu80], conceptualised the principle of *elementarism* and pre-supposed that complex stimuli can be reduced to indivisible elementary, local sensory experiences. In the same vein, the *atomist* stream explained the complex perception mechanisms as being a combination of simple sensory responses.

By opposing the structuralist theory, Wertheimer, Koffka and Köhler elaborated the form theory (*Gestalttheory*) in the twenties [ESS97]. Years earlier, Christian von Ehrenfels, in a paper "On Gestalt Qualities" [Smi88] pointed out that a melody is still recognisable when played in different keys, even though none of the notes are the same. Ehrenfels argued that if a melody and the notes that comprise it are so independent, then a *whole* is not simply the sum of its parts, but corresponds to a synthesised *whole effect*, or Gestalt. The existence of *global attribute* (*gestaltqualität*) was then advanced. According to Wertheimer, the latter is instantaneously perceived and prior to any combination process of elementary events. The Gestalt psychologists

radically rejected the notion of atomism (or elementarism) and proposed *holism* as alternative, emphasising the role of emergent properties and the importance of context. The more general concept of perceptual organisation which described objects as organised entities, as opposed to a combination of elementary structures, represents the basic principle of the Gestalt theory. In this theory, the form is the fundamental element and several laws and concepts explain the mechanisms of perception.

> *The fundamental "formula" of Gestalt theory might be expressed in this way. There are wholes, the behaviour of which is not determined by that of their individual elements, but where the part-processes are themselves determined by the intrinsic nature of the whole. It is the hope of Gestalt theory to determine the nature of such wholes.*
>
> Max Wertheimer (1924)

The most basic rule of Gestalt is the law of "Prägnanz". This law states that humans experience their environment by having a tendency towards *good forms*. The complementary Gestalt laws include:

- **Similarity**: if several stimuli are presented together, there will be a tendency to perceive objects in a way that similar stimuli are grouped together (figure 1.7(a)).
- **Proximity**: elements that are closed together (spatially or temporally) are more likely to be grouped together in a single object (figure 1.7(b)).
- **Good continuation**: the human perceptual organisation tends to prefer a continuity between stimuli in contrast to abrupt and discontinuous changes (figure 1.7(c)).
- **Closure**: closure denotes the human tendency to complete familiar objects or forms that have gaps in them (figure 1.7(d)).
- **Common fate**: the common fate law states that elements moving in the same *direction* are perceived together.

Although these principles were generally described first in relation to vision, they are equally applicable to audition. It is known that music played a major role in the emergence of the Gestalt school. Early research has focused on the perception of rhythm, pitch, melody, consonance and timbre.

However, perceptual grouping in more complex music is not simply a matter of linking different stimuli together. Rather, it involves a process whereby these stimuli

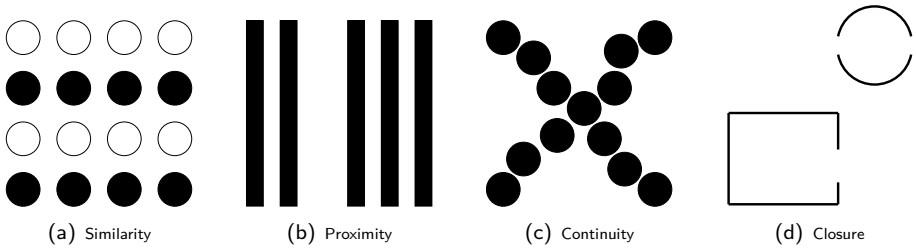


Figure 1.7: The Gestalt principles of similarity, proximity, continuity and closure.

are fragmented into their separate attributes, followed by a process of synthesis or grouping in which the different attribute values are recombined. Auditory scene analysis (ASA) formalises these principles and is introduced in the next section.

1.2.2 Auditory Scene Analysis (ASA)

By analogy with the perception of a visual scene, auditory scene analysis considers the auditory field as a *landscape* of sounds or streams. The approach proposes principles to describe the processes required of the human auditory system as it analyses mixtures of sounds. Auditory scene analysis assumes at first that mixtures are broken into small elements which are then grouped into sources using perceptual cues [Bre90]. These rules are, to a certain extent, derived from the Gestalt theory but completed by attributes specific to audio and musical signals such as the harmonicity, common onset times or modulation [DC95].

Bregman makes a distinction between two types of auditory grouping, namely *primitive grouping* and *schema-driven grouping*. *Primitive grouping* is driven by the incoming acoustic data, and is probably innate. In contrast, *schema-driven grouping* employs the knowledge of familiar patterns and concepts that have been acquired and learned through experience of acoustic environments.

Rather than objecting the Gestalt principles, auditory scene analysis has been built upon. It also defines a scope for the computer *implementation* of auditory perception algorithms. In Computational Auditory Scene Analysis (CASA), the aim is to take into account several grouping rules as starting hypotheses during the analysis of audio signals. These hypotheses are confirmed or refuted according to prior contextual knowledge about the signal. The best hypotheses are then selected for further processing [Ell96].

For instance, the *similarity* principle can be verified in that sounds similar in timbre or that change smoothly in frequency are likely to be emanating from the same

source. The principle of *proximity* refers to distances between auditory features with respect to their common onsets, pitches, and loudnesses.[2] Features that are grouped together have a small distance between each other, and a long distance to elements of another group. Likewise, the principle of *good continuation* identifies smoothly varying frequencies, loudnesses, or spectra with a changing sound source whereas abrupt changes indicate the appearance of a new source. The Gestalt principle of *closure* refers to a tendency to complete perceptual forms. Indeed, it has been shown that listeners are able to perceptually restore parts of a quieter sound that have been masked by a louder sound. This process is known as *auditory induction* [Deu82]. Finally, the principle of *common fate* groups frequency components together, when similar changes occur synchronously, such as synchronous onsets, glides, or vibratos.

1.2.3 Music and semantics

Associating a semantic for music signals assumes that music is pertaining to the meanings of words and concepts and supposes that there exist *musical languages*. Undoubtedly, music conveys a message, but an universal musical referential does not exist as such. At first sight, speech and musical signals are similar so that researchers in digital music processing were well equipped to start with. They share the same subjective descriptors, such as the pitch, loudness and timbre. The same techniques can be used to analyse either of them. However, their differences lie at a higher cognitive level.[3]

In speech recognition, the process is to map the acoustic signal into its corresponding semantic internal representation, i.e. words and sentences. In automatic music transcription, the process is to determine the pitches of the notes, their starting and ending times as well as the instrument being played. But in contrast to speech, these objects do not refer to any concept outside the musical world. This led Meyer [Mey56] to conclude that music was not a form of knowledge.

So how could the nature of the musical message be defined if music is not a form of knowledge? In the case of spoken speech, speaker and listener share the same dictionary that they both need to know in order to understand each other. On the other hand, a speech signal is also the vector of emotions and side information such as exclamation, interrogation or hesitation carried by the prosody. This information is almost invariant across a language, so that people from the same cultural background

[2] the *proximity* principle is equally applicable to the grouping of partials into harmonic objects [Bre90]. An example of such grouping is given in appendix B

[3] if we exclude the fact that music is polyphonic by nature as opposed to spoken speech

are able to interpret the message and the associated emotional information. In the case of music, the melody, the rhythm or the timbre can constitute some of the various manifestation of the musical message but the effect they generate on us is purely emotional. Do I need to be musician to appreciate music? Do I need to know the dictionary, the rules that composes music? I don't think so, musicians and non-musicians will appreciate it, but in different ways because music has this powerful ability to convey emotions.

So music is a emotional experience. Defining a semantic for emotions is a highly subjective and difficult process. However, there is a tremendous need in our modern society to describe musics in terms of words, concepts and relationships. The main goal of research in Musical Information Retrieval (MIR) is to associate a semantic to music. Practically, this corresponds to extract acoustic signal descriptors from the signals that can be related to emotions.

1.2.4 Categorisation

Categorisation is the basic cognitive process whereby sensory *information* gathered from the outside world is arranged into classes or objects. At the physical level, a stimulus is represented in a continuous, infinitely variable domain. As an example, an acoustic signal sampled at 20 kHz, perceived by the ears which has approximately a dynamic of 20 bits (0–120 dB) corresponds to a raw information rate of 400 kbits/s. At a higher cognitive level, what the brain interprets as a meaningful object (that can be the notes, the beat or phonemes in the case of speech) is *encoded* at a much lower rate. In the case of spoken speech, at the phonetic level, roughly 32 phonemes[4] can be distinguished to represent all the sounds in a language. At an average rate of 10 phonemes per second, the information rate is reduced to 50 bits/s. At a higher lexical level, for an average of 2.5 words per second and 20000 words in a dictionary, this information rate drops to 37 bits/sec. The process of *categorisation* is depicted in figure 1.8. The gap of Smolinsky qualifies the drop in information rate between the physical and symbolic representations of a signal.

The fundamental problems linked to the categorisation process are:

1. Recovering the physical world from its internal representation.

2. Naming and tagging the objects.

[4] 40–45 in the case of English

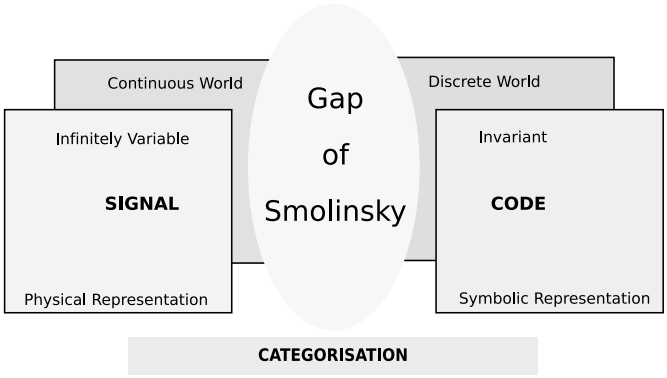


Figure 1.8: The categorisation process.

3. Defining an equivalence relationship between physical external world and internal semantic world.

Musical signals have a meaning, or more precisely, are given a meaning. Trying to define the nature of the message conveyed by a song, or a piece of music is a highly subjective and context-dependent process. There cannot be a unique and universal answer as cultural background varies from one individual to another. From there arises the difficulty of designing automatic musical genre classification, for example.

1.3 Audio texture and timbre

We focus in this section on the notions of texture and timbre of sound. Textures and timbre are subjective concepts referring to the quality of sound. In addition to the loudness and pitch, it is used to uniquely distinguish between sounds.

Definitions of timbre tend to be enigmatic and indicate what it is not rather than what it is. For instance, according to the American Standards Association (1960, p. 45),

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

It is further stated that:

> *Timbre depends primarily on the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus*

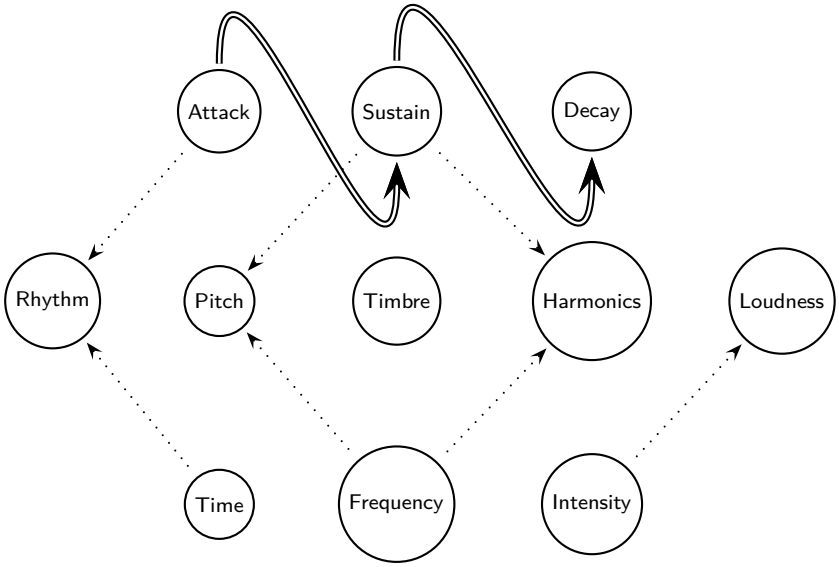


Figure 1.9: Timbre and other dimensions of sound.

	Loudness	Pitch	Timbre	Duration
Pressure	+++	+	+	+
Frequency	+	+++	++	+
Spectrum	+	+	+++	+
Duration	+	+	+	+++
Envelope	+	+	++	+

Table 1.1: Dependence of subjective qualities of sound on physical parameters. + = weakly dependent ++ = moderately dependant +++ = strongly dependent [Ros89].

This definition is rather confusing and tend to indicate that timbre is in fact not well defined. In contrast to other descriptors of sound, timbre has no defined associated physical quantities. Whereas intensity can be expressed in decibels, frequency in Hertz, timbre has no corresponding physical attribute. In consequence, timbre is often seen as a multi-dimensional attribute of sound [Plo70]. In figure 1.9, common subjective and physical components of sound are shown. In addition, table 1.1 relates the subjective qualities of sound to the physical parameters that are used to characterise the signal structure. In particular, note the dependence of timbre upon spectrum, frequency and envelope respectively.

Although timbre is not well-defined in terms of the physical attributes that con-

tribute to it, it has some characteristic properties that can be highlighted. In [RW82], the notion of timbral constancy and invariance is discussed. In particular, it is mentioned that sound sources can be reliably identified by humans in a wide variety of circumstances. Take the example of a speech signal over the telephone network. The bandwidth only covers the frequencies up to 4 kHz (corresponding to roughly a fifth of the maximum perceptible auditory field) but humans perform remarkably well at recognising familiar voices in such conditions. Similarly, a saxophone sound is still perceived as being a saxophone sound whether it is played from a vinyl recording, over an AM radio band, from a compact disc or in a concert hall. Thus the question arises as to the physical correlates of this constancy, and especially as to a physical invariant of timbre.

The search for timbre-invariant highly motivates research in timbre and texture modelling. Because the timbre of a sound can be recognised under various conditions, it is advanced that there should exist particular properties at the signal level, whose contribution to whole led the definition of an invariant.

1.3.1 Timbre correlates

Timbre can be seen at the same time as a low-, mid- and high-level descriptor of sound. Low-level when it is associated with the physical structures of sound, mid-level when it is associated with musical cues, such as pitch and harmonicity and high-level when it is used to describe the texture or *feel* of a sound. For instance, a sound can be qualified by the terms *shrill*, *rough*, *mellow* or *warm*... Such words describe real and consistent differences in our responses to musical sounds but it is no simple matter to relate these sensations to the signal's physical structure.

Timbre is therefore thought to be described by a number of features and their combination. The difficulty of characterising an invariant for the timbre is a direct consequence of this multiplicity of aspects. Further, the wide range of physical structures of the *class* of musical sounds tremendously increases the amount of parameters to be studied and taken into account. For these reasons, research in the field independently focused on one aspect of timbre at a time.

The spectral characteristics of audio signals remain the foundation of the conventional understanding of musical sounds by humans. Early published works from Helmholtz [Hel54] dealt with the quality and colour of pure tones. Helmholtz considered that the majority of the timbral information was contained within the quasi-stationary part of a sound. These conclusions have been drawn from experimental

studies and it was then believed that spectrum and harmonic series would define timbre. This view is partially acceptable, especially if one only considers the class of pseudo or quasi-stationary signals. However, simple experimental observations reveal the existence of other cues participating in the perception of timbre.

Specifically, the other contributors are related to time. It has been shown that transients, onsets attack and decay [Ber63] on the one hand and vibrato and tremolo [SC64] [Mar99] on the other hand are important timbre correlates.

In the following, the spectral and temporal correlates of timbre are studied in more detail. Prior to that, the dependence of timbre upon phase is discussed.

1.3.1.1 Timbre and phase

Helmholtz believed that the ear was phase insensitive. In other words, if the spectral representation of two sounds have similar harmonic representations but different patterns in phase relationships, a listener will be unable to perceive the difference between the two sounds, even though they may have different waveforms. Helmholtz stated that "the quality of the musical portion of a compound tone depends solely on the number and relative strength of its partial simple tones, and in no respect on their differences in phase". Although he found that the effect of phase on the timbre of pure tones to be negligible, he admits an exception for non-musical sounds.

However, this characteristic of the auditory system is still discussed and it has been shown that indeed, modifying the phase between harmonically related partials could affect the timbre of sound up to a certain degree. Recently, it has been experimentally found in [DR02] that phase coupling phenomena between partials of sustained portions of sound can be used to distinguish between musical instrument families.

1.3.1.2 Timbre and spectrum

It is generally accepted that the main physical characteristics of sound used to define the timbre are related to the spectral representation of the signal. Over this restricted temporal span, periodic signals can be decomposed as a Fourier series and it was believed that the spectrum of these harmonic series would define timbre. Helmholtz [Hel54] was convinced that "certain general rules will result for the arrangement of the upper partials which answer to such species of musical quality as are called *soft, piercing, braying, hollow or poor, full or rich, dull, bright, crisp, pungent* and so on". He then distinguished between tones without upper partials or tones with inharmonic upper partials as affecting the quality of tones.

For sustained tones, the most important of these correlates are the harmonic content as well as the number and relative intensity of the upper harmonics present in the sound. Further, the odd/even relation between harmonics is also an important characteristic in the quality of sounds, for example the clarinet. Similarly, some musical sound sources have overtones whose frequencies are not exactly a whole integer multiple of the fundamental frequency, for example the piano. These peculiarities in the harmonic content of instrument sound obviously affect the quality of tones.

1.3.1.3 Timbre and formant structure

It is important to discuss the formant structure of sound, especially for musical instrument sounds. Formants are defined as frequency regions of the spectra of higher energies that are virtually independent of the pitch. Saldanha [SC64] mentioned that:

> [...] *it is believed that the strengthening of the partials in the formant region is due to the resonance of some part of the musical instrument being played or to the resonance of the body of air enclosed within the instrument.*

For musical signals, the formant theory [Fle34] complements the classical theory of Helmholtz and holds that:

> [...] *the characteristic tone quality of an instrument is due to the relative strengthening of whatever partials lie within a fixed or relatively fixed region of the musical scale.*

In [RW82], it is further noted that:

> *Sound spectrograms suggested that, for a given intensity, the spectrum has a formant structure; that is, it varies with frequency so as to keep a roughly invariant spectral envelope.*

In the same vein, Schumann [Sch29] propounded several laws of perception of wind instrument timbre in relation with the stream segregation and Gestalt principles. His theory served as the basic foundations for explaining the production of wind instrument timbre through their typical formant areas. Reuter [Reu97] recalled the following:

> *The principle of formant areas (*Formantstreckengesetz*) postulates that the formants of musical instruments are fixed areas of the spectrum which are independent of the fundamental frequency.*

> *The principle of formant intervals (*Formantintervallgesetz*) states that the partial with the strongest amplitude located in the first (or lowest) formant area and the partial with the strongest amplitude located in the second (or higher) formant area result in an interval which is typical for the respective instrument.*

Reuter [Reu97] outlined that the validity of these principles have been experimentally confirmed for reed- and double-reed instruments, and that they seem to be valid for brass instrument timbres. On the other hand, he mentioned that these principles could not be extrapolated to the case of string instruments. However, Benade [Ben76] showed that violin sounds exhibit sharp spectral resonances corresponding to the combined influences of violin air and wood resonances. He stated that:

> *This overall curve* [i.e. the loudness curve or the spectral envelope] *has an interpretation that is very similar to that for the vocal-tract curves* [...]

This corroborates what is known for speech signals. Formants are created by the natural resonant frequencies of the vocal tract or resonating body. Further, it is known that voiced sounds are uniquely characterised by their formant structures and that the first three formant frequencies are usually sufficient to identify vowels.

In music, the vocal tract contribution can be transposed to the guitar body or the tube of a brass instrument, for example. It can then be argued that modelling the formant structure of an instrument can serve to uniquely characterise it. This also means that formant structures can be used as descriptors in musical instrument identification systems.

The principles governing the *formant theory of musical instruments* constitute the foundations of our research. In this thesis, the *formant theory of musical instruments* is also termed as the *formant theory of timbre*.

For periodic tones, timbre depends upon spectrum. However, instruments can still be recognised from very poor recordings and at different sampling rates, thus refuting the theory stating that spectral envelope is the unique timbre correlate. Furthermore, room acoustic and equalisation theories teach that a reverberant room can affect the spectral envelope of sound at levels as high as 20 dB for some frequencies. This room transfer function also depends upon the physical location of the listener. In other words, although the spectral information *received* by each listener is different, the identity of the sound can still be perfectly retrieved by them.

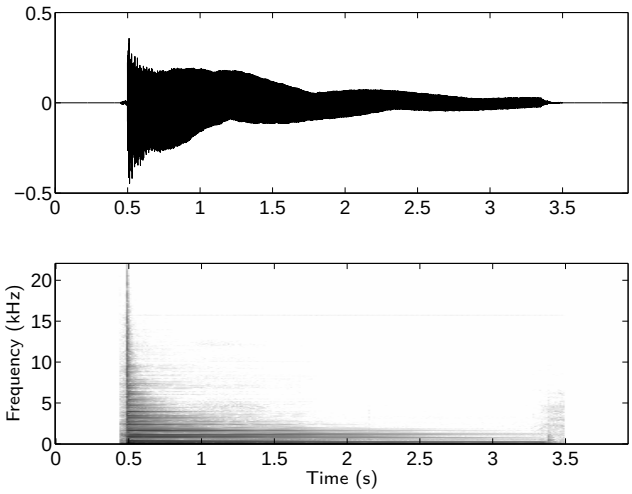


Figure 1.10: Plucked guitar string waveform (top) and corresponding spectrogram (bottom).

1.3.1.4 Timbre and time cues

Figure 1.10 shows the attack and decay of a plucked guitar string. The plucking action gives it a sudden attack characterised by a rapid rise to its peak amplitude. The decay is long and gradual by comparison. The ear is sensitive to these attacks and uses them to identify the instrument producing the sound. In [SC64] and [Ber63], it has been experimentally shown that the onset of a sound is psycho-acoustically important. Removing the initial segment of notes played by various instruments impairs the correct recognition of the instruments. Further, transients are to a certain extent, unique. For instance, the transient of a low pitch piano sound cannot be *cut* and transposed to synthesise a higher pitch piano note. This particularity is similar to the *coarticulation* phenomenon encountered in speech. It is therefore difficult to properly separate transients from steady-state portions of sound as to model transients for sound synthesis purposes.

In his experiments, Saldanha [SC64] evaluated the relative importance of harmonic structure, frequency and vibrato as timbre cues in the absolute judgement of musical tones. He showed that for ten instruments in the database, the average correct identification rates dropped from 47% to 32% when initial transients have been removed from the samples. More detail about several experimental studies assessing human ability for the task to identify musical instruments will be given in section 2.1.

Other time-related timbre correlates are concerned with the variations in amplitude

and pitch of tones as a function of time, the tremolo and the vibrato. If the harmonic content of a sustained sound from a voice or wind instrument is reproduced precisely, the ear can readily detect the difference in timbre because of the absence of vibrato. For instance, Saldanha [SC64] conducted experiments involving sounds played with and without vibrato and showed that the presence of vibrato helped to increase the correct identification by 3%. Vibrato and tremolo are present to some extent in voice or musical instrument sounds. However, in most cases, they depend more on the playing style and can rarely be seen as an intrinsic property of an instrument.

1.3.1.5 Timbre, pitch and loudness

Musical instruments have a *natural* frequency range that can be defined by the range of notes that are played in *realistic* conditions. Many instruments have different registers and it is known, for example, that the clarinet has a different timbre whether it is played on the low or high pitch range. Playing a tone at a higher pitch is not a simple matter of shifting a spectral envelope up to the higher frequencies and involves more complex phenomena. In the same vein, an instrument tone played very loudly will sound different than a soft version of the same tone. For instance, producing a very low pitch note on a wind instrument with the same loudness as a higher pitch tone can be achieved if the musician blows stronger, thus providing more energy to the system. This affects the *mechanical response* of the instrument and therefore the corresponding structure of the sound.

Timbre dependence upon pitch and loudness are fundamental attributes that need to be taken into account when one wishes to build automated systems for recognising musical instruments. These aspects will be covered more in depth in chapters 5 and 6.

1.3.2 Timbre space representation

Research in timbre space representation directly results from the characterisation of the physical correlates of timbre. Due to the multi-dimensional nature of timbre, it is of great interest to represent sound objects in a lower-dimensional space whose principal components can be labelled in a perceptually meaningful way. This suggests that by pointing out a point in this timbre space, one could hear the timbre corresponding to these coordinates and subsequently moves continuously on a graded and labelled timbre scale.

Probably the first to study such representation was Wessel [Wes78] who derived a 2-dimensional timbre space from perceptual data for *compositional control of timbre*. Using multi-dimensional scaling algorithms (MDS), each sound object is represented by a point in a geometric representation generated from input data. In essence, the technique consists of gathering perceptual data corresponding to all pairs in a set of stimuli and then to select the directions that best fit these data. In particular, Wessel observed a clear correlation between the width of a sound's spectrum and one axis of his timbre space. This led him to define *brightness* as a measurement of the energy distribution among the sound harmonics. As Wessel has shown, this dimension is the one that can best articulate stream segregation [Bre90]. The other axes of the space were either related to the attack rate or to the extent of synchronicity among the various components.

More recently, attempts to build timbre space representation using Principal Component Analysis (PCA) have been reported in [dPLY04] and [CHM97]. In [dPLY04], the timbre palette produced by a single instrument over the whole pitch range is studied. The work is restricted to the quasi-stationary part of clarinet sounds, excluding the attack, decay and transitions between consecutive notes. The timbre space is built by considering three principal components of the amplitude curves obtained from a sinusoidal analysis stage [MQ86]. It is shown that timbre classes are a function of the intensity and that the lower register of the clarinet exhibits in general much more richness in timbre variation than the higher register. However, it has been found difficult to relate the principal components other than the first one to any perceptual judgements of feeling. Furthermore, the universality of the approach and in particular its extension to deal with several musical instrument sounds are not guaranteed.

This type of application decidedly provides efficient ways to validate the pertinence of timbre models. For instance, the partitioning of a 2 or 3 dimensional Euclidean space regarding the perceptual contrast between sound objects or sound families would denote a certain success in the approach. The extension to the musical instrument identification problem would be rather straightforward. However, in practice, these methods have shown little ability to be extended to tackle musical instrument identification problems.

1.3.3 Timbre perception

To summarise, little is known about the precise physical characteristics that tend to contribute to the perception of timbre. Tackling the problem from a different

perspective, the quality of a sound can often be described in terms of the feeling and sensations they induce. And what better way have humans found to describe their feelings that by words.

Early in 1877, and from a radically experimental point of view, Helmholtz [Hel54] extensively used verbal attributes to describe his research. He stated, for example that "the peculiar quality of tone commonly termed *powerty* as opposed to *richness*, arises from the upper partials being comparatively too strong for the prime tone." Further, his research and observations led him to conclude that: "simple tones [...] have a very soft, pleasant sound, free from all roughness [...] and dull at low pitches", and that "musical tones [...] are more harmonious [...]. Compared with simple tones they are rich and splendid, while they are at the same time perfectly sweet and soft if the higher upper partials are absent."

Later in the seventies, von Bismarck [vB74] attempted to extract from the timbre percept those independent features which can be described in terms of verbal attributes. In his experiments, listeners were asked to rate 35 artificial sounds on 30 scales whose endpoints were characterised by pairs of opposite attributes such as dark–bright or smooth–rough. The conclusion was that 4 scales were considered nearly sufficient to describe timbre. In particular, he found that the scale dull–sharp was most preferred among all listeners.

Chapter summary

This chapter introduced the basic concepts linking acoustic and physical signals to human perception. Previous research showed that the ear does not have the same response for all sounds. The non-linear properties of the ear frequency response have been highlighted. The concepts of absolute threshold of hearing, frequency and temporal masking, that are used in numerous applications in audio compression and audio signal analysis, have been introduced.

Rather than providing a thorough review of the psycho-acoustic principles driving the perception of sounds by humans, the first part of this chapter introduced the fundamental and basic notions that motivate the design of psycho-acoustic models in digital audio signal processing. An example of such a model is given in appendix A.

Further, the mechanisms through which the pitch of a complex stimulus is perceived have been discussed. The different theories of pitch processing by humans helped us to introduce the two most influential theories in audio perception. Their

fundamental principles have been outlined in section 1.2.

In section 1.3.1, the acoustic and perceptual properties of timbre have been discussed. The concept of timbre correlates which is important for understanding why a particular feature or descriptor will be used throughout this work has been introduced. In particular, principles of the *formant theory* of timbre, which is the main theme of this research, have been described. Finally, a distinction between spectral and temporal attributes of timbre has been made.

The next chapter is concerned with automatic musical instrument identification. After having summarised experimental results assessing human ability to identify and recognise musical instrument tones, a description of techniques and systems encountered in the literature is given. An approach to tackling the problem of modelling the timbre of an instrument is then proposed.

2. Musical instrument identification

Musical instrument identification systems are used to classify sounds according to salient properties of the signal. These systems attempt to reproduce how humans can recognise and identify the sounds populating their environment. Indeed most people are able to discern a piano tone from a violin tone or able to distinguish different voices or singers.

It is important to point out here the importance of learning in the identification of sounds by humans. Take the set of musical instruments described in figure 2.1. Although the distinction between string and wind instrument sounds can be easily made by most human beings, the ability to distinguish between single and double-reed instrument tones, or a violin and a cello tones, involves a certain form of learning. That is to say: like humans do, machines learn.

This chapter is concerned with musical instrument identification. In section 2.1, several perceptual experiments on the perception of timbre by humans are summarised. They will serve to define a *ground-truth* in terms of performance for evaluating computer systems. In particular, the relative importance of temporal cues in the identification of sounds is made apparent.

Next, principles of computer models and algorithms for musical instrument identification are introduced in section 2.2. System evaluation protocols, limitations and applications of these algorithms in wider MIR frameworks are discussed. Several approaches and systems encountered in the literature for the purpose of identifying and classifying musical instrument sounds are described in section 2.3.

In section 2.4, an approach to tackle the problem of modelling timbre is proposed. A system is described. It is suggested that it preserves the coherence between physical and perceptual aspects of timbre. In particular, it is suggested that it conforms to the principles governing the *formant theory of timbre*.

2.1 How well do humans perform?

Musical instrument identification systems attempt to reproduce how humans can recognise and identify the musical sounds populating their environment. When designing such systems, it is important to directly compare its performance with humans, and this for equivalent tasks and under similar conditions. This procedure is widespread in most MIR applications and more generally in perceptually related system design as very often, human performance are the only *baseline* performance available.[1] It is a fact that in speaker verification, for example, modern computer systems outperform human at recognising voices recorded in good acoustic conditions. On the other hand, humans are still much better than computers at separating sources from a mixture (e.g. the *cocktail-party* problem [Che57]).

As far as the identification of orchestral musical instrument is concerned, several experiments have been conducted throughout the years. Due to the nature of the considered sounds and the difficulty for ordinary people to distinguish between all orchestral musical instruments, experiments often involve the participation of experienced listeners such as musicians, composers or music students.

Berger [Ber63] focused on the relative importance of transient/onset and steady-state portions of sound in the identification of isolated tones produced by wind instruments. Martin [Mar99] conducted experiments of identification of isolated notes among 27 instruments to validate his computer models. More recently, Srinavasan [SSF02] carried out several experiments using sets of instruments that have been used to evaluate computer models in other research works.

In the following, the most relevant studies for our research are summarised. The conclusions reached thereafter suggest that in many cases, human identification rates are lower than expected.

2.1.1 Experiments involving isolated tones

In [SSF02], experiments have been conducted using isolated notes extracted from the McGill sound database [OW87]. The study consisted of evaluating the performance of 88 experienced listeners at recognising orchestral musical instruments notes out of musical context. The importance of a training session prior to the classification has also been evaluated. Performance were compared to similar experimental studies available in the literature ([SC67], [MK98]) and to computer systems using identical sets of instruments. In figure 2.2, the results of various experiments involving

[1] and is more precisely what systems are intended to be compared to

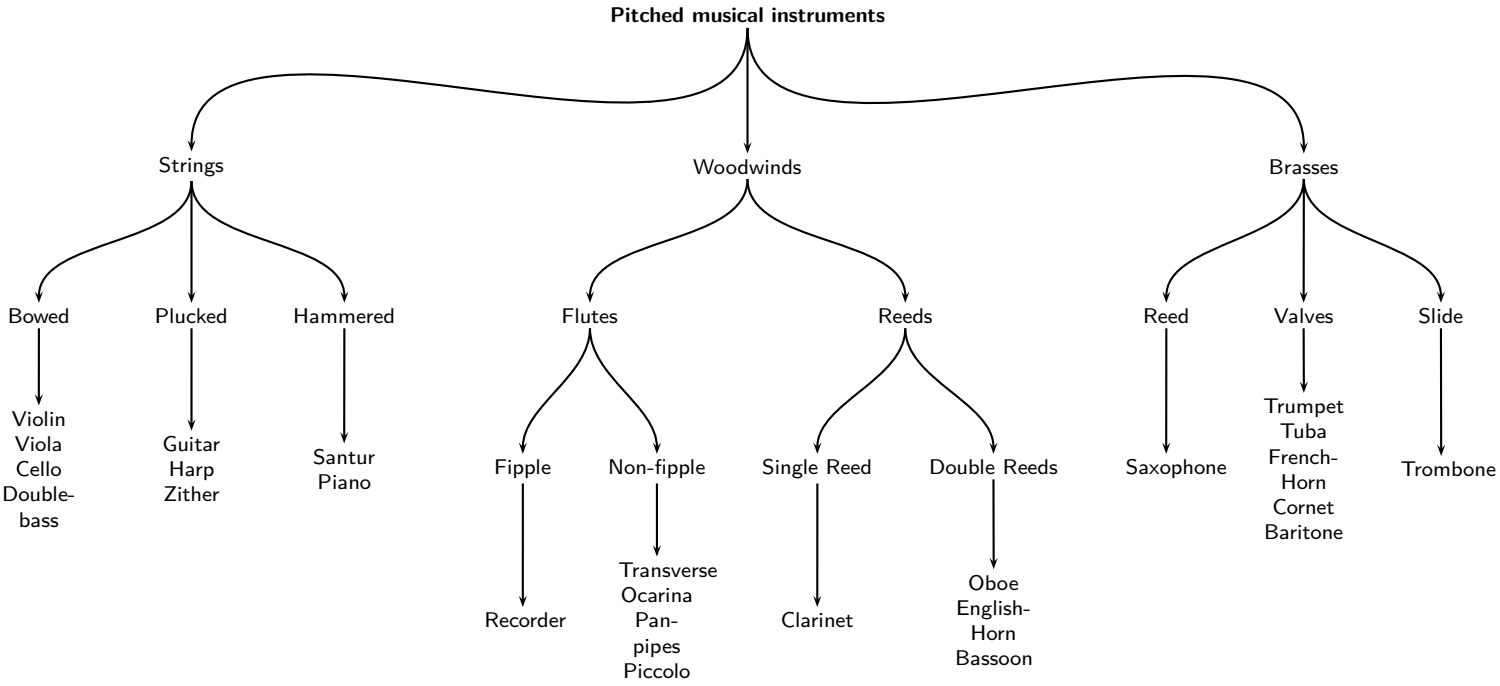


Figure 2.1: The classical taxonomic classification of pitched musical instruments.

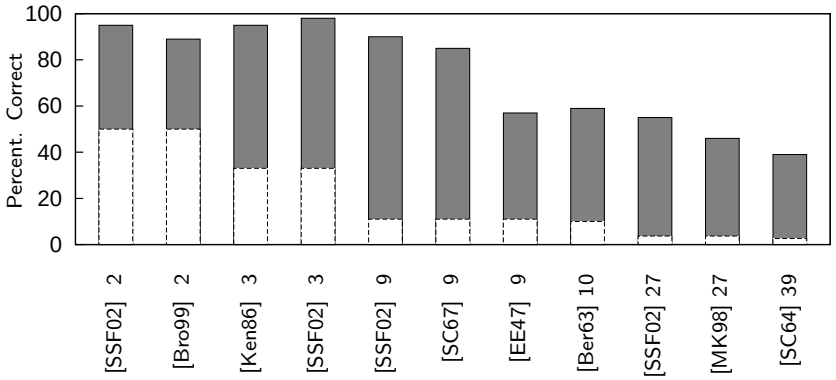


Figure 2.2: Comparative performances in terms of average correct identification rates for several experimental studies. In abscissa are given the bibliographical reference for each system as well as the number of instruments retained for the experiments. White boxes correpond to the average identification rates that would be obtained by random guessing. Adapted from [SSF02].

different types and number of instruments are shown. Although number to number comparison might not be relevant here, it can be noticed that human performance decreases as the number of instruments in the database increases. Next, for equal number of instruments, the listeners having participated in the experiments in [SSF02] performed better than in previous experiments. Consequently, we will consider these performances as baseline performance for the evaluation of the system that will be presented in this thesis.

2.1.1.1 Instrument identification

Tests were performed with sets of 2, 3, 9 and 27 instruments respectively by presenting segments of 4–7s in duration. For each instrument, the entire available pitch range has been used. The listeners were asked to identify isolated notes and had to give an answer even if they did not recognise the instrument. Overall, for the experiments involving 2, 3, 9 and 27 instruments, the average correct identification rates were 94.5%, 97.6%, 90.2% and 55.7% respectively.

In table 2.1 is reported the confusion matrix for the experiment involving 9 instruments. Correct identification rates range from 99% (flute) to 83% (trumpet). Most important confusions concern the violin and cello tones. This type of confusion is expected since both instruments differ only by size and have very similar timbres. Next, trumpet and sax tones, which both belong to the brass family, are confused 8% of the time. Finally, 6% of the bassoon samples were mis-identified as being trombone. This is slightly more surprising since bassoon is a double-reed excited wood-wind instru-

	flute	oboe	clarinet	bassoon	sax	trumpet	trombone	violin	cello	n.c.
flute	**99**									1
oboe		**92**	3				3			2
clarinet		4	**87**	4			5			
bassoon		2	2	**84**		3	6			3
sax		2			**97**					1
trumpet				5	8	**83**	2			2
trombone			4				**94**			2
violin								**84**	17	-1
cello								9	**91**	

Table 2.1: Human performance for the task of recognising isolated notes amongst 9 instruments. Presented instruments are shown in rows while answers given by the listeners are shown in columns. The correct identification rates are reported in percentages, the column *n.c* represents the difference to 100. From [SSF02].

ment while trombone is a slide brass instrument. Contrary to what could have been expected, experienced musicians show difficulties at recognising instruments tones out of context.

2.1.1.2 Instrument family identification

When evaluating humans and computer system performance, it is also interesting to analyse the correct inter-families identification or confusion rates as there is often a tendency for the error of identification for a given instrument to be made in terms of other instruments within the same group. In table 2.2, two typical confusion matrices for two different experiments involving 27 different instruments are summarised. A first interesting point is that very similar trend can be observed for both experiments: firstly, tones belonging to the bowed strings and flute families are very well identified (in both experiments, each family is correctly identified more than 98% of the time) and rarely confused with other families of instruments. Secondly, there appears to have been a persistent mis-identification of reed instruments tones whose confusions rates are scattered throughout the other instruments. In particular, note the difficulty to distinguish saxes from reed and brass sounds respectively. Note also that only 65% of the sax sounds were correctly identified.

Experiments involving isolated notes correspond to the bottom line of human performance evaluation as tones are presented out of any musical context. It has been previously discussed that musical context can provide essential clues for the correct

	strings	brasses	dble reed	clarinet	flutes	saxes
strings	**99**					
brasses		**90**	4	2		3
db reed		4	**76**	7		12
clarinet		2	12	**71**		14
flutes					**99**	
saxes		5	13	15		**65**

(a) [SSF02]

	strings	brasses	dble reed	clarinet	flutes	saxes
strings	**98**					
brasses		**86**	8	3		2
db reed	2	6	**74**	11	3	6
clarinet		4	13	**73**		11
flutes					**99**	
saxes	–	–	–	–	–	–

(b) [MK98]

Table 2.2: Confusion matrices corresponding to instrument families identification experiments. In rows are shown the instrument families while in columns are shown the answers given by the listeners.

identification of instruments. However, under such controlled acoustic environment, the effect of extraneous timbre correlates are attenuated, thus resulting in a better picture regarding the contribution of each of these correlates to the *whole*. Note that experimental evaluations of our system using a database isolated notes will be reported in chapter 5.

2.1.2 The importance of transient information

Berger [Ber63] investigated the ability of 30 music students at identifying wind instrument tones having, among other alterations, the first and last half seconds removed. The corresponding confusion matrix is presented in table 2.3. A similar experiment using non-altered entire tones has also been conducted and results in 59% correct identification for similar instruments.[2] The following conclusions can be drawn. First, with the only knowledge of the sustained and quasi-periodic portion of sounds, the flute tones are almost never correctly identified (1 out of 30) as opposed to 13 out of 30 for non-altered samples.[3] More importantly, they are confused with instruments belonging to the brass family for 20 subjects out of 30. Second, oboe tones have been correctly identified by 28 out of 30 students. This can be explained by the fact that the oboe is the only double-reed instrument in the database and exhibits a characteristic and harmonically rich penetrating tone. Next, one could intuitively expect the tenor and alto saxes to have similar timbres, so that confusions would equally spread between both of them. However, whereas tenor sax tones were mostly confused with

[2] note that these performance are worse than the ones reported in [SSF02]

[3] note the listeners low performance in Berger's experiment compared to the one reported by Srinavasan (see table 2.1)

clarinet for 25 listeners, the alto sax tones were confused with the french horn for 11 students. The brass instruments were overall mainly confused with instruments from the same group. Finally, the alto sax tones attracted most confusions, just as if these tones corresponded to an *average* timbre of the considered set of wind instruments. Contrasting with the 59% correct identification for non-altered samples, the tones with missing starting and ending portions produced a mean listener score of 35% correct.

By comparing these figures to the ones reported in Srinivasan's experiment, the conclusion that correct identification rates are considerably higher when onsets/transient are present can be drawn. This confirms that essential cues for the correct identification of musical instrument sounds are conveyed by the attack and to a certain extent by the end of tones.

2.2 Computer models

The increasing available computer power for both research and end-user applications highly contributes to the growth of MIR related systems design. Techniques which required extensive computational power such as machine learning algorithms, for example, can now run in acceptable and *human* time-scales. Grey [Gre77] decidedly pointed out three decades ago that

> [...] *these recent improvements in timbre research are largely the result of technological advances in the use of digital computers* [...]

This is remarkably valid today. He continued by stating

> [...] *which has given the investigator powerful new means for the analysis and synthesis of complex, time-variant musical instrument tones* [...], *and for the analysis and presentation of complex, multi-dimensional data structures of the type that may be collected from studying the perception of timbre.*

However, the problem of analysing the human timbre perception mechanisms has been transposed to the problem of automatically recognising musical instruments, ideally in complex mixtures and under a wide variety of acoustic conditions. Modern research considers the timbre correlates as *features* or *descriptors* that are fed into a machine learning algorithm for building an instrument model.

	flute	oboe	clarinet	sax tenor	sax alto	trumpet	cornet	french horn	baritone	trombone	n.a
flute	**1(13)**	2		1	6	5	4			4	7
oboe		**28(29)**									2
clarinet	1	1	**20(27)**	4	3						1
tenor sax			25	**2(24)**	1						2
alto sax				3	**4(9)**		1	11	5	5	1
trumpet	8				6	**2(11)**	3	4	1	3	3
cornet		1				12	**15(23)**				2
french horn	1			2	3			**5(12)**	6	6	7
baritone			1	1	2	3	2	4	**7(16)**	3	7
trombone	2	1		5	3			1	5	**9(14)**	4

Table 2.3: Confusion matrix in Berger's experiment involving tones having their starting and ending portions removed [Ber63]. Performance for non-altered instrument tones are reported in brackets. In rows are shown instrument names while in columns are shown the answers given by the listeners.

A look at the approaches and techniques encountered in the literature reveals that researchers in the area investigate several techniques and their possible combination to achieve this goal. Incidentally, this plethora of systems sometimes makes the problem more complex for the researchers themselves to tackle. Firstly, in terms of reproducibility, due to the lack of common databases and the differences in the experimental protocols,[4] it is difficult to objectively compare performance between systems. Next, in terms of complexity, certain algorithms sometimes involve a tremendous amount of processing levels (in terms of features or statistical classifiers) thus increasing the difficulty to reproduce a system's behaviour. Finally, in terms of the *a-posteriori* interpretation of the results. This is especially valid if one wishes to relate the performance of a system to the concept of timbre. The route that is often taken in the field is oriented towards systems that perform well as opposed to study the perception of timbre and its correlates through the design of computer models.

2.2.1 But "can one hear the shape of a drum?"

The task of building systems able to identify musical instruments is embodied in a wider and fundamental problem in science: the inverse problem. In 1966, Mark Kac asked [Kac66] the following question "Can one hear the shape of a drum?". Beyond this question lays an essential question about the bijectivity between the mechanisms of sound production and the observation of the sound spectrum that can be output from a microphone. The problem can be stated as follows: given the frequency representation of a drum sound, would it be possible to infer the shape of the drum at the origin of this sound? The problem can be legitimately transposed in our case to: given a frequency representation of a musical instrument sound, would it be possible to infer the physical mechanisms that yielded the creation of this sound? A solution, even partial, would assuringly help us to solve the musical instrument identification problem as characterising the timbre of an instrument corresponds to a certain extent to characterise the physical mechanisms of this instrument.

2.2.2 An ill-posed problem and its algorithmic approximations

We shall describe here the underlying scope in which research in musical instrument identification lies. The two concurrent approaches can be considered:

- **Timbre model**: this approach find its direct origins in the study of the mechanism of perception of timbre by humans and can be seen as a direct algorithmic

[4] such as the length of testing samples or the number of instrument identities, for example

transposition. In essence, the multi-dimensional nature of timbre is transposed into a multi-feature, multi-descriptor based system which represents the sensory information received by the ear and processed at the early stages of the perception chain. This multitude of information is then recombined using a machine learning algorithm whose task is to mimic to a certain extent the remaining process yielding the perception of timbre.

- **Instrument model**: indeed, the physical mechanisms yielding the production of sound by musical instruments are different by nature. Plucking a string, blowing in a saxophone mouth-piece or pressing a piano key involve different physical principles. There also exists a taxonomy (see figure 2.1, for example) of such physical mechanisms based on their likely similarities that can lead to instrument classification. These particularities are undoubtedly carried by the signal and at the same time independent of considerations in terms of human perception and sensitivity.

Depending on the chosen approach, the methodologies and techniques that are considered are different. The instrument and physical acoustic modelling approaches are often too complex to solve so that timbre modelling systems are often preferred. However, knowledge in physical acoustics and of the mechanisms of sound production by instruments can bring essential clues for the pertinent choice of features to build the models from. For instance, several modern techniques in digital signal processing can be related to a semi-physical representation of the mechanisms of sound production. The linear predictive model, for example, is widely used to model the mechanisms of speech production. Its application for musical instrument identification purposes will be investigated in chapter 3.

In the more general context of timbre or texture representation of sound, it can be argued that the choice between these techniques is more influenced by the types of sounds considered than being solely dependent upon a desired research orientation. All musical sounds have a timbre. In particular, modern music makes extensive use of electronically generated audio textures, mimicking or not the timbre of *acoustic* instruments. In this case, it becomes difficult to relate a signal with the physical mechanisms involved during its creation.[5] The classical timbre modelling approach is therefore more appropriate to deal with such musical sounds. A similar reasoning can be applied in the case where audio effects are applied to an instrument sound. An artificial reverberation effect modifies the signal waveform and spectral characteristics

[5] although computer music uses physical modelling principles such as oscillators, modulators, etc

in a way that it can become impossible to trace back its physical origins. At another extreme, a distortion effect applied to an electric guitar sound radically affects the nature of the observation so that almost none of the *original* characteristics of the sound are preserved.

Amid the various justifications accompanying the choice of either of the two approaches, questions about the pertinence of using an algorithm to model complex perception mechanisms can be raised. In other words, how difficult is it to *a-posteriori* trace back or infer an instrument identity from a multi-layered system biased at several levels? In essence, even if each descriptor is independently meaningful for representing one aspect of timbral information, to which extent is it guaranteed that the combination of several of these descriptors with such or such machine learning algorithm will result in an *optimum* timbre model? For instance, it can be shown that efficient systems are in fact optimised both at the feature and classifier levels. In the same vein, the question of whether concatenating feature data with different physical units into a single feature vector is more appropriate than independently building the models for each feature data can be raised.

2.2.3 Systems evaluation

Systems are usually evaluated using sets of audio excerpts representing the considered classes of instrument. Databases include recordings from various acoustic environments, different brands of instruments and several playing styles. However, these precautions does not necessary yield the definition of a ground-truth and an optimum system in terms of performance. Each investigator evaluates his system with the data sets at his disposal. As a consequence, different systems and their performance have to be compared with care as most of the time, the corresponding instruments differ from one experiment to another. Moreover, the methodologies used for the evaluation procedures are often different.

Amid the generally acceptable system competences, an ideal algorithm would have the following characteristics:

- **Generalisation**: generalisation is concerned with the fact that given a system trained using a specific dataset, the performance of the algorithm using any subsequent unknown data will statistically correspond to what have been previously achieved. In other words, it is desirable that the experiments that can be conducted closely represents what could be achieved with any other dataset.

- **Robustness**: an ideal system should recognise different instances of the same sound as emanating from the same source. Different instances comprises different recordings conditions, different playing styles and to a certain extent different pitches and qualities.

- **Meaningful behaviour**: a system is expected to behave in an *understandable* way and as close to human performance as possible, especially in terms of confusions between instruments or families of instruments.

- **Reasonable computational requirements**: likewise such systems can be easily included as a module into wider MIR frameworks without significantly affecting the responsiveness and latency of the overall structure. Note that the computational requirements needed for training the models are not critical since this operation is often performed off-line.

- **Modularity**: it should be relatively easy to update the models with new samples that have been correctly recognised or taken from a complementary database, and this at any given time. Likewise, it is is expected that new instruments can be added to the system without retraining all the models in the database.

2.2.4 Limitations

Ideally, we would like to be able to characterise and recognise all the instruments or sources in poly-timbral mixtures and complex music pieces. Moreover, just as humans can do, an ideal system should be able to retrieve the number of *sound objects* present in the signal, their identities[6], as well as to segment in time a piece of music in accordance with the presence or not of each source. Figure 2.3 illustrates what an ideal musical instrument identification system would achieve.

However, such problem is complex and difficult to tackle. In contrast to the other components of sounds, timbre is not carried by a salient and identifiable acoustic property of a signal. Further, in contrast to rhythm, which can be related to the organisation in time of series of energy bursts[7], or pitch, harmony and melody which can be directly related to the signal's frequency composition, timbre can only be related to a multitude of physical characteristics whose precise contributions to the *whole* are unclear.

[6] note that humans are still able to discern between sound objects, even if their labels or identities are not precisely known

[7] by rhythm, we mean the calculation of a beat/tempo, knowing that the problem of retrieving the rhythm for non-percussive pieces is much more complex to tackle

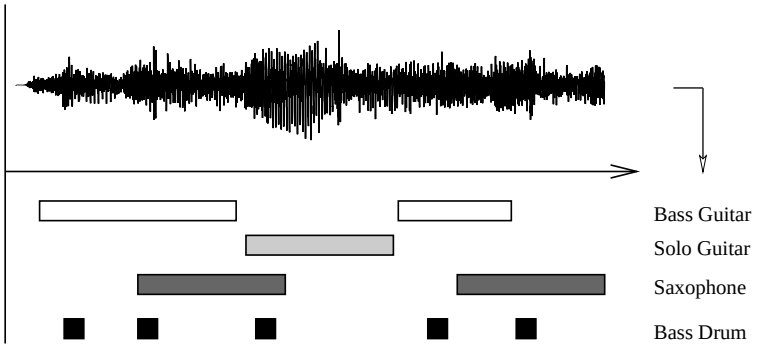


Figure 2.3: An ideal musical instrument identification system.

A piece a music can be seen as a mixture of several mono-timbral signals. The problem of separating a mixture into the original sources from its only consideration is difficult to tackle. Several approaches and techniques are dedicated to the source separation problem. However, in the case of a single observation (also known as one microphone source separation), only techniques setting priors on the signal composition and structure have been shown to give exploitable results. In the case of stereo signals, DUET-type algorithms [YR04] can be successfully employed under the constraint of dealing with linear instantaneous mixtures. However, these techniques introduce artifacts[8] in the separated signal thus limiting their use as a front-end to a musical instrument identification system. In figure 2.4 is depicted the diagram of a complete system using a source separation algorithm as front-end.

For these reasons, current research for musical instrument identification in polyphonic mixtures tends generally to avoid the use of source separation algorithms. Several systems are reported to use feature binary masks [EB04], or simply to build models of instrument mixtures. In this case, each considered class and their corresponding models consists of a mixture of timbres. Although their use in concrete applications is possible, one should note that systems' complexity is growing exponentially with the number of instruments since all possible combinations of pairs, triplets, or more of instruments have to be considered. More specifically, these techniques can address problems of classifying jazz trios, quartets and other small musical formations [ERDa].

Finally, a common point for almost all supervised systems is that it is impractical to

[8] research in the field focuses mainly on the increase of intelligibility rather than on the conservation of the timbral information

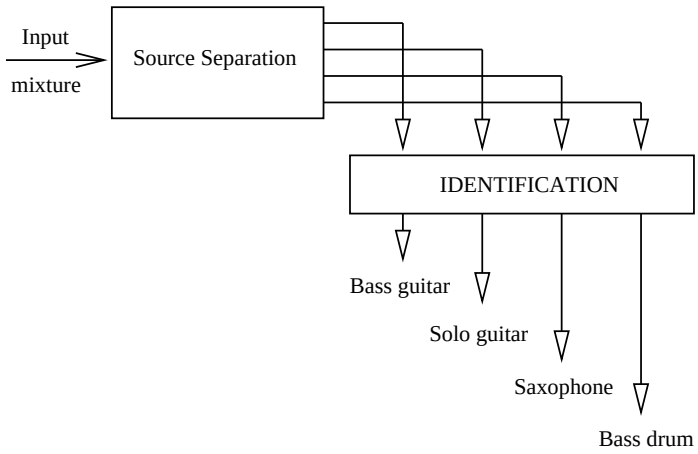


Figure 2.4: Musical instrument identification system using a source separation module as front-end.

build models for all available instruments or all available timbres in order to recognise any possible components in any random musical piece. Again, in contrast to pitch or rhythm that are independent of the nature of the source, timbre characterises the source itself. This intrinsically sets the limit on the universality of the supervised approach.

2.2.5 Applications

Automatic audio source identification systems can provide interesting modules for MIR systems. In terms of classification, they have great potential for use in database search interfaces: retrieving pieces of music containing a given instrument, looking for similarities in different musical extracts or classifying pop songs in accordance with the lead singer voice are few examples of possible applications.

The standardised MPEG-7 musical description format [Cas02] defines a number of competences that a system should have for describing any kind of sound file or sound stream. Through the use of *metadata* or *tags*, the content of sound can be described in a *humanly readable* way. These informations can then be used to manipulate, organise, classify and retrieve sounds from databases. One of these applications is directly related with timbre. For example, it is desirable to automatically determine if at a given time, a given instrument is played or not. In a more general framework, web-based search engines can be developed to retrieve *music that sounds the same* or to recommend users with similar music in a given style. Metadata-based applications

can be used for multitude of purposes, provided that the *tags* can be extracted in an automated and robust manner.

Another important aspect is concerned with assisted data labelling. MIR systems are often prone to errors. By using user feedback, a musical instrument identification system could be used to assist the user during a hand-labelling operation of songs and music pieces.

Musical instrument identification systems are in essence very similar to the ones that have been developed for speaker verification purposes. A direct extension of this work is concerned with artist classification using singing voice information. Examples of such approaches can be found in [KW02] and [OPGB05] using source-adapted source separation algorithms.

For compression and low-rate object-oriented coding purposes, one could think about designing a system that could automatically adapt a compression algorithm to the timbral nature of the signal being played. Furthermore, under the condition of having developed a timbre model, high-level object coding algorithms could allow encoding high-level musical information, such as pitch, loudness and timbre, which are then reconstructed at the receiver end.

Beyond the design of musical instrument identification systems, numerous applications in audio related fields can be envisaged. As artistic and compositional tools, such systems can be used to evaluate similarities between sounds in order to drive automatic accompaniment systems for live electro-acoustic and electronic performances. In this particular case, the process is not aimed at identifying instruments precisely but to classify sounds by groups according to their timbral similarities.

2.3 Existing approaches

Automated systems for identifying musical instruments started to be developed in the late nineties. Since then, a great deal of research has been carried out on the topic.

Most of the techniques presented here belong to the class of *supervised learning*, where the purpose is to derive rules from an existing labelled dataset to classify one unknown sample. In essence, the identities of the instruments as well as some characteristic sounds are known *a-priori*. The aim is to infer from the available pairs labels/sounds a mathematical relationship whereby one unknown sample, after the identification process, will be assigned a label taken from the database.

In table 2.4, typical performance for six systems are reported. As can be drawn from their analyses, problems arise when one tries to compare the percentages. Con-

System		Correct Identification	Number of instruments
Marques	[MM99]	70.0%	8 (0.2 seconds test)
		83.0%	8 (2 seconds test)
Brown	[Bro99]	94.0%	2
Brown	[BHM01]	84.0%	4
Martin	[MK98]	71.6%	14
Agostini	[ALP01]	92.8%	27
		95.3%	20
Essid	[ERD04]	79.1%	10

Table 2.4: Recognition performance for six systems using mono-timbral excerpts.

sidered instruments, their number and the duration of the audio samples used for the training and testing are usually different. Implementation of the feature extraction function and/or the classifier may also differ from one system to another. This obviously affects the objectivity of number to number comparisons.

However, general trends in the system's behaviours can be exploited and used for the design of new classification algorithms. Cross-comparison with experimental studies on timbre perception, both in terms of correct instrument and family identification rates also contributes to evaluate the performance of a system.

For clarity purposes, the existing approaches will be grouped depending on their orientation. As a result, descriptor-based modelling and instrument modelling techniques will be distinguished. State of the art systems in mono-timbral instrument recognition use solo recordings of sustained isolated notes or excerpts extracted from *melodic* phrases.

2.3.1 Descriptor-based modelling

Descriptor-based modelling systems include a feature extraction module. This stage is concerned with the choice and calculation of relevant acoustic features used to model the timbre of a sound. These features are then fed into a machine learning algorithm to obtain a condensed and representative model of each class. The types of extracted features clearly reflect the conclusions drawn from studies on timbre perception. Within this type of algorithms, one can encounter mono-feature systems, multi-feature systems and data-mining approaches.

2.3.1.1 Mono-feature systems

In mono-feature systems, one single type of feature is retained to build the instrument models. Although it can be argued that timbre cannot be efficiently modelled by one type of feature, these systems allow a better understanding of the interaction feature/classifier.

Brown [Bro99] used speaker recognition techniques for classifying between oboe and saxophone. Using cepstral coefficients based on a constant-Q transform, 94% correct identification was reported. In a later study [BHM01], her system classified between oboe, saxophone, flute and clarinet. The most successful feature set was the frequency derivative of 22 constant-Q coefficients measuring the spectral smoothness. A performance of 84% correct identification was reported using a standard GMM classifier. In [MM99], Marques described a system capable of recognising between 8 different instruments (bagpipes, clarinet, flute, harpsichord, organ, piano, trombone and violin). Using 16 Mel Frequency Cepstral Coefficients (MFCC) and a Support Vector Machine (SVM) as classifier, 70% correct identification was reported for 0.2 second test samples and 83% for 2 seconds audio samples. Krishna [KS04] studied the use of a particular set of linear predictive coefficients, the Line Spectrum Frequencies (LSF).[9] Using a mixture of 54 Gaussians, performance of 87.3% can be achieved for the classification of one excerpt among 14 instruments. It has been further shown that the LSF performed better than the MFCC for a similar task. Eggink [EB03] first evaluated the performance of a technique designed to identify instruments in artificial poly-timbral mixtures. Prior to feature extraction, the fundamental frequency f_0 is calculated. A binary mask is then determined to select spectral descriptors based on the overtones frequencies. With this system, average instrument identification for tones extracted from the McGill database [OW87] was 66% for 5 instruments, the flute, the clarinet, the oboe, the violin and the cello.

2.3.1.2 Multi-feature systems

Multi-feature systems are a direct extension of the multi-dimensional aspects of timbre. In this approach, timbre is modelled by a mixture of spectral, harmonic and temporal descriptors.

In [MK98], Martin used a large set of 31 features including the pitch, spectral centroid, attack asynchrony, ratio of odd-to-even harmonic energy (based on the first six partials) and the strength of vibrato-tremolo calculated from the output of a

[9] although using the same type of feature, Krishna's work and the one presented in this thesis have been independently carried out

log-lag correlogram. A kth Nearest Neighbours (k-NN) classifier was used within a taxonomic hierarchy after having applied a Fischer discriminant analysis [McL92] on the feature data set in order to reduce the required number of training samples. For 1023 isolated tones over the full pitch range of 14 instruments, 71.6% correct accuracy for the identification of individual instruments has been reported. Agostini described in [ALP01] a system using the mean and the standard deviation of 9 features derived from a STFT, including the spectral centroid, spectral bandwidth, harmonic energy percentage, inharmonicity and harmonic energy skewness. The three last parameters were calculated for the first four partials. The best results have been achieved using a Quadratic Discriminant Analysis (QDA) classifier (92.8% for 27 instruments and the maximum 95.3% for 20 instruments), followed by a SVM (69.7% for 27 instruments), a Canonical Discriminant Analysis (CDA) (66.7% for 27 instruments) and finally a k-NN classifier (65.7% for 27 instruments).

2.3.1.3 Data-mining approaches

These techniques consist of optimising a whole system both at the feature and classifier levels. In essence, a consequent number of features is extracted from the waveforms. Next, the principle is to maximise the system's performance in terms of correct identification rates by selecting, for each class, the feature set allowing the best discrimination between the other classes in the database. These approaches usually involve iterative and trial-and-error procedures.

Fujinaga [Fuj98] used a Genetic Algorithm (GA) to select the best feature set among 352 descriptors. These descriptors consisted of spectral statistical moments extracted from steady-state segments of musical instrument tones. He then used a GA to find an optimum set of feature weights to build the models. His system allowed 50.3% of correct classification of one unknown tone among 39 instruments. In the same vein, Peeters ([Pee03], [PR03]) used a feature selection technique based on the maximisation of the Fisher discriminant ratio in a GMM framework. It has been shown that performance can be increased by 15% for the identification of one isolated note among 28 instruments. Next, Essid [ERD04] explored the use of class-pairwise feature selection techniques [HT98], the principle being to automatically select the feature set among a large amount of descriptors that optimally discriminates between each possible pair of instruments. In [ERD04], a GMM was used to build the instrument models from the selected features. Between 77.8% and 79.1% correct identification were reported using a one vs one classification scheme as opposed to 73.9% when the classical maximum *a-posteriori* (MAP) rule was used. More recently, a system

using Support Vector Machines (SVM) has been described in [ERDb] that helped to improve the performance up to 92% for test samples of 5 seconds.

2.3.2 Instrument modelling

These techniques put the emphasis on the learning of characteristic properties of sound production by musical instruments. Starting from a mathematical assumption about the signal content, the process consists of adapting this model to a training set, evaluating the relevant parameters during a training stage and using it to identify new excerpts. As an example, a log-power spectrum plus noise model in an independent subspace analysis framework has been used in [VR04]. In [VR04], it is assumed that instruments can play a finite number of notes lying on a semitone scale. The short-term log-power spectra are represented as a non-linear sum of weighted typical notes spectra plus background noise. Training the models using isolated notes, 90% correct identification has been achieved for a database of 5 instruments and for testing samples of 5 seconds extracted from commercial recordings.

Other approaches are concerned with acoustic features extracted from the amplitude envelope (e.g. attack time and energy) or from the output of a sinusoidal analysis stage (e.g. partial frequencies and amplitudes, harmonicity or inharmonicity factors [Jen02] [RW82]). However the difficulty to accurately attain these features from *realistic* recordings such as melodic phrases limit the extension of these models for the classification of large musical databases.

2.4 A mixed approach

Mixed models combine the two approaches described in section 2.3.1 and 2.3.2. On the one hand, a prior is set on the mechanisms of sound production and on the signal structure, whereas on the other hand, features are extracted and used to build the models. As an example, the use of synchronous and asynchronous deviations of the phase of the partials for instrument identification purposes is investigated in [DR02]. It is suggested that these features may help to distinguish between instruments or families of instruments. The techniques presented in this thesis belong to this category.

In this section, a computer implementation of the *formant theory of timbre* is proposed. A complete framework for identifying and classifying musical instrument sounds is described. The process yielding the building of *instrument* and *database* models is decomposed into three distinct processing layers. We suggest that this

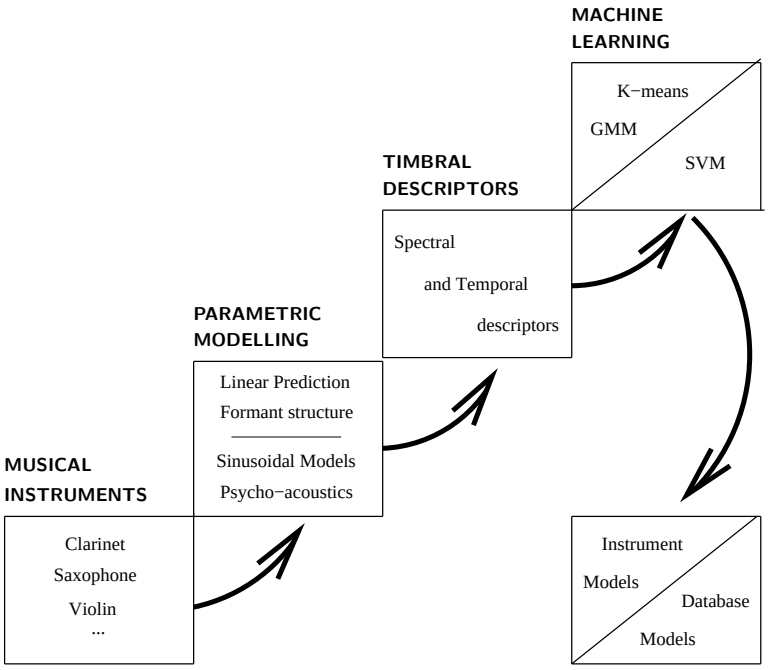


Figure 2.5: Schematic diagram of the system. Three processing layers yielding the building of instrument and database models.

approach preserves the consistency between physical and perceptual aspects of timbre. A schematic diagram is given in figure 2.5.

2.4.1 Modelling the formant structure

Our approach relies on the fact that spectral envelopes and formant structures can be estimated using linear predictive models. In chapter 3, it is described how this model, widely used for modelling the mechanisms of speech production, can be transposed to the case of sounds created by musical instruments.

Drawing knowledge from research in speech coding, the use of a particular set of linear predictive coefficients, the Line Spectrum Frequencies (LSF) is investigated in a musical instrument identification context. It is argued that the localised spectral sensitivity and inter-frame correlation properties they exhibit can serve to determine characteristic spectral shapes for each instrument in the database. Following the principles of the formant theory of musical timbre described in section 1.3.1.3, it is argued that these characteristic spectral shapes can serve to uniquely characterise musical instruments.

The principle behind a mono-feature system is that timbre mostly depends on one

type of acoustic *information*. This approach has clear origins within the speech community and more particularly the speaker identification/verification field. Its use can be justified by the fact that the mechanisms of speech production across speakers are similar. However, this is not totally valid for musical instruments since the physical mechanisms of sound production by musical instruments differ from one source to another. Nevertheless, musical tones share common characteristics, especially in terms of harmonicity and spectral energy distribution so that it can be argued that a single type of feature might be able to capture at the same time global and salient timbre properties across sounds. Early research works in musical instrument identification, such as the one presented in [Bro99] explicitly used speaker verification techniques to identify mono-timbral recordings.

It has been mentioned in section 1.3.1.5 that the timbre of an instrument depends to a certain extent on the pitch. It can therefore be argued that a pitch dependent information is carried at the feature level. For this reason, we propose to use the pitch as a prior for both the instrument modelling and identification phases. The principle is to build instrument models for two frequency registers.[10] The strategy is applied at the database level. Corresponding experiments are summarised in section 5.5.

Experimental research works on timbre perception highlighted the importance of temporal properties in the identification of sounds by humans. As the uniqueness of the spectral envelope cannot be absolutely guaranteed, particularly when dealing with instruments belonging to the same family, the ear often relies on information carried by the attack and onset of sounds. As a consequence, the consideration of temporal features in automatic identification systems can be envisaged as means to include such signal characteristics into the models. These aspects will be covered in detail in section 3.3 where an overview of the available techniques for incorporating temporal descriptors in the systems is given.

2.4.2 Perceptual cues

The perception and interpretation of timbre rely on the low-level processing that takes place in the ear. For this reason, it can be argued that the consideration of psychoacoustic knowledge would better fit the mechanisms of sound perception. Another contribution of our research is to propose a perceptually-motivated analysis/synthesis sinusoidal model that can be used prior to feature extraction stage. After having recalled the principles of analysis/synthesis sinusoidal model, we describe in section

[10] all throughout this thesis, the term register will be used to specify different frequency bands, independently of any musical connotations

3.5 how the ISO/MPEG psycho-acoustic model described in appendix A can be used to select relevant partials in the spectra. A psycho-acoustic version of the LSF, the Perceptual LSF (PLSF) is introduced in section 5.7.

2.4.3 Building instrument and database models

Numerous techniques for data learning and classification problems are encountered in the literature. The use of three machine learning algorithms is investigated in chapter 4. Each of them corresponds to a particular interpretation of the overall modelling process, namely *identification* and *classification*.

The first one consists of learning characteristic feature vectors for each instrument in the database using a K-means algorithm. A minimum distance measure is then used to classify an unknown excerpt among the instruments in the database. Details about this approach are provided in section 4.2. In contrast to this deterministic method, probabilistic classification can also be considered through the use of Gaussian Mixture Models (GMM) in which the feature space is continuously partitioned. GMM are introduced in section 4.3. These two methods will be used to build *instrument models*.

Support Vector Machines (SVM), which have found recently numerous applications for data classification problems, are described in section 4.4. They will be used to build *database models*.

The processes of determining characteristic spectral shapes for each instrument using K-means and GMM on the one hand and classifying spectral envelopes on the other hand are illustrated in chapter 4.

Chapter summary

Experimental results assessing experienced listeners' ability for the task to identify musical instrument notes have been summarised in section 2.1. Conclusions suggested that human identification rates are lower than expected. Previous research showed that onset and attack segments of notes were particularly important for the distinction between instrument sounds.

In section 2.2, principles of computer models and algorithms for musical instrument identification have been introduced. Examples of application of these systems have been given. Several systems encountered in the literature for identifying and classifying orchestral musical instruments sounds have been presented in section 2.3.

The problem has been defined from two angles: the descriptor-based approach and the modelling of the mechanisms of sound production.

Finally, we proposed in section 2.4 an approach to tackle the problem of identifying and classifying musical instrument sounds. The system is composed of three processing layers: parametric modelling, acoustic timbral descriptor extraction and machine learning algorithms. We suggested that this approach preserves the consistence between physical and perceptual aspects of timbre. Further, the distinction between *instrument* and *database* models has been made.

3. Acoustic timbral descriptors

In this chapter, the acoustic descriptors used in our system are detailed. The task of extracting acoustic timbral descriptors from the waveforms is addressed from two complementary perspectives.

The first approach is concerned with a direct computer implementation of the *formant theory of timbre*. In section 3.2, the linear predictive model of sound production is considered as means of modelling both spectral and formant structures of sound. Specifically, the use of a particular set of features, the Line Spectrum Frequencies (LSF), is investigated in a musical instrument identification context.

The perception and interpretation of timbre rely on the low-level processing that takes place in the ear. For instance, it has been highlighted in section 1.1 that the ear frequency response was not linear along the amplitude and frequency scales so that raw acoustic features extracted from the signals might not exactly correspond to what exactly the ear perceives. For this reason, it can be argued that the consideration of psycho-acoustic knowledge would better fit the mechanisms of sound perception. In this second approach, we propose to include perceptual principles at the feature extraction level. In section 3.5, after having recalled the definition of the Mel-Frequency Cepstral Coefficients (MFCC), we propose to calculate features after a psycho-acoustic masking determined using the ISO/MPEG model [ISO] is applied on the signal.

These two approaches offer complementary points of view to the problem of identifying musical instruments. On the one hand, the first approach focuses on the retrieval of physical principles of sound production and signal composition so that psycho-acoustics does not influence the process of discriminating between instrument textures. On the other hand, including psycho-acoustic considerations at the feature level offers a realistic computer implementation of the mechanisms of sound perception by humans.

3.1 Acoustic front-end

A non-negligible advantage in the identification of musical instrument sounds over speech sounds is that corresponding recordings are often noise-free. Current research for real-life speaker identification and verification applications focuses on the control of background noise introduced by the surrounding environment or the distortion introduced by ordinary microphones such as the ones encountered in mobile phones. Systems able to identify a speaker in *clean*, homogeneous and dry acoustic environments yield excellent performance.

For musical signals, the problem is posed in other terms since musical recordings are generally high-quality audio signals. However, the various electronic processing chains used in recording studios introduces significant alterations that are carried by the acoustic features. Differences in microphones and room frequency responses or in the type and amount of audio effects that can be added as post-processing can make the sounds of a particular musical instrument have rather different feature distributions. Although being two realisations in sound of the same instrument, these differences can affect the generalisation power of the models to the point of resulting in an unusable system [LR03]. Obviously, this problem is mostly tackled by the machine learning algorithm which, under the constraint of having a reasonable amount of the various sounds that an instrument can create will generally produce models with reasonable robustness properties.

Various channel normalisation techniques are used prior to the feature extraction stage. Their role is to *normalise* different audio recordings by attenuating any bias that could be introduced by different microphones or recording conditions. A simple example is concerned with the amplitude normalisation of waveforms prior to the analysis in a way that signal energies lay in similar ranges. Similarly, eventual DC-bias that can be introduced by the recording electronic chain can be removed by subtracting the long-term mean from the signal or by applying a high-pass filter. A typical pre-processing chain commonly used in audio signal analysis is depicted in figure 3.1. Note that the pre-emphasis block aimed at increasing the relative contribution of the high frequency content is optional. However, it is commonly used before a linear predictive analysis is applied on the signal.

Audio signals can be considered as stationary over short periods of time. This is due to the physical principles involved in the mechanisms of sound production, for example in speech, and to the nature of the human/instrument mechanical interaction, for example in music. In practice, it is typical to consider frames of 30–40 ms

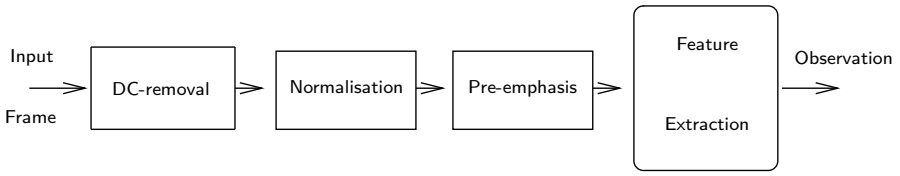


Figure 3.1: A typical pre-processing chain composed of a DC-removal, amplitude normalisation and pre-emphasis blocks.

in duration for the analysis. Furthermore, an overlap of 5–20 ms between adjacent analysis windows can be used to accurately capture the signal physical properties' evolution with time.

After the pre-processing stage, features are calculated within each frame. The grouping of all these descriptors constitute *the observation* and is used as input to the machine learning algorithm.

3.2 Spectral envelope descriptors

The spectral characteristics of audio signals remain the foundation of the understanding of musical sounds by humans. The harmonicity or inharmonicity degrees and the spectral envelopes are examples of features attainable from the calculation of the short-term spectral energy distribution.

In this section, the emphasis is on the modelling of spectral envelopes and formant structures. After having recalled the theoretical principles of the linear predictive model, principles yielding the calculation of the LSF are detailed. The inter-frame correlation and localised spectral sensitivity properties they exhibit are highlighted.

3.2.1 The Linear Predictive (LP) model

The linear predictive model is widely used to represent the mechanisms of speech production and is fundamental in the design of speech compression algorithms. Its theoretical principles are described in this section.

3.2.1.1 Theoretical principle

In the linear predictive model [MG76], each current sample of signal $s(n)$ is estimated or predicted from the p previous weighted samples as:

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i s(n-i)$$

where $\tilde{s}(n)$ is the estimate and p the order of the prediction.

The prediction error or residual signal which is the difference between the original sample value and its prediction can be mathematically written as:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{3.1}$$

The principle behind Linear Predictive Coding (LPC) is to determine the order p and the set of coefficients a_i in order to minimise the energy of the prediction error $e(n)$.

By calculating the Z-transform of the expression 3.1, the transfer function $A(z)$ of a system taking as input the signal $s(n)$ and outputting $e(n)$ can be obtained. Mathematically,

$$\begin{aligned} E(z) &= \mathscr{Z}[e(n)] = \mathscr{Z}[s(n) - \sum_{i=1}^{p} a_i s(n-i)] \\ &= S(z) - S(z)\sum_{i=1}^{p} a_i z^{-i} \\ &= S(z)(1 - \sum_{i=1}^{p} a_i z^{-i}) = S(z)(1 - P(z)) \end{aligned} \tag{3.2}$$

Defining the transfer function as follows:

$$A(z) = \frac{E(z)}{S(z)}, \tag{3.3}$$

it finally comes:

$$A(z) = 1 - P(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{3.4}$$

$P(z)$ is called the predictor filter, $A(z)$ the inverse linear predictive filter, or whitening filter and $e(n)$ the prediction error or residual signal. The filtering operation is depicted in figure 3.2.

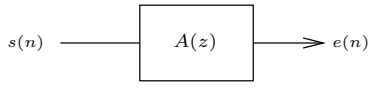


Figure 3.2: Inverse LP analysis filtering. $A(z)$ is the frequency response of a FIR filter whose coefficients are determined in order to minimise the energy of the residual signal $e(n)$.

3.2.1.2 Determination of the filter polynomial coefficients

In this section, we present the technique known as the "autocorrelation LPC" method for calculating the filter polynomial coefficients. The analysis is performed on a frame basis by considering segments of a few tens of milliseconds in duration.

The optimum set of coefficients $\{a_i\}_{i=1,...,p}$ is the one that minimises the energy of the residual for each considered frame, and for a given prediction order. This can be performed by setting the partial derivatives of the energy of the residual E for the whole frame of length N

$$\begin{aligned} E &= \sum_{n=0}^{N-1+p} e^2(n) \\ &= \sum_n \left[s(n) - \sum_{i=1}^{p} a_i s(n-i) \right]^2, \quad \text{with} \quad s(n-i) = 0 \quad \text{if} \quad n-i < 0 \end{aligned}$$

with respect to a_i to zero. In other words,

$$\frac{\partial E}{\partial a_i} = 0, \qquad i = 1, ..., p$$

Subsequently,

$$\begin{aligned} \frac{\partial E}{\partial a_i} &= 2 \sum_n \left[\left[s(n) + \sum_i a_i s(n-i) \right] s(n-j) \right] \\ &= 2 \left[\sum_n s(n)s(n-j) + \sum_i a_i \sum_n s(n-i)s(n-j) \right] = 0, \qquad i, j = 1, .., p \end{aligned}$$

which can be rewritten as:

$$\underbrace{\sum_n s(n)s(n-j)}_{} + \sum_i a_i \underbrace{\sum_n s(n-i)s(n-j)}_{} = 0 \tag{3.5}$$

The first term in Eq. (3.5) corresponds to the short-term cross-correlation coefficients while the second term represents a shifted version of the cross-correlation. Defining

$$\begin{cases} R_j &= \sum_n s(n)s(n-j), & j &= 1, ..., p \\ R_{i-j} &= \sum_n s(n-i)s(n-j), & i, j &= 1, ..., p \end{cases}$$

Eq. (3.5) can be rewritten

$$R_j + \sum a_i R(i-j) = 0$$

This defines a set of equations known as the Yule-Walker equations. It can be written in a matrix form:

$$\begin{pmatrix} R_0 & R_1 & \dots & R_{p-1} \\ R_1 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & R_1 \\ R_{p-1} & & R_1 & R_0 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_1 \\ \vdots \\ \vdots \\ R_p \end{pmatrix}$$

In order to calculate the vector $\mathbf{a} = \{a_i\}_{i=1,\dots,p}$, the Toeplitz matrix R has to be inverted. This can be performed by using a Gaussian elimination decomposition. However, several efficient algorithms such as Durbin or Levinson-Durbin are often preferred [MG76].

3.2.1.3 A basic model of sound production

The linear predictive model is widely used for modelling the mechanisms of speech production. A simplified model of sound production is depicted in figure 3.3. In this model, the all-pole frequency response of the IIR filter $H(z) = 1/A(z)$ models the vocal tract transfer function while the residual signal $e(n)$ represents the glottal excitation.

In the basic model of speech production, the excitation is modelled either by an impulse train at period T_0 if the frame is considered voiced, or by a white noise if

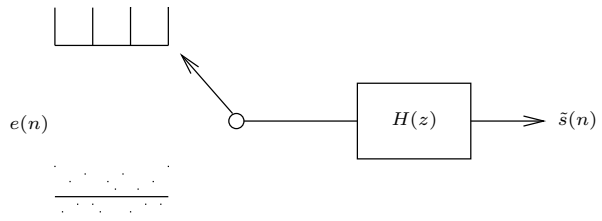


Figure 3.3: A simplified model of speech production. The excitation $e(n)$ obtained after filtering by $A(z)$ is modelled either by an impulse train, either by a random sequence, depending on the voiced or unvoiced nature of the frame. The modelled excitation is then fed to the synthesis filter $H(z)$ to create the synthesised frame of signal $\tilde{s}(n)$.

the frame is unvoiced. This implementation is used in very low-bit rate coders such as the LPC-10 [Tre82] which encodes a 8 kHz sampled speech signal at 2.4 kbits/s. Although such implementation produces low-quality speech due to the simple form of the modelled residual, most modern speech coders use in one way or another the linear predictive technique, primarily for redundancy reduction purposes.

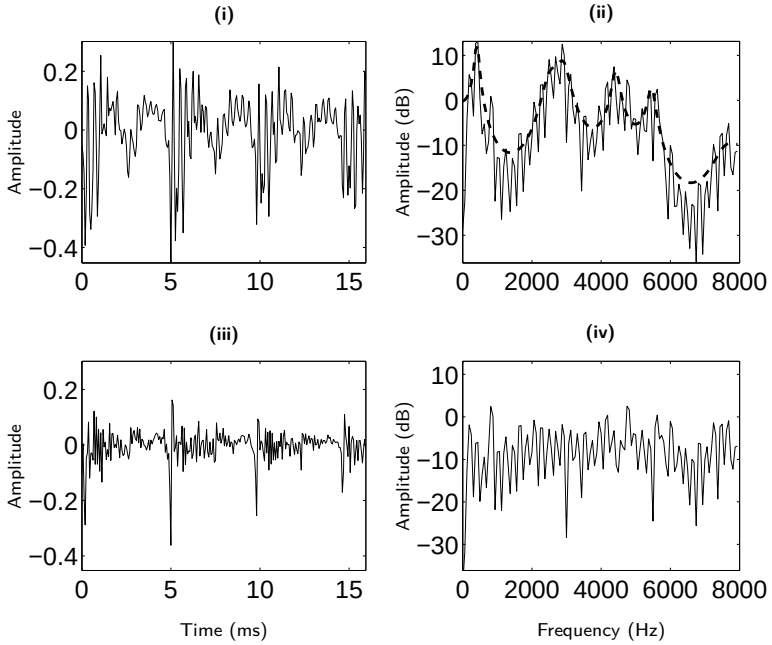
The effects of LP analysis filtering for both voiced and unvoiced speech frames are illustrated in figure 3.4. A 12th order linear prediction has been applied in both cases. Note that after filtering, the spectra are flattened. The effect is more pronounced in the case of the voiced sound (figure 3.4(a)) which exhibits a clear formant structure. Note that the time-domain residual signal (figure 3.4(a)(iii)) exhibits regularly spaced *impulses* corresponding to the glottal excitation. The time-period between two impulses corresponds to the fundamental frequency value of the sound. For this reason, it is said that the linear predictive technique can be used to deconvolve from speech signals the respective contributions of the excitation and the vocal tract. Note also that the frequency response of the LP synthesis filter gives an accurate estimate of the signal's formant structure (figure 3.4(a)(ii)) and short-term high-pass spectral envelope (figure 3.4(a)(iv)).

In figure 3.5 is illustrated the effect of the LP analysis for two musical instrument sounds, the flute and the clarinet. The frames have been extracted from steady-state portions of isolated notes. A 12th order linear prediction has been applied. Similarly to the case of speech, the spectra are flattened after filtering. Note the clear formant structure these two instrument sounds exhibit. Observe also the noisy nature of the two time-domain residual signals (figures 3.5(a)(iii) and 3.5(b)(iii) respectively) in contrast to the speech voiced sound residual signal shown in figure 3.4(a)(iii).

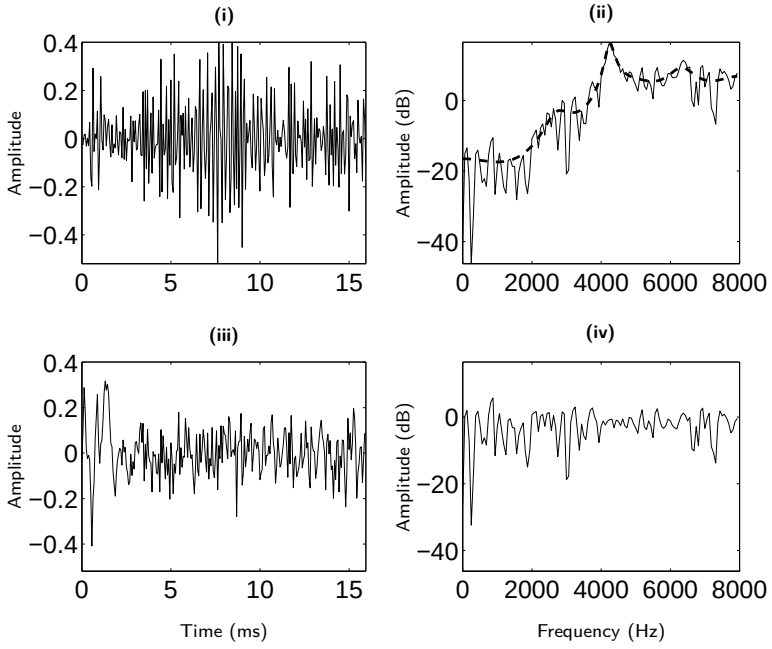
3.2.2 The Line Spectrum Frequencies (LSF)

Several features are directly or indirectly derived from the linear predictive filter polynomial coefficients. It can be distinguished the PARCOR (PARtial CORrelation) or reflection coefficients corresponding to intermediary variables in the calculation of the filter polynomial coefficients using the Durbin-Levinson algorithm. They also correspond to the ratio between adjacent sections in the tubular model of speech production mechanisms [Cal00].

A particular set of linear predictive coefficients, the Line Spectrum Frequencies, derived from the Line Spectrum Pairs, have been introduced by Itakura [Ita75] for efficient scalar and Vector Quantisation (VQ) of the short-term spectral envelope

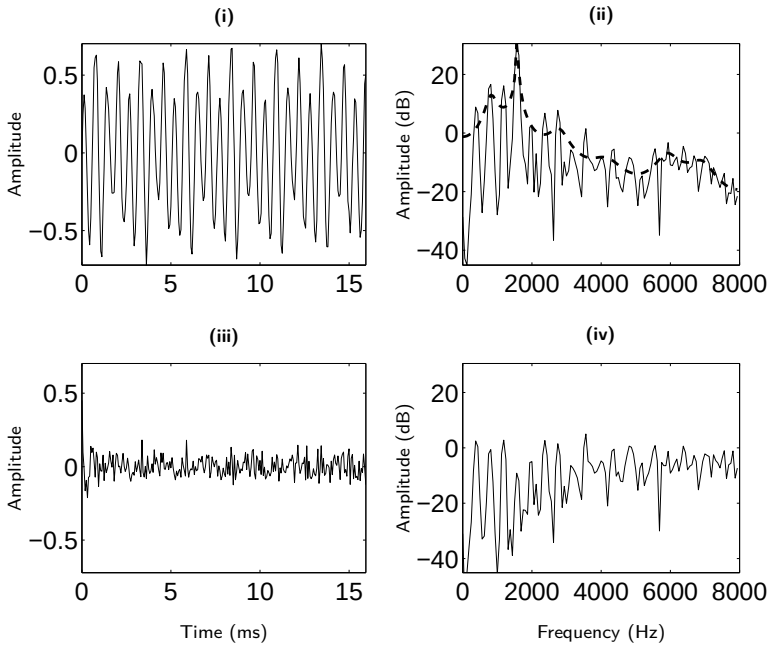


(a) voiced frame

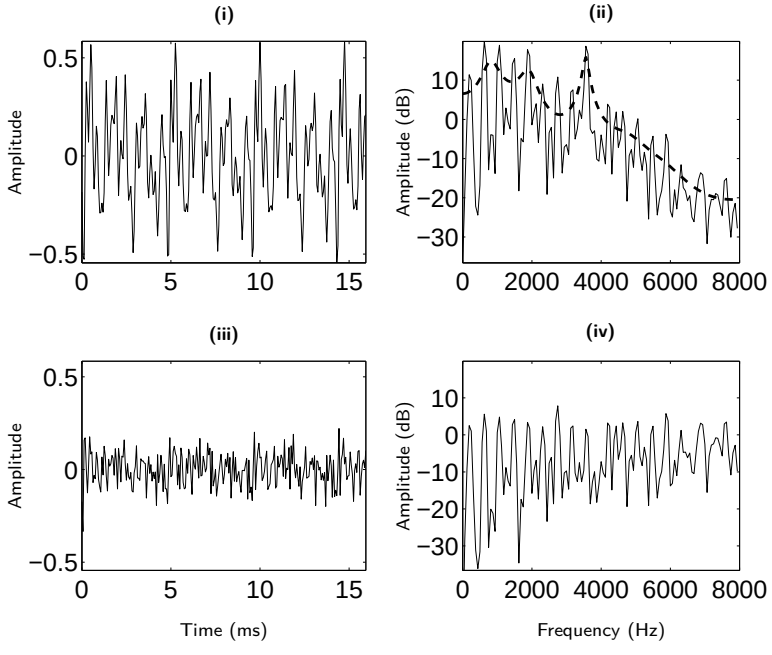


(b) unvoiced frame

Figure 3.4: Inverse LP filtering and spectral flatness. Illustration for (a) voiced and (b) unvoiced speech frames. The prediction order is $p = 12$. Spectral envelope estimates using linear predictive analysis are represented by dashed-lines. Plots (i) and (ii) are the original time-domain frames and their corresponding magnitude spectra respectively. Plots (iii) and (iv) are the residual time-domain signals obtained after filtering and their corresponding spectra respectively. After filtering (iv), the short-term correlation between samples is removed and the spectra are flattened.



(a) flute



(b) clarinet

Figure 3.5: Inverse LP filtering and spectral flatness. Illustration for two (a) flute and (b) clarinet frames extracted from steady-state portions of sounds. The prediction order is $p = 12$. Spectral envelope estimates using linear predictive analysis are represented by dashed-lines. Plots (i) and (ii) are the original time-domain frames and their corresponding magnitude spectra respectively. Plots (iii) and (iv) are the residual time-domain signals obtained after filtering and their corresponding spectra respectively. After filtering (iv), the short-term correlation is removed and the spectra are flattened.

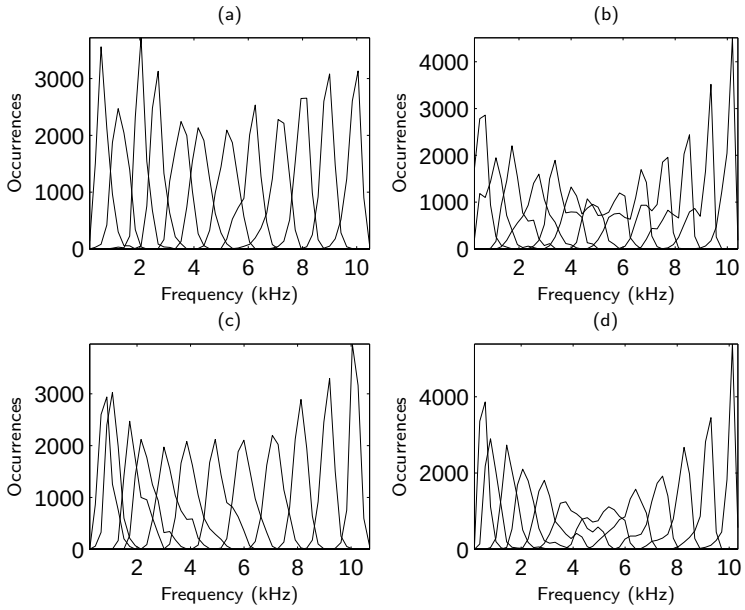


Figure 3.6: Distribution histogram for 12 LSF parameters calculated from 2 minutes of monophonic recordings. (a) cello, (b) clarinet, (c) flute and (d) piano.

parameters in speech coders [PA93]. Their use in a musical instrument identification context constitutes one of the foundations of our approach.

We recall that in a source-filter configuration, the short segment of a signal is assumed to be generated as the output of an all-poles filter $H(z) = 1/A(z)$, where $A(z)$ is the inverse filter given by:

$$A(z) = 1 + a_1 z^{-1} + ... + a_p z^{-p}, \tag{3.6}$$

where p is the order of the LPC analysis and $\{a_i\}_{i=1,...,p}$ the filter coefficients.

The Line Spectrum Pairs (LSP) are the roots of two polynomials $P(z)$ and $Q(z)$ defined as:

$$\begin{cases} P(z) = A(z) + z^{-p+1} A(z^{-1}) \\ Q(z) = A(z) - z^{-p+1} A(z^{-1}) \end{cases} \tag{3.7}$$

Assuming that $H(z)$ is a stable filter, it can be shown that the roots of P and Q lie on the unit circle, are interleaved, distinct and that exactly two of the zeros are at $+1$ and -1. Their corresponding angular frequencies are called the Line Spectrum Frequencies and lie in the range $]0\ \pi[$. Representing at the same time the short-term

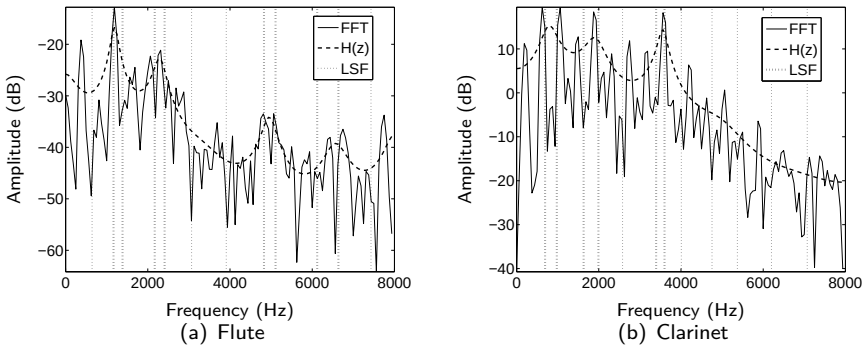


Figure 3.7: Magnitudes of the STFT, LP filter frequency responses, and LSF representations (vertical dashed lines) for two frames of steady-state segments of (a) flute and (b) clarinet sounds having similar pitches. Note that two close LSF values characterise a peak in the spectra.

spectral envelope and energy distribution, it can be assumed that the LSF are good candidates for modelling the spectral envelopes and formant structures of sounds. A computationally efficient algorithm to calculate these roots using Chebyshev polynomials is provided in [KR86].

Figure 3.6 shows the distribution of 12 LSF parameters that have been extracted from two minutes of monophonic recordings for four instruments, the cello, the clarinet, the flute and the piano. It can be noticed that the coefficients are ordered along the frequency axis – the range]0 π[has been mapped to a frequency scale in Hz – and their respective distributions exhibit characteristic bell shapes. In figure 3.7 are shown two short-term spectra (flute and clarinet) with their corresponding LP spectral envelope representations for a prediction order $p = 12$. On the other hand, the LSF are represented by vertical dashed lines. Note that two close LSF parameters characterise a formant in the spectra (they are shown in bold-dashed lines) and that the closer the LSF values, the stronger the peak amplitudes.

Another interesting property of the LSF is the strong inter-frame correlation they exhibit. This inter-frame correlation is illustrated in figure 3.8 where a melodic clarinet excerpt is considered. The time-domain signal is shown on top. The frequency responses of $H(z)$, that have been determined using a 12th order linear predictive analysis, are plotted at the bottom in vertical lines. The corresponding LSF representations are plotted on the same graph in horizontal dashed lines. Observe the formant with the lowest frequency at $t \approx 0.3$ s and note how the spaces between its associated pairs of LSF coefficients shrink as the amplitude of the peak increases.

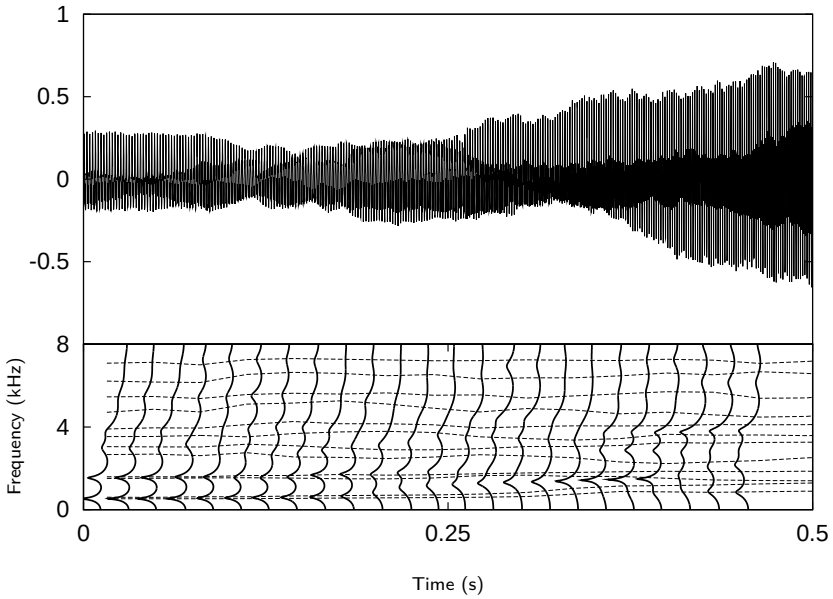


Figure 3.8: Clarinet melodic phrase excerpt time-domain representation (top), spectral envelopes estimated using a 12th order linear predictor (vertical lines) and corresponding LSF representation (horizontal dashed lines) as a function of time (bottom). Note at $t \approx 0.3$ s how the spaces between pairs of LSF coefficients shrink as the amplitude of the peak with the lowest frequency increases.

3.3 Temporal features

This section is concerned with temporal descriptors. They have been shown to be important timbre correlates in sections 1.3.1.4 and 2.1.2.

Following early perceptual experiments conducted by Berger [Ber63], recent research works [Ero01] [ELR$^+$05] attempted to evaluate the importance of temporal signal properties in the identification of sounds by machines. It is a fact that the onset of a hammered or plucked string note is different in nature than that of a flute, for example, so that intuitively, extra-information about the signal temporal behaviour can be considered to better model the signal's characteristics and therefore to improve the system's performance.

In [Ero01], several cepstral features derived from linear predictive coefficients have been studied in parallel with *time-related* features. The transient/steady-state separation was performed using an energy-based criterion. More precisely, it was assumed that the steady-state began when the signal energy reached its average RMS-energy level for the first time. This way, two sets of features could be extracted from both types of signals. It was found that 12 MFCC extracted from the steady-state portions of isolated tones performed slightly better than the same features extracted from the

onset database: a rough 4% increase in average performance was observed for the task of identifying one excerpt among 29 instruments.

In [ELR+05], musical phrases extracted from commercial recordings were considered. The segmentation transient/steady-state was performed using two onset detection algorithms [BDA+05]. The strategy first consisted of classifying frames into two databases as a function of their *transientness degree* [ELR+05]. Next, a large amount of features was extracted from the resulting datasets. For each database, a feature selection algorithm based on Fischer's Linear Discriminant Algorithm (LDA) has been used together with a pairwise classification strategy. For each database, the set of 40 features discriminating the best between all possible pairs of instruments was automatically selected. Using an average class separability criterion, it has been found that features selected from the databases of transient signals yielded a better discrimination power than the ones extracted from non-classified and non-transient databases respectively. However, when applied to a practical classification problem, the segmentation transient/non-transient prior to feature extraction/selection stage did not show any significant improvements in terms of average correct classification rate.

In these two experiments, the non-improvement in terms of correct identification rates can be explained as follows. Non-transient (i.e. stationary) segments are often longer in duration than transient segments. Assuming that the sustained portion of a note has consistent and homogeneous characteristics, the features extracted from non-transient signals are in fact redundant over several temporal windows. They result therefore in *statistically insignificant* information – since non-novel – for building the models. In contrast, transient signals are shorter in duration and different by nature. Added to the fact that the onset detection algorithm is prone to errors and may therefore considers stationary signals as being transient, it can be argued that the resulting transient database is in fact a non-redundant snapshot of the non-segmented database. Thus the non-improvement in terms of correct average identification rates.

It can further be noticed that the use of automated feature selection algorithms in [ELR+05] makes the interpretation of the results difficult, especially in terms of the relevance of the descriptors used to model the non-stationary part of signals. For instance, since similar features are extracted from both types of signals, and that for each, the best set of 40 features is chosen, it is not possible to independently evaluate the contribution of a particular type of features during the modelling process.

To summarise,

- Extracting features from onsets have shown to result in slightly improved system's performance compared to when features were extracted from the steady-state segments of sounds. This has been interpreted and explained and we noted the database of onsets obtained after segmentation was in fact a non-redundant snapshot of the whole database.

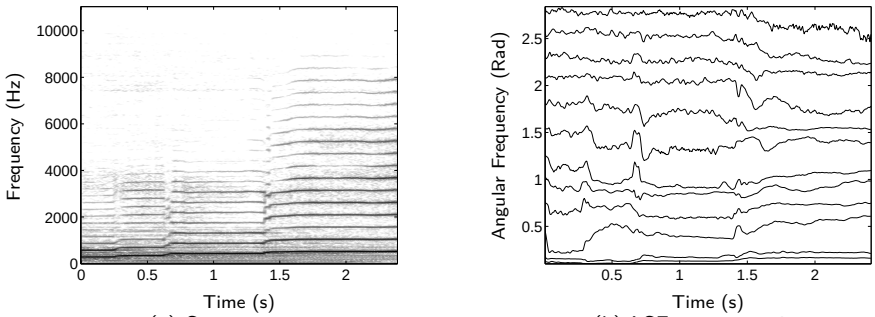
- The main difficulty in extracting temporal features resides in the fact that robust automated pre-processing techniques for onsets or transients detection are difficult to design, especially in the case of pitched musical phrases. As a result, it is difficult to accurately quantify the increase in performance when transient and steady-state features are independently considered. However, when using isolated notes, the problem is easier to tackle. Experiments will be conducted towards this direction in section 5.6.1.

- We note that the experimental works reported in this section considered pitched musical sounds for the experiments. Although presenting non-linearity properties, the onsets of pitched musical sounds created by wind, brass or string instruments are much less singular than that of a piano or a plucked string sound. This can explain why, by using a differentiated transient/steady-state sound modelling strategy prior to the feature extraction stage, no significant improvement could be observed.

In order to include temporal considerations in the models, a more general approach to the problem can be envisaged. It consists of characterising the transitional spectral information and considering the derivative of the feature vectors as a function of time. This is commonly termed as calculating the *delta* coefficients of the feature vector. The usual procedure is to append them to the original descriptors in the feature vector prior to the modelling process.

Since the features are calculated at regular time intervals, they do not have analytical form. Therefore, the derivative can only be approximated by a finite difference. However, a first order finite difference is intrinsically noisy so that Furui [Fur81] proposed to fit a first order polynomial to each time series of feature coefficient. In other words, considering that x is a feature vector, the delta coefficients are calculated using:

$$\delta_x[n] = \frac{\sum_{k=-K}^{K} kx[n-k]}{\sum_{k=-K}^{K} k^2} \tag{3.8}$$



(a) Spectrogram (b) LSF representation

Figure 3.9: (a) Spectrogram and (b) LSF representations for a saxophone melodic phrase. A 12th prediction order has been used. Changes between notes can be spotted at $t = 0.6$ s and $t = 1.4$ s on both representations.

The augmentation of the LSF feature vectors with their delta as commonly performed with cepstral coefficients in speaker/speech recognition will be investigated in chapter 6.

In figure 3.9 are shown both the spectrogram of a saxophone solo excerpt (figure 3.9(a)) and its corresponding LSF representation (figure 3.9(b)). Onset locations corresponding to the note changes ($t = 0.6$ s and $t = 1.4$ s) can be spotted on both plots. It can be argued that this transitional information can be captured by considering the delta of the LSF feature vectors.

3.4 Pitch, vibrato and tremolo features

These features try to capture some other frequency and time-related characteristics of instrument sounds. It has been highlighted that the pitch influences the timbre to a certain degree so that it can be considered as one fundamental parameter of sound that an identification system should take into account. In practice, pitch values can be used to adapt the analysis window lengths prior the Fourier transform, in order to set an appropriate frequency resolution, to select features as a function of the signal harmonic structure, or to build models of instruments for several pitch ranges. The latter strategy will be explored in chapter 5.

On the other hand, features such as vibrato and tremolo have been studied by Martin [Mar99] and considered in more recent research works [Ero01] [ELR+05]. Although these descriptors correspond to characteristic properties of sound, their consideration in an automatic musical instrument identification system has not been shown to be essential. Furthermore, it can be argued that vibrato and tremolo are

captured in some ways by the spectral envelope features and their evolutions with time. As a consequence, these features will not be studied more in detail here and not used in our system.

3.5 Psycho-acoustic considerations

The use of psycho-acoustic models is wide spread in the audio signal processing field so that there exists strong justifications in using psycho-acoustic considerations for an efficient acoustic description of waveforms. Early research works in speech/speaker verification already used psycho-acoustic considerations at the feature level. Likewise modern systems for automated musical instrument identification incorporate such principles.

For this purpose, several methodologies can be envisaged. This can be performed at a pre-processing or front-end level by using, for example, an auditory filterbank or *log-lag* correlogram [MK98] prior to the feature extraction stage. Another approach consists of *including* psycho-acoustic clues at the feature level by warping the linear frequency axis onto a Mel or Bark scale prior to the spectral feature calculation. These approaches differ in their complexity and in the number of perceptual principles they take into account.

In this section, we investigate various methods to include psycho-acoustic knowledge for the purpose of modelling timbre. In section 3.5.1, we outline several techniques encountered in the literature. Next, we propose to extend the field of application of the sinusoidal decomposition described in section 3.5.2 by introducing a psycho-acoustically motivated sinusoidal spectral analysis/synthesis technique based on the ISO/MPEG psychoacoustic model [ISO] described in appendix A.

3.5.1 Background

In this section, we briefly outline the existing techniques encountered in the literature.

3.5.1.1 Frequency warping

Frequency warping consists of transforming one spectral representation in a given frequency scale to another representation on a new, psycho-acoustically relevant frequency scale.

This approach has been experimented by Eronen [Ero01]. He used a warping function to map the calculation of the autocorrelation function from the frequency to the Bark scale [HL01] prior to the calculation of various LPC coefficients. He

compared the performance of the LP polynomial coefficients and PARCOR that were transformed into cepstral coefficients to their respective warped versions. Experiments involving isolated notes showed that for prediction orders ranging between 3 to 30 (for signals sampled at 44.1 kHz), the WLPC and warped reflection coefficients performed slightly better than their non-warped counterpart for both the identification of instruments and instrument families. Specifically, an average increase of 8% in terms of correct instrument identification and 10% for instrument families have been achieved.[1]

Note that the use of a warped version of the LSF, the WLSF, will be investigated in chapter 5.

3.5.1.2 Auditory filterbanks

Using auditory filterbanks as pre-processing allows to decompose a signal into band-limited signals whose frequency content can be interpreted as following the decomposition that takes place in the inner ear (see section 1.1). In particular, auditory filterbanks implemented in the frequency domain[2] are widely used in audio signal analysis and modelling.

For instance, the determination of the MFCC involves the calculation of the total energy of frequency filters regularly spaced in the Mel-frequency scale. Likewise, Essid [ERDb] proposed to apply a triangular octave filterbank for the calculation of spectral features he named the Octave Band Signal Intensities (OBSI).

Principles of spectral auditory filterbanks are illustrated in the next section. In particular, details about the MFCC are given.

3.5.1.3 The Mel-Frequency Cepstral Coefficients (MFCC)

MFCC are widely used in speech recognition and speaker verification systems. They constitute the classical feature in audio spectral pattern recognition problems. For this reason, they have been the first feature to be studied in a musical instrument identification context [Bro99]. We briefly recall in this section how the MFCC can be calculated from a frame of audio signal. This procedure is based on the implementation proposed in [Sla98].

For a given frame, the short-term magnitude spectrum is calculated using a FFT. Next, a *perceptual* triangular filterbank having approximately equal bandwidth in the

[1] for a database containing 29 instruments, 33% correct instrument identification and 66% correct instrument family identification have been achieved

[2] as opposed to band-pass time-domain auditory filterbanks

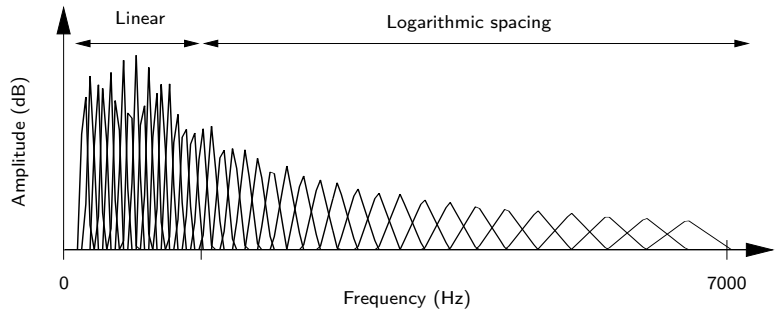


Figure 3.10: Triangular filterbank used to calculate the MFCC coefficients [Sla98]. The non-uniform weighting comes from the fact that each filter is given unit weight. Each filter has approximately equal bandwidth in the Mel-frequency scale.

Mel-frequency scale is applied in the frequency domain. The filters are spaced linearly for the low-frequencies (13 filters) up to roughly 1000 Hz and logarithmically afterwards (27 filters). The upper and lower frequencies of each filter are the centre frequencies of the adjacent filters respectively. The filterbank used in [Sla98] is depicted in figure 3.10.

The total energies in the 40 bands are then calculated, yielding 40 coefficients also called the *filterbank coefficients*. Next, the log-energy outputs are cosine transformed, yielding the Mel-cepstral coefficients. In practice, a discrete cosine transform is used. Assuming that $\{f_i\}_{i=1,\ldots,N_f}$ are the filterbank coefficients with N_f being equal to the total number of filters, the MFCC c_i are calculated using:

$$c_i = \frac{1}{N_f} \sum_{j=1}^{N_f} \log f_j \cos\left[\frac{\pi i}{N_f}(j - \frac{1}{2})\right], \quad i = 0, \ldots, N_f - 1 \tag{3.9}$$

In practice, c_0, which represents the average power of the spectrum is discarded. On the other hand, only the first coefficients (typically 12–16) are usually considered for building the feature vectors. The MFCC will be used as a reference to compare our system to in chapter 5.

3.5.2 Sinusoidal modelling

This section discusses the sinusoidal analysis/synthesis model based on the short-term Fourier transform. This model will be used to include psycho-acoustic considerations at the pre-processing level. In the following sections, a complete analysis/synthesis framework based on the STFT is described.

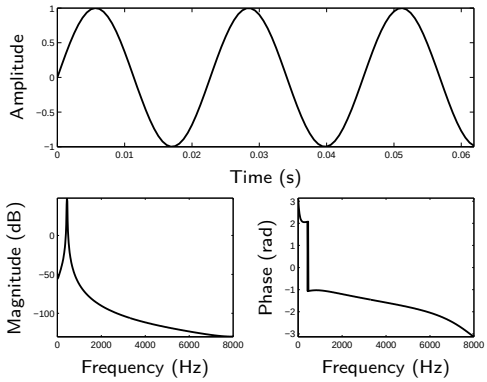


Figure 3.11: Sinusoidal signal time-domain waveform (top) and frequency representation using the STFT. Magnitude (bottom left) and phase (bottom right) after Hanning windowing.

Sinusoidal modelling techniques are related in some ways to the *phase vocoder* originally presented by Flanagan in 1966 [FG66] with an application in speech transmission. They have since found numerous applications in high quality time-scaling and pitch-shifting of audio signals [Dol86] [Lar99] or in speech processing [MQ86]. They are also at the origins of the Sinusoidal Modelling Synthesis (SMS) framework [Ser97].

3.5.2.1 Theoretical principle

Sine waves, or pure tones, are an important class of sound waves as they convey the notion of frequency and its dual, the time period. A sine wave is a deterministic periodic signal whose time evolution $x(t)$ is entirely defined by the knowledge of three parameters: amplitude, frequency and phase. A sinusoidal waveform can be mathematically expressed as:

$$x(t) = A\sin(2\pi ft + \phi) = A\sin(wt + \phi) = A\sin\Psi(t)$$

where A is the amplitude, f the frequency, $w = 2\pi f$ the pulsation (in $rad.s^{-1}$) and ϕ the initial phase (in rad). In figure 3.11 are shown the waveform, magnitude and phase representations of a sinusoidal signal.

Sinusoids are important in a variety of ways. Firstly, they are fundamental in physics. Systems that resonate or oscillate produce quasi-sinusoidal motions (e.g. simple pendulum or LC oscillator). Another reason is that complex exponentials are eigenfunctions of linear time-invariant (LTI) systems, meaning that they are important

for the analysis and characterisation of linear filters and for the estimation of filter frequency responses. Similarly, they constitute a set of orthogonal basis functions that can be used to analyse and decompose signals (e.g. using the Fourier series expansion or the Fourier transform). More importantly, from a computer music and signal processing points of view, the human ear acts in some ways as a spectrum analyser: as has been discussed in chapter 1, the cochlea physically splits sounds into their near sinusoidal components. In other words, by looking at spectra, which display the *amount* of energy corresponding to each sinusoidal basis present in a signal, we are looking at a representation very similar to that the brain receives on hearing. Moreover, it has to be noted that results from psycho-acoustic experiments, such as the tables used in psycho-acoustic models [ISO], are determined using, among others, pure tones stimuli.

In sinusoidal modelling, a discrete time valued signal $x(n)$ is approximated by a linear sum of evolving sinusoids as:

$$x(n) \approx \tilde{x}(n) = \sum_{q=1}^{Q(n)} A_q(n) \cos \Psi_q(n) \tag{3.10}$$

where $Q(n)$ is the maximum number of partials at the time n, $A_q(n)$ the instantaneous amplitude of the partial q and $\Psi_q(n)$ its instantaneous phase. The additive components in this model are assumed to vary on a time scale longer than the sampling period, meaning that the parameters can be estimated at a subsampled rate.

The approximation symbol in the equation above implies that the sum of partials model does not represent exactly the original signal. Therefore, a residual term $r(n)$ can be added in order to reconstruct perfectly $x(n)$:

$$x(n) = \tilde{x}(n) + r(n) = \sum_{q=1}^{Q(n)} A_q(n) \cos \Psi_q(n) + r(n)$$

The signals $\tilde{x}(n)$ and $r(n)$ contain different musically meaningful information of $x(n)$: the sum of partials captures characteristics such as the spectral envelope, the harmonicity, the loudness or the pitch whereas the residue represents a mixture of impulsive (e.g. transient and strong onsets during the attack of a piano note) and/or highly uncorrelated noise (such as the friction sound created by a bow). Since a sum of slowly varying sinusoids is rather ineffective for modelling noisy signals, a common strategy consists of classifying and separating the two components during the analysis stage. Similar to a partially filled or partially empty bottle, the estimation of one of

the two signals automatically determines the content of the other one.

The determination of instantaneous sinusoidal parameters is commonly performed by calculating STFT over sliding windows. The local maxima in the magnitude spectra are interpreted as corresponding to sinusoidal components.

Mathematically, let us define $X(k,n)$ as being the STFT of $x(n)$. It is a function of both the time and frequency indices n and k and is defined as:

$$X(k,n) = \sum_{m=n}^{n+N-1} x(m)h(n-m)e^{-2j\pi\frac{km}{N}} \tag{3.11}$$

where $h(n)$ is a finite length analysis window zero-valued outside the interval [n n+N-1] and N the length of the FFT. Setting $\omega_k = 2\pi k/N$ in Eq. (3.11) yields:

$$X(k,n) = \sum_{m=n}^{n+N-1} x(m)h(n-m)e^{-j\omega_k m} \tag{3.12}$$

The Fourier transform being invertible, a general re-synthesis equation is given by

$$x(n) = \sum_{m=n}^{n+N-1} g(n-m)\frac{1}{N}\sum_{k=0}^{K-1} X(k,m)e^{j\omega_k m} \tag{3.13}$$

where $g(n)$ is the synthesis window zero-valued outside the interval [n n+N-1]. In the absence of modifications on $X(k,n)$ and under certain conditions on the shape of the windows $h(n)$ and $g(n)$, $x(n)$ can be perfectly recovered using Eq. (3.13).

In the case where $X(k,n)$ is not modified, one should ensure that the synthesis leads to perfect reconstruction. For this purpose an OverLap-and-Add strategy (OLA) is used. In our implementation, analysis and synthesis windows on the one hand, and analysis and synthesis window lengths on the other hand are similar. In this case, the condition for perfect reconstruction is that the sum of the squared windows regularly spaced by the hop-size equals unity. The reader is referred to [ABL02] for more details about the other cases.

The analysis/synthesis procedure using sliding STFT is the fundamental principle of sinusoidal modelling. An illustration is given in figure 3.12. After segmentation, the frames are weighted by the analysis window. The STFT is then applied. The processing is performed in the frequency domain. After the inverse STFT, the synthesised signals are weighted by the synthesis window. An overlap-and-add strategy ensures minimal distortion at frames boundaries during the reconstruction. Note that the sum of the analysis windows (shown in solid line) regularly spaced by the hop-size

equals unity.

3.5.2.2 Sinusoidal analysis

We recall that the analysis is performed on a STFT frequency representation where local maxima in the magnitude spectrum are interpreted as being sinusoidal components. For the rest of this development, the FFT of the input windowed frame $x(n)$ of length N calculated at a given time index m will be noted $X(k) = |X(k)|e^{j\phi_k}$, $k = 0, ..., N-1$, where N is the length of the FFT.

The STFT displays the signal energy frequency distribution evaluated at each frequency bin $k = 0, ..., N-1$ with a resolution $\delta f = f_s/N$, where f_s is the sampling frequency and N the length of the observation.

Within the short-term magnitude spectra, the frequency bin k is considered as a local maximum if

$$|X(k)| > |X(k+1)| \quad \text{and} \quad |X(k)| > |X(k-1)|, \quad k = 1, ..., N/2-1$$

As the Fourier transform evaluates the energy distribution at discrete equally spaced frequencies, the frequency value of a local maximum corresponds to the frequency value of its corresponding bin index. Let us consider the case of a sinusoidal signal at *true* frequency f_0. If k_0 is the index of the local maximum in the corresponding short-term spectrum, its frequency is:

$$f_{k_0} = k_0 \frac{f_s}{N}$$

However if the original frequency f_0 is not a multiple of f_s/N, f_{k_0} and f_0 are different.

High quality audio signal are usually sampled at 44100 Hz, providing a spectral resolution of 21.5 Hz if a 2048 points FFT is used. Depending on the application, it is often necessary to attain more accurate frequency values.[3] Several spectral interpolation techniques can be used to counterbalance the STFT finite resolution. For instance, a widely used algorithm is based on the observation that the STFT log-magnitude of a sinusoidal signal windowed by a Gaussian window can be approximated by a sampled parabola [SS87]. Therefore, by fitting a parabola to a local maximum and its two closest neighbours in the log-magnitude spectrum, it is possible to determine more accurately the *true* amplitude and frequency of the sinusoid.

[3] this is especially true if one wishes to determine the fundamental frequency of musical notes on the Western equally tempered pitch scale, for example

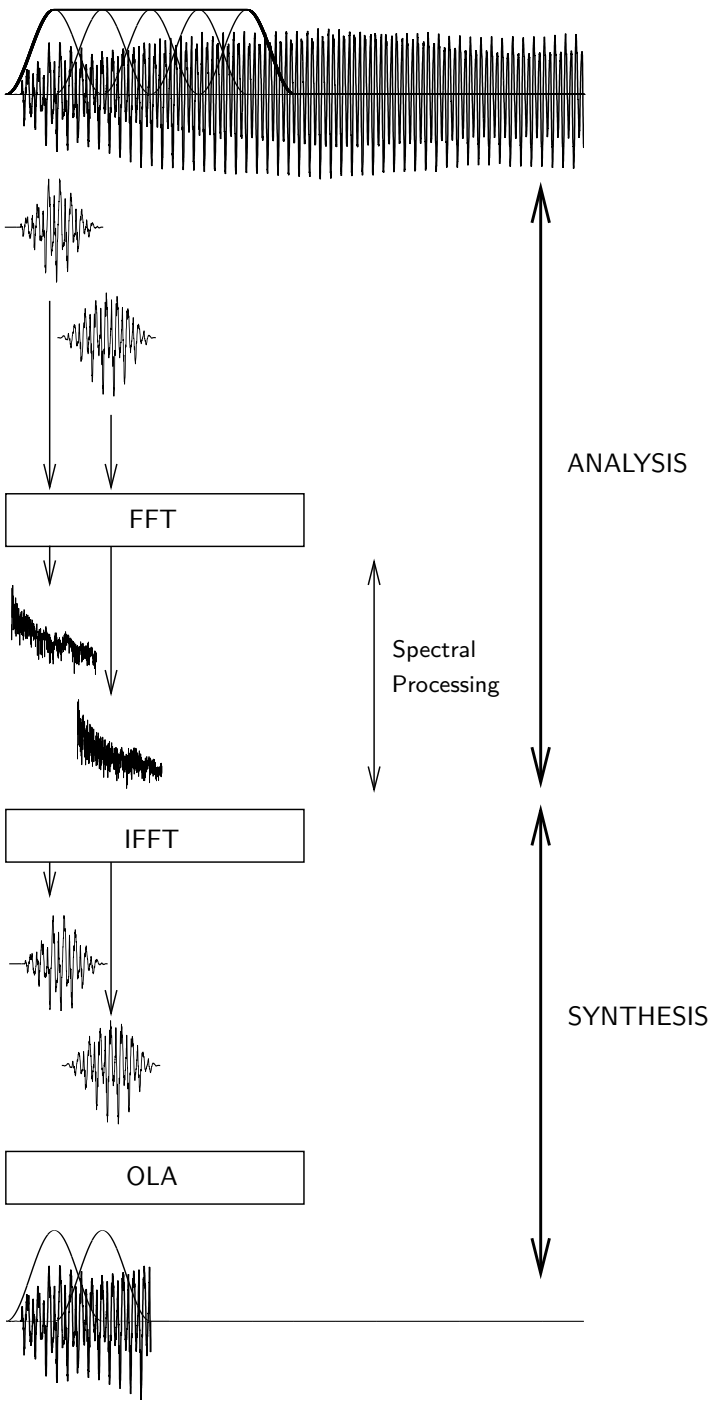


Figure 3.12: Principle of sinusoidal modelling using sliding STFT. Illustration for two time-domain grains. After segmentation, the frames are weighted by the analysis window. The STFT is then applied. The processing is performed in the frequency domain. After the inverse STFT, the synthesised signals are weighted by the synthesis window. An overlap-and-add strategy ensures minimal distortion at frames boundaries during the reconstruction.

This approximation can be extended to the case of windows with non-Gaussian but bell-shaped short-term magnitudes. Defining

$$X_{dB}(k) = 20\log_{10}|X(k)|, \quad k = 0, ..., N$$

as the signal short-term magnitude in dB and k_0 as the bin frequency of a local maximum, the process consists of fitting a parabola through the three samples with amplitudes:

$$A_{-1} = X_{dB}(k_0 - 1), \quad A_0 = X_{dB}(k_0), \quad A_1 = X_{dB}(k_0 + 1)$$

The frequency difference between the estimated actual frequency and the bin frequency values in FFT bins can be expressed by:

$$\delta_k = \frac{1}{2}\frac{A_{-1} - A_{+1}}{A_{-1} - 2A_0 + A_{+1}}$$

The new interpolated bin index becomes:

$$\tilde{k}_0 = k_0 + \delta_k$$

and its corresponding interpolated amplitude is:

$$\tilde{A}_0 = A_0 - \frac{\delta_k}{4}(A_{-1} - A_{+1})$$

Finally, the phase is linearly interpolated at $\tilde{k}_0$ using the phase values of the two neighbouring bins.

Figure 3.13 illustrates the effects of spectral interpolation for the determination of the frequencies and amplitudes for three sinusoids. The spectra exhibit peaks at the same discrete frequencies for the three considered signals (pure tones with frequencies $f_0 = 610$ Hz, $f_0 = 625$ Hz and $f_0 = 640$ Hz for the left, middle and right plots respectively). A peak-picking algorithm based on a local maximum detection would consider the three signals as having the same frequency (i.e. 625 Hz). After quadratic interpolation using the two closest neighbours, more accurate frequency and amplitude values can be determined and the three signals can now be distinguished in the frequency domain.

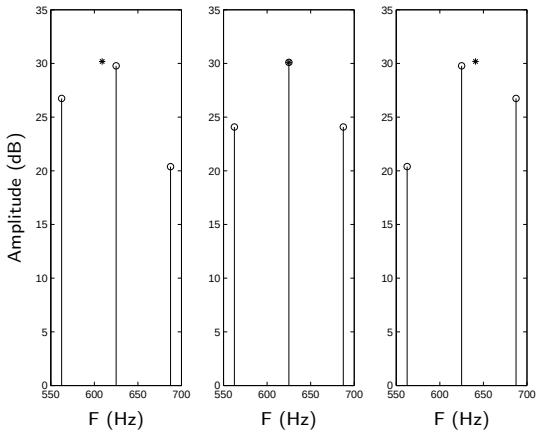


Figure 3.13: Examples of spectral quadratic interpolation for three sinusoids: $f_0 = 610$ Hz (left), $f_0 = 625$ Hz (middle) and $f_0 = 640$ Hz (right). Signals were sampled at 8 kHz, FFT on 128 points. Circle markers represent the magnitudes of the FFT while the stars show the amplitudes and frequencies of the sinusoidal components after quadratic interpolation. Without interpolation, the bin frequencies of the three local maxima are equal. The STFT spectral resolution of $\delta f = f_s/N = 8000/128 = 62.5$ Hz is not fine enough to separate the three frequencies.

3.5.2.3 Sinusoidal synthesis

After the analysis stage, a series of local maxima has been selected. Their amplitudes, frequencies and phases have been interpolated. The synthesis of the selected sinusoidal components can be performed either in the time or frequency domain. The time-domain synthesis is a direct implementation of Eq. (3.10) where the parameters $A_q(n)$ and $\Psi_q(n)$ are linearly interpolated between successive frames [ABL02].

An approach based on the IFFT is presented in this section: the principle is first to *reconstruct* the complex short-time FFT vector from the information extracted during the analysis stage and second to use the inverse FFT together with the OLA method to recover the time-domain grain. The IFFT/OLA based synthesis has the advantage in automatically interpolating the parameters at frame boundaries. Failing to do so would otherwise introduces artifacts in the synthesised signals.

Let us first consider the case of a pure sinusoidal signal $x(n)$ with frequency f_0. During the analysis stage, $x(n)$ is partitioned into overlapping frames, windowed by the analysis window and the STFT is calculated prior to the spectral analysis. Next, the *true* instantaneous amplitude, frequency and phase of the local maximum are estimated for each frame within the short-term magnitude representation. These three interpolated parameters are then used to rebuild the short-term FFT vector.

An important property of the Fourier transform is known as the convolution the-

orem and holds that a multiplication operation in one domain (time or frequency) is identical to a convolution in the dual domain. In other words, the Fourier transform of the windowed signal is the Fourier transform of the signal (theoretically defined on an infinite support) convolved with the Fourier transform of the analysis window. Let us consider a sinusoidal signal with amplitude A. The magnitude of its Fourier transform exhibits theoretically a Dirac located at $f = f_0$ with an amplitude $A/2$. The resulting Fourier transform is the modulated version of the window Fourier transform at frequency f_0.

The following FFT vector reconstruction implementation is based on the modulation of the theoretical FFT of the windowing function by the interpolated sinusoidal frequency. The STFT vector $X(k) = |X(k)|e^{j\phi(k)}$, $k = 0, ..., N-1$ is reconstructed using the interpolated instantaneous sinusoidal parameters $\tilde{f}_0$, $\tilde{A}_0$ and $\tilde{\phi}_0$. In the following, the case is illustrated for a Hanning window.

The kernel function of the Hanning window can be constructed using three Fourier transform of the rectangular window $W_r(\theta)$ [Por96]:

$$W(\theta) = 0.5W_r(\theta) - 0.25W_r(\theta - \frac{2\pi}{N-1}) - 0.25W_r(\theta + \frac{2\pi}{N-1})$$

where

$$W_r(\theta) = D(\theta, N)e^{-j0.5\theta(N-1)}$$

and $D(\theta, N)$ being the Dirichlet kernel defined as:

$$D(\theta, N) = \frac{\sin(0.5\theta N)}{\sin(0.5\theta)}$$

Given $\tilde{f}_0$ and $\tilde{A}_0$ being the interpolated frequency and amplitude of the sinusoid, the magnitude of the Fourier transform magnitude $|X(k)|$ is approximated by:

$$|\tilde{X}(k)| = 2\frac{\tilde{A}_0}{N}W\left(\theta(k) - \theta_0\right), \quad 0 \leq k \leq N/2 - 1 \tag{3.14}$$

where θ_0 is the angular frequency of the sinusoidal component defined by:

$$\theta_0 = 2\pi\frac{\tilde{f}_0}{f_s}$$

The function $W(\theta)$ is in fact evaluated for each angular frequency bin

$$\theta(k) = \frac{2\pi k}{N} - \theta_0, \quad 0 \leq k \leq N/2 - 1$$

On the other hand, the phase vector $\phi(k)$ is reconstructed using $\tilde{\phi}_0$ as:

$$\phi(k) = \tilde{\phi}_0, \quad 0 \leq k \leq N/2 - 1 \tag{3.15}$$

Finally, prior to the inverse FFT, the full symmetric vector has to be reconstructed. Knowing that real signals have symmetric FFT, it can be written:

$$\begin{aligned} |X(\frac{N}{2} + k)| &= |X(\frac{N}{2} - k)|, \quad 0 \leq k \leq \frac{N}{2} - 1 \\ \phi(\frac{N}{2} + k) &= -\phi(\frac{N}{2} - k), \quad 0 \leq k \leq \frac{N}{2} - 1 \end{aligned}$$

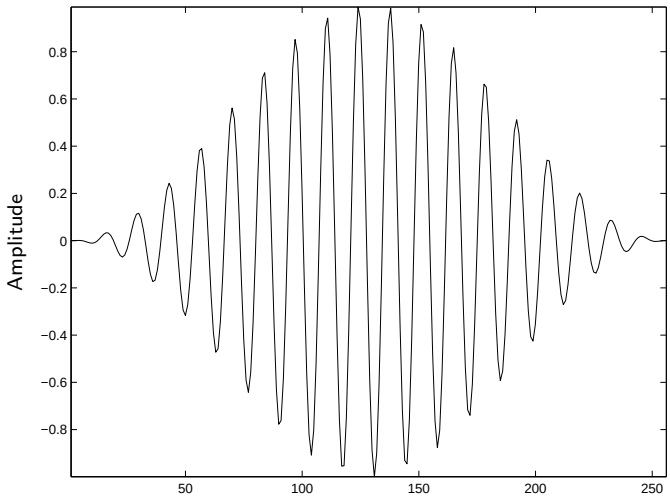
Figure 3.14 illustrates the use of the kernel modulation technique for the reconstruction of a FFT vector composed of a single sinusoid. The original signal (sampled at $f_s = 8$ kHz) is segmented into fixed length overlapping frames (FFT on 256 points with a hop-size of 64 samples). Each time-domain grain is then weighted by a Hanning window (figure 3.14(a)) before the FFT analysis. The frequency, amplitude and phase of the local maximum are determined using a peak-picking algorithm. Next, the interpolated values are used to reconstruct the FFT magnitude vector on the one hand (using equation Eq. (3.14)) and the FFT phase vector on the other hand (using Eq. (3.15)) as shown in figure 3.14(b).

The overlap-and-add strategy used in the implementation ensuring perfect reconstruction after synthesis, errors that are introduced in the re-synthesised signals are only due to the limited accuracy of the interpolation method. It has to be noted that Eq. (3.15) is not theoretically valid as the phase value of the sinusoidal component is assigned to all the FFT bins. However, a single sinusoid magnitude spectrum exhibits significant energy only at the bins located around the analysed frequency. Therefore, artifacts are not introduced during the inversion operation.

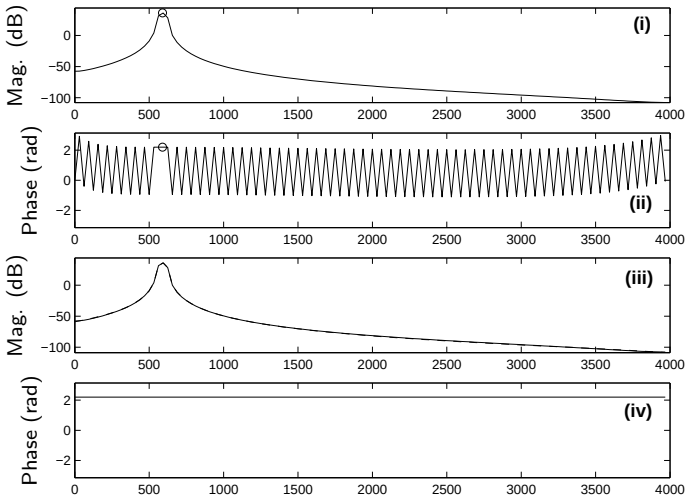
In the following, this sinusoidal synthesis technique is extended to the case of audio signals composed of several sinusoidal components. Prior to the synthesis, it is assumed that peak-picking and quadratic interpolation algorithms have been applied to estimate accurate amplitudes, frequencies and phases of the selected sinusoidal components.[4] Two possible approaches can be used for the synthesis of multiple sinusoidal components:

- first, by using an iterative scheme within each frame where a single sinusoidal grain is synthesised at a time, each one resulting from an inverse FFT operation.

[4] the concept of partials selection is not expanded here but in section 3.5 where a psychoacoustically motivated selection of the local maxima will be proposed

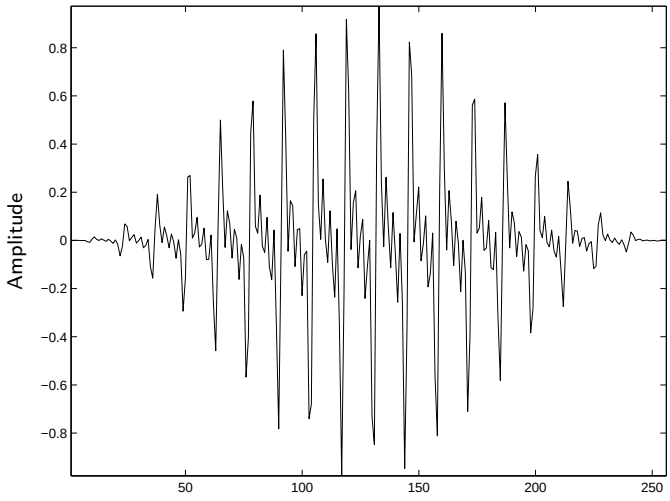


(a) Sinusoidal grain weighted by a Hanning window

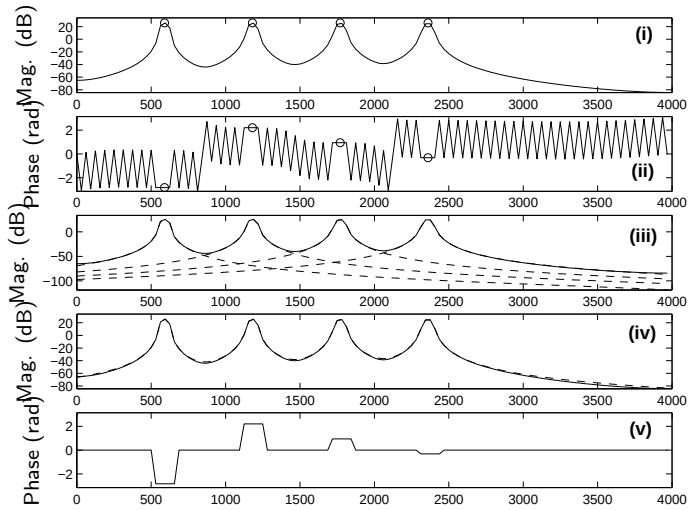


(b) Spectral analysis and reconstruction

Figure 3.14: Single sinusoid spectral reconstruction ($f_0 = 590$ Hz, $f_s = 8$ kHz). (a) Original sinusoidal grain weighted by a Hanning window. (b) Details about the spectral analysis/reconstruction stage: the FFT magnitude (256 points) in dB is shown in (i) while the FFT phase component (in rad) is plotted in (ii). Circle markers represent the estimated amplitude, frequency and phase parameters after quadratic interpolation. On (iii) are plotted the original FFT magnitude (line) and the result of the Hanning kernel modulation (dashed) determined using Eq. (3.14) (note that both plots are very similar so that they are not distinguishable). The reconstructed phase vector is shown in (iv).



(a) Harmonic grain weighted by a Hanning window



(b) Harmonic synthesis

Figure 3.15: Harmonic sinusoidal spectral reconstruction ($f_0 = 590$ Hz, 4 harmonics, $f_s = 8$ kHz). (a) Original harmonic grain weighted by a Hanning window (b) Details about the two spectral reconstruction methods: the FFT magnitude (256 points) in dB is shown in (i) while the FFT phase component (in rad) is plotted in (ii). Circle markers represent the estimated amplitudes, frequencies and phases parameters after quadratic interpolation. On (iii) are plotted the original FFT magnitude (line) and the amplitudes of the four modulated Hanning kernel corresponding to the four sinusoidal components. Their corresponding phase vectors are calculated as detailed in figure 3.14. On (iv) is plotted the original FFT magnitude (line) and the sum of the four modulated Hanning kernels (dashed). The corresponding recomposed phase vector is shown in (v).

The addition of the different grains is then performed in the time domain. This case is a direct extension of the single sinusoidal grain synthesis detailed in figure 3.14. Figure 3.15(b)(iii) illustrates the process where the modulated kernels are independently considered during time-domain grain synthesis.

- second, by reconstructing the whole FFT vector using the total number of modulated kernels. In this case, only one inverse FFT operation per frame is necessary.[5] The *addition* is performed in the frequency domain. More precisely, the magnitudes of the modulated kernels are added but the phases are *composed*. Figures 3.15(b)(iv)/(v) illustrate this case where only one FFT vector is reconstructed to synthesise the harmonic time-domain grain. This strategy has been used in our implementation.

3.5.3 Frequency masking

While general audio coders use psycho-acoustic models for bit allocation and quantisation purposes, it is interesting to exploit the psycho-acoustic knowledge for direct manipulation of the audio signal in the frequency domain.

As an extension of the sinusoidal analysis/synthesis model described in the previous section, we propose to include psycho-acoustic principles in the frequency domain prior to the feature calculation. In essence, the ISO/MPEG psycho-acoustic model described in appendix A is used to select the most relevant partials in the frequency domain. They are then used to resynthesise a time-domain signal. This design saves on the need to specify the number of sinusoids to be extracted *a priori*. This section is concerned with a detailed description of the algorithm.

Its general principle is depicted in figure 3.16. The input frame is firstly processed by a FFT. Next, a psycho-acoustic mask is applied and relevant partials, i.e. the ones having their amplitudes above the mask, are selected from the spectra. The selected sinusoidal components are then used to synthesise the output frame using the kernel modulation technique described in section 3.5.2.3. The subtraction of the latter from the original signal yields the residual signal. The hop-size during the analysis/synthesis is chosen to ensure perfect reconstruction in the case where the spectral representations stay unaltered.

[5] note that this saves a considerable amount of processing compared to the iterative method since on average 30–150 sinusoids are needed to accurately synthesise a musical instrument sound

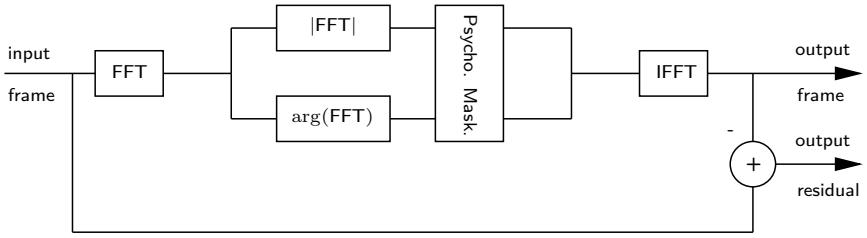


Figure 3.16: Illustration of a sinusoidal analysis/synthesis loop with partial selection stage using psycho-acoustic masking.

3.5.3.1 Using the absolute threshold in quiet

This experiment focuses on the use of the Absolute Threshold of Hearing (ATH) when manipulating signals in the frequency domain. The absolute threshold in quiet corresponds to the faintest sound energy level audible by the human ear (see chapter 1). Within a short-term magnitude spectrum, only the sinusoidal components located above this threshold are retained and used for the synthesis. In our implementation, the following analytical ATH function [LAM] has been used:

$$\begin{aligned} Ath(f) \quad &= 3.64f^{-0.8} - 6.8\exp(-0.6(f-3.4)^2) \\ &\quad +6\exp(-0.15(f-8.7)^2) + 0.0006f^4 \end{aligned} \tag{3.16}$$

where f is expressed in Hz and Ath in dB SPL. Note that this analytical formula replaces the one proposed in [TSS82] commonly used in the literature. A corresponding plot is shown in figure 3.17.

In practice, the analysis stage is decomposed as follows:

- First the input signal is segmented into overlapping frames (for $f_s = 22.05$ kHz, frames of N=1024 samples in length with an overlap of $N/4$ has been used to ensure perfect reconstruction). The segments are then weighted by a Hanning window prior to the STFT calculation. Next, all the local maxima in the magnitude spectrum are found by peak-picking and stored in a list.
- The psycho-acoustic model maps the absolute threshold in quiet to the current STFT frame by normalising the maximum amplitude to 96 dB SPL.
- Next, the local maxima located above the absolute threshold in quiet level are retained for the synthesis. The others are discarded.

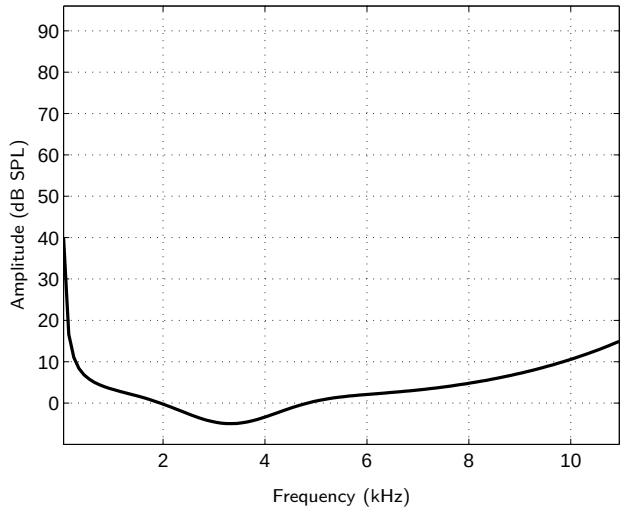


Figure 3.17: Absolute threshold of hearing in dB SPL as a function of frequency calculated using Eq. (3.16). 96 dB SPL is used as reference level.

- Each of these local maxima amplitude is then interpolated using the quadratic scheme described in section 3.5.2.2. Their corresponding interpolated frequencies and phases are used for the synthesis.

In figure 3.18 is illustrated the whole process for a clarinet solo phrase. The number of retained sinusoidal components as a function of time is shown in figure 3.18(b). Dots correspond to the number of local maxima that have been found in the spectra while the horizontal line delineates the average number of partials kept for the synthesis. Note that on average, 65 sinusoids are retained for this excerpt.

As this stage, the following conclusions can be drawn regarding the algorithm behaviour:

- Segments with small SNR (t = 0–0.02s in figure 3.18) are modelled using more sinusoids than average. This is a common drawback when one tries to decompose noisy signals using a sum of sine components. However, in this particular case, this segment of sound is inaudible and do not have to be considered. A combined energy and pitch detector can overcome this situation.

- Transients corresponding to note onsets are characterised by high energy frequency content. They are modelled using more components than average. They correspond to peaks located at $t \approx 0.03$s and $t \approx 0.5$s. A more efficient implementation of sinusoidal modelling algorithm would consider the transients

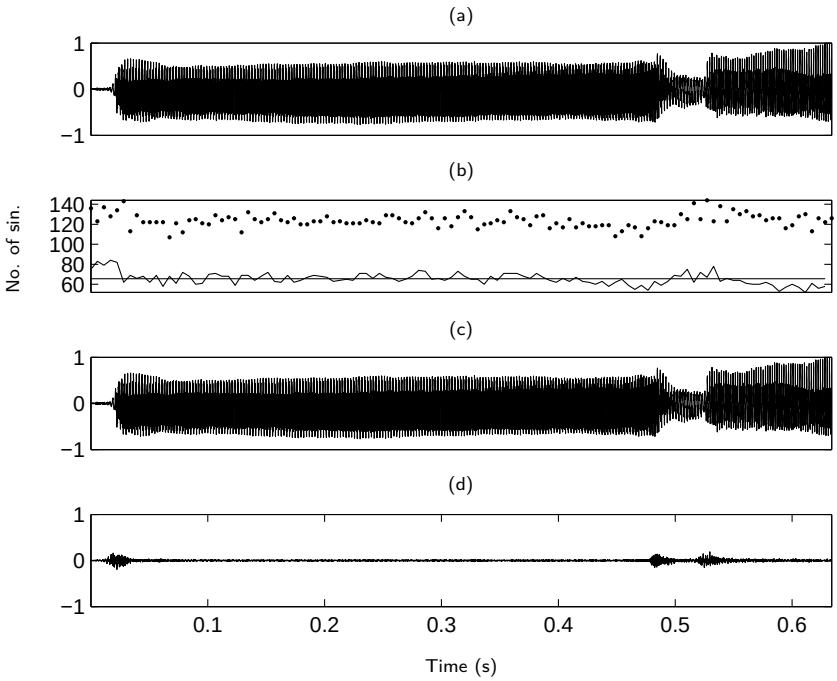


Figure 3.18: Sinusoidal modelling using the absolute threshold in quiet to select the relevant local maxima in the short-term spectrum. The signal shown in (a) is a solo clarinet phrase. In (b) are shown the number of retained sinusoidal located above the absolute threshold of hearing. Dots correspond to the number of local maxima that have been found by peak-picking. Plot (c) corresponds to the synthesised time-domain signal while the residual signal is shown in (d).

sections as being part of the residual signal. Another drawback of modelling transients with sinusoids is that the attacks are *smoothed* since the energy distribution is averaged over the whole analysis window when the spectra are calculated.

- In purely harmonic and sinusoidal sections, the number of retained components stays around 60. The algorithm performs well for pitched sounds and the simple use of the absolute threshold in quiet significantly reduces the number of components needed to model the signal.

The synthesised signal is shown in figure 3.18(c) while the residual after subtraction in the time domain is shown in figure 3.18(d). The synthesis quality is very good, without any perceivable artifacts or loss in tone colour. However, the quality of similarly resynthesised signals using trumpet solo phrases revealed artifacts located at the note onsets: attacks were smoothed and difference between original and modelled signals were noticeable.

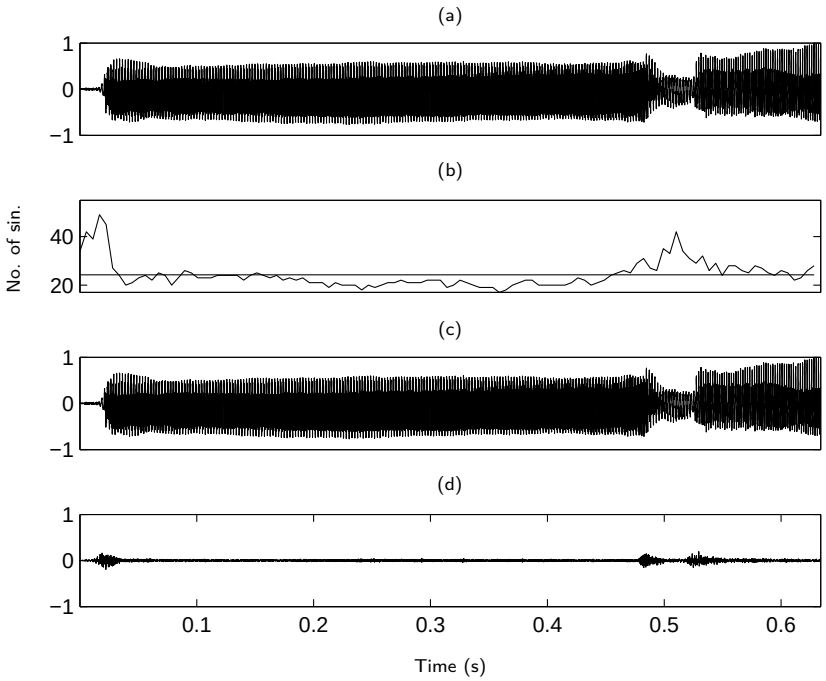


Figure 3.19: Sinusoidal modelling using a MPEG-1 layer II psycho-acoustic model. The signal shown in (a) is a solo clarinet phrase. In (b) is shown the number of retained sinusoidal located above the global masking threshold. Plot (c) corresponds to the synthesised time-domain signal while the residual is shown in (d).

3.5.3.2 Using the complete psycho-acoustic model

The global masking threshold calculated using the complete psycho-acoustic model described in appendix A is now used to select the relevant components. In this case, the frequency masking of both tonal and non-tonal as well as the absolute threshold of hearing in quiet are considered for the global masking curve calculation.

The principle of the decomposition is similar to the one described in section 3.5.3.1. Corresponding results are presented in figure 3.19 for the same clarinet melodic phrase.

The general trend for the number of extracted sinusoidal components as a function of time is preserved (same behaviour at transient, harmonic and noisy sections) but the average number of partials is now more than halved (25 against 65 for the previous experiment involving the absolute threshold of hearing). Despite this reduction, the quality between the synthesised signal is very good for this sound. Note the increase in the number of extracted sinusoids noticeable at $t \approx 0.5$s and compare to figure

3.18(b). The attack of a note is characterised by a significant burst of energy in the high frequencies which move across the masking threshold, thus resulting in a higher number of sinusoids.[6]

This framework will be used during the feature extraction stage in order to include psycho-acoustic knowledge in our system prior to the LSF calculation. Details about this novel feature, the Perceptual LSF (PLSF), will be given in section 5.7.

Chapter summary

The acoustic descriptors that will be used to build our system have been described in this chapter.

First, it has been shown in section 3.2 how linear predictive models can be used to represent signals' short-term spectral envelopes and at the same time to capture the signal formant structure. Principles of the deconvolution between the contribution of the excitation and resonating body for the mechanisms of sound production have been introduced. We have recalled how the linear predictive analysis/synthesis filter coefficients can be calculated using the autocorrelation method. The emphasis has been on the description of the LSF as means of spectral envelope and formant structure descriptors.

Second, the importance of temporal information in the perception of timbre has been highlighted. The difficulty in automatically attaining such features has been emphasised in section 3.3. In particular, a review of existing approaches encountered in the literature did not show significant improvements in overall performance when features were extracted individually from transient/onset and steady-state segments of signals.

Third, several methods for including psycho-acoustic at the feature and pre-processing levels have been described in section 3.5. After having reviewed one method to calculate the MFCC, we proposed to include psycho-acoustic knowledge in the analysis/synthesis sinusoidal model described in section 3.5.2. The strategy consists of perceptually selecting relevant partials in the spectra. For this purpose, a complete frequency domain analysis/synthesis framework has been described in section 3.5.3. A detailed description of the perceptual LSF calculation will be provided in chapter 5.

[6] at a cost of increased computation compared to existing techniques [BDA+05], one can use the number of psycho-acoustically selected sinusoids as a function of time as an onset detection function

4. Machine learning algorithms

The search for an invariance and constancy in timbre is central to the building of musical instrument models. It has been mentioned in the introduction of this thesis that the identification of a sound object by humans can be achieved in a wide variety of acoustic circumstances. In the same vein, a trumpet sound is recognised as being a trumpet sound, independently of the instrument brand or playing style: these different realisations in sounds and waveforms share the same and unique identity.

To a certain extent, building a timbre model consists of characterising this invariance. From a set of multi-dimensional acoustic descriptors, machine learning algorithms are used to define mathematical rules from which unknown sample identities can later be inferred when compared to the models stored in the database.

This chapter discusses the building models that can serve to identify and classify instruments. It is described how the principles governing the *formant theory of timbre* can be implemented at the machine learning level. For this purpose, *generative* and *discriminative* methods are studied. The distinction between these two types of techniques is emphasised in section 4.1.1. Next, the theoretical principles of the K-means, Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) are recalled in sections 4.2, 4.3 and 4.4 respectively. Through the use of the K-means and GMM algorithms, the interpretations of learning characteristic formant structures and building *instrument models* are proposed. This approach is a central point of this work. The use of Support Vector Machines (SVM) for classifying spectral envelopes and building *database models* is further investigated. After having introduced the theoretical principles of each of the three methods, the processes of building instrument and database models are interpreted and illustrated.

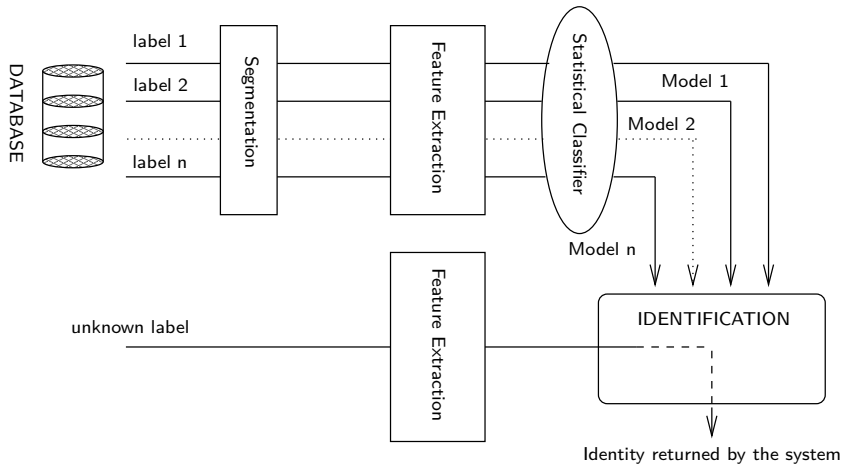


Figure 4.1: General principle of a supervised system for identification and classification.

4.1 Supervised learning

In *supervised learning*, examples of inputs and corresponding outputs are given. The aim is to predict the output for future inputs having seen only a restricted number of training examples. In essence, a representation of the data to models and their corresponding labels or identities are known *a-priori*.[1] In our case, the associated labels are instrument names, such as violin, flute or saxophone.

4.1.1 Generative vs discriminative methods

For building models, two types of algorithms will be considered. On the one hand, the problem of data classification will be tackled using generative methods such as K-means or GMM. In this case, each model for each class is built independently of any knowledge about the other classes in the database. The identification process consists of minimising a similarity measure (K-means) or maximising an *a-posteriori* likelihood of the unknown feature distribution knowing the models (GMM).

Strictly speaking, these two algorithms can be seen as learning techniques since the models are explicitly and individually learnt from a set of training data. On the other hand, discriminative methods, such as SVM, consider the whole database in order to determine optimum boundaries between the various classes.

[1] as opposed to *unsupervised learning* for which the classes are discovered from the data themselves

It is important to point out the differences between both identification and classification methods and especially through the methodologies they involve. By definition, a classifier needs to have the knowledge of the whole database to work on since the process consists of discriminating between each class's feature distribution against the others. One obvious drawback of the approach is that models have to be re-trained when new classes are added to the database, thus infringing the modularity principle (see section 2.2). Their main advantage is that they generally yield better systems' performance. On the other hand, extending the database when a generative method is used is simply a matter of training the models for the newly added families.

These two types of systems correspond to two different interpretations and implementations of the same problem.

4.1.2 Principle

In figure 4.1 is illustrated the general principle of a supervised learning system. Independently of the algorithm's final application, a common architecture consists of two distinct phases, namely the *training* and *testing* phases.

- during the **training phase**, models are built for each individual labelled instrument class using a set of data representative of the different realisations of the same identity. In practice, models are trained using a set of acoustic descriptors extracted from waveforms. In the case where generative methods are used, models can be seen as condensed representations of the classes intra-variability. In the case where discriminative methods are used, the principle is to define rules for optimally separating the classes between each other.

- during the **testing phase**, *unknown* excerpts (i.e. the samples that have not been used to train the models) are presented to the system. They are similarly pre-processed, and the same type of features are extracted than during the training phase. This set of features is then compared to all the models in the database. The system returns the identity of the presented excerpt based on a *similarity* criterion.

By building a model, the principle is to generalise a system behaviour that have been trained using a limited database for any future unseen inputs. Burges [Bur98] mentioned that:

> *Roughly speaking, for a given learning task, with a given finite amount of given data, the best generalisation performance will be achieved if the right balance is struck between accuracy attained on that particular training set, and the "capacity" of the machine, that is, the ability of the machine to learn any training set without error.*

As the outputs for the training database are known, the system is evaluated using a subset of the whole database that are unknown to the system. These performances are then generalised to any future inputs.

4.2 The K-means

The K-means algorithm is an iterative clustering method whereby a data set can be approximated by a finite number of codevectors. Being one of the simplest parametric machine learning algorithms, it can be used in numerous applications, spanning the fields of audio and video compression[2], speaker identification and verification. Its use in a musical instrument identification context is another fundamental aspect of this research.

In contrast to the other techniques presented in this chapter, the K-means algorithm relies on the use of distance measures, both for the learning and identification phases.

4.2.1 Theoretical principle

The principle of the K-means algorithm is to cluster a n-dimensional space into K distinct regions in terms of a chosen distance, each region having one single representant. In essence, after a K-means optimisation, the feature data set is represented by a fewer number of data, called the centres or codevectors.[3]

The training or optimisation phase involves an iterative scheme that can be decomposed as follows:

1. Each region has a centre which is the *mean* all the data points in that region.
2. Each data point is assigned to the region whose centre it is closest to.

These two steps are alternated and repeated until a stop criterion is met, i.e., when there is no further change in the assignment of the data points. In that case, the algorithm reaches a local minimum.

[2] the K-means algorithm is also called Vector Quantiser (VQ), for example in speech coding [PA93]

[3] or also centroids or prototypes

Let us consider a training data set $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of N n-dimensional feature vectors and let us define a metric distance $d(\mathbf{x}, \mathbf{y})$ in $\mathbb{R}^n$. Initially, K n-dimensional codevectors are given or randomly determined. The aim is to optimise the definition of these prototypes in such a way that they reflect the statistical distribution of the training data set.

Given an initial dictionary $\mathscr{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ of K n-dimensional codevectors, the following two-stage iterative algorithm is used:

- **Nearest Neighbour:** the regions or clusters $R_i, i = 1, ..., K$, also called *Voronoi* regions, are defined by:

$$R_i = \{\mathbf{x}_n \in \mathscr{X} : d(\mathbf{x}_n, \mathbf{c}_i) \leq d(\mathbf{x}_n, \mathbf{c}_j); i \neq j\}, \qquad 1 \leq i, j \leq K$$

- **New Prototypes:** at each iteration and for each cluster R_i containing l_i vectors, the new prototype $\mathbf{c}_i$ is calculated using:

$$\mathbf{c}_i = \frac{1}{l_i} \sum_{k=1}^{l_i} \mathbf{x}_k, \quad 1 \leq i \leq K$$

The choice of a relevant distance d used for the new prototype calculation and the nearest neighbours determination is one of the main algorithm parameters. A commonly used metric is the Euclidean distance defined for two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ by:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} ||x_i - y_i||^2 \tag{4.1}$$

The two-stage procedure described above is then repeated until the average total distortion D^* defined as

$$D^* = \frac{1}{N} \sum_{i=1}^{N} \min_{1 \leq k \leq K} d(\mathbf{x}_n, \mathbf{c}_k) \tag{4.2}$$

does not significantly change between two successive iterations, so that the algorithm reaches a local minimum.

In the following section, we describe how the K-means can be used to build models representative of a training data set. Details about the training (instrument modelling) and the testing (instrument identification) procedures are given.

4.2.2 Training phase

During the training phase, a dictionary $\mathscr{C}$ of K prototypes is optimised for each instrument in the database. In our system, the method based on the LBG (Linde-Buzo-Gray, or generalised Lloyd) algorithm described in [LBG80] and [GG99] has been considered.

The algorithm requires an initial codebook obtained by the splitting technique: starting with one codeword (the mean of the entire data set), each vector in the dictionary is iteratively *split* into two vectors until the closest inferior power of two of K is reached. Finally, the maximally populated cluster is split into two and the iterative two-stage optimisation is performed. The process is repeated until the desired number of centroids is reached. In our implementation, splitting is performed by adding a small perturbation $\epsilon = 0.01$ proportional to the standard deviation of the regions to the vector coordinates in each direction. The procedure is detailed below:

1. Given a training sequence $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of N observations, ϵ and η to be small numbers.

2. Let $K = 1$ be the initial number of codevectors, $l = 0$ be the iteration index and $\mathbf{c}_1$ the first codevector defined as the mean of the entire dataset $\mathscr{X}$:

$$\mathbf{c}_1^* = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

   The average total error calculated using the Euclidean distance defined in Eq. (4.1) can be written as:

$$D^* = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{c}_1^*||^2$$

   D^* corresponds to the sum of the distances of each training data to the first codevector.

3. The splitting technique is then applied. For $k = 1, ..., K$, the new codevectors are calculated using:

$$\begin{aligned} \mathbf{c}_k^{(0)} &= (1+\epsilon)\mathbf{c}_k^* \\ \mathbf{c}_{K+k}^{(0)} &= (1-\epsilon)\mathbf{c}_k^* \end{aligned}$$

   and the number of codevectors is now doubled, yielding $K = 2K$.

4. The codevectors are now optimised. The average error is $D^{(l)} = D^*$.

   i. Each training vector is assigned to its closest centroid. In other words, each vector $\mathbf{x} \in \mathscr{X}$ is approximated by the closest codevector $\mathbf{c}_{k^*}^l \in \mathscr{C}$:

   $$\mathscr{Q}(\mathbf{x}_i) = \mathbf{c}_{k^*}^l, \quad 1 \leq i \leq N$$

   where

   $$k^* = \arg \min_{1 \leq k \leq K} ||\mathbf{x}_i - \mathbf{c}_k^l||^2, \quad 1 \leq i \leq N$$

   Note that $\mathscr{Q}$ can be seen as a quantiser.

   ii. The codevectors are then updated using:

   $$\mathbf{c}_k^{(l+1)} = \frac{\sum_{\mathscr{Q}(\mathbf{x}_i)=\mathbf{c}_k^{(l)}} \mathbf{x}_i}{\sum_{\mathscr{Q}(\mathbf{x}_i)=\mathbf{c}_k^{(l)}} 1}, \quad 1 \leq k \leq K$$

   In other words, each new codevector k is the mean of all the feature data in the kth region.

   iii. The iteration counter is updated: $l = l + 1$

   iv. The new total average error becomes

   $$D^{(l)} = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathscr{Q}(\mathbf{x}_i)||^2$$

   v. If $(D^{(i-1)} - D^{(i)})/D^{(i-1)} > \eta$, the algorithm has not met the stop criterion and the steps **i–iv** are repeated.

   vi. A local minimum is reached and the final optimised codevectors are:

   $$\mathbf{c}_k^* = \mathbf{c}_k^l, \quad 1 \leq k \leq K$$

   The final total average error becomes

   $$D^* = D^{(l)}$$

5. Steps (3) and (4) are repeated until the desired number of codevectors is obtained.

After this optimisation procedure, the training data set $\mathscr{X}$ of N feature vectors is represented by a dictionary $\mathscr{C}$ of K codevectors.

4.2.3 Identification phase

The I instruments in the database are represented by their codebooks $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_I$ containing K codevectors each. In the following, it is assumed that the unknown instrument to identify is represented by an observation $\mathscr{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\}$ of feature vectors.

The use of K-means classifiers for pattern recognition and identification commonly involves the minimisation of a similarity measure between the unknown observation and the models. The choice of a relevant *metric* ideally reflects the subjectivity inherent to the task of evaluating similarities between sets of acoustic features. In contrast to other machine learning algorithms presented in this thesis, the choice of a proper metric is essential for the system's robustness during the identification procedure.

For this purpose, two types of similarity measure can be distinguished: the minimum distance similarity measure and the similarity evaluation between distributions. In both cases, the identity of an unknown excerpt is the identity corresponding to the model in the database that minimises the measure between the observation and all the models in the database. Mathematically, a minimum distance classifier allows to retrieve the identity of the unknown observation by finding I^* such that

$$I^* = \arg \min_{1 \leq i \leq I} \mathscr{D}(\mathscr{Y}, \mathscr{C}_i) \tag{4.3}$$

where $\mathscr{D}(\mathscr{Y}, \mathscr{C}_i)$ is a similarity measure between the observation $\mathscr{Y}$ and one codebook $\mathscr{C}_i$ in the database. The identity of the observation is the one of I^*. Since the observation $\mathscr{Y}$ is usually composed of several feature vectors, $\mathscr{D}(\mathscr{Y}, \mathscr{C}_i)$ is usually expressed as

$$\mathscr{D}(\mathscr{Y}, \mathscr{C}_i) = \frac{1}{M} \sum_{j=1}^{M} \min_{1 \leq k \leq K} d(\mathbf{y}_j, \mathbf{c}_{k,i}) \tag{4.4}$$

where $\mathbf{c}_{k,i}$ corresponds to the kth codeword of the ith instrument model. In the following, various metrics that are commonly used in pattern recognition problems are reviewed.

4.2.3.1 Euclidean distance

Let us consider for simplification $\mathbf{y}$ as being an observation vector and $\mathbf{c}$ as being one codevector of a particular instrument model. Using the Euclidean distance defined in

Eq. (4.1), $d(\mathbf{y}, \mathbf{c})$ can be written:

$$d(\mathbf{y}, \mathbf{c}) = \sum_{i=1}^{n} ||y_i - c_i||^2 \tag{4.5}$$

4.2.3.2 Mahalanobis distance

A similar approach has been considered in [SRRJ87] in a speaker verification context. In this case, the *Mahalanobis* distance $d_m(\mathbf{y}, \mathbf{c})$ which uses inverse covariance weighting is defined as:

$$d_m(\mathbf{y}, \mathbf{c}) = (\mathbf{y} - \mathbf{c})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{c}) \tag{4.6}$$

where $\mathbf{R}$ is the pooled intra-instrument covariance matrix. Under the assumption that the covariance matrix $\mathbf{R}$ is diagonal, the distance can be rewritten as:

$$d_m(\mathbf{y}, \mathbf{c}) = \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{c}_i)^2 v_i^{-1} \tag{4.7}$$

with v_i being the elements of the diagonal covariance matrix.

4.2.3.3 Similarity measurement between two codebooks

In order to take advantage of the whole observation during the identification stage, the K-means algorithm is used to build a codebook $\tilde{\mathscr{C}} = \{\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \ldots, \tilde{\mathbf{c}}_K\}$ which statistically represents $\mathscr{Y}$. The identification process consists of evaluating the similarity between two distributions represented by two dictionaries. To this effect, the definition of a similarity metric between two codebooks $d(\tilde{\mathscr{C}}, \mathscr{C})$ is introduced in the following. Considering that:

1. Running the K-means algorithm twice on the same data set does not ensure that after each run the codewords are similarly ordered.[4] In other words, it is not guaranteed that
$$\sum_{k=1}^{K} d(\tilde{\mathbf{c}}_k^{t_1}, \tilde{\mathbf{c}}_k^{t_2}) = 0 \tag{4.8}$$
where t_1 and t_2 indicate two runs of the algorithm on the same training data.

2. In the case where $\tilde{\mathscr{C}}$ contains K times the same codevector identical to one codevector of $\mathscr{C}$, the measure should verify $d(\tilde{\mathscr{C}}, \mathscr{C}) = 0$.

[4] note that this is not valid if an initialisation using splitting is performed

We propose the following similarity measure between the codebook $\tilde{\mathscr{C}}$ and one codebook $\mathscr{C}_i$ taken from the database:

$$d(\tilde{\mathscr{C}}, \mathscr{C}_i) = \sum_{k=1}^{K} \left[\min_{1 \leq k \leq K} d(\tilde{\mathbf{c}}_k, \mathbf{c}_{i,k}) \right] \tag{4.9}$$

in which d is the Euclidean distance defined in Eq. (4.1). The similarity measure verifies: $d(\tilde{\mathscr{C}}, \mathscr{C}) \geqslant 0$ and $d_{\mathscr{C},\mathscr{C}} = 0$ but is not symmetric.

Finally, the identity of the unknown codebook is retrieved by finding I^* such that:

$$I^* = \arg \min_{1 \leq i \leq I} d(\tilde{\mathscr{C}}, \mathscr{C}_i) \tag{4.10}$$

4.2.4 Learning using K-means

In this section, we interpret and illustrate the process of learning characteristic spectral shapes using K-means. The LSF are the considered features throughout.

The K-means is extensively used for the vector quantisation of the LSF in speech coders [PA93]. The localised spectral sensitivity property they exhibit (section 3.2.2) make them suitable to be used together with an averaging iterative algorithm.

In the case of musical instrument identification, an interesting interpretation of the modelling process is concerned with the determination of *average* or *characteristic* spectral shapes of each instrument. In [RW82], it is outlined that:

> [...] *but since different instruments had different average spectra, it was believed that this difference in average spectrum was utterly responsible for timbre differences.*

Transposing this assumption to a computer algorithm, it can be argued under the condition that the training data set contains sufficient and typical waveforms, the K-means optimisation can result, to a certain extent, in the definition of such characteristic spectral shapes. Relating this interpretation to the *formant theory of timbre* (see section 1.3.1.3), it can reasonably be advanced that a dictionary of codevectors can be used to characterise an instrument and therefore its timbre.

The advantages of the method in terms of learning power and data reduction (as opposed to non-parametric techniques such as the k-NN, for example) can also be highlighted. Similar techniques have been experimented in speaker identification frameworks, for example in [RS86].

During the training phase, the models corresponding to a dictionary of codevectors are built for each instrument in the database. For a number of codewords specified *a-*

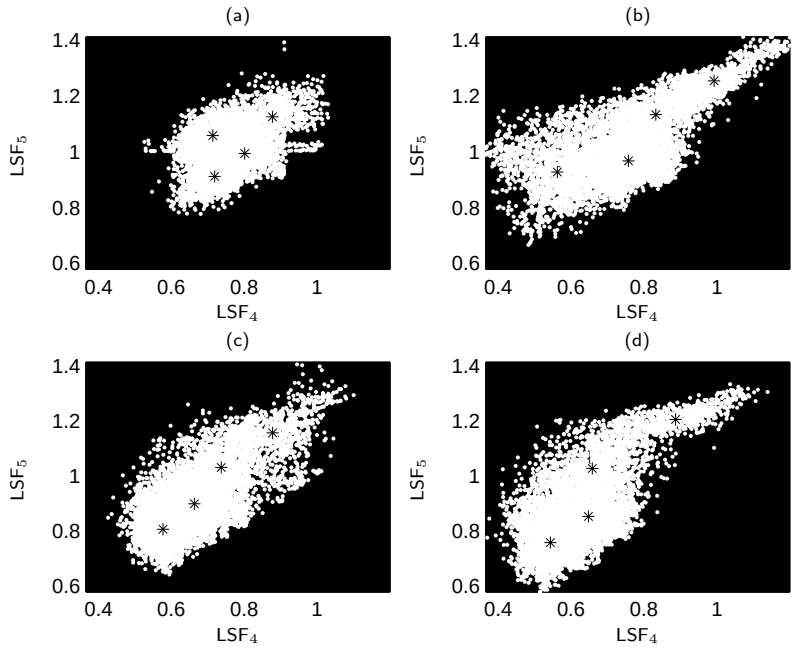


Figure 4.2: Optimised codebooks using the iterative K-means algorithm. Data (white dots) are two-dimensional LSF feature vectors calculated using two minutes of monophonic recordings. Black stars represent the four optimised codevectors. (a) cello, (b) clarinet, (c) flute and (d) piano.

priori[5], the K-means is run until a local minimum is reached. The process is illustrated in figure 4.2. For ease of representation, two-dimensional LSF feature vectors (white stars) extracted after 12th order linear predictive analyses have been applied on the waveforms are considered. The process is represented for four instruments, the cello, the clarinet, the flute and the piano. Black stars correspond to four codevectors obtained after optimisation using K-means. They optimally represent the feature data by minimising the total error between the data and their respective codewords.

It has been mentioned in section 3.2.2 how the LSF parameters can be related to an IIR filter frequency response. Therefore, any given LSF vector can be associated with a unique spectral envelope. Hence, codevectors generated from K-means optimisation using LSF can be regarded as characteristic spectral shapes of the data training set. As an illustration, 12 spectral envelopes corresponding to codebooks of 12 LSF vectors are plotted in figure 4.3 for the same four considered instruments as in figure 4.2.

[5]note that this number is usually empirically determined

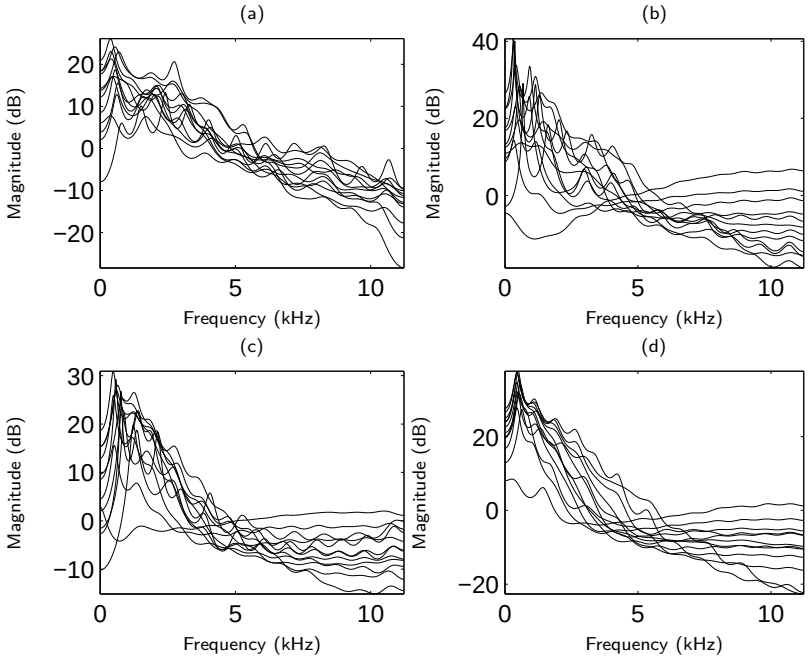


Figure 4.3: 12 IIR filter frequency responses evaluated from 12 optimised codevectors using LSF and the K-means algorithm. Features extracted from two minutes of monophonic recordings have been used to train the models. Instruments are (a) cello, (b) clarinet, (c) flute and (d) piano.

4.3 Gaussian Mixture Models (GMM)

In contrast to the K-means which clusters the n-dimensional space into regions with *hard* boundaries, a Gaussian Mixture Model assigns a belonging factor for each data point to each Gaussian in the mixture. As a result, a GMM can be seen as a *soft* or *fuzzy* vector quantiser. It is assumed that each training data vector is generated from a pool of Gaussians having fixed mixture weights.

The parameter estimation of a GMM as well as the identification procedure are reviewed in this section. This system is the one described in [RR95] in a speaker identification framework.

4.3.1 Principle

A GMM models the probability density function of an observed n-dimensional feature vector $\mathbf{x}$ by a multivariate Gaussian mixture density

$$p(\mathbf{x}|\Lambda) = \sum_{k=1}^{K} w_k \Phi_k(\mathbf{x})$$

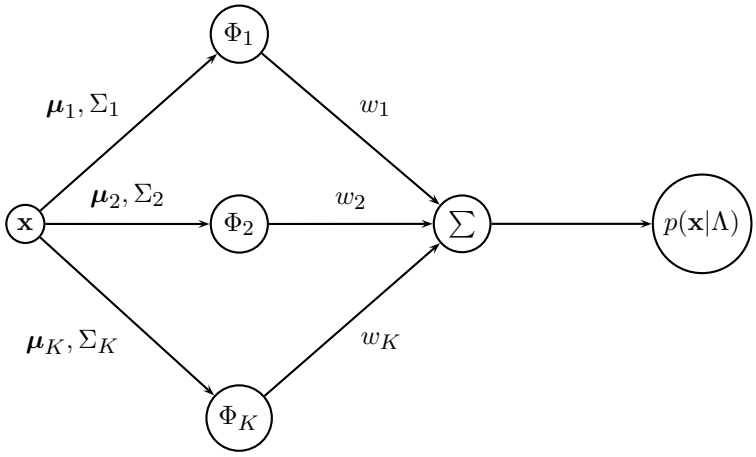


Figure 4.4: Graphical representation of a mixture of K Gaussians.

where K is the number of Gaussian components and w_k, $k = 1, ..., K$, the mixture weights with the constraint $\sum_{k=1}^{K} w_k = 1$. Further, each component density Φ_k, $k = 1, ..., K$, is a function of the form

$$\Phi_k(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k))$$

In a recognition system, each instrument in the database is represented by a GMM Λ entirely defined by the mean vectors $\boldsymbol{\mu}_k$, covariance matrices Σ_k and weights w_k noted

$$\Lambda = \{\boldsymbol{\mu}_k, \Sigma_k, w_k\}, \quad k = 1, \ldots, K$$

It is further assumed that each Gaussian in the model has a diagonal covariance matrix. The use of diagonal covariance matrices provides a good compromise between modelling power and algorithm complexity compare to a full covariance GMM.

4.3.2 Training phase

The aim of the training phase is to determine the model parameters Λ for each instrument in the database. The initialisation of the mean vectors $\{\boldsymbol{\mu}_k\}_{k=1,...,K}$, is performed using the K-means algorithm described in section 4.2 while the Gaussian mixture parameters are determined using the iterative Expectation Maximisation (EM) method as detailed in [RR95].

Given a set of N training feature vectors for one instrument, the objective is to

estimate the parameters of the GMM Λ that would best represent the feature data set distribution. This can be performed by maximising the likelihood $\mathscr{L}$ of the training data $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ given the model Λ:

$$p(\mathscr{X}|\Lambda) = \prod_{i=1}^{N} p(\mathbf{x}_i|\Lambda) = \mathscr{L}(\Lambda|\mathscr{X})$$

Finding the optimum model Λ^* can be mathematically written:

$$\Lambda^* = \arg\max_{\Lambda} \mathscr{L}(\Lambda|\mathscr{X})$$

The difficulty to attain Λ^* depends on the form of $p(\mathbf{x}|\Lambda)$. In the case of multivariate Gaussian mixture models, the EM algorithm originally presented by Dempster et al. in 1977 [DLR77] is used. The iterative procedure can be summarised as follows:

1. Choose an initial model Λ by initialising the means using a K-means algorithm, the variances and weights to unity.
2. Determine a new model $\tilde{\Lambda}$ so that $p(\mathscr{X}|\tilde{\Lambda}) > p(\mathscr{X}|\Lambda)$.
3. Repeat the step above until $p(\mathscr{X}|\tilde{\Lambda}) - p(\mathscr{X}|\Lambda)$ is above a certain threshold or if the required number of iterations has been reached.

The reader is referred to [DLR77] and [Bil97] for a detailed description of the EM algorithm, and especially of the equations ensuring a monotonic increase in the model's likelihood value.

4.3.3 Identification phase

The I instruments in the database are represented by their GMM $\Lambda_1, \Lambda_2, ..., \Lambda_I$. The identity of an unknown excerpt is the identity corresponding to the model that maximises the *a-posteriori* probability for the given observation sequence $\mathscr{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\}$. It can be mathematically written:

$$I^* = \arg\max_{1 \le i \le I} p(\Lambda_i|\mathscr{Y}) \tag{4.11}$$

Due to Bayes's rule

$$p(\Lambda_i|\mathscr{Y}) = \frac{p(\mathscr{Y}|\Lambda_i)p(\Lambda_i)}{p(\mathscr{Y})}, \quad i = 1, ..., I$$

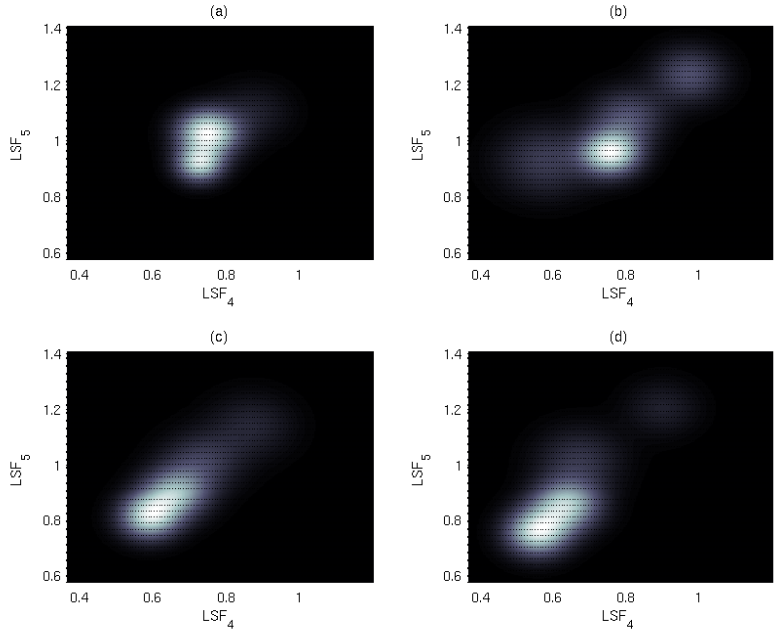


Figure 4.5: Two-dimensional density modelling using a mixture of four Gaussians. Training data set correspond to two minutes of monophonic recording. (a) cello, (b) clarinet, (c) flute and (d) piano.

and assuming equally likely instruments $p(\Lambda_i) = 1/I$, it comes

$$p(\Lambda_i|\mathscr{Y}) = \frac{p(\mathscr{Y}|\Lambda_i)}{Ip(\mathscr{Y})}, \quad i = 1, ..., I$$

Finally, $p(\mathscr{Y})$ is the same for all instrument models. Therefore, Eq. (4.11) becomes

$$I^* = \arg \max_{1 \leq i \leq I} p(\mathscr{Y}|\Lambda_i)$$

The probability $p(\mathscr{Y}|\Lambda_i)$ can be expressed in the case of i.i.d. by:

$$p(\mathscr{Y}|\Lambda_i) = \prod_{j=1}^{M} p(\mathbf{y}_j|\Lambda_i)$$

4.3.4 Learning using GMM

We describe in this section the process of building instrument models using a GMM. The EM algorithm implementation used is based on the one available in [VOI]. When the EM algorithm is used to determine *optimum* instrument models, the following three steps have to be considered:

- **Choice of** K: the determination of an optimum number of Gaussian components is usually experimentally driven. Choosing too few components produces models which are not specific enough to characterise the inter-instrument variability. On the other hand, choosing too many components can decrease the performance by introducing singularities in the models.

- **Algorithm initialisation**: the system has to be initialised prior to the EM algorithm with an initial model Λ_0. The K-means is used to perform a clustering of the training data into K classes. The mean vectors $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ for the K Gaussians are firstly determined. They correspond to the codewords obtained after the K-means optimisation. Next, the K initial mixture weights are set to unity and each Gaussian component is initialised such that it has unit variance in each direction.

- **EM algorithm**: while training a nodal variance GMM, the variances can become very small and degrade the classification performance by introducing singularities in the model. This is particularly true for a mixture model with a large number of components compared to the number of training data.[6] In order to avoid such a situation, a prior is set on the variances at each EM iteration:

$$\tilde{\sigma}_i^2 = \begin{cases} \tilde{\sigma}_i^2 & \text{if } \tilde{\sigma}_i^2 \geq \tilde{\sigma}_{min}^2 \\ \sigma_{min}^2 & \text{if } \tilde{\sigma}_i^2 < \sigma_{min}^2 \end{cases}$$

Figure 4.5 illustrates the two-dimensional feature vector PDF modelling using a mixture of four Gaussians. Data is extracted from a 12th order LP analysis and consists of the same LSF feature vectors that have been considered in figure 4.2. The two-dimensional feature space is modelled according to a continuous density distribution generated by the mixture of four Gaussians. Note that in order to avoid singularities during the EM optimisation, a variance limiting prior ($\sigma_{\min}^2 = 0.01$) has been set at each iteration of the EM algorithm: the EM optimisation stopped when the relative increase of the log likelihood $\log p(\mathscr{X})$ between each iteration fall below 0.1%.

[6] this is actually similar to the case where running a K-means algorithm, only a few data points are associated to a given centroid

4.4 Support Vector Machines (SVM)

Support Vector Machines are becoming increasingly popular for data classification and pattern recognition problems. Their use in a musical instrument identification context has been studied in [MM99] and [GR04] for pitched and percussive musical instruments respectively. They belong to the class of discriminative methods and can be used for classification and regression problems. In numerous cases, it has been observed that SVM generalisation performance either match or were significantly better than that of competing methods. Their application for classification tasks is described in this section.

4.4.1 Principle

Support Vector Machines perform classification by constructing an n-dimensional hyper-plane that optimally separates a labelled data set into two distinct classes. The optimum hyper-plane is the one that maximises the margin between the feature data corresponding to each class.

Given a training data set $\mathbf{x}_k, k = 1, \ldots, m$ and a vector of labels $\mathbf{y}$ such that $y_k \in \{-1, +1\}, k = 1, \ldots, m$, a SVM searches for the hyper-plane $\mathbf{w}.\mathbf{x} + b = 0$ verifying:

$$y_k(\mathbf{x}_k.\mathbf{w} + b) - 1 \geq 0, \quad k = 1, \ldots, m \tag{4.12}$$

For an instance $\mathbf{x}$ to be classified, the decision function is:

$$f(\mathbf{x}) = \text{sgn}(\mathbf{x}.\mathbf{w} + b), \tag{4.13}$$

The reader is referred to [Vap95] and [CL01] for more information about the SVM theoretical principles.

In figure 4.6 are illustrated two cases: the case of linearly separable data (figure 4.6(a)) and the case non-linearly separable data (figure 4.6(b)). Circled markers are the support vectors. A SVM finds a hyper-plane that maximises the margin m between the two classes (figure 4.6(a)). In the case of non-linearly separable data, rather than fitting a non-linear curve as shown in figure 4.6(b), SVM use a kernel function to map the data into a higher-dimensional where the classes become linearly separable [Bur98].

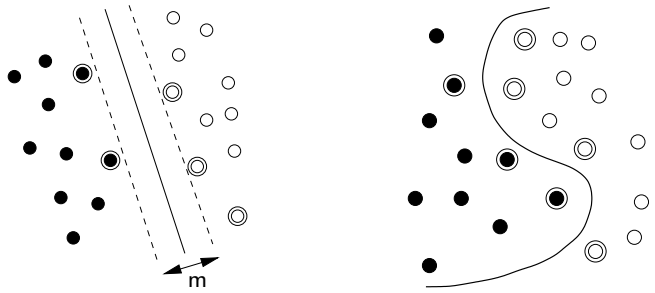


Figure 4.6: Classification using SVM. Illustration of (a) a linearly and (b) a non-linearly separable cases. Circled markers are the support vectors. Rather than fitting non-linear curves to the data, SVM use a kernel function to map the data into a higher-dimensional space where a hyper-plane is used to separate the classes.

4.4.2 Extension to multi-class problems

Any classification problem involving several classes can be decomposed as a combination of elementary binary classifiers. A *one-against-one* approach [CL01] has been used in our system. Specifically, it consists of training a SVM for each class against all the others in the database. A total of $I(I-1)/2$ binary classifiers, where I is the number of classes, are constructed. During the identification, each observation is successively classified by all the binary classifiers to reach a final decision.

4.4.3 Testing and classifying

During the identification stage, each individual frame $\mathbf{y}$ of $\mathscr{Y}$ is tested against the model which returns a possible identity. The final identity of the observation $\mathscr{Y}$ is the one that has been the most often retrieved over the M tested frames.

4.4.4 Classification using SVM

In this section, an example of classification is given. Several kernels that can be used are presented.

In order to be able to separate non-linearly separable data, a kernel function k is used to map the n-dimensional input vector into a higher dimensional space where the classes becomes linearly separable. There exists several kernels to be used with SVM, including [CL01]:

- Linear kernel:

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y} \tag{4.14}$$

- Polynomial kernel:

$$k(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^T \mathbf{y} + a)^d, \quad \gamma > 0 \tag{4.15}$$

  where γ and a are the kernel parameters and d the polynomial degree respectively.

- Radial Basis Function (RBF):

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma ||\mathbf{x} - \mathbf{y}||^2), \quad \gamma > 0 \tag{4.16}$$

  where γ is set to $1/I$, I being the number of instruments in the database.

In figure 4.7, a classification problem in the binary case using three different kernels is illustrated. Two sets of two-dimensional data belonging to two classes are represented by dots and plus markers respectively. After a SVM optimisation, a hyperplane optimally separating two regions (white and black respectively) is determined. Note that depending on the kernel being used, the boundary separating the two classes is *adapted* to the feature data distributions.

Chapter summary

In this chapter, the theoretical principles of three machine learning algorithms have been outlined. Our primary aim was not to perform a thorough review of the available methods but to limit the scope of applications associated with each of them. In particular, the distinction between techniques used for identification problems (i.e. instrument modelling) and for classification tasks (i.e. database modelling) has been made.

The emphasis has been on the K-means and on the interpretation of learning characteristic spectral shapes for each instrument. This process has been related to the *formant theory of timbre* and it has been argued that this approach could serve to model salient spectral characteristics of musical instrument sounds, and to a certain extent, their timbres. Several similarity measures have been presented. We have introduced a codebook to codebook similarity measure designed to take advantage of the whole observation prior to the identification phase.

The principles of GMM have been recalled. The combination MFCC/GMM being a classical approach in audio pattern recognition, it will be used as a reference to compare our system to.

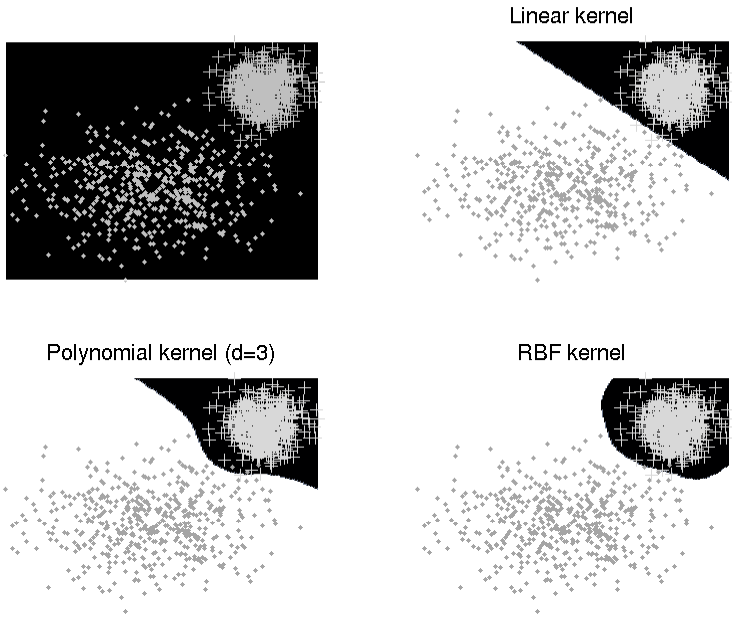


Figure 4.7: Illustration of using a linear, polynomial and RBF kernels for a binary classification task. Clockwise from top left: two-dimensional data classified into two classes (dots and plus markers respectively), classification using a linear kernel, classification using a polynomial kernel of degree 3 and classification using a radial basis function respectively.

Finally, SVM have been presented. They will be used for classifying and building database models. It has been described how a binary classification problem can be transposed to deal with multi-class problems. The properties of several kernels have been illustrated.

In the next two chapters, it is described how acoustic timbral descriptors and machine learning algorithms can be combined to build automated musical instrument identification systems. Our approach is evaluated for the tasks of identifying and classifying isolated notes and melodic phrases respectively.

5. Recognition of isolated notes

This chapter is concerned with the performance evaluation of the supervised systems whose three processing layers have been described in chapters 3 and 4. Now, the task of identifying and classifying isolated notes, i.e. notes taken independently of any *musical* context, is considered.

The use of isolated notes provides an interesting experimental environment for several reasons. Firstly, with regard to existing perceptual studies on the perception of timbre by humans, such as the ones described in section 2.1, which mostly considered isolated tones as stimuli. Considering isolated notes for computer simulations allows a direct comparison between humans and algorithm performance. Secondly, because a database of isolated notes provides various ways of fine tuning experimental protocols.

In section 5.1, information about the database used in the experiments are given while in section 5.2, the feature extraction procedure is detailed.

Our base system is presented in section 5.3. It consists of a mono-feature system, whereby a single type of feature is used to build models of instruments. In essence, these experiments attempt to confirm the theory stating that timbral information can be efficiently modelled using sound spectral envelopes and formant structures. Specifically, the Line Spectrum Frequencies are the considered features throughout. Results of a comparative study involving several acoustic descriptors are summarised in section 5.3.3.

In section 5.4, this base system is compared to a classification approach using Support Vector Machines (SVM). This will serve to define an upper limit in terms of achievable performance.

Next, the emphasis is on the study of pitch and timbre in automated musical instrument identification algorithms. After having illustrated the dependence of our models of instrument upon pitch, we propose in section 5.5 to use the pitch as a prior for both the modelling and identification phases. This strategy avoids the comparison

Instrument	Instances	No. of notes
bassoon	*pp*, *mf*, *ff*, Vib.	235
oboe	*pp*, *mf*, *ff*	199
clarinet	Eb, Bb	364
flute	Bass, Alto	278
sax	Alto	286
trombone	Bass, Tenor	307
trumpet	Vib, no Vib.	310
cello	*pp*, *mf*, *ff*, Vib.	472
viola	*pp*, *mf*, *ff*	438
violin	*pp*, *mf*, *ff*	403
TOTAL		3292

Table 5.1: Instances and number of notes per instrument in the database that have been retained for the experiments. Abbreviations *pp*, *mf*, *ff* and Vib. stand for *pianissimo, mezzo-forte, fortissimo and vibrato* respectively. Bass, Tenor and Alto are different instrument frequency ranges. Eb and Bb are two different clarinet registers.

between features and models being too *distant* in pitch.

In section 5.6, we propose a computer implementation of the perceptual experiments conducted by Berger [Ber63]. It is shown that our approach saves on the need to rely on a pre-segmentation onset/steady-state segments of sound and naturally incorporates specific acoustic information about the attacks of notes.

Finally, a perceptually motivated feature extraction algorithm is presented in section 5.7. In essence, we propose to use the standard ISO/MPEG psycho-acoustic model described in appendix A to select relevant partials in the spectra. The latter are then used to resynthetise the waveform from which the LSF are extracted. The performance of this novel feature, the PLSF, are evaluated.

5.1 Database

The choice of instruments to be considered for the experiments is important since it is in practice not possible to evaluate a system using all the existing musical instruments. For this reason, a subset of instruments is often chosen, generally among the class of orchestral musical instruments, because several databases are freely or commercially available to researchers. As a result, although the generalisation of the system's performance to other instruments has to be performed with care, general trends and particular system properties can nevertheless be studied in such limited scale experiments.

In this thesis, we will consider a subset of orchestral acoustic instruments, belong-

ing to the bowed string, wind and brass families, reflecting both the choice of instruments made by other researchers and the amount of data at our disposal. Recordings are of reasonable quality, originally sampled at 44100 Hz, and of sufficient number to represent the classes they correspond to. On average, notes are 5 seconds in duration. Details about the database are given in table 5.1 where in particular, the number of notes and the type of variations for each class are reported. The database consists of a combined subset of the IR–IOWA [The] and the RWC [RWC] isolated notes collections. It contains 3292 tones recorded at different loudnesses (*pianissimo*, *fortissimo*, ...), for several playing styles (vibrato, no vibrato), thus covering a wide range of timbre variation among each considered class. Samples have been recorded in anechoic chambers and no audio effects have been applied on the recordings.

The following 10 instruments have been retained for the experiments: bassoon, clarinet, flute, oboe, sax, trombone, trumpet, cello, violin and viola. Care has been taken to include several instruments within each family so that intra- and inter-families correct and incorrect identification rates can be studied. Moreover, an extensive pitch range is also represented with tones having their MIDI numbers ranging between 32 and 92 (corresponding roughly to 52 Hz and 1660 Hz respectively).

High-quality audio signals are commonly sampled at 44100 Hz. Such relatively high frequency resolution can result in a considerable and unnecessary amount of processing during the feature extraction and instrument modelling stages. Added to the fact that the contribution of the very high frequency bands in the perception of sounds is relatively weak (see section 1.1.2), all the signals were re-sampled at 22050 Hz prior to any processing. Similar resampling operations are commonly performed in the literature.

5.2 Feature extraction

Features extraction is performed on signals sampled at 22050 Hz. Silence and low level segments are firstly discarded. Only frames having an average energy level above -90 dB are retained. Next the pre-processing chain depicted in figure 3.1 is used: the DC bias is removed using a first-order IIR high-pass filter of frequency response $H(z) = (1 - z^{-1})/(1 - 0.999z^{-1})$. Amplitudes are then normalised to the 0 dB level over the whole tone duration, and a pre-emphasis ($H(z) = 1 - 0.97z^{-1}$) is used to increase the relative high frequency energy components. This step is particularly useful when LP-based models are used as it helps the algorithm to better pick the envelope high frequency structure. Figure 5.1 shows the frequency responses of both

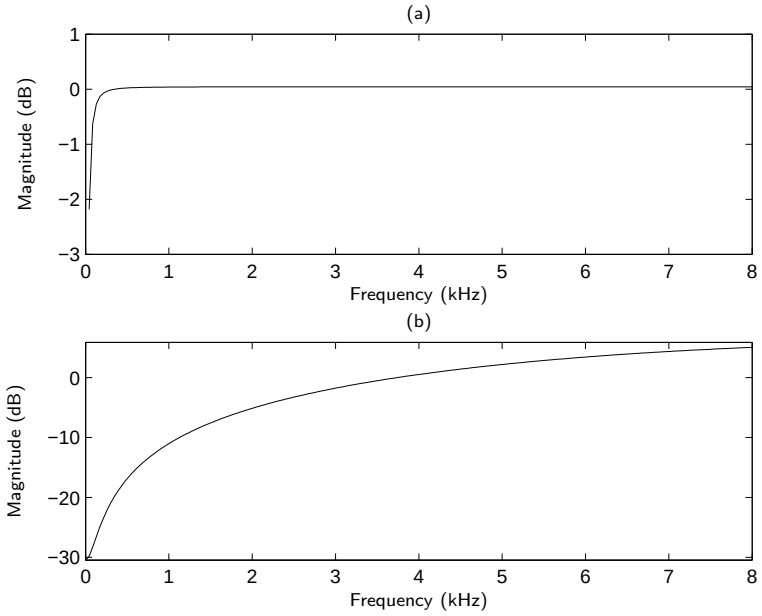


Figure 5.1: Frequency responses of (a) the high-pass filter $(1 - z^{-1})/(1 - 0.999z^{-1})$ and (b) the pre-emphasis filter (bottom) $1 - 0.97z^{-1}$ used to pre-process the waveforms prior to the feature extraction.

notch and pre-emphasis filters respectively.

Features are then extracted every 17 ms within frames of 23 ms in duration that have been previously weighted by a Hanning window. The LSF are calculated using the technique described in [KR86] after having determined the LP filter coefficients using the Levinson-Durbin algorithm (see section 3.2.1).

5.3 Instrument modelling

This section is concerned with the description and evaluation of the base system [CDS05b]. It consists of extracting a single type of features which are then used either with a K-means algorithm, either in a GMM framework to build instrument models, as described in sections 4.2.4 and 4.3.4 respectively. In the following, a first series of experiments evaluates the system's performance as a function of the prediction order and the type of classifiers used. Several similarity measures between the unknown feature data set and the models in the database are further tested.to compare the systems, similar experiments have

In practice, models have been trained for each instrument using 50% of the avail-

able data while 50% were retained for the testing. Experiments were repeated 3 times with different training and testing sets randomly chosen. The identification rates were accordingly averaged. Note that as opposed to [CDS05b], training and testing data sets are similar across experiments.

Comparative performance are summarised in table 5.2. The rows in each table show different prediction orders, ranging from 8 to 40 while the columns show the number of clusters, ranging from 8 to 64 (for both GMM and K-means). Note that the standard deviations reported in tables are given for information only and have to be interpreted with care, especially in terms of statistical confidence since only three runs have been performed in all our experiments. However, they are typical of the numbers that have been found during experiments. Consequently, they will not be reported in all tables.

5.3.1 Instrument identification

The performance of four systems using the LSF are reported in table 5.2. For the GMM-based classifier, total average correct identification rates range between 44% and 62.5%. The best performance is obtained by using 24 LSF and a mixture of 16 Gaussians. The corresponding confusion matrix is shown in table 5.3. Individual correct identification rates range from 48.2% for the flute class to 80.3% for the bassoon. Most important confusions involve the pairs flute–bassoon (28.1% of the tested flute samples were identified as being bassoon), sax–violin (18.1%) and trumpet–viola (15.5% of the trumpet samples were identified as viola). Note that the confusion viola–violin (19.2% of the viola samples were mis-identified as being violin) is expected since viola and violin sounds have similar origins in terms of physical mechanisms of sound production.

When using the K-means and the Mahalanobis minimum similarity measure described in section 4.2.3.2, the average correct identification rates increased to a maximum of 78.8% (table 5.2(b)). This performance is obtained by using 16 LSF and dictionaries of 32 codewords for the models.

Overall, the best performance are obtained using the minimum distance classifier and the codebook to codebook similarity measure based on the Euclidean distance. Total average correct identification rates are reported in tables 5.2(c) and 5.2(d) respectively. Using the codebook to codebook distance, 24 LSF and 32 clusters yields 83.2% correct identification. The corresponding confusion matrix is reported in table 5.4. Individual correct identification rates range from 71.4% (clarinet) to 98.1%

(a) GMM

		No. of Gaussians			
		8	16	32	64
LSF order	8	59.4	60.3	60.0	58.0
	16	48.0	50.6	62.3	62.0
	24	56.0	**62.5**$\pm$2.1	55.4	60.3
	32	61.6	61.0	62.0	58.0
	40	44.0	56.0	55.0	56.7

(b) K-means - Mahanalobis similarity measure

		No. of clusters			
		8	16	32	64
LSF order	8	75.5	78.4	77.7	77.4
	16	76.7	75.8	**78.8**$\pm$1.8	75.6
	24	73.0	78.0	73.9	72.5
	32	69.2	72.7	69.9	67.3
	40	66.2	71.8	67.1	69.2

(c) K-means - Minimum distance similarity measure

		No. of clusters			
		8	16	32	64
LSF order	8	76.9	76.6	76.0	68.0
	16	74.7	74.4	75.6	76.0
	24	78.0	79.0	**83.0**$\pm$1.3	81.3
	32	77.0	80.7	80.1	79.0
	40	76.0	77.0	78.0	75.2

(d) K-means - Codebook to codebook similarity measure

		No. of clusters			
		8	16	32	64
LSF order	8	78.9	77.7	80.1	69.0
	16	75.1	74.0	81.8	77.7
	24	81.2	82.9	**83.2**$\pm$1.2	81.6
	32	77.5	81.1	81.8	80.9
	40	78.8	76.8	80.4	76.8

Table 5.2: Percentages of correct identification as a function of the prediction order (showed in rows) and the number of clusters (showed in columns) for four systems. Experiments have been repeated three times and involved similar training and testing data sets. Baseline performance are 10% in the case of random guesses.

(trombone). These systems show clear improvements compared to the previous ones using a GMM and the Mahalanobis distance respectively, thus revealing the advantage of using Euclidean-based similarity measure between sets of LSF vectors. Note that the difference in performance between the two systems (tables 5.2(c) and 5.2(d)) is not statistically significant. However, at this stage, a decision has to be taken so that the codebook to codebook similarity measure will be chosen for all experiments involving K-means, except when explicitly mentioned.

As a comparison, the performance of a *conventional* system using MFCC as features and a Gaussian Mixture Model as classifier are presented in table 5.5. Note that the first MFCC parameters c_0 have not been considered for building the feature vectors (see section 3.5.1.3). The best performance are achieved when 12 MFCC and 32 GMM are used. The corresponding confusion matrix is shown in table 5.6. This configuration allows slightly more than 74% of the tested samples to be correctly identified. Note that in contrast to [KS04], the MFCC performed better than the LSF when used with a GMM.

5.3.2 Family identification

In this section, we analyse the results from another angle. Confusion matrices for family identification are shown in tables 5.7(a), 5.7(b) and 5.7(c) respectively. Note that the organisation of the instruments into families corresponds to the one that has been used for the perceptual experiments carried out by [SSF02] and reported in section 2.1.1.2. Note that although being single-reed instruments, the sax and the clarinet are considered as two distinct instrument families.

For the GMM classifier (table 5.7(a)), the average correct rate in terms of family identification is 63.6%. The string family is the most correctly identified, with 80.6% of the cello, viola and violin samples being recognised as being strings. On the other hand, the flute and clarinet are correctly identified 48.2% and 56% of the time respectively. These rather poor performances are however better than random guesses. Note also that these three families only contain one instrument (clarinet, flute and sax) so that a mis-identification can only be in favour of another family. In terms of inter-family confusion, the wind instruments are highly confused with the strings. In particular, sax are mis-identified as being strings 32.8% of the time. This type of confusion is representative of the *non-meaningful* errors that an automated system should avoid.

The use of the K-means significantly improves the performance and nearly 85%

	bassoon	clarinet	flute	oboe	trombone	trumpet	sax	cello	viola	violin
bassoon	**80.3**	1.7	10.3				3.4	4.3		
clarinet	7.2	**56.0**	13.7	1.7			8.5	4.9	3.6	4.4
flute	28.1	3.6	**48.2**	0.7			2.2	10.8	2.2	4.2
oboe	3.0	5.1		**52.5**		5.1		2.0	14.1	18.2
trombone	2.5				**75.8**	4.2			10.8	6.7
trumpet		0.6		9.0	1.9	**62.6**	0.6	0.6	15.5	9.2
sax	4.9		5.7				**56.6**	8.4	6.3	18.1
cello	5.5	0.4	2.5		0.8	0.4	9.5	**70.3**	4.2	6.4
viola		4.1	2.3	2.7	3.2	3.2	5.9	8.3	**51.1**	19.2
violin		3.4	2.9	3.5		0.2	7.8	3.2	7.4	**71.6**

Table 5.3: Confusion matrix using a GMM-based classifier. 24 LSF and a mixture of 16 Gaussians have been used. This system exhibits rather poor performances with 62.5% average correct instrument identification. Identities of the presented samples are in rows while identities returned by the system are in columns. Baseline performance are 10% for 10 instruments in the database.

	bassoon	clarinet	flute	oboe	trombone	trumpet	sax	cello	viola	violin
bassoon	**91.5**	0.9	3.4		2.6			1.6		
clarinet	0.5	**71.4**	10.4	1.1	0.5	1.1	2.7	0.5	7.1	4.7
flute		7.9	**82.7**	2.2			1.4	0.7	0.1	5.0
oboe		1.2		**79.5**	2.0	7.1			4.1	6.1
trombone				0.6	**98.1**	1.3				
trumpet	2.0	1.3	0.6	3.9	7.1	**77.4**			3.9	3.8
sax		1.4	2.1	0.7	0.7		**85.3**	2.8	1.4	5.6
cello			2.1		1.7		0.8	**91.9**	1.7	1.8
viola	0.5	3.2	1.8	3.2	0.3		0.5	3.7	**78.1**	8.7
violin		2.1	1.0	1.0	1.0	1.2	2.4	6.9	8.3	**76.1**

Table 5.4: Confusion matrix using a K-means classifier and the codebook to codebook similarity measure based on the Euclidean distance. 24 LSF and 32 codewords have been used, yielding an average correct instrument identification rate of 83.2%. Identities of the presented samples are in rows while identities returned by the system are in columns. Baseline performance are 10% for 10 instruments in the database.

		No. of Gaussians			
		8	16	32	64
MFCC order	8	55.7	59.3	62.1	58.0
	12	67.8	71.1	**74.1**$\pm$1.9	70.4
	24	66.3	68.9	67.1	68.7
	32	55.6	58.4	55.0	51.0

Table 5.5: Percentages of correct identification as a function of the number of MFCC coefficients (showed in rows) and the number of Gaussians (showed in columns). Experiments have been repeated three times and concordantly averaged. Baseline performance are 10% in the case of random guesses.

of the test samples are now correctly identified as belonging to their respective family (table 5.7(b)). Specifically, the strings are on average the most correctly identified (92.4%), followed by the brasses (92%) and the double-reeds (85.5%). The clarinet is the least correctly identified with 71.4% of the tested samples but still 15% better than if a GMM was used. In this configuration, the winds are much less confused with the strings, with 12.3% of the clarinet samples recognised as being string. However, overall, the strings are still attracting the most confusions.

For the conventional system using 12 MFCC and 32 Gaussians (table 5.7(c)), 77.3% correct family identification can be achieved. Overall, this system exhibits performance between the ones shown in tables 5.7(a) and 5.7(b). Nevertheless, one can notice an improvement by 4% for the clarinet compared to the LSF/K-means system.

5.3.3 Comparison with other acoustic descriptors

In order to illustrate the advantages of using the LSF as features when a K-means is used, experiments involving other acoustic descriptors have been conducted. Specifically, this comparative study involved the Linear Predictive Coefficients (LPC), the reflection coefficients[1] (or PARCOR), the LSF and the Mel-Frequency Cepstrum Coefficients (MFCC). The performance have been evaluated as a function of the prediction order or number of MFCC parameters. Average individual and family identification rates are reported in figures 5.2(a) and 5.2(b) respectively.

In all experiments, a dictionary of 32 codewords has been used to build the instrument models and similarly to the previous experiments, three runs have been performed with different training and testing data sets. Note however that they are the same as in section 5.3.

[1] they are calculated using the Schur recursion [Eur00]

	bassoon	clarinet	flute	oboe	trombone	trumpet	sax	cello	viola	violin
bassoon	**85.5**	0.9	2.6	2.6	0.9				4.3	3.2
clarinet		**75.8**	4.4	7.7		1.6	7.1	1.1	1.1	1.2
flute		11.5	**64.7**	1.4		7.2	2.2	0.7	10.1	2.2
oboe		10.1	6.1	**63.6**		3.0		1.0	12.1	4.1
trombone	1.3		1.3	0.6	**88.5**	3.8		1.3	3.2	
trumpet				9.0	1.3	**75.5**	0.6	1.9	5.8	5.9
sax		4.9	5.6	1.4			**72.0**	4.2	7.7	4.2
cello		1.3	0.8			0.8	0.8	**82.6**	10.6	3.1
viola		2.3	4.6	2.3			2.3	14.1	**61.2**	13.2
violin		4.4	2.9	1.0		2.7	1.0	6.2	10.7	**71.1**

Table 5.6: Confusion matrix for a *conventional* mono-feature system using 12 MFCC as feature and a mixture of 32 Gaussians. This system yields 74.1% correct identification. Rows correspond to the identities of the presented samples while the identities returned by the system are shown in columns. Baseline performance are 10% for 10 instruments in the database.

(a) 24 LSF and 16 GMM

	strings	brasses	dble reed	clarinet	flute	sax
strings	**80.6**	2.6	3.9	2.6	2.6	7.7
brasses	21.4	**72.3**	5.8	0.3		0.3
db reed	19.3	2.6	**67.9**	3.4	5.2	1.7
clarinet	21.4		8.9	**56.0**	13.7	8.5
flute	17.2		28.8	3.6	**48.2**	2.2
sax	32.8		4.9		5.7	**56.6**

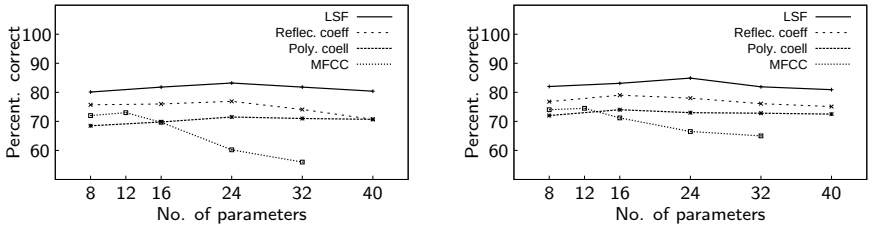
(b) 24 LSF and 32 Codevectors with codebook to codebook similarity measure

	strings	brasses	dble reed	clarinet	flute	sax
strings	**92.4**	1.4	1.6	1.8	1.4	1.2
brasses	3.9	**92.0**	3.3	1.3	0.6	
db reed	5.1	5.9	**85.5**	1.1	1.7	
clarinet	12.3	1.6	1.6	**71.4**	10.4	2.7
flute	5.8		2.2	7.9	**82.7**	1.4
sax	9.8	0.7	0.7	1.4	2.1	**85.3**

(c) 12 MFCC and 32 GMM

	strings	brasses	dble reed	clarinet	flute	sax
strings	**90.9**	1.2	1.1	2.7	2.8	1.4
brasses	9.1	**84.6**	5.5		0.7	0.3
db reed	12.4	2.0	**75.9**	5.5	4.4	
clarinet	3.4	1.6	7.7	**75.8**	4.4	7.1
flute	13.0	7.2	1.4	11.5	**64.7**	2.2
sax	16.1		1.4	4.9	5.6	**72.0**

Table 5.7: Confusion matrices corresponding to instrument family identification experiments. (a) 24 LSF and 16 GMM, (b) K-means with the codebook to codebook similarity measure based on the Euclidean distance and (c) 12 MFCC and 32 GMM. Percentages of average correct family identification are 63.6%, 84.9% and 77.3% respectively. Families of the presented samples are in rows while answers returned by the system are in columns.



(a) Instrument identification

(b) Family identification

Figure 5.2: Comparative performance for various acoustic descriptors, including the linear predictive coefficients, the reflection coefficients, the LSF and the MFCC. (a) average correct instrument identification and (b) average correct family identification. Experiments have been conducted with 10 instruments in the database.

It can be observed in figure 5.2(a) that for all the considered prediction orders, performance of the LSF are superior than that of the other linear predictive parameters and better than the MFCC respectively. For a prediction order of 24, the use of the LSF offers a gain of 6% over the PARCOR and more than 11% over the polynomial coefficients. Note the similar evolution of the performance as a function of the prediction order for the linear predictive parameters. In contrast, the best correct identification rate is obtained when 12 MFCC are used (73%). Note that this is consistent with what has been found in section 5.3 where the combination of 12 MFCC and 32 GMM yielded the best performance (74.1%).

Likewise, the total average performance in terms of family identification is increased by 6% and 9% over the PARCOR and LPC on the one hand and by 6% over the MFCC on the other hand (figure 5.2(b)).

At this stage, the following interpretations can be drawn. Using the LSF, information about the formant structures are directly input to the K-means, that is, the differences between pairs of LSF parameters, which can be related to high frequency regions within spectra (see figure 3.7), are captured by the models. On the other hand, although containing *spectral* information, the PARCOR and LPC on the one hand, and the MFCC on the other hand do not implicitly carry information about the formant structures. For this reason, it can argued that the LSF/K-means combination provides a meaningful and consistent implementation of the *formant theory of timbre*.

5.3.4 Summary

This first series of experiments has focused on the evaluation of use of the pair LSF/K-means in a musical instrument identification context. Overall, the LSF used together with the K-means performs better than with a GMM classifier. It has also been highlighted that the use of two similarity measures based on the minimisation of the Euclidean distance results in better performance than if a Mahalanobis distance was used. This confirms to a certain extent that the Euclidean distance is particularly well suited to be used with the LSF, and in particular that it provides a meaningful perceptual approach to the similarity evaluation between spectral envelopes. Note that in contrast to [KS04], it has been shown using our database that the combination LSF/GMM yielded worse performance than the combination MFCC/GMM.

Next, a preferred prediction order yielding the best performance has been isolated. In essence, the system's performance degrade as the prediction order increases beyond the optimum. This can be interpreted as follows: using high prediction orders tends to model too much spectral information such as salient partials or overtones that the spectra can exhibit[2] thus introducing *spectral singularities* in the models. Therefore, modelling more closely the signal frequency distribution will result in a pitch dependent set of features. In contrast, lower prediction orders focus more on the modelling of the spectral envelopes, thus preserving the formant structures which is the central point of our approach.

We further showed that the LSF performed better than the reflection coefficients, the polynomial coefficients and the MFCC respectively, for any prediction order, both in terms of correct instrument and family identification. This comparative study helps us introduce one of the fundamental principles of our approach. In particular, that the process of averaging spectral envelopes and therefore formant structures for each instrument through the use of the LSF and K-means provides a meaningful and consistent implementation of the *formant theory of timbre*.

The performance of this system is comparable to human ability to recognise tones in an isolated context. Srinivasan [SSF02] and Martin [Mar99] reported rates of 83.3% and 86% in terms of correct family identification for 12 and 27 instruments in the database respectively. The three databases being different, number to number comparisons have to be interpreted with care since a choice between 27 or 12 instruments identities is undoubtedly more prone to confusion than a choice between 10 instruments. It can be further noticed that humans usually make expected and

[2] theoretically, an infinite prediction order is the short-term spectrum itself

meaningful confusions, especially in terms of instruments having similar mechanisms of sound production. In contrast, the computer system presented in this section behaves sometimes in an unexpected way. As an example, very few confusions made by humans involve wind and string families (2% in total in both experiments [SSF02] [Mar99]) while it is the source of most of the confusions for the computer system (8% in average).

For these reasons, extra information about instrument sounds has to be taken into account. In section 5.5, we show how a sound attribute such as the pitch of a tone can efficiently be used as a prior for both the modelling and identification stages.

Prior to that, experiments involving the building of database models are conducted. In the following section, we evaluate the performance of a classification approach using the LSF as timbre descriptors and a SVM as machine learning algorithm. This system will serve to compare our learning approach to a classification problem involving the database as a whole.

5.4 Database modelling

This series of experiments aimed at comparing our instrument modelling approach to a classification approach involving a discriminative algorithm. We recall that these two problems are different and that they yield different system properties, especially in terms of modularity (see section 2.2). Further, the interpretations of the learning/classification processes are also different. In particular, we will speak in this case of classifying spectral envelopes, as opposed of learning characteristic spectral shapes and formant structures. It is also expected that for similar types of features, such approach yields better performance than a generative one.

For this purpose, the use of Support Vector Machines (SVM) for building a database model is investigated. The SVM theoretical principles have been introduced in section 4.4. In particular, it has been mentioned in section 4.4.2 how a binary classifier can be extended to perform a multi-class classification. A Radial Basis Function (RBF) has been used in the experiments since it offers the most flexibility in terms of boundary separating the two classes distributions (see figure 4.4.4). Normally, the parameter γ can be optimised in order to maximise the system's performance but for fairness to the other considered algorithms, its value has been set to $1/I$, where $I = 10$ is the number of instruments in the database. Note finally that prior to the SVM optimisation, data are normalised to lay in the range [-1 +1] [CL01].

In order to classify spectral envelopes, 24 LSF have been considered. Three

experiments have been conducted using the same training and testing data sets as in section 5.3. Performance were accordingly averaged.

5.4.1 Instrument classification

The corresponding confusion matrix is shown in table 5.11. This system yields 86.6% correct identification for 10 instruments in the database, 3% more on average than using a learning approach using K-means. Individual correct performance ranges from 100% for the trombone to 76.7% for the viola. When compared to table 5.4, most significant improvements concern the clarinet (plus 11%), the flute (plus 6%) and the trumpet (plus 6%). Other noticeable increase in performance involve string instruments, and in particular the violin (plus 5%). Note finally that none of the correct identification rates, except for the oboe, is worse than the ones that can achieved using the system described in section 5.3.

An analysis of the results from a family classification point of view is proposed in the following section.

5.4.2 Family classification

In table 5.9 are reported the average correct family classification rates. 88.6% correct family classification can be achieved in this configuration, an overall improvement of 1.5% compared to a learning approach using a K-means and the codebook to codebook similarity measure (see table 5.7(b)). Improvements can be noticed for the flute (plus 6%), the strings and the double-reed (plus 2%).

5.4.3 Comparison with other acoustic descriptors

Similar to section 5.3.3, we have compared the performance of several spectral envelope descriptors for the same classification task involving SVM. Average correct classification rates are reported in figure 5.4.3. Best average performance are achieved using the LSF (86.6%), followed by the PARCOR (83.0%), the LPC (80%), and finally the MFCC (75.7%). Note that this order differs from the one found when similar experiments using a generative approach have been conducted (see figure 5.2). Note also the performance of the PARCOR (83%) which are comparable to what can be achieved with the LSF/K-means approach (83.2%). Overall, using a SVM, performance are increased for the four considered types of features compared to figure 5.2.

	bassoon	clarinet	flute	oboe	trombone	trumpet	sax	cello	viola	violin
bassoon	**91.9**	0.5	1.7	0.5	0.9		0.9	2.6		1.0
clarinet		**82.4**	2.2	4.4		0.3	1.1	0.8	3.0	5.8
flute	0.4	7.9	**88.0**	0.4			0.4	0.7	1.8	0.4
oboe	3.0	5.1	0.5	**79.3**		1.5	1.0	1.0	3.0	5.6
trombone					**100**					
trumpet		1.0		2.6		**86.5**	0.2	1.6	4.9	3.2
sax		1.1				0.7	**85.6**	2.1	5.6	4.9
cello				0.2			1.7	**93.9**	0.6	3.6
viola	0.5	3.5	1.0	1.4		1.9	0.3	1.6	**76.7**	13.1
violin		1.8	0.5	1.8		0.3	0.8	2.8	10.6	**81.4**

Table 5.8: Confusion matrix using a SVM as classifier. 24 LSF have been used, yielding 86.6% correct identification. Identities of the presented samples are in rows while identities returned by the system are in columns. Baseline performance are 10% for 10 instruments in the database.

	strings	brasses	dble reed	clarinet	flute	sax
strings	**94.8**	0.7	1.3	1.8	0.5	0.9
brasses	4.9	**93.3**	1.3	0.5		0.1
db reed	6.6	1.2	**87.4**	2.8	1.0	1.0
clarinet	9.6	0.3	4.4	**82.4**	2.2	1.1
flute	2.9		0.8	7.9	**88.0**	0.4
sax	12.6	0.7		1.1		**85.6**

Table 5.9: Confusion matrix showing correct instrument families classification rates. 24 LSF and a SVM have been used to build a model of the database. 88.6% correct family classification can be achieved in this configuration. Rows correspond to the true families while columns correspond to the answers given by the system.

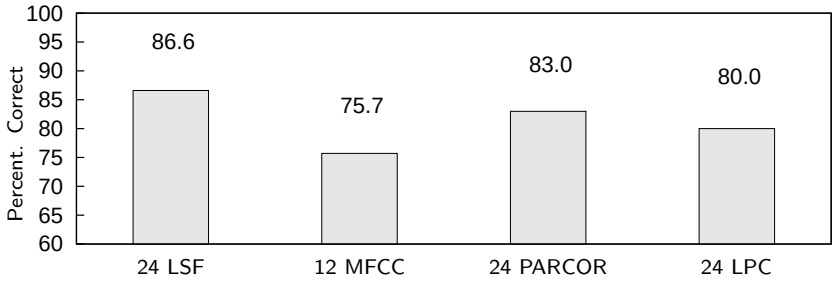


Figure 5.3: Performance of several acoustic descriptors for a classification task using SVM.

5.4.4 Summary

The central point of this section was to evaluate a system based on the classification of spectral envelopes. Using a SVM as classifier and LSF as descriptors, 86.6% and 88.6% correct instrument and family classification have been achieved respectively. It has been shown that for an identical type of feature, a classification approach to the problem yielded better performance than a generative one.

In the following section, after having illustrated the dependence of generative models upon pitch, we propose to improve this approach by using pitch as a prior for both modelling and identification stages.

5.5 Learning with a pitch prior

Note that in this section, both the terms pitch and fundamental frequency are indifferently used to describe the fundamental frequency of a pitched sound.

It has been mentioned in section 1.3.1.5 that the timbre of an instrument depends

to a certain extent on the pitch. At the acoustic level, two tones having very low and very high pitches created by the same instrument can correspond to two rather distinct physical mechanisms of sound production. For instance, producing a very low pitch note on a wind instrument with the same loudness as a higher pitch tone can be achieved if the musician blows stronger, thus providing more energy to the system. This affects the *mechanical response* of the instrument and therefore the corresponding structure of the sound. Further, the corresponding harmonic structures are rather different (and especially regarding the number of partials) in a way that the overall formant structure can be different along the pitch scale.

Our database of isolated notes cover a wide frequency range, some of them being somewhat atypical of the instrument's *nominal* playing range. The corresponding pitch dependent information is carried by the features and *averaged* during the modelling phase. Such extremes yield a non-optimal instrument model since very different spectral information are averaged in one single model. For instance, the spectral envelopes that can be obtained by averaging two spectral envelopes distinct in pitch yield a *blurry* model that might not be anymore interpreted as characteristic formant structures of the instrument.

Figure 5.4 illustrates this effect. Plots (a), (c) and (e) illustrate the process of averaging two spectral envelopes corresponding to two notes with close fundamental frequencies (G2, 98 Hz and A2, 110 Hz), while plots (b), (d) and (f) illustrates the same averaging process with two spectral envelopes corresponding to two notes with distant fundamental frequencies (G2, 98 Hz and C5, 523.25 Hz). Note that the formant structure shown in (e) is very similar to the ones shown in (a) and (c) respectively. In contrast, the formant structure shown in (f), obtained after having averaged (b) and (d), is different from (b) and (d). In particular, the first formant frequency has been shifted to a value between the ones shown in (b) and (d) respectively.

Automated musical instrument identification systems encountered in the literature often take into account the instantaneous pitch value as a feature appended to the other descriptors of sound [MK98]. Therefore, the models have some *knowledge* about the correspondence between feature vector and fundamental frequency. The pitch dependency upon timbre has also been explored at the classifier level in [KGO03], where a f_0-dependent machine learning algorithm has been proposed. It has been shown that on average, performance at individual-instrument level improved by 4%, from 75.73% to 79.73% for 19 instruments in the database.

In order to respect the approach that is presented throughout this thesis (and

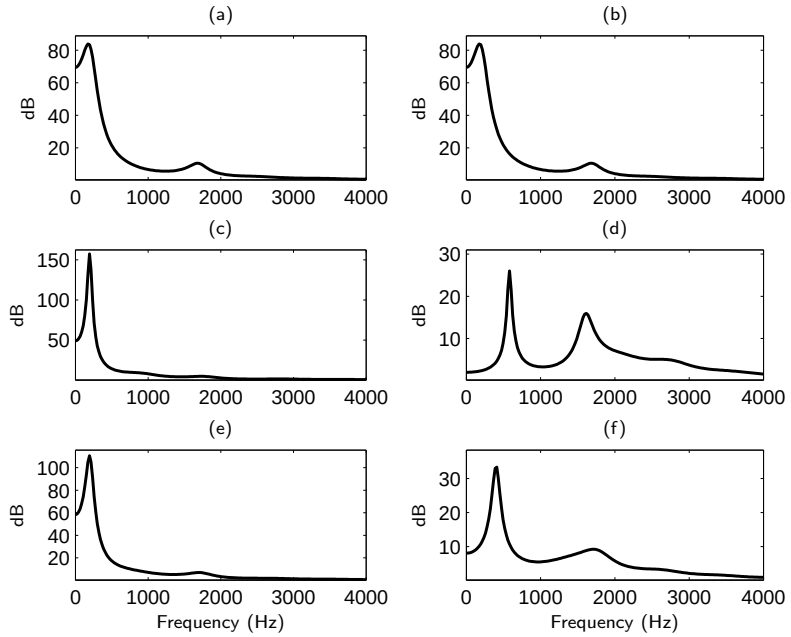


Figure 5.4: Averaging spectral envelopes corresponding to different fundamental frequencies. Plots (a) and (b) are the same representations of a spectral envelope estimate obtained after a 24th prediction order has been applied on a frame of cello sound (G2, 98 Hz). In (c) and (d) are plotted two spectral envelope estimates for the notes A2 (110 Hz) and C5 (523.25 Hz) respectively. Plots (e) and (f) are the *averaged* spectral envelopes after having calculated the means in the LSF domain of (a) and (c) on the one hand, and plots (b) and (d) on the other hand. Note the similar formant structure in (e) than in (a) and (c). Note also the different formant structure in (f) than in (b) and (d) respectively.

particularly the interpretation of averaging spectral envelopes), we propose to use the pitch as a prior for both the instrument modelling and identification phases. The strategy is applied at the database level.

In the following, a pitch detection algorithm is firstly described. Next, experiments that have been conducted are presented. Results from experiments involving the same database as in previous experiments are finally summarised.

5.5.1 Pitch detection

The fundamental frequency determination of a mono-timbral steady-state audio signal can be considered as trivial provided an adequate strategy is used. To this effect, various techniques including spectral (STFT-based) and temporal (autocorrelation, AMDF [Tre82]) methods can be used. Our first implementation used a modified AMDF function [Tre82]. However, it has been informally found that this algorithm was not reliable for the full pitch range covered by the notes in our database. For

this reason, the YIN [dCK02] algorithm has been preferred. It has been chosen for the good trade-off between simplicity of implementation and accuracy.

For an input frame $s(n)$ of N samples, the difference function D is first calculated such that:

$$D(k) = \sum_{n=1}^{N} (s(n) - s(n+k))^2, \quad k = 0, \ldots, N \tag{5.1}$$

where k represents the lags between samples. Note that Eq. (5.1) is similar to the detection function used in [Tre82]. An unknown period may be found in D by searching for the values of k for which $D(k)$ is minimum. However, it has been shown in [dCK02] that secondary dips due to the signal's formant structure can be deeper than the one corresponding to the period, thus introducing errors in the determination of the period value. As a consequence, the cumulative mean normalised difference function has been proposed.

The cumulative mean normalised difference function D' [dCK02] is obtained by dividing each value of D by its average over shorter-lag values:

$$D'(k) = \begin{cases} 1 & \text{, if } k\text{=0} \\ D(k)/[(1/k)\sum_{j=1}^{k} D(j)] & \text{otherwise} \end{cases} \tag{5.2}$$

A frame s having a definite pitch has local minima in the function D'. In practice, the smaller lag value k whose minimum $D'(k)$ is below a threshold fixed *a-priori* corresponds to the period in samples of the frame. The process is illustrated in figure 5.5 for one frame of 512 samples extracted from the steady-state portion of an oboe note. The lag value below the threshold (horizontal dashed line) in figure 5.5(c) corresponds to the time-period k_0 in samples of the frame. In the case where no local minima are found or no local minima has a value below the threshold, the frame is considered as being unpitched.

Accurate pitch values can be determined if the signal's period is a multiple of the sampling period. However, in other cases, the estimate can be biased up to half the sampling period [dCK02]. For this reason, the quadratic interpolation technique described in section 3.5.2.2 is applied to obtain a better estimate $\tilde{k}_0$ of k_0. The corresponding fundamental frequency in Hz can then be calculated as $f_0 = \frac{f_s}{\tilde{k}_0}$ where f_s is the sampling frequency.

For each frame of each tone waveform in the database, a pitch value is determined. To this effect, frames of 2048 samples have been considered. In practice, the search for local minima in $D^{'}$ has been performed from a lag value of 12 samples, thus

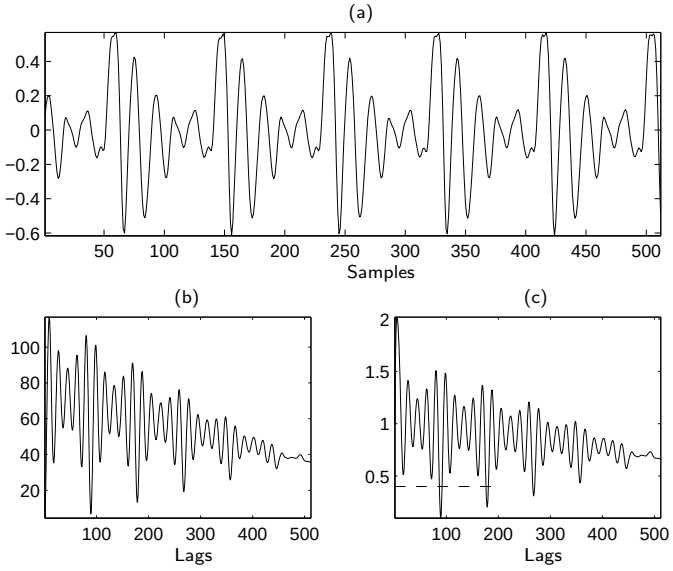


Figure 5.5: Example of pitch detection using the YIN pitch detector algorithm. (a) Time-domain frame extracted from the steady-state portion of an oboe note. (b) Difference function D, Eq. (5.1) and (c) Cumulative mean normalised difference function D', Eq. (5.2). The horizontal dashed line corresponds to the absolute detection threshold. In (c), the smaller lag corresponding to a local minimum whose amplitude is below the threshold is the signal's sample-period. In this case, $k_0 \approx 90$.

corresponding to a maximum pitch value of roughly 1837 Hz for $f_s = 22050$ Hz. The absolute threshold value has been set to 0.4. Next starting and ending pitch values corresponding to the attack and decay/end of the notes[3] were discarded for the calculation of a mean pitch value for the whole tone. Note that it cannot be ruled out that errors in the pitch calculation occur in the process. However, as will be seen in section 5.5.2, we propose to apply a strategy at the database level that does not rely on accurate pitch values.

Finally, the fundamental frequency is transformed into a value on the MIDI note scale and used as a label for the waveform. In figure 5.6 are shown the frequency distribution of the tones in the whole database as well as the frequency range per instrument that have been determined using the YIN pitch detector.

5.5.2 Using two frequency registers

In this experiment, two frequency registers are considered, 32–66 and 67–92 in MIDI number respectively. In the following, they will be called low and high-register (LR and MR respectively). This choice reflects at the same time the pitch distribution of

[3] typically the first and last 6 values

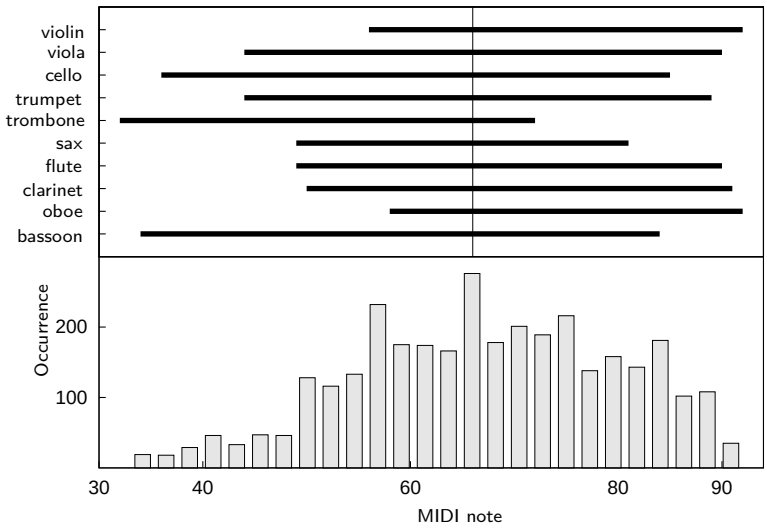


Figure 5.6: Pitch distribution in MIDI notes of the 3292 tones in the database. Pitch range per instrument (top) and histogram for the whole database (bottom). The vertical solid line delineates the boundary when two registers are used.

each instrument in the database and the fact that the ten instruments are relatively well represented in each register.

Experiments have been conducted as follows. Each note, for each instrument in the database was firstly classified into one of the two registers using the pitch detector described in section 5.5.1. Next, in each register and for each instrument, 50% of the files have been used for training the models while 50% were kept for the evaluation. This way, the number of notes for building and evaluating the system are the same as in section 5.3. Three runs have been performed with different training and testing data sets and the performance were accordingly averaged.

In table 5.10 are shown the number of notes, for each instrument in the *new* database. For each register and for each instrument, 24 LSF have been extracted to build the feature vectors. Models were trained by running a K-means, and 32 code-words were determined. The choice of these parameters conformed to the conclusions drawn from experiments reported in section 5.3.

During the identification, the pitch value of the note to identify drives the choice of the frequency register to compare the observation to. Similarity measures between the unknown feature data and all the models in the corresponding register are calculated. Similar to section 5.3, the codebook to codebook similarity measure has been used.

The percentages of average correct identification, for each instrument and for

Instrument	LR (32–66)	HR (67–92)
bassoon	106	129
oboe	31	168
clarinet	138	226
flute	102	176
sax	157	129
trombone	256	51
trumpet	132	178
cello	360	112
viola	186	252
violin	92	317

Table 5.10: The 3292 notes in the database classified into 2 registers.

each register are shown in figure 5.7. On average, 81.8% and 82.7% of the tones can be correctly recognised for the low- and high-register respectively. Note that these rates are calculated for different numbers of tones for each instrument and for each register.

Note the correct identification rate for the trombone in the high-register (80.4%) compared to the low-register (96.9%). As a result, the global performance for this instrument (94.0%) dropped when compared to the base system (98.1%, see table 5.11). This particular example shows how too few data used for training the models can affect the overall system's performance. This fact is also verified for the oboe, whose performance dropped by 3.4% from 79.5% to 76.1% due to the low performance in the low register (32.3%).

In table 5.11 are compared the average correct identification rates for the base system and the one using pitch as a prior. Noticeable improvements are concerned with the clarinet, (plus 10.8%), the trumpet (plus 6.4%) and the flute (plus 3.7%), three instruments that are well represented in the two registers.[4] Performance for the bassoon are also improved by 3.6%. Note also the improved correct identification rates for all the string instruments. Overall, 85.2% average correct identification is achieved, an improvement of 2% over the original system (see section 5.3) that did not consider the pitch as a prior.

In tables 5.12 and 5.13 are shown the confusion matrices for instrument and family identification respectively. Using the pitch as a prior yields 87.5% correct family identification, 2.5% more than the base system (see table 5.7(b)) but slightly less than if a classification approach using SVM is chosen (88.6%, see table 5.9).

[4] one can relate the improvement for the clarinet with the fact that this instrument has a different timbre depending on the register

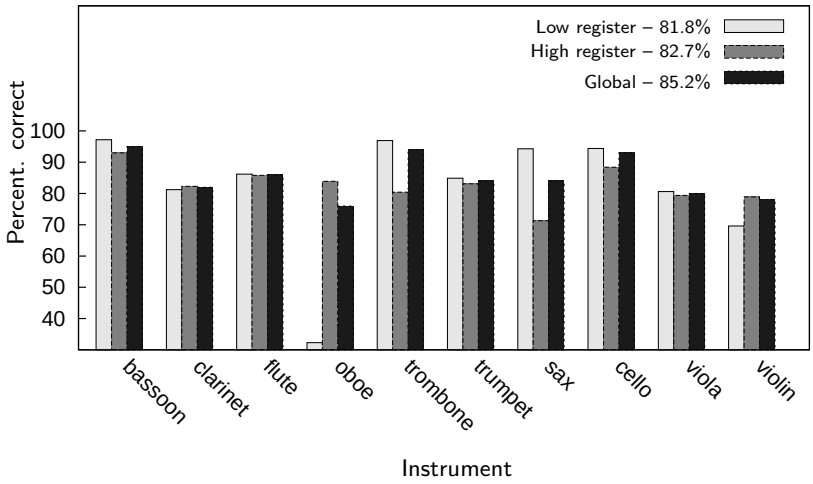


Figure 5.7: Average correct identification rates for each instrument and for each of the two registers. Individual correct identification rates for the whole database (Global) are also reported.

Instrument	Base system	Two registers	Diff.
bassoon	91.5	**95.1**	+3.6
clarinet	71.4	**82.2**	+10.8
flute	82.7	**86.4**	+3.7
oboe	**79.5**	76.1	-3.4
trombone	**98.1**	94.0	-4.1
trumpet	77.4	**83.8**	+6.4
sax	**85.3**	84.6	-0.7
cello	91.9	**92.8**	+0.9
viola	78.1	**79.3**	+1.2
violin	76.1	**78.1**	+2.0
TOTAL	83.2	85.2	

Table 5.11: Comparison between the base system and a system using the pitch as prior with two registers.

One can notice that the strings and the flute attract much less confusions than in the base system (on average 3.1% and 0.4% against 6.2% and 2.7% for the base system). The performance for the clarinet are improved by more than 10% (note however that this family contains only one instrument).

Cross-register experiments have been conducted in order to demonstrate the fact that training and testing models using different registers yielded a non-optimal system. For this purpose, each testing sample in each register is tested against the models in the other register. In table 5.14, rows correspond to the register in which the tested samples belong to, while columns indicate the register for which the tested samples

	bassoon	clarinet	flute	oboe	trombone	trumpet	sax	cello	viola	violin
bassoon	**95.1**		2.4			2.5				
clarinet		**82.2**		8.3		2.6	3.0		3.9	
flute		1.5	**86.4**	2.3	5.5	2.1	2.2			
oboe		2.0		**76.1**	2.0	8.8	6.6	4.5		
trombone				1.0	**94.0**			5.0		
trumpet		7.2			5.9	**83.8**				3.1
sax		4.3		2.7			**84.6**	5.0	1.0	2.4
cello			1.4				0.6	**92.8**	3.8	1.4
viola				2.8	4.2			5.5	**79.3**	8.2
violin		2.1	1.2		3.1			7.9	7.6	**78.1**

Table 5.12: Confusion matrix when the pitch is used as a prior. Two registers have been considered. Models consist of 24 LSF and 32 codewords. This system yields 85.2% average correct instrument identification rate, an increase of 2% over the base system. Identities of the presented samples are in rows while identities returned by the system are in columns. Baseline performance are 10% for 10 instruments in the database.

	strings	brasses	dble reed	clarinet	flute	sax
strings	**94.9**	2.4	0.9	0.7	0.9	0.2
brasses	4.1	**91.2**	0.5	3.6		
db reed	2.3	6.3	**85.8**	1.0	1.4	3.2
clarinet	3.9	2.6	8.3	**82.2**		3.0
flute		7.6	2.3	1.5	**86.4**	2.2
sax	8.4		2.7	4.3		**84.6**

Table 5.13: Confusion matrix showing correct instrument families classification rates. Two frequency registers have been used. 87.5% correct family identification can be achieved in this configuration. Rows correspond to the true families while columns correspond to the answers given by the system.

	LR 32–66	HR 67–92
LR 32–66	**81.8**	78.8
HR 67–92	75.2	**82.7**

Table 5.14: Cross-register experiments. Average correct identification rates. In rows are shown the register of the presented tones and in columns the register of the models.

are compared to. The average correct performance of the system are reported in diagonal[5] while the other percentages correspond to cross-register experiments.

It can be observed that training and testing using similar registers (diagonal in table 5.14) yields the best performance so that the best overall system can be designed by choosing the register models in accordance with the pitch of the unknown tone. Next, training and identifying using *different* registers yields worse performance. For instance, training using LR and testing using HR results in 78.8% of the samples to be correctly identified while training using HR and testing using LR results in 75.2% of the samples to be correctly identified.

However, note that each instrument class in each register does not contain the same number of notes, so that the percentages reported in table 5.14 have to be compared with care. By combining the individual performance in terms of correct instrument identification rates for the two cross-register experiments, 79.7% average correct identification has been achieved. This is, in comparison, roughly 6% worse than training and testing using *corresponding* registers (85.2%).

[5]this means that the register of the models is chosen accordingly to the register of the tested sample

5.5.3 Summary

The aim of this series of experiments was to take into account a global characteristic of sound when building the instrument models. Specifically, it has been interpreted that sound spectral envelopes and formant structures are dependent on the pitch so that the determination of a single model of characteristic spectral envelope covering the whole frequency range might not be optimal. Therefore, we have presented a strategy at the database level[6] whereby the pitch is used as a prior for the modelling and testing phases.

Using the pitch as a prior significantly improved the correct identification rates for the clarinet, the trumpet, the flute and the bassoon. On the other hand, performance for other instruments such as the oboe or the trombone dropped. This has been explained by the fact that there were not enough data in the low and high-register respectively for building robust models. Note however that these models are robust enough to avoid confusing the system. Overall, improvements of 2% and 3% in terms of correct instrument and correct family identification rates have been observed compared to the base system described in section 5.3.

The boundary between registers has been chosen in such a way that all the instruments in the database and in each register were represented, thus preserving 9 possible confusions for each identification decision, similar to the base system. However, it can be argued that using two registers is not optimum since very similar formant structures, especially around the boundary between registers, are represented in the low- and high-register models. Nevertheless, it can also be argued that in this configuration, formant structures corresponding to very low and very high pitches are not averaged, thus resulting in noticeable improvements in terms of average correct identification rates.

The purpose of these experiments was to illustrate the dependency of our model upon pitch. Cross-register experiments have been conducted and it has been confirmed that the models trained using the LSF depend to a certain extent on the pitch. It has been subsequently shown that the timbre modelling approach presented herein is indeed sensitive to pitch. This provides a possible justification for using melodic phrases instead of isolated notes when one wishes to identify melodic phrases, as will be shown in chapter 6.

[6] as opposed to methods that are applied at the feature or classifier levels, e.g. [KGO03]

Require: Ga and Gr, attack and release rates respectively
  Initialisation: $Env \leftarrow 0$, $i = 0$, $Tmp = 0$
  for each new sample $x(i)$ **do**
    Calculate: $Tmp \leftarrow |x(i)|$
    if $Env \leq Tmp$ **then**
      $Env \leftarrow Ga * Env$
      $Env \leftarrow Env + (1 - Ga) * Tmp$
    else
      $Env \leftarrow Gr * Env$
      $Env \leftarrow Env + (1 - Gr) * Tmp$
    end if
  end for

Figure 5.8: A simple envelope follower algorithm.

5.6 Differentiated transient/steady-state instrument sound modelling

Following the work of Berger [Ber63] on the importance of transients and onsets in the identification of sound by humans, we propose in this section to implement a computer simulation of his experiments using the musical instrument identification system described in section 5.3. A background review of two systems considering such segmentation prior to the feature extraction stage has been given in section 3.3 for the case of isolated notes and melodic phrases respectively.

In the following, a technique for automatically segmenting the isolated note waveforms into transient/steady-state segments based on an envelope follower algorithm is described. Next, we propose to evaluate the performance of a system where similar features are independently extracted from the transient and steady-state segments of tone respectively.

5.6.1 Transient/steady-state segmentation

Envelope following algorithms are fundamental in the audio processing field for measuring the power of an audio signal as a function of time. An envelope follower algorithm averages the signal power over a time-lag long enough for the instantaneous oscillations in power value not to significantly affect the overall power estimate and short enough to determine a reasonably accurate estimate of the envelope.

A simple envelope follower algorithm [Mus] is presented in figure 5.8. The algorithm has two main parameters, the attack and release rates, which control how the

follower responds to *transients*.[7] In our implementation, the parameters Ga and Gr were set to:

$$\begin{aligned} Ga &= \exp -\frac{1}{f_s t_a} \\ Gr &= \exp -\frac{1}{f_s t_r} \end{aligned} \tag{5.3}$$

where f_s is the sample rate in Hz, and $t_a = 5$ ms and $t_r = 30$ ms the attack and release times respectively.

This envelope estimate obtained using the algorithm described in figure 5.8 is however *noisy* and still contains high frequency component. Further, as we are interested in the calculation of the time envelopes that roughly models the slower modulations in amplitude that the signal exhibits, a low-pass filtering operation with cut-off frequency of 10 Hz is applied as post-processing to smooth the shape of the envelope as a function of time.

The onset time location is then estimated using a derivative of the envelope. For the same reasons mentioned in section 3.3, the derivative is calculated by fitting a first order polynomial to avoid obtaining a noisy estimate of the instantaneous envelope derivative. In figure 5.9 is illustrated the segmentation process for two flute and viola isolated notes. The peak location in the derivative of the envelope having the maximum amplitude (figure 5.9(c)) gives a rough estimate of the onset location. The corresponding value is then used to segment the waveform into transient and steady-state segments respectively. Specifically, the signal segment located before the peak (left arrow in figure 5.9(d)) is considered as the onset signal while the segment located after (right arrow in figure 5.9(d)) is considered as the steady-state segment of the tone. Note that this algorithm only gives a rough estimate of the onset location. As can be seen in figure 5.9(d), the flute onset signal contains samples from the steady-state segment of the tone.

On average, it has been found that the onsets segments produced by this technique lasted between 100 ms and 200 ms, thus corresponding to roughly 4 and 10 feature vectors per tone.

5.6.2 Experiments

Experiments were conducted as follows: all the files in the original database were segmented into onset/transient and steady-state segments respectively and two new

[7] note that the term transients is used in this context to qualify the typical high frequency oscillations that audio waveforms exhibit and not only the start/attack of notes

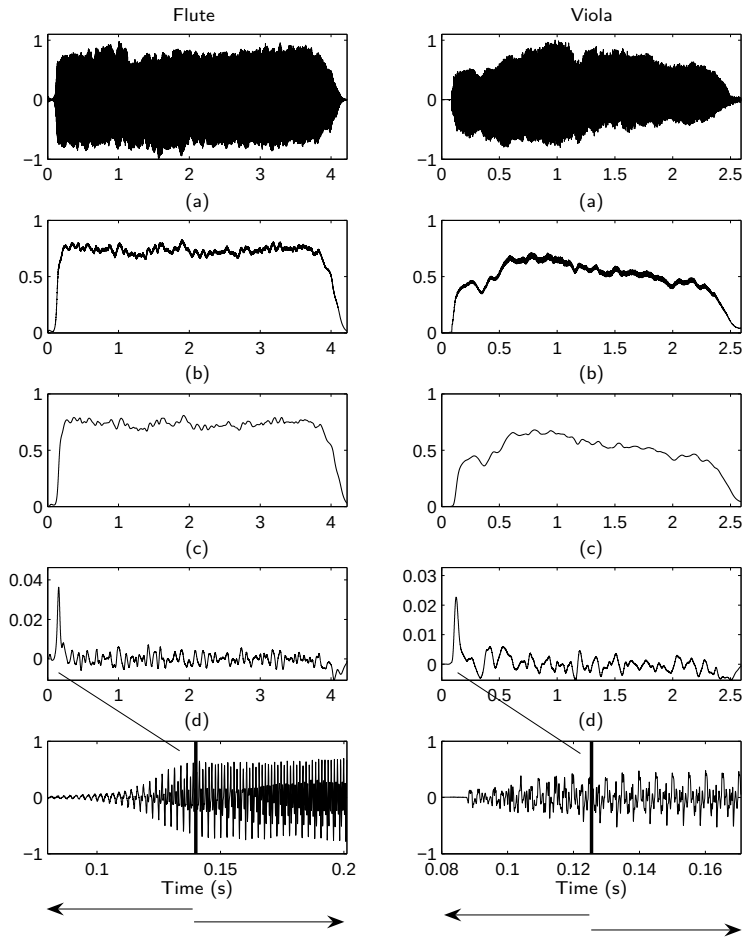


Figure 5.9: Transient/steady state segmentation using an envelope follower algorithm for two flute (left) and viola (right) isolated notes. (top) time-domain signals (a) estimate of the envelope using the envelope follower algorithm described in figure 5.8 (b) envelope estimate after low-pass filtering with a 10 Hz cut-off frequency (c) derivative of the envelope estimate using the polynomial fitting algorithm described in section 3.3 using a linear window of 9 samples (d) magnified view around the onset location.

databases were built. Next, each of these databases was considered independently for the descriptors extraction, the models calculation and the performance evaluation.

Features were extracted in a similar way as for the base system, consisting of one vector of 24 LSF for each frame. Models were built using 32 codewords and the minimum distance measure described in section 4.2.3.1 was used to identify each unknown excerpt during the testing phase. The reason for using this measure is that the onset database contains very short excerpts. As a consequence, the use of the codebook

Instrument	Onset database	Steady-state database	Un-segmented database
bassoon	84.3	**92.6**	91.5
clarinet	**74.3**	68.8	71.4
flute	72.2	80.6	**82.7**
oboe	**86.2**	66.6	79.5
trombone	89.4	98.1	**98.1**
trumpet	80.8	**85.8**	77.4
sax	**90.0**	81.8	85.3
cello	79.4	**92.0**	91.9
viola	54.5	75.8	**78.1**
violin	49.0	57.8	**76.1**
TOTAL	76.0	80.0	**83.2**

Table 5.15: Comparative performances when feature and models are trained on separate onset/transient databases. Percentage of correct identifications. 24 LSF have been used for training while the models consisted of 32 codewords. Standard deviations are reported in parentheses.

to codebook similarity measure is not possible since it involves the calculation of a dictionary of 32 codewords from the observation to be identified. However, it has been shown in section 5.3 that the performance of these two similarity measures in this particular configuration (24 LSF and 32 codevectors) were statistically similar. It can therefore be argued that the comparison between the systems' performance involving the three different databases is meaningful. Three runs have been performed with different training and testing samples but with the same file *combinations* as in section 5.3.

Comparative performance in terms of correct identification rates for each instrument are reported in table 5.15. In the last column are recalled the performance of the base system in which features were indifferently extracted over the whole note.

On average, training the models using the steady-state portions of tones as opposed to using the onset database yields better performance, with an increase of 4% in average identification rate. However, this is in comparison 3% less than if similar models were trained on the un-segmented database. A clear difference in identification rates is concerned with the string family. A drop of roughly 12% for the cello, 21% for the viola and 18% for the violin in average performance between steady-state and onset database can be observed. For these instruments, models are more advantageously built using the steady-state and un-segmented databases than the onset database. This can be explained by the fact that the cello, the viola and the violin have very similar mechanisms of sound production. Therefore, their note onsets have similar physical characteristics – resulting from the same mechanical bowing action

– that do not contain specific enough information about the instruments for building robust models. Using the database of steady-state samples bring extra-information about the spectral envelopes that help to better discriminate between the three instruments sounds. The fact that these instruments mainly differ by size and that they have slightly different nominal playing ranges obviously yield different spectral envelope models that are, in consequence, more representative of each class. Combining these two facts into one model (i.e. using the un-segmented database) is useful since the overall performance increased. As a result, it can be concluded that the models built using the un-segmented database contain to a certain extent information of both the attack and steady-state segments of sound.

On the other hand, it is more interesting to use the onset database for the clarinet, oboe and sax. Using the onset database yields 74.3%, 86.2% and 90% correct identification rates, an improvement over both the steady-state and un-segmented databases. Note for these three instruments the average performance between onset and steady-state databases when the un-segmented database is used: it can be argued that the models built from the un-segmented database contain essential information about the transients.

Overall, building models from the onset database still yields 76.0% correct identification which corresponds to reasonably high performance considering the duration of the considered signals (typically few hundreds milliseconds). This way, we confirm the conclusion drawn in section 3.3 where it has been highlighted that the onset database was in fact a non-redundant snapshot of the whole database. In consequence, this database contains a small amount of non-redundant information about the spectral envelopes of the steady-state segments that are advantageously exploited by the models. Note that this is particularly valid in the current case since isolated tones have a constant pitch over their whole duration. This has been informally verified by listening to the various samples: although being very short in duration, it has been found that the onset segments had a clear and definite pitch.

Finally, note the different behaviour of this system compared to Eronen's one (see section 3.3), in which 12 MFCC extracted from the onset database performed slightly better than the same features extracted from the steady-state database. Using the LSF and K-means, better performance are achieved with the un-segmented database, thus justifying the use of our generative approach in a melodic context, as will be investigated in chapter 6.

5.6.3 Summary

In this section, we proposed to evaluate the performance of our system in a differentiated transient/steady-state instrument sound modelling framework. At the same time, the aim was to present a computer implementation of the perceptual experiments conducted by Berger [Ber63] dealing with the importance of onset/transient in the identification of sound by humans.

On the one hand, it has been shown that on average, training the models using the steady-state database yielded better performance than if they were trained using the database of onsets. This has been shown to be particularly true for the bowed string family for which the spectral envelopes extracted from the steady-state segments of tones exhibit more discriminating power than the ones extracted from the onset segments. The consideration of the steady-state segments of tones improved the performance for most of the other instruments except for the clarinet, oboe and sax for which building models from their onset database offered improvements in terms of performance.

On the other hand, considering the un-segmented database improved the average performance so that it can be concluded that models consisting of average spectral envelopes determined using a K-means algorithm carry essential information about the waveforms' transient structure. It has been proved that this information is useful since the performance increased when features were extracted over the whole tone duration. This provides a justification for using this generative approach in a melodic context, as will be shown in chapter 6.

One can argue that the LSF are not appropriate for modelling transients and onsets. However, as has been mentioned in section 3.3, the onsets of pitched musical sounds created by wind, brass or string instruments are much less singular than the transient of a piano or a plucked string sound. In particular, one can observe in figure 5.9(d) the non-percussive nature of the flute and viola onset segments. For the flute, the onset segment is periodic and similar to the steady-state. For the viola, the onset segment is noisy. This is due to the initial friction of the bow on the strings. As the LSF model in some ways spectral envelopes, it can be argued that they can model *steady-state* (low-pass spectral envelopes) or *noisy* (high-pass spectral envelopes) onsets without relying on formal temporal considerations.

5.7 Psycho-acoustic considerations

In this section, we propose to include psycho-acoustic considerations for building the models. This corresponds to the second approach to the timbre modelling problem introduced in chapter 3.

The work of Eronen [Ero01] has been reported in section 3.5. In particular, he showed how a frequency warping into the Bark scale improved the performance of his system over non-warped versions of the LP coefficients.[8] In this series of experiments, the use of a similar warping method prior to the LSF calculation is investigated.

Next, we propose a novel feature based on a psycho-acoustically motivated partial selection algorithm applied prior to the LSF calculation. Principles of the calculation of these descriptors using the ISO/MPEG psycho-acoustic model (see appendix A) are described in the following section.

5.7.1 Perceptual LSF (PLSF) calculation

We have described in section 3.5.3 how a frequency masking technique could be used to include psycho-acoustic considerations into the sinusoidal analysis/synthesis framework. As a result, we propose to use the instantaneous global frequency masking curves to calculate a psycho-acoustic version of the LSF coefficients, the Perceptual LSF (PLSF).

The PLSF are determined as follows. Signals are processed on a frame basis. For each frame, the global masking threshold (see appendix A) is calculated and used to select the relevant local maxima in the spectrum.[9] Their amplitudes, frequencies and phases are interpolated using the quadratic and linear schemes described in section 3.5.2.2 and used as input to the kernel modulation algorithm (see section 3.5.2.3) to synthesise a time-domain frame. Next, a LP analysis is applied and the LSF are calculated using the standard technique described in section 3.2.2. The whole process is illustrated in figure 5.10 for a single frame of cello sound. Compare figures 5.10(b) and 5.10(d) and note the missing spectral information at $f \approx 5000$ Hz and $f \approx 8000$ Hz corresponding to sinusoidal components that have not be retained after the psycho-acoustic spectral transformation.

[8] note he used cepstral versions of the LP and warped LP coefficients respectively

[9] the ones whose magnitudes are above the global masking curve

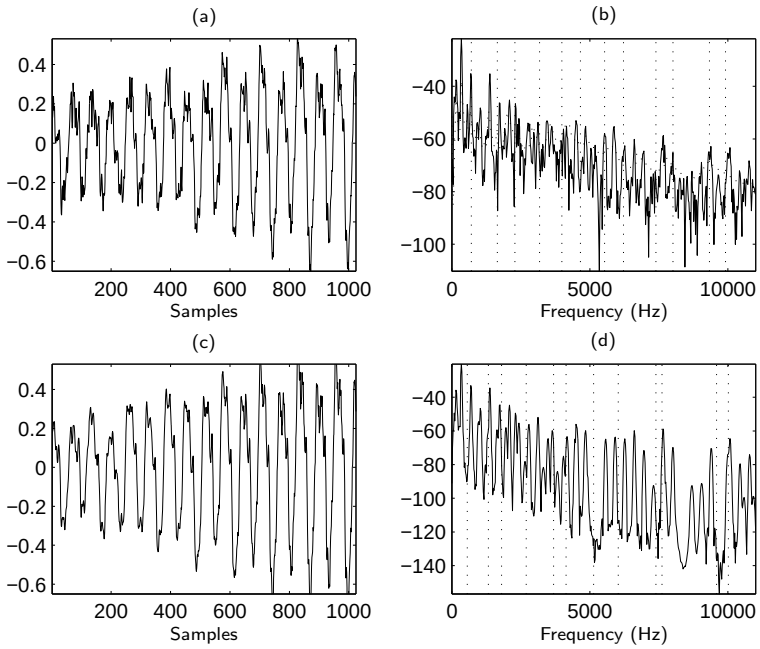


Figure 5.10: Illustration of psycho-acoustic motivated LSF calculation. (a) Original time-domain frame (b) Corresponding magnitude spectrum and LSF representation (vertical dashed lines) for a 12th prediction order (c) Time-domain frame after spectral psycho-acoustic decimation and resynthesis (d) Corresponding magnitude spectrum and LSF representation for a 12th prediction order.

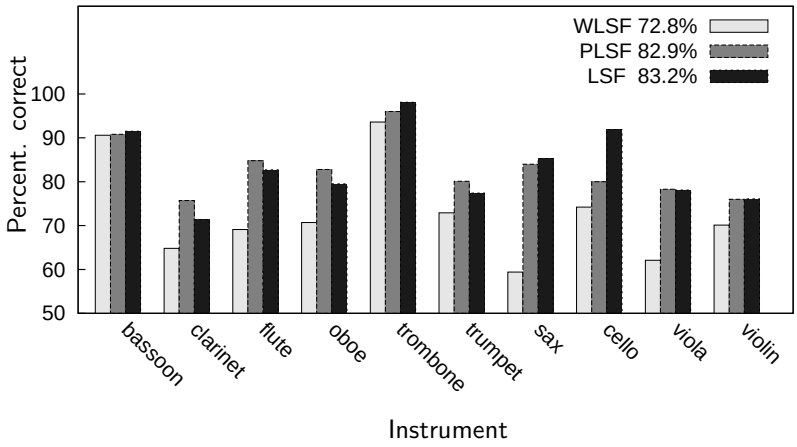
5.7.2 Experiments

In these experiments, two perceptual features, the Warped LSF (WLSF) and the Perceptual LSF (PLSF) are considered with the K-means and the SVM for building the instrument and database models. A number of 24 parameters has been chosen in all cases in order to compare the performance with the results summarised in sections 5.3 and 5.4.

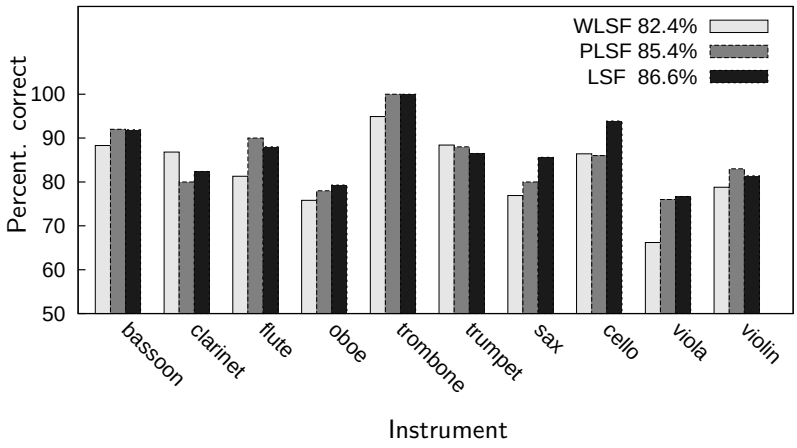
The WLSF are calculated as follows. First, the warped LP coefficients are determined using the code provided in [War]. A warping factor of 0.6461, which corresponds approximately to a transformation from frequency to Bark scale [Ero01] [HL01] for $f_s = 22050$ Hz, has been used. Next, the warped LP coefficients are transformed to WLSF using the procedure described in section 3.2.2.

In figure 5.11 are reported the systems' comparative performance. Figure 5.11(a) is concerned with the K-means while figure 5.11(b) with the SVM. Corresponding performance using the LSF (see tables 5.4 and 5.8) are recalled in both figures.

For the K-means (figure 5.11(a)), 72.8%, 82.9% and 83.2% average correct identification are achieved for the WLSF, the PLSF and the LSF respectively. Note that



(a) K-means



(b) SVM

Figure 5.11: Comparative performance between perceptual and standard LSF. Average correct identification rates for 10 instruments in the database. (a) K-means (b) SVM. WLSF and PLSF stand for Warped LSF and Perceptual LSF respectively.

the performance of the WLSF are comparable to what can be achieved when 12 MFCC and a K-means are used (73% average correct identification, see section 5.3.3). For the PLSF, improvements in terms of correct identification rates concern the clarinet (plus 4%), the oboe (plus 3%) and the flute (plus 3%). On the other hand, the performance for the cello and trombone drop by nearly 12% and 2% respectively. As a consequence, the average performance of the system using the PLSF are almost similar to the ones obtained with a system using the standard LSF as descriptors.

For the SVM (figure 5.11(b)), average performance are improved for the two perceptual descriptors. On average, 82.4%, 85.4% and 86.6% correct identification have been observed for the WLSF, the PLSF and the LSF respectively. One can notice the difference in relative performance between the WLSF and the LSF depending on the classifier used. Specifically, the drop in average correct identification rates is more pronounced in the case of the K-means (roughly 9% for the WLSF compared to the LSF) than in the case of the SVM (roughly 4% for the WLSF compared to the LSF). Further, for the two classifiers, the LSF are better than the PLSF, that are themselves better than the WLSF.

Although it was argued the consideration of psycho-acoustic features would better fit the mechanisms of sound perception by humans, experiments did not show significant improvements by considering perceptually derived LSF coefficients. This can be analysed as follows. Applying a psycho-acoustic transformation can be interpreted as transposing the sound physical formant structure into a perceptually relevant scale. However, in this new scale, it is not guaranteed that the process of averaging spectral envelopes and evaluating a similarity measure between observation and models using an Euclidean distance is appropriate. For instance, it is known in speech coding that the Euclidean distance is relevant for measuring a *distance* between two LSF vectors. This can explain the better performance of the LSF over the WLSF when a K-means is used.

Experiments involving the PLSF confirm this hypothesis. In this case, the formant structures are much less altered since the pre-processing consists of removing sinusoidal components that are often located in the high frequencies, where formants are rarely present. In the worse case, a missing sinusoidal component in the lower frequency can affect the formant structure, but in a much smaller scale than in the case of the WLSF. Thus resulting on average in the mid-performance of the PLSF compared to the WLSF and LSF respectively.

Finally, experiments involving SVM revealed the capacity of the adaptation of the algorithm. SVM are much less sensitive to the type of features being used since they

dynamically adapt the boundaries between feature data distributions.

5.7.3 Summary

In this series of experiments, we proposed to include psycho-acoustic knowledge prior to the feature extraction stage. After having described how the PLSF were calculated, we compared their performance with the WLSF and the standard LSF for the same identification and classification tasks.

We have related the performance of the system WLSF/K-means with the ones that can be achieved using the MFCC (see section 5.3.3). The drops in performance for the WLSF and PLSF have been interpreted and it has been concluded that the Euclidean distance was not suitable for evaluating similarity measures into perceptual scales. On the other hand, using a SVM resulted in relatively uniform performance for the three considered features, thus revealing the SVM adaptive properties.

Chapter summary

In this chapter, we have presented various applications of our system for the identification of isolated tones taken out of musical context. The use of this database helps to limit the amount of musical acoustic information that could intervene for a proper interpretation of the results and system behaviour. Our approach finds strong justifications in the *formant theory of musical timbre* whose fundamental principles have been exposed throughout this thesis. Not only was the aim to propose a system able to identify musical instrument sounds, we described a framework in which the relative importance of the timbral characteristics could be verified and reproduced.

In section 5.3, the base system which served as reference for our research has been presented. The use of the LSF as spectral envelope and formant structure descriptors has been presented in a musical instrument context. We justified the use of the K-means for building the models by considering at the same time principles governing the formant theory of timbre and various research works in speech coding and speaker identification/verification. Preliminary experiments consisted of finding a preferred prediction order and number of codewords for the models that maximised the average performance. Next, the performance of several similarity measures between observations and the models have been evaluated. We have experimentally found that a minimum distance classifier based on the isotropic Euclidean distance yields the best performance together with the codebook to codebook similarity measure. It has also been found that the use of Mahalanobis distance and the MAP rule used

in the GMM was inappropriate with the type of considered features. Finally, it has been experimentally shown that our system outperformed the classical MFCC/GMM approach for the same learning task. To summarise, this base system yielded more than 83% average correct identification among 10 instruments identities, the database totalling 3292 samples while training and testing databases being considered in a 50/50 ratio. Although the number of considered instruments in our database is limited, we advanced that the performance of this system in terms of instrument family identification is comparable to what humans can achieve.

Various acoustic descriptors can be used for the purpose of modelling *spectral envelopes*. In section 5.3.3, experiments comparing the performance of the LPC, PARCOR, LSF and MFCC respectively have been conducted. For similar algorithm parameters (prediction order and number of codewords), the use of the LSF improved the performance both in terms of correct instrument and correct instrument family identification. This has been justified by the fact that since the LSF exhibit localised spectral sensitivity properties, they were particularly suitable to be used together with an averaging machine learning procedure such as the K-means.

In section 5.4, our learning approach has been compared to a classification approach involving SVM. The use of the SVM helped to increase the average performance by more than 3% up to 86.6% for the LSF. Experiments involving other acoustic descriptors also showed improvements compared to experiments involving K-means. It has been argued that this database model yielded an upper limit in terms of achievable performance for the LSF-based mono-feature system described in this thesis.

In section 5.5, a study on the dependence of the models upon pitch that were obtained after a K-means optimisation has been conducted. Our argument was that although the formant structure of an instrument sound is thought to be theoretically invariant to the pitch, the LSF being extracted are in practice dependent on the position of the spectral envelope along the frequency axis. It was therefore believed that averaging a set of spectral envelopes distant in pitch would result in non-optimum and *blurry* models. Consequently, it has been shown that models can advantageously be built for different instrument frequency registers. For this purpose, we have isolated two frequency registers in the database for which two corresponding models were trained using identical procedures. Experiments have quantified the improvements in terms of correct identification rates when pitch was used as prior for the clarinet, trumpet, bassoon and flute. At the same time, cross-register experiments showed that performance degraded as the *distance* in pitch between unknown samples and

models increases. Using the pitch as prior and the same database, the average correct identification rates increased by 2% up to 85.2% compared to the base system.

In section 5.6, it has been verified that a single model built for each instrument and using features extracted from un-differentiated frames types would nevertheless contains information about the onset and attack segments of tones. To this effect, a segmentation onset/steady-state algorithm based on a spectral envelope follower has been used. Next, identical features were extracted from the two newly created databases. Overall, the average identification rates are improved when models are built from the steady-state database and the un-segmented database respectively compared to when models are built only from the transient database. It has been concluded that our approach confirmed to a certain extent the experiments conducted by Berger [Ber63] and that models trained from the LSF and K-means naturally incorporate specific acoustic information about the attacks and onsets of notes.

The second approach we proposed to build timbre models consisted of using psycho-acoustics at the pre-processing level prior to the LSF calculation. To this effect, we have described in section 5.7 a perceptually motivated pre-processing technique based on the ISO/MPEG psycho-acoustic model described in appendix A. We compared this feature, the PLSF, with the WLSF and the standard LSF using similar parameters as in previous experiments. When using K-means, it has been found that performance dropped by 9% with the WLSF and less by 2% with the PLSF. We explained this by the fact that although the spectral envelopes obtained after warping into the Bark scale on the one hand, and after psycho-acoustic spectral selection on the other hand were perceptually relevant, the processes of modelling and averaging were more relevant when the standard LSF and the Euclidean distance were used.

In the next chapter, the generative and discriminative approaches described in this chapter are extended to deal with realistic musical recordings.

6. Instrument identification and classification in a melodic context

Chapter 5 was concerned with the evaluation of systems aimed at identifying musical instrument notes taken out of any musical context. It has been mentioned how such a controlled musical environment could serve to conduct finely tuned experiments. To this effect, a computer system conforming to the formant theory of musical timbre has been built. The dependence of this system on pitch has been quantified and a solution whereby different models were built for two frequency registers has been proposed. It has also been shown how such a system could advantageously take into account the information carried by the onsets of musical instrument sounds without necessarily relying on a pre-segmentation processing module.

Musical sounds encountered in real-life MIR applications are melodic and poly-timbral by nature. It is expected that an automated system designed for identifying musical instruments or evaluating the timbral similarity between sounds is able to deal with the various musical melodies that acoustic instruments can produce. As stated in the introduction of this thesis, the problem of identifying instruments in polyphonic mixtures is beyond the scope of this research. We will therefore restrict our study to the task of identifying melodic excerpts having continuous varying pitches and various temporal properties such as, for example, the number of notes in the considered audio samples.

In essence, this approach is a direct extension of the previous research work involving isolated notes. For instance, it is assumed that under the condition that a system is reasonably robust at identifying isolated notes, its timbre modelling power could be directly used for identifying melodic phrases.

In this chapter, several systems designed for the task of identifying mono-timbral melodic phrases are described. In section 6.2, we propose to train the models using

isolated notes while melodic excerpts of various durations have to be identified. We then show in section 6.3 that training models using melodic phrases is more appropriate for identifying melodic phrases since they intrinsically carry essential information about the transitions between notes and the expressiveness of musical phrases. An attempt to include time consideration and to append the delta of the acoustic timbral descriptors to the original feature vector is made. Similar to chapter 5, our learning approach to the musical instrument identification problem is compared to the classification approach using Support Vector Machines (SVM).

In section 6.4, we propose to combine generative and discriminative methods to *classify musical instrument models*. It is shown how the SVM training computational load can be reduced at a reasonable expense in terms of the system's performance.

6.1 Database

The difficulty in conducting experiments involving melodic excerpts resides in the fact that such databases are not available as such to researchers. The preliminary step was therefore to gather as much data from various sources as possible. We shall stress here the difficulty of finding mono-timbral solo phrases within *real* classical musical pieces. As a consequence, our database is now restricted to 6 instruments: the clarinet (Cl.), the oboe (Ob.), the flute (Fl.), the cello (Ce.), the violin (Vl.) and the piano (Pn.). Piano sounds have been included because they have percussive attacks. Following the conclusions drawn in section 5.6, it is expected that models built using K-means capture these characteristics.

For each instrument, the data set contains 600 seconds of sounds originating from 10 different sources. Similarly to the experiments involving isolated notes reported in chapter 5, all samples are downsampled to 22050 Hz prior to any further processing. In our experience, features extracted every 34.5 ms over frames of 46.43 ms gave the best results, so this segmentation has been used in all the experiments presented in this chapter.

6.2 From isolated notes to melodic phrases

This section is concerned with the description of a system in which models are built from isolated notes [CS06].[1]

[1] note that results from similar experiments have been succinctly reported in [KS04]

Instrument	No. of notes	Pitch range
clarinet	120	50 – 90
oboe	100	58 – 90
flute	111	60 – 90
cello	188	36 – 81
violin	192	55 – 90
piano	265	28 – 92
TOTAL	976	

Table 6.1: Instances and number of notes per instrument in the isolated note database that have been retained for the experiments. The pitches have been determined using the YIN algorithm described in section 5.5.1.

This approach can be referred to the work presented in [VR04] where sounds produced by each instrument in the database were modelled using a weighted sum of log-power spectra plus noise quantised on a semi-tone frequency scale. This model was firstly trained on a database of isolated notes and then used to identify mono-timbral melodic excerpts of 5 seconds in duration. In our approach, assuming that the K-means yields the definition of characteristic spectral shapes, the process can be seen as determining a condensed timbral representation of each instrument in the database, at the expense of losing the frequency resolution of the method in [VR04].

Similar to [VR04], the training database consists of a subset of the RWC collection [RWC]. Information about this database can be found in table 6.1. Note that 265 piano tones have now been added. The pitch for each instrument note has been determined using the YIN algorithm described in section 5.5.1.

6.2.1 System description

In this experiment, we propose to evaluate the performance of K-means and SVM at the task of identifying melodic excerpts. The combination of 16 LSF/40 codewords has been chosen for all the experiments involving K-means [CS06].

During the evaluation phase, segments between 2 and 5 seconds in duration were presented to the system which returned their identities. Note that the segments to be tested might contain different numbers of notes, truncated attacks or steady-state portions as well as soft and loud passages. This corresponds however to typical data that can be encountered in real-life applications of the system.

In order to compare the performance of this system with the one presented in section 6.3, half of the database of melodic phrases is considered for the evaluation, that is, 300 seconds of data are kept for testing the system.

6.2.2 Experiments

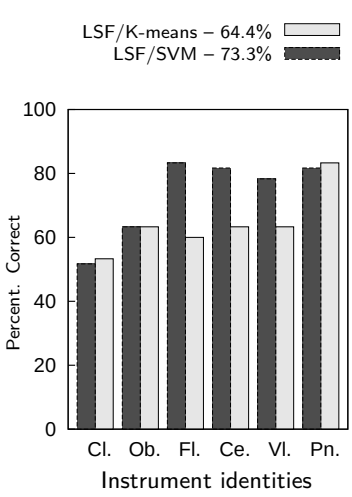
Results are summarised in figure 6.1(a) where the average correct identification rates are reported for the two machine learning algorithms. For the K-means, and except when explicitly mentioned, the codebook to codebook similarity measure has been used as a measure of distortion between observations and models. Similar to section 5.4, the Radial Basis Function (RBF) has been used in all the experiments involving SVM. The parameter γ has been set to $\gamma = 1/6$ and the feature data was normalised to lay in the range [-1 +1] [CL01].

In this configuration, the correct average identification rates were 64.4% and 73.3% for the K-means and SVM. The corresponding confusion matrices are shown in figures 6.1(c) and 6.1(d) respectively. Although above random guesses (16.7%), these figures are worse than human performance (98% and 85% average correct identification rates have been reported in [SSF02] for 3 and 9 instruments in the database respectively) and below what could be achieved when isolated notes are used for both training and testing (83.2% and 86.6% for the K-means and SVM respectively in experiments reported in sections 5.3 and 5.4 for 10 instruments).

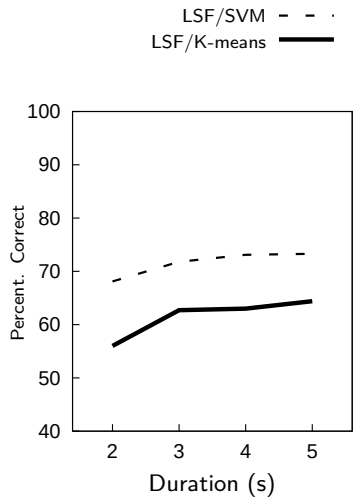
For the K-means, performance range from 53.3% for the clarinet and to 83.3% for the piano. Note how the K-means performs slightly better than the SVM for this instrument. This can be explained by the fact that being the only *percussive* instruments in the database, the piano model contains a significant amount of characteristic high-pass spectral envelopes. These characteristics are well-captured by the similarity measure during the identification phase when compared to the other instrument models which do not exhibit high-pass characteristic spectral shapes [CS06]. This confirms to a certain extent the conclusions drawn in section 5.6 where it has been shown that models built using K-means implicitly incorporate specific acoustic information about the attacks of notes.

Most important confusions involve the pairs violin–cello (20%), cello–violin (15%) and clarinet–oboe (15%). Note that each of these confusions concerns instruments belonging to the same families (strings and reeds respectively, if we consider single and double-reed instruments as belonging to the same group).

Overall, using a SVM helps to increase the identification rates for the flute, cello, and violin classes by 23%, 18% and 15% respectively compared to the K-means. The corresponding confusion matrix is reported in figure 6.1(d). Most important confusions are concerned with the pairs clarinet–oboe (21.7% of the presented clarinet excerpts have been mis-identified as being oboe) and cello–violin (16.7%). The same



(a)



(b)

	Cl.	Ob.	Fl.	Ce.	Vl.	Pn.
Cl.	**53.3**	15.0	10.0		10.0	11.7
Ob.	5.0	**63.3**	13.3		11.6	6.7
Fl.	11.7	8.3	**60.0**	10.0	10.0	
Ce.	10.0	1.7	8.3	**63.3**	15.0	1.7
Vl.	6.7	8.3		20.0	**63.3**	1.7
Pn.	13.3		1.7		1.7	**83.3**

(c)

	Cl.	Ob.	Fl.	Ce.	Vl.	Pn.
Cl.	**51.7**	21.7	10.0	8.3	6.7	1.7
Ob.	15.0	**63.3**	6.7			15.0
Fl.	10.0	6.7	**83.3**			
Ce.				**81.7**	16.7	1.7
Vl.	15.0			6.7	**78.3**	
Pn.	6.7	3.3	8.7			**81.7**

(d)

Figure 6.1: (a) Comparative performance for the task of identifying melodic phrases of 5 seconds in duration using models trained from isolated notes. Results for a K-means and SVM respectively. For each instrument, 60 excerpts have been tested. (b) Total average correct identification rates as a function of the duration of the tested samples, ranging from 2 to 5 seconds. (c) Confusion matrix when K-means is used (64.4% average correct identification). (d) Confusion matrix when a SVM is used (73.3% average correct identification). Rows correspond to the true instrument identities while columns correspond to the answers given by the system.

conclusions about these confusions as in the case where a K-means is used can be drawn. Finally, note the similar performance for the clarinet and the oboe for the K-means and SVM respectively.

In figure 6.1(b) are shown the total correct identification rates for the considered systems as a function of the duration of the testing samples. Note that for the experiments involving segments of 2 and 3 seconds respectively, the minimum distance measure described in section 4.2.3.1 has been used instead of the codebook to codebook similarity measure: in few cases, silences in the tested samples resulted in too little observation to run a K-means. One can noticed that the longer the observation, the better the average performance. Overall, the influence of the duration of the tested sample on the average performance is more limited in the case of the SVM than in the case of the K-means.

6.2.3 Summary

The system presented in section 6.2.2 has been designed to identify melodic excerpts using models trained using isolated notes. It has been found that average performance increased by nearly 9% to 73.3% using a classification approach compared to a system using a K-means and the codebook to codebook similarity measure. We have noticed that the K-means performed particularly well with piano sounds. For similar acoustic features, it has been confirmed that a discriminative approach performed better than a generative one. It has been further noticed that the two systems exhibited meaningful behaviours in terms of family confusions.

Drawing on the conclusions reached in chapter 5, it can be argued the following. First, that the database of isolated notes contains very low and very high pitched notes that might not yield optimum models. Second, that solo phrases exhibit temporal properties at note changes that models trained using isolated notes do not capture.

For these reasons, we propose in the next section to train models using melodic phrases.

6.3 Using melodic phrases for training

This series of experiments focuses on the use of melodic phrases for both training and testing. It is argued that timbral melody and expression can advantageously be taken into account to build more robust instrument models. Research works dealing with mono-timbral melodic phrases have been reported in [BHM01], [EB04] and [ERDb].

6.3.1 The importance of context

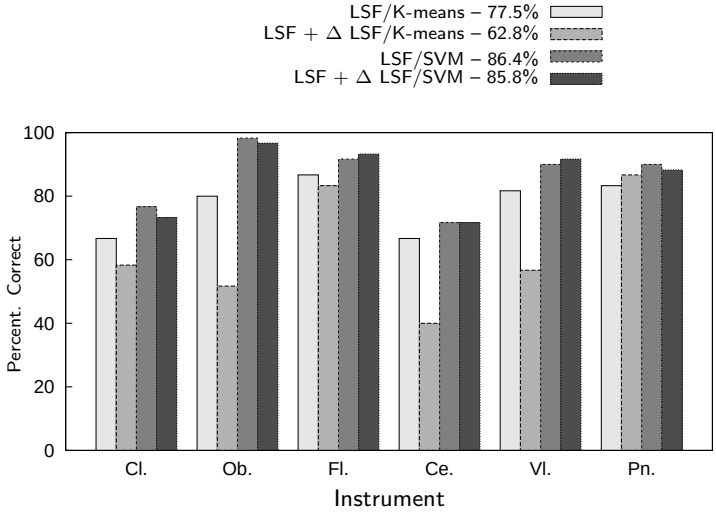
The presence of musical context in the identification of sounds by humans is essential. However, little perceptual research has been carried out on the topic so that it is difficult to accurately quantify the corresponding contribution to *the whole*. This musical context can be related to successive note attacks, to various playing techniques characterising the instruments or to the overall expressiveness carried by the phrases.

These aspects play a crucial role in identifying musical sounds. For instance, it has been informally observed that when asking professional musicians to recognise isolated notes out of musical context, confusions between a high pitched violin note and a high pitched oboe sound, among others, are not rare. By training the models using isolated notes, the essential information present in musical phrases is not taken into account, thus resulting in low identification rates during experiments, as has been shown in section 6.2.2.

6.3.2 Experiments

This series of experiments involves the same machine learning algorithms than in section 6.2. Models have been trained using 50% of the available data in the melodic phrase database (300 seconds for each instrument) while the other 50% were retained for the system evaluation. Note that the testing data set is the same as in section 6.2. Feature vectors were built by extracting 16 LSF from the training data set. The influence of the delta LSF coefficients is also investigated. It can be argued that they model the LSF fluctuations with time, and to a certain extent the transitional information that characterise musical phrases (see figure 3.9). They have been calculated using the polynomial fitting technique described in section 3.3. A linear window of length $K = 9$ has been used (see equation 3.8).

In figure 6.2(a) are summarised the average performance when solo phrases are used for both training and testing. The base system using a K-means and 40 codevectors is able to correctly identify 77.5% of the tested samples, an increase of 13% over the system using isolated notes for training. Individual correct identification rates range from 66.7% for the cello and clarinet to 86.7% for the flute. The corresponding confusion matrix is shown in figure 6.2(b). The most important confusions involved the pairs cello–violin (21.7%) and oboe–flute (15%). In terms of instrument families identification, bowed strings (cello and violin) are recognised 92% of the time and reeds (oboe and clarinet) 78% of the time. The global performance of this system is



(a)

	Cl.	Ob.	Fl.	Ce.	Vl.	Pn.
Cl.	**66.7**	8.3	11.7		8.3	5.0
Ob.		**80.0**	15.0	1.7		3.3
Fl.	3.3		**86.7**	3.3		6.7
Ce.	8.3		3.3	**66.7**	21.7	
Vl.	5.0			13.3	**81.7**	
Pn.	11.7		5.0			**83.3**

(b)

	Cl.	Ob.	Fl.	Ce.	Vl.	Pn.
Cl.	**76.7**		8.3	3.3	11.7	
Ob.		**98.3**	1.7			
Fl.	1.7		**91.7**			6.7
Ce.	1.7		6.7	**71.7**	16.7	3.3
Vl.	8.3			1.7	**90.0**	
Pn.	8.3		1.7			**90.0**

(c)

Figure 6.2: (a) Individual correct identification rates for the 2 considered systems using melodic phrases for both training and testing. (b) Confusion matrix for a system using a K-means and 32 codewords (77.5% average correct identification) (c) Confusion matrix for a system using a SVM (86.4% average correct identification). Rows correspond to the true instrument identities while columns correspond to the answers given by the system.

comparable to the one described in [BHM01] where 77% average correct identification has been achieved using ten cepstral coefficients and a Gaussian mixture model. Finally, note that the addition of the delta coefficients to the feature vectors degraded the performance which dropped by nearly 15% to 62.8%.

Using a Support Vector Machine allows 86.4% of the testing samples to be correctly recognised, an improvement of nearly 9% compared to the system using K-means. The corresponding confusion matrix is shown in figure 6.2(c). The best individual correct identification rates are achieved for the oboe class (98.3%), followed by the flute class (91.7%). Again, among the 6 instruments in the database, the cello samples are the worst classified (71.7%). Most important confusions involve the pairs cello–violin (16.7% of the cello samples were recognised as being violin) and clarinet–violin (12%). Finally, bowed strings and reeds families are identified 90% and 88% of the time respectively. It can also be noted that appending the delta did not significantly change the average percentage of correct identification (85.8% against 86.4%).

The overall performance is comparable to the systems described in [VR04] and [EB04] where respectively 90% and 84% correct identification were achieved for 5 instruments.

6.3.3 Summary

This series of experiments showed that training models using melodic phrases substantially improves the performance when one wishes to recognise melodic excerpts. By implicitly taking into account the note transitions that characterise melodic phrases, an increase of 13% in terms of average correct identification rates has been achieved for the K-means and SVM systems over the systems described in section 6.2.

Further, we have shown that using a K-means, the consideration of the delta LSF significantly degraded the performance. This can be interpreted as follows. By concatenating these two types of features, the two different information about the LSF and the delta LSF are averaged in a single model. As a result, the models do not correspond to characteristic spectral shapes (i.e. formant structures and high-pass spectral envelopes for the transients and onsets) of each instrument, thus explaining the drop in performance.

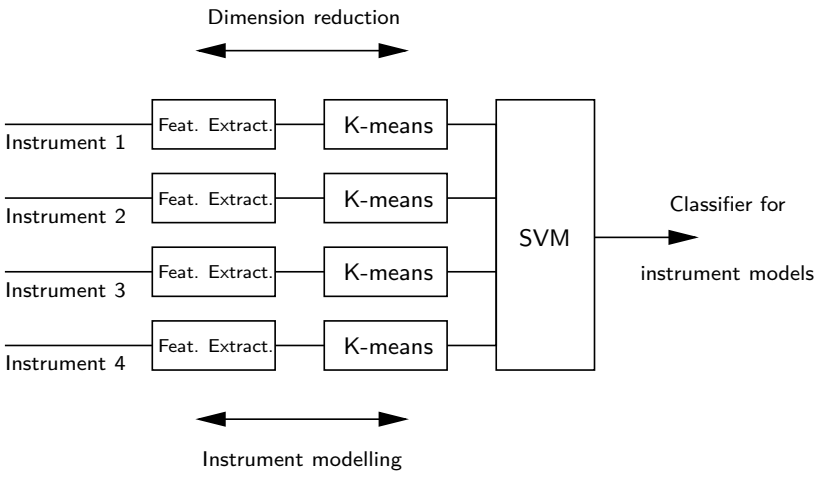


Figure 6.3: Principle of classifying instrument models using K-means and SVM. For each class, a K-means is performed on the extracted features. A multi-class SVM is then trained on the resulting codevectors to build the classifier.

6.4 Classifying instrument models

In this section, we propose to combine the generative and discriminative approaches in a single system [CDS05a]. In practice, a K-means is performed in order to obtain a condensed representation of the data intra-variability prior to the SVM optimisation.

Although it is known that the bigger the training set, the better the model, intra-class feature data are redundant by nature. For instance, the information contained in successive frames of a steady-state portion of note are highly correlated, especially when the LSF are used as spectral envelope descriptors (see figure 3.8). In the same vein, it has been shown in section 5.6 that reasonable system's performance can be obtained when features were extracted from the onset database only. It has been interpreted that the onset database was in fact a non-redundant snapshot of the un-segmented database.

For this reason, it can be argued that using K-means prior the SVM training can at the same time reduce the data dimension and serve to build instrument models. In the following, this operation is referred to as *cluster modelling*. This condensed and meaningful information is then input to the SVM for building a classifier for instrument models. The overall principle is depicted in figure 6.3.

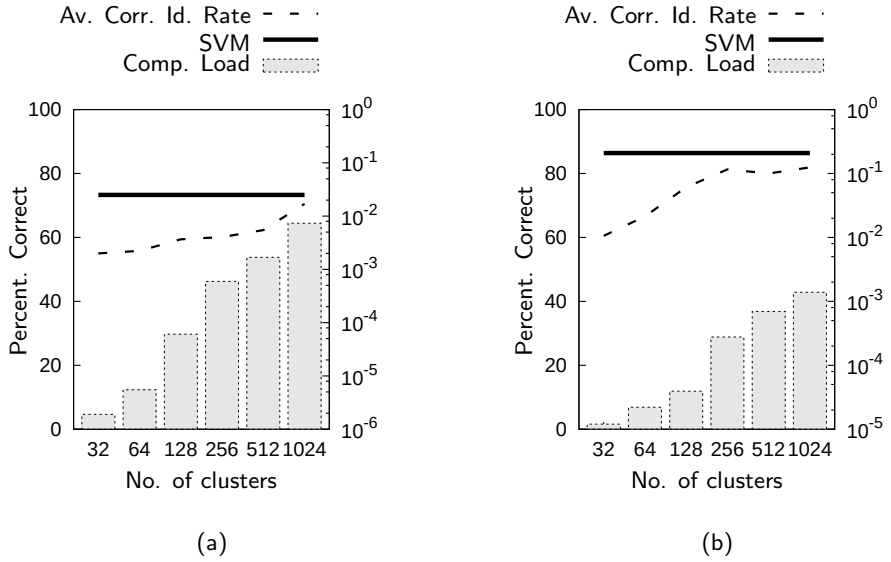


Figure 6.4: Total average correct identification rate as a function of the number of clusters when a cluster modelling stage is applied prior to the SVM optimisation (dashed line). (a) Training using isolated notes (b) Training using melodic phrases. The bold horizontal lines delineate the systems' performance when models are built without cluster modelling (73.3% and 86.4% respectively). Vertical bars represent the relative computational load in arbitrary unit needed for the SVM training.

6.4.1 Experiments

The experiments have been conducted as follows. Features were extracted from the same training data set as in sections 6.2 and 6.3. After the cluster modelling stage, the amount of data was reduced to 32, 64, 128, 256, 512 and 1024 codevectors respectively. A SVM was then trained and the same testing data set as in sections 6.2 and 6.3 was used to evaluate the performance.

In figure 6.4 are shown the systems' performance as a function of the number of clusters when an intermediate cluster modelling stage is used. Figures 6.4(a) and 6.4(b) correspond to experiments in which models were trained using isolated notes and melodic phrases respectively. The number of clusters is shown in abscissa while the percentages of correct identification are reported in ordinate (dashed lines). The vertical bars represent the relative computational load in arbitrary unit needed for the training. Values have been normalised so that 1 in the scale corresponds to the computational load of the systems yielding 73.3% and 86.4% correct identification respectively (see figures 6.1(d) and 6.2(c)). Corresponding performance are delineated by the horizontal solid lines.

When models are built from isolated notes (figure 6.4(a)), performance range from

55.0% for 32 clusters to 70.5% for 1024 clusters. At the same time, the computational load can be reduced by a factor 10^6 to 10^2 respectively. One can note the significant degradation in performance as the number of clusters decreases from 1024 to 256 compared to figure 6.4(b).

When models are built from melodic phrases (figure 6.4(b)), the training set is originally composed by roughly 8690 feature vectors for each instrument. Average correct identification rates range from 60.5%, for 32 clusters, up to 81.9%, for 1024 clusters. The best performance is achieved while roughly 11% of the original amount of data remains. At the same time the computational load can be reduced by a factor 10^5 (32 clusters) to 10^3 (1024 clusters) respectively. One can note that for 32 clusters, the performance are better if the K-means and the Euclidean similarity measure are used (77.5% for the generative approach against 60.5% for the mixed K-means/SVM system). This reveals to a certain extent the validity of our generative approach to the musical instrument identification problem.

Finally, one can note the graceful degradation in performance for the two systems as the number of clusters decreases.

6.4.2 Summary

In this section, we proposed to combine generative and discriminative approaches in a single system. In practice, a K-means is run prior to classification using SVM. The advantage of this strategy is twofold: first, it provides a meaningful approach to the data compaction problem by reducing a feature data set into a model of instrument. Second, the use of a cluster modelling stage significantly reduces the SVM training time. We have interpreted the process as *classifying musical instrument models*.

It has been experimentally shown that when models were trained using melodic phrases, the computational load for the SVM training can be significantly reduced at the expense of losing 4% in terms of correct identification. On the other, by training the models from isolated notes, a minimum drop of 7% in performance has been observed compared to a system without cluster modelling stage.

Mixed approaches, such as the one presented in this section, can provide a possible solution for dealing with huge databases of sounds, music pieces and songs. For instance, the automated classification of songs by genres, artists, bands or groups involves a tremendous amount of processing at the machine learning level. It can be argued that the design and building of such spectral or timbral similarity algorithms can be first performed using a mixed approach. Having noticed the graceful degra-

dation in performance they exhibit, the systems can be modified and improved while having reliable information about their *true* performance. The final and optimised versions of the algorithms can finally be built without cluster modelling stage.

Chapter summary

In this chapter, our system has been evaluated for the task of identifying musical phrases in a realistic musical context. Considering excerpts extracted from commercial recordings, the performance of both generative and discriminative approaches to the musical instrument identification problem have been evaluated and compared to other systems encountered in the literature.

Drawing on the conclusions reached in chapter 5, we have justified the use of the generative approach in a melodic context. In sections 6.2 and 6.3, we showed how well our learning approach performed when models were built from isolated notes and melodic phrases respectively. It has been concluded that training the models using isolated notes, the essential information present in musical phrases was not taken into account, thus resulting in lower identification rates during the experiments.

By using isolated notes for training the models, 64.4% and 73.3% correct identification rates have been achieved for the K-means and SVM. By training the models using melodic phrases, performance increased by 13% to 77.5% and 86.4% for the K-means and SVM respectively.

In section 6.4, we proposed a mixed approach to instrument identification problem by combining K-means and SVM in a single system. The process has been interpreted as *classifying instrument models*. The advantages in terms of computational load reduction have been highlighted. Specifically, it has been shown that the SVM training can be reduced by a factor 1000 at the expense of losing 4% in terms of correct identification when models were trained using melodic phrases. We suggested that such mixed approach can provide a possible solution for the design of computationally extensive algorithms such as the ones encountered in MIR for organising music databases.

Conclusion and perspectives

In this chapter, the work presented in this thesis is summarised. The achievements and contributions to the research field are highlighted. Limitations inherent to the use of supervised approaches for musical instrument identification problems – and especially in the case where the identification of poly-timbral mixtures is desired – are addressed. Research perspectives towards a better understanding of timbre on the one hand and the design of novel systems on the other hand are proposed. Extensions to other MIR applications such as texture modelling for music similarity measures are further discussed.

Thesis summary

This thesis investigated the use of computer algorithms for building automated musical instrument identification systems. Tackling the problem from a low-level perspective, a complete system that can be used as a module for MIR applications has been described. In essence, we proposed an approach unifying both physical and perceptual aspects of timbre that conforms to previous research on the perception of musical instrument textures by humans. The *formant theory of timbre* has been the main theme throughout.

Chapter 1 was concerned with the introduction of the basic principles driving the mechanisms of hearing and perception of sound by humans. This chapter also served to define the fundamental aspects of psycho-acoustics that have been considered when the ISO/MPEG psycho-acoustics model has been used in the experiments. Details about various timbre correlates and in particular about the *formant theory of timbre* have been given.

In chapter 2, several perceptual experiments assessing human performance in the task of identifying orchestral musical instrument sounds taken out of musical context have been reported and commented. Next, a literature review of various systems that

have been built to identify and classify musical instrument sounds has been provided. Finally, our approach to tackle the problem has been introduced.

In chapter 3, the various acoustic timbral descriptors used in our system have been introduced. First, we have recalled the fundamental principles of the source-filter LP model. We have justified why it can be transposed to model the mechanisms of sound production by musical instruments. In particular, we have presented the LSF which are derived from the LP coefficients. Interesting properties such as the inter-frame correlation and localised spectral sensitivity they exhibit are advantageously exploited when models are built using iterative averaging algorithms such as the K-means. Second, we introduced and described a perceptually motivated feature extraction method based on a signal sinusoidal analysis/synthesis model and proposed to use psycho-acoustic masking models as a way to select perceptually relevant sinusoidal components in the spectra.

Chapter 4 focused on the building of *instrument* and *database* models. To this effect, a distinction between *generative* and *discriminative* techniques has been made. We have subsequently isolated the tasks of identifying and classifying, the two of them being used for automatically organising musical instrument sound databases. Through the use of the K-means and GMM algorithms, the interpretation of learning characteristic formant structures has been proposed. This approach has been one of the central points of this work. Through the use of SVM, the interpretation of classifying spectral envelopes has been advanced.

The experimental part covered chapters 5 and 6. In chapter 5, a first series of experiments served to validate various hypotheses about timbre correlates that have been advanced *a-priori*. Using a database of isolated notes, a preferred prediction order and number of codewords for the models that maximised the average performance have been determined. We then showed that using the LSF yielded better performance than using other acoustic descriptors, and in particular better than the filter polynomial and reflection coefficients on the one hand and the MFCC on the other hand. The overall performance in terms of instrument family identification of the base system have been found to be comparable to what humans can achieve for similar instrument families. Using a database of isolated notes that allowed us to finely tune experimental protocols, we have been able to characterise the dependence of our system on pitch and proposed the fundamental frequency as a prior to both instrument modelling and identification phases. Cross-register experiments allowed us to confirm the dependence on pitch of the system.

Next, we showed that extracting similar features independently from the onset and

steady-state segments of sound slightly decreases the overall system performance. It has been then concluded that the models consisting of average spectral envelopes determined using a K-means algorithm were able to capture the essential information about the notes' onset structure, thus justifying their use to identify melodic phrases.

Finally, we proposed to include psycho-acoustic considerations at the feature level. For this purpose, the calculation of a perceptual version of the LSF, the PLSF has been introduced. The performance were compared to the standard LSF on the one hand and to the WLSF. We have related the performance of the WLSF with the ones that can be achieved with the MFCC and concluded that a similarity measure based on the Euclidean distance was inappropriate to deal with perceptual scales. Experiments involving the PLSF confirmed this hypothesis.

In chapter 6, we have evaluated the system performance at identifying and classifying melodic phrases extracted from commercial recordings. Following the conclusions drawn in chapter 5, we have justified the use of the generative approach in a melodic context and it has been shown how well it performed in realistic recording conditions. It has been found that training models on melodic phrases yields better performance than training models using isolated notes. We have also confirmed that for similar considered features, a classification approach using SVM yielded better performance than a learning approach using K-means.

Thesis achievements

This thesis proposed an approach to the musical instrument identification problem specifically designed to preserve the coherence between physical and perceptual considerations of timbre. Each layer composing our system has been related to the principles governing the *formant theory of musical timbre*.

The use of the LSF as spectral envelope and formant structure descriptors has been presented in a musical instrument context. Drawing on research in speech coding, we have justified their use with the K-means for building acoustic musical instrument models. The combination LSF/K-means as well are the interpretation of the learning process are novel in this particular research field.

We have shown that choosing too low or too high prediction orders degraded the system's performance. This can be explained as follows: although small order LP analyses can estimate spectral envelopes, they do not accurately capture the detailed information carried by the formant structure. On the other hand, high prediction orders tend to model too much spectral information such as salient partials that the

spectra can exhibit. This way, we confirmed that formant structures were important features for distinguishing between acoustic musical instrument sounds. It has been further advanced that they are important timbre correlates.

Although research works in speaker recognition use the LSF as acoustic descriptors, state-of-the-art systems are often based on the combination MFCC/GMM. In our experience, it has been found that the pair MFCC/GMM on the one hand and MFCC/K-means yielded worse performance that the combination LSF/K-means. This can be interpreted as follows: in speech, the mechanisms of speech production across speakers are similar so that the formant structure cannot be considered as being a salient characteristic of a speaker.[2] By contrast, one principle of the formant theory of musical timbre states that the formant structure is a salient characteristic of the instrument, so that it can be used to uniquely characterise it. This can explain the good performance of the combination LSF/K-means to deal with orchestral musical instruments. This leads us to advance that the combination LSF/K-means can efficiently be used to discriminate between speech and acoustic musical instrument sounds, if speech is considered as a *class of instrument* on its own. Indeed we have shown in [CDS05b] that the system presented in this thesis yielded good performance in discerning singing voice sounds from acoustic musical instrument sounds.

We have shown that pitch influences LSF so that building models over different frequency registers improved the system's performance. These experiments lead to the following concept of pitch abstraction or dependence in musical instrument identification system design. The spectral envelopes of a musical instrument sound – as have been used here – are related in some ways to the pitch, and in a similar manner, timbre is related to the pitch. On the one hand, an ideal system should model the timbre independently of the pitch whereas on the other hand, another model should be able to quantify the timbral variation as a function of pitch.

Limitations

The challenge in timbre modelling is to capture salient properties characterising the tone colour of sound. Since timbre is believed to be a multi-dimensional attribute, a computer approach involving a multi-feature extraction stage should theoretically better fit this reality. Corresponding systems abound in the literature and are prevalent directions for investigation for many researchers within the MIR community. In

[2] note that this remark is not valid if one considers speakers with vocal tracts of different sizes, i.e. males vs females vs children

particular, the multi-descriptor based approaches including automatic feature selection algorithms belongs to this category. However, the *a-posteriori* interpretation of the results, that is, trying to understand why a particular set of features is performing better than others is often impossible. Although constituting the most promising approaches for classification tasks, they also mark the upper-limit in terms of performance achievable using supervised systems. That is: systems that *learn* the timbre[3] from a set of acoustic descriptors can see their performance increase if the database as a whole is considered, or in other words, if a discriminative approach involving all the instrument feature distributions in the database is chosen. The experiments involving SVM conducted in chapters 5 and 6 illustrates this fact. However, strong limits are set on the practical use of such systems, especially in terms of modularity, since the addition of a new instrument in the database necessitates a new training and the determination of new database models.

Our approach tackles the problem from a different learning perspective. In essence, the acoustic timbral descriptors have been chosen *a-priori* so that each of them tend to model one particular aspect of timbre. These hypotheses were then experimentally verified or refuted. Spectral envelope descriptors, and in particular the LSF, have been found to be *an* essential contributor to modelling the timbre of sound. However, it is known that they only represent one of the various facets of timbre.

The challenge in this research field is to relate our understanding of timbre, from its perceptual attributes to its acoustic correlates, to a system biased at several layers. In other words, what one believes to be an appropriate timbral descriptor *a-priori* might not necessarily contribute to improve the system's performance. Experiments conducted in section 6.3 using the *delta* LSF illustrated this. Hence the narrow relation between the type of acoustic features and machine learning algorithms.

The consideration of psycho-acoustic principles at the feature level has shown its limitation. Although improvements at the instrument-level have been noticed in experiments involving isolated notes, none of the proposed perceptual versions of the LSF (WLSF and PLSF) have shown to significantly improve the average correct identification rates.

While our system showed performance comparable to what humans can achieve, it still performs worse on average in terms of exact instrument identification. Another key point worth mentioning is concerned with the confusions between instruments and families of instruments that can occur. We have stressed that computer systems sometimes make non-meaningful and un-expected confusions so that a wind instru-

[3] note the term *timbre* might not be appropriate for these types of systems

ment sound can be mis-identified as a bowed string instrument sound, and this at a non-negligible rate. This fact has been observed in our experiments. It can therefore be expected that since such systems are by nature intended to be compared to human performance, there is obviously room for improvement.

Looking back at figure 2.3, where an ideal musical instrument identification system is depicted, one can see the tasks remaining to be tackled. This thesis investigated a possible solution for the identification of mono-timbral orchestral instrument sounds. Other research works focus on the identification of percussive sounds [GR04] [HDG03]. The challenging extension of any research work in musical instrument identification is concerned with the identification of poly-timbral mixtures. Indeed, another aim of musical instrument identification systems is to model this ability of humans to differentiate between two sound objects played in unison. It has been mentioned in section 2.2.4 how a supervised approach to a pattern recognition problem could be used to deal with polyphonic mixtures. In essence, models have to be trained for all possible combinations of all possible instruments. This approach is perfectly suitable for particular and specific tasks such as the identification of jazz ensembles, duos or trios but inherently lacks generalisation properties, as has been shown in [ERDa].

Perspectives

Research perspectives towards the *formant theory of timbre* are twofold. First in [PA93], it is mentioned that the use of weighted LSF distance measures during the similarity measure calculation improved the codebook design in terms of introduced quantisation noise. Likewise a similar weighting has been proposed in [Pal88] for speech recognition purposes. By extending this concept to musical instrument identification, one could think about designing a similarity measure in which the contribution of each LSF parameter is individually weighted. In essence, these weights could be determined using an empirical approach, by using genetic algorithms, for example, in a similar way to that described in [Fuj98]. Second, from a perceptual perspective, one can attempt to quantify the precise contribution of the formant structure to timbre in an analysis/synthesis framework. For instance, by using a sinusoidal or LP model, an algorithm can be designed to alter the sound formant structure on a short-term basis. Modifications such as formant bandwidth expansion, formant emphasis, formant flattening or formant shifting could be studied in such framework. Finally, alterations of the quality of sounds could be explored using listening tests.

At the signal level, a joint feature extraction between formant structure and resid-

ual signal in a LP framework can be envisaged. It has been mentioned in section 3.2.1 that using LP analysis, the contributions of the resonating body and excitation can be characterised. Musical instruments have different mechanisms of sound production so that it can be argued that descriptors extracted from the residual signals might characterise the different excitations, or different families of excitations (e.g. reeds, double reeds, winds or bowed strings). Likewise a transient/sinusoidal signal separation in a sinusoidal modelling framework can be employed for similar reasons: to characterise the tonal musical property of the sound on the one hand, and the noisy or transient characteristics of the signal on the other hand. These two orientations for future research involve the calculation of acoustic features from two types of signals. From there arise concerns about the strategy for merging several features in a single system. The usual procedure encountered in the literature is to perform the fusion at the feature level, by concatenating the descriptors in a single vector.[4] This approach has shown its limitation so that automated feature selection are now preferred. On the other hand, the use of data fusion techniques at the classifier level, which is a research field on its own, shouldn't be ruled out.

When modelling timbre, our system and the ones encountered in the literature *lose* time consideration. In other words, the temporal organisation of the various acoustic events is not *represented* at the model level. This approach is understandable if one considers timbre as a global attribute of sound. However, we showed in our experiments that, for example, onset and steady-state segments of tones have different characteristics, so that they each contribute to a particular aspect of timbre. Instead of *averaging* them in one single model, one could think of independently and explicitly modelling them. Similar to speaker recognition, the incorporation of Dynamic Time Warping (DTW) or Hidden Markov Models (HMM) can constitute a possible orientation for future research.

A prevalent direction in modern research in musical signal processing is concerned with the consideration of mid-level musical knowledge during the signal analysis. As a consequence, the waveforms are not only seen as *audio signals* but musical priors are set on their spectral and temporal structures. For instance, such priors can consist of the consideration of the Western equally tempered frequency scale when one wishes to analyse the pitch of a sound, or of the harmonic content when one wishes to decompose a signal into harmonic objects or again of the temporal organisation of musical events such as notes succession, beats and rhythm.

The harmonic matching pursuit algorithm described in appendix B can fit this

[4] as has been performed with the delta LSF in section 6.3

purpose. By considering a signal as a sum of harmonic *plus* sinusoidal components *plus* noise, information about harmonicity, inharmonicity and spectral harmonic energy can be isolated and used as input to a musical instrument identification system. It can further be advanced that such technique could serve the determination of multiple pitches in polyphonic mixtures, thus constituting in some ways a possible approach to the poly-timbral musical instrument identification problem.

Automated classification of songs by genres, artists, bands or groups is more and more needed and constitutes to a certain extent the natural prolongation of our research. In particular, *spectral similarity* evaluation has found numerous applications in musical genre classification and automatic playlist generation. Also termed as *texture similarity* evaluation, a common approach to the problem essentially consists of transposing a supervised algorithm to a classification problem involving much more complex sounds than the ones used in this thesis. In essence, acoustic features are extracted from entire musical pieces, or songs, that have been previously organised into classes of interest.

At the expense of losing the temporal organisation of sound objects that characterises music, this approach relies on the predominance of characteristic sound textures in the mixtures. It is assumed that this predominance is reflected at the acoustic descriptor level and captured by the models.

Semantic Interaction with Music Content (SIMAC)

Part of this research work has been carried out within the SIMAC [SIM] project, a European funded network aimed at investigating novel methods and algorithms for the automatic manipulation of audio and musical contents. Concentrating at the same time on both low- (acoustic) and high- (metadata) level aspects of music, the SIMAC project proposes to apply complementary approaches to tackle the problem of archiving, classifying and retrieving musical information. Our contribution towards an integration of our system in such wide framework relied on the proposal of novel acoustic features for this research field and computationally efficient methods for acoustic timbral modelling (see section 6.4). We have also accompanied our work by an extensive evaluation of the system's performance for the task of identifying and classifying orchestral musical instrument sounds.

Publications

This thesis yields the publication of the following articles in national and international conference proceedings:

- **"Linear Predictive Models for Musical Instrument Identification"**

  Nicolas Chétry and Mark Sandler

  accepted for publication in Proc. ICASSP, 2006, Toulouse, France.

- **"Identification of Monophonic Instrument Recordings using K-means and Support Vector Machines"**

  Nicolas Chétry, Mike Davies and Mark Sandler

  in Proc. Digital Music Research Network Conference, 2005, Glasgow, UK.

- **"Musical Instrument Identification using LSF and K-means"**

  Nicolas Chétry, Mike Davies and Mark Sandler

  in Proc. AES 118th Convention, 2005, Barcelona, Spain.

- **"An Efficient Two-Stage Implementation of Harmonic Matching Pursuit"**

  Chris Duxbury, Nicolas Chétry, Mark Sandler and Mike Davies

  in Proc. EUSIPCO, 2004, Vienna, Austria.

In addition:

- **"Embedding Side Information into a Speech Codec Residual"**

  Nicolas Chétry and Mike Davies

  submitted for publication in Proc. EUSIPCO, 2006, Florence, Italy.

- **"The Formant Theory of Timbre: Principles and Applications to Musical Instrument Identification"**

  Nicolas Chétry and Mark Sandler

  in preparation for submission in J. Acoust Soc. Am..

Appendices

A. Overview of the MPEG-1 layer II psycho-acoustic model

Psycho-acoustic models are of common use in modern audio and musical signals processing algorithms. Fundamental principles of psycho-acoustics and mechanisms of hearing have been introduced in chapter 1. In particular, the notions of absolute threshold of hearing, frequency and temporal maskings have been recalled.

This appendix is concerned with the description of the MPEG-1 layer II psycho-acoustic model. After having briefly recalled the scope in which such models are used, the process yielding the calculation of the global masking curve for a frame of signal is detailed step by step.

A.1 Principle

The aim of perceptual models is to give a quantitative indication about the frequencies that are perceived and about the ones that are masked. The principle of psycho-acoustic driven encoding schemes is optimum in terms of quantisation noise shaping. For instance, common audio codecs tend to minimise the quantisation noise, which is uniform in all the frequency bands.[1] However, it may happen that this noise, introduced by the quantiser, has a level close to the signal level and becomes audible, as illustrated in figure A.1(a).

In perceptual codecs, the use of psycho-acoustic knowledge helps to shape the quantisation noise level in a way that it always stays below the signal level that have to be encoded (figure A.1(b)). Traditionally performed after a subband analysis, the process consists of injecting as much noise in each frequency band as possible by allocating in each band a number of bits determined by the psycho-acoustic model.

[1] this is case in the PCM quantiser in which samples are uniformly quantised to 16 bits

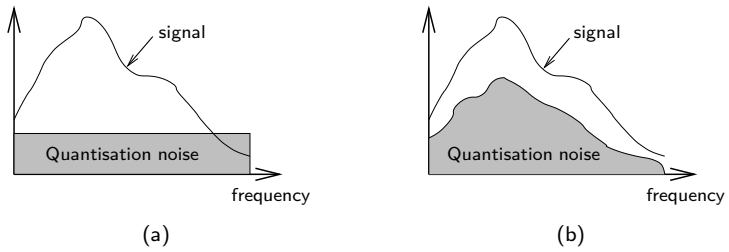


Figure A.1: Diagrams showing the effects of shaping the quantisation noise in perceptual codecs. (a) Uniform quantisation noise and (b) Quantisation noise shaped according to the signal instantaneous frequency content.

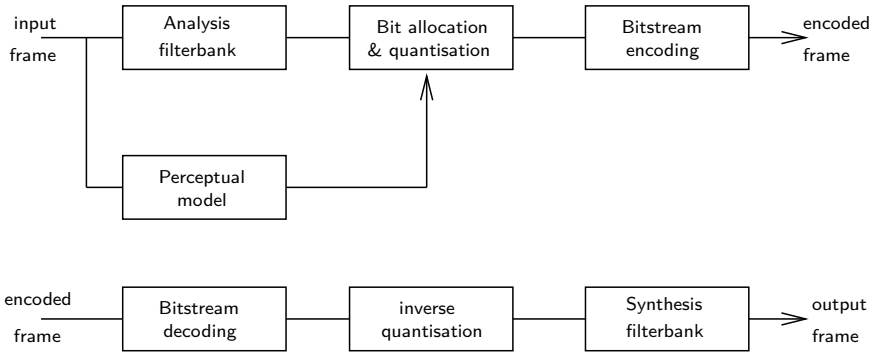


Figure A.2: Block diagram of a general perceptual encoder (top) and decoder (bottom).

In figure A.2 is depicted a basic block diagram of a perceptual coder/decoder system. It consists of the following:

- **Analysis/synthesis filterbank**: it is used to decompose the input frame according to a time-frequency representation. The outputs of the analysis filterbank are subsampled version of the time domain input samples calculated for each frequency band. The analysis/synthesis filterbank system is usually designed in order to respect a quasi-perfect reconstruction constraint.

- **Perceptual model**: the perceptual model, by determining first the individual masking thresholds for both tonal and non-tonal components, and then by taking into account the absolute threshold of hearing, outputs the global masking threshold. The latter determines the maximum level of inaudible noise that can be injected in each subband.

- **Bit allocation and quantisation**: given a targeted bit rate, the global masking

threshold is used to optimally allocate the number of bits for the quantisation of the *parameters* in each subband. Depending on the codec, the encoded parameters are either the time/frequency domain samples, or their transformed version after a MDCT.

A.2 The MPEG-1 layer II psycho-acoustic model

The MPEG standard (1 and 2) (layer I, II or III) is by far the most spread algorithm for audio compression purposes. It can operate at several bit rates (fixed and variable), at different sampling rates, and in different spatial configurations (monophonic, stereo and joint-stereo). More information about the codec can be found in the standard [ISO] and in [Pan95].

In its first version, the MP3 codec allows to compress signals at rates up to 6:1 without perceptible loss in quality: in such a configuration, the bandwidth of a 16 bits stereo signal sampled at 48 kHz can be reduced to 256 kbits/s. Note however that the Sony ATRAC (Mini-Disc), and its successor the ATRAC-3, AAC and Ogg-Vorbis codecs include similar psycho-acoustic models.

In this section, the MPEG-1 layer II psycho-acoustic model is presented. In MPEG, the bit allocation of the $S = 32$ subbands is calculated on the basis of the signal-to-mask ratio (SMR) calculation in each subband. It measures the difference between the signal level and the noise level that can be added in each subband. It is therefore necessary to determine for each subband its maximum signal level and its corresponding minimum masking threshold. The determination of the SMR is based on the following steps:

1. Calculation of the FFT for the spectral analysis
2. Calculation of the acoustic pressure level in each subband
3. Determination of the tonal (sinusoid-like) and non-tonal (noise-like) components
4. Selection of the components used for the masking threshold calculation
5. Calculation of the individual masking thresholds
6. Calculation of the global masking threshold
7. Determination of the minimum masking threshold in each subband
8. Calculation of the SMR in each subband

In the following, we will assume that the signal $x(n)$ is sampled at $f_s = 22.05$ kHz. For the illustrations, one frame of trumpet sound is considered.

A.2.1 Power Spectral Density (PSD) calculation

The layer II psycho-acoustic model is based on the 1024 point power density spectrum (PSD) $X(k)$ calculated for the input frame $x(n)$ weighted by a Hann window $w(n)$:

$$X(k) = 20\log_{10}\left|\frac{1}{N}\sum_{n=0}^{N-1} w(n)x(n)e^{-2\pi jkn/N}\right|, \qquad k = 0, ..., \frac{N}{2} - 1 \qquad \text{(A.1)}$$

where $w(n)$ is the Hann window defined as

$$w(n) = \frac{1}{2}\sqrt{8/3}(1 - \cos(2\pi n/N)), \qquad n = 0, ..., N-1 \qquad \text{(A.2)}$$

The PSD is then normalised to the reference level of 96 dB SPL.

A.2.2 Sound pressure level calculation

The sound pressure level L_{sb} in each subband s is computed by:

$$L_{sb}(s) = \max_k X_s(k) \qquad \text{(A.3)}$$

where $X_s(k)$ is the sound pressure level of the kth spectral line within the subband s.

A.2.3 Determination of the tonal and non-tonal components

The tonality has an influence on the masking threshold. The tonal (more sine-like) and non-tonal (more noisy-like) components are then determined using the PSD $X(k)$. This step starts with the determination of the local maximum, then extracts the tonal components and estimates the intensity of the non-tonal components within each critical band.

A spectral line is labelled as local maximum if

$$X(k-1) < X(k) \geq X(k+1), \qquad k = 2, ..., 500$$

The bandwidth of the critical bands varies from 100 Hz at low frequency to 4000 Hz at high frequency. To determine if a local maximum may be a tonal component,

a frequency range δf around the local maximum is examined:

$$\delta f = \begin{cases} 86.133 \text{ Hz} & \text{if } 0 \text{ kHz} < f \leq 2.756 \text{ kHz} \\ 129.199 \text{ Hz} & \text{if } 2.756 \text{ kHz} < f \leq 5.512 \text{ kHz} \\ 258.398 \text{ Hz} & \text{if } 5.512 \text{ kHz} < f \leq 10.336 \text{ kHz} \end{cases}$$

A local maximum is labelled as a tonal component when its sound pressure level is 7 dB above the spectral lines over the considered window δf. In other words, k corresponds to a tonal component when:

$$X(k) - X(k+j) \geq 7 \quad \text{dB}$$

where

$$j = \begin{cases} -4, +4 & \text{for } 4 < k < 128 \\ -6, ..., -2, +2, ..., +6 & \text{for } 128 \leq k < 256 \\ -12, ..., -2, +2, ..., +12 & \text{for } 256 \leq k < 500 \end{cases}$$

Next, the new sound pressure levels of the tonal components are calculated by averaging the PSD over a small neighbourhood of the local maxima:

$$X_t(k) = 10 \log_{10} \left[10^{X(k-1)/10} + 10^{X(k)/10} + 10^{X(k+1)/10} \right]$$

Once the tonal components are determined, the non-tonal components are calculated from the remaining not examined components. Within each critical band (26 are considered in layer II for $f_s = 22.05$ kHz), the spectral lines powers are summed to form a new unique non-tonal component whose new index k is the index of the nearest spectral line to the geometric mean of the critical band.

A.2.4 Decimation of tonal and non-tonal masking components

This stage is used to reduce the number of maskers which are considered for the calculation of the global masking threshold. Firstly, only the components (tonal and non-tonal) above the absolute threshold of hearing in quiet LT_q are retained. LT_q is determined from psycho-acoustic experiments. Secondly, no more than one tonal component is selected within a distance of 0.5 Bark. In practice, the highest power component over the window is kept and the others are discarded from the list.

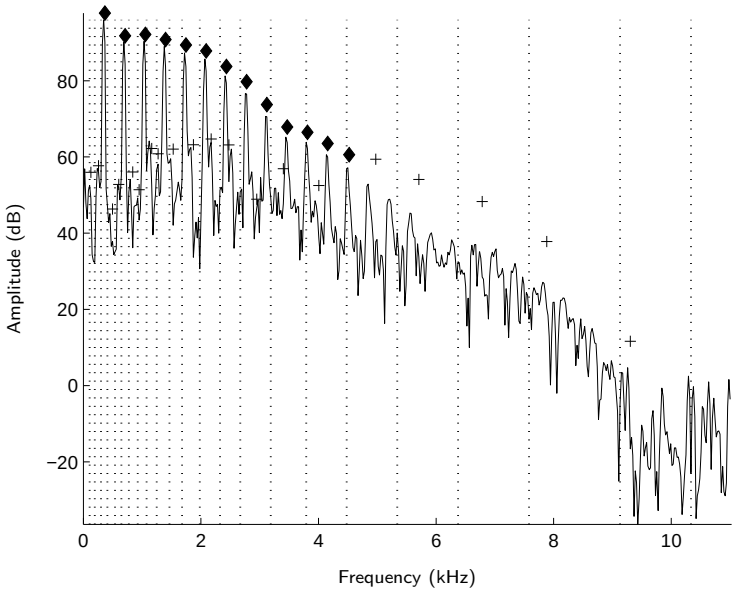


Figure A.3: PSD (solid line) normalised at 96 dB SPL, tonal components (diamonds markers), non-tonal components (plus markers) and critical band boundaries (dotted vertical line).

A.2.5 Individual masking thresholds calculation

In MPEG, only a subset of the $N/2$ samples are considered for the global masking threshold calculation. In layer II, their number is 132 and they are referenced in tables. Every tonal and non-tonal component is assigned the frequency value taken from the tables that closely corresponds to the frequency of the original spectral line $X(k)$. The individual masking thresholds for both tonal and non-tonal components are then individually calculated using:

$$\begin{aligned} LT_{tm}[z(j), z(i)] &= X_{tm}[z(j)] + av_{tm}[z(j)] + vf[z(j), z(i)] \quad \text{dB} \\ LT_{nm}[z(j), z(i)] &= X_{nm}[z(j)] + av_{nm}[z(j)] + vf[z(j), z(i)] \quad \text{dB} \end{aligned}$$

where LT_{*m} and LT_{*m} represent the individual masking thresholds at critical band rate z in Bark of the masking component at the critical band rate of the masker z_m in Bark. The values in dB can be either positive or negative. The term X_{*m} is the sound pressure level of the masking component with the index number j at the corresponding critical band rate $z(j)$. The term av_{*m} is given by:

$$\begin{aligned} av_{tm} &= -1.525 - 0.275 * z(j) - 4.5 \quad \text{dB} \\ av_{nm} &= -1.525 - 0.175 * z(j) - 0.5 \quad \text{dB} \end{aligned}$$

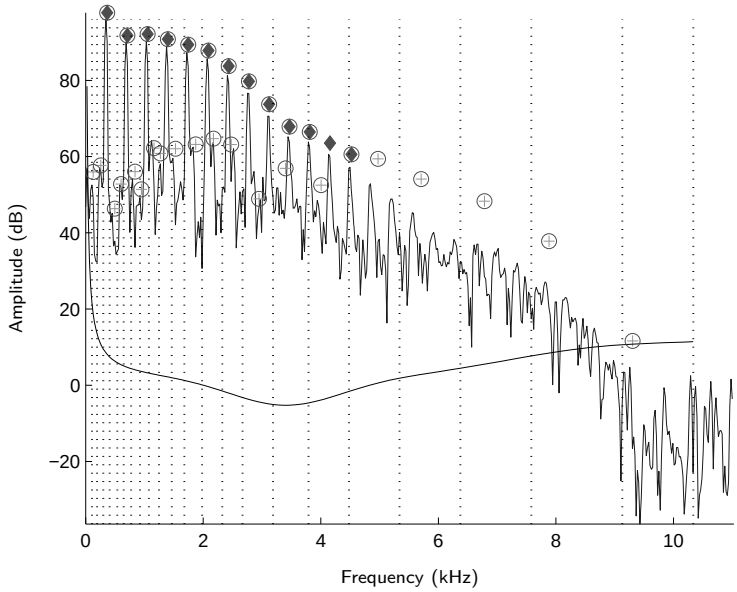


Figure A.4: Circled markers correspond to the retained components for the calculation of the individual masking thresholds, the others are discarded. The line at the bottom marks the level of the absolute threshold of hearing in quiet. Note that the analytical formula in section 3.5.3.1 has been used in our implementation instead of the table given in the standard [ISO].

The masking function vf of a masker is characterised by different lower and upper slopes, which depend on the distance in Bark $dz = z(i) - z(j)$ to the masker. It is given by:

$$vf = \begin{cases} 17(dz+1) - (0.4X[z(j)]+6) & \text{dB for } -3 \leq dz < -1 \text{ Bark} \\ (0.4X[z(j)]+6)dz & \text{dB for } -1 \leq dz < 0 \text{ Bark} \\ -17dz & \text{dB for } 0 \leq dz < 1 \text{ Bark} \\ -(dz-1)(17-0.15X[z(j)]) - 17 & \text{dB for } 1 \leq dz < 8 \text{ Bark} \end{cases}$$

where $X[z(j)]$ is the sound pressure level of the jth masking component in dB. Finally, the masking is no longer considered if $dz < -3$ Bark or $dz \geq 8$ Bark.

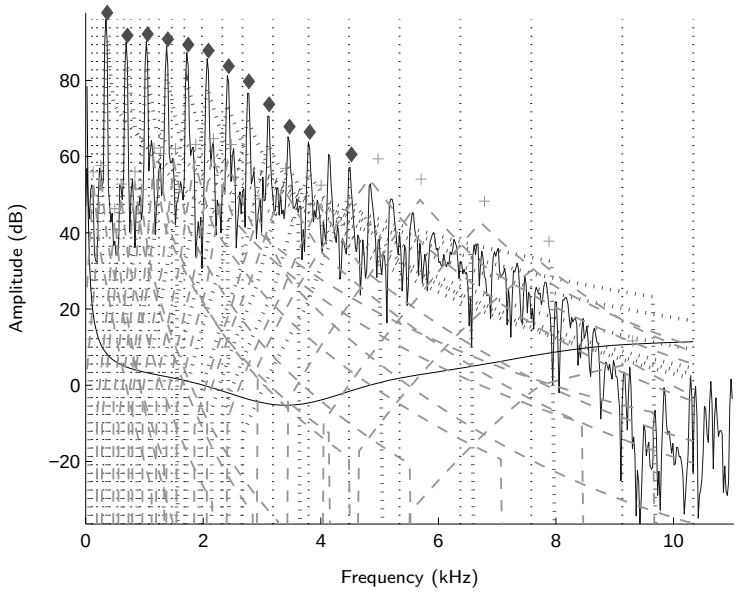


Figure A.5: Individual masking thresholds for tonal (dash-dotted line) and non-tonal (dashed line) components.

A.2.6 Global masking threshold calculation

The global masking threshold LT_g is calculated by summing the powers corresponding to the individual masking thresholds and the threshold in quiet. Mathematically,

$$LT_g(i) = 10\log_{10}(10^{LT_q(i)/10} + \sum_{j=1}^{m} 10^{LT_{tm}[z(j),z(i)]/10} + \sum_{j=1}^{n} 10^{LT_{nm}[z(j),z(i)]/10})$$

The total number of tonal maskers is given by m, while the total number of non-tonal maskers is given by n.

A.2.7 Minimum masking threshold determination

The minimum masking level $LT_min(n)$ for each subband n is determined by:

$$LT_{min}(n) = \min_k LT_g(i) \quad \text{dB}$$

where k is the index of the spectral line within the frequency band n.

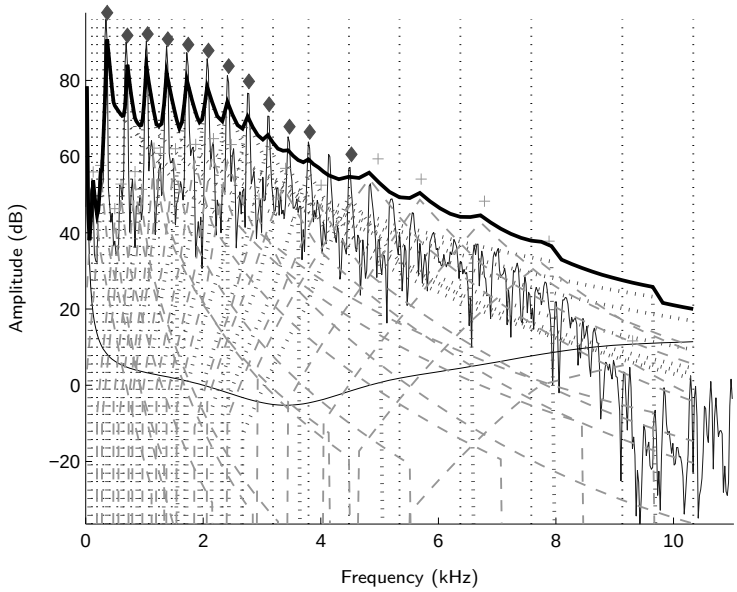


Figure A.6: Global masking threshold (bold solid line) calculated as the sum of the individual masking thresholds with the absolute threshold in quiet.

A.2.7.1 Signal-to-mask ratio calculation

For every subband, the signal-to-mask ratio is computed using:

$$SMR_{sb}(n) = L_{sb}(n) - LT_{min}(n) \quad \text{dB}$$

The signal-to-mask ratio represents the level of noise that it is possible to inject in each subband. In most codecs, the SMR is used to drive a bit allocation scheme responsible for the parameters quantisation. The operation is usually performed in a transform domain (e.g. MDCT, MCLT). Starting from a total number of available bits (fixed for each frame in the case of fixed bit rate transmission or variable and taken from a bit reservoir in the case of a VBR scheme), the algorithm firstly allocates the bits to the subbands having small SMR before sharing the remaining bits between the other subbands.

B. A two-stage implementation of Harmonic Matching Pursuit

In this chapter, the sinusoidal analysis/synthesis technique presented in section 3.5.2 is extended for the analysis and processing of audio signals created by musical instruments. In particular, a novel approach to the iterative atomic decomposition problem is proposed. This technique provides a computationally efficient and meaningful approach to the harmonic grouping principle encountered in auditory scene analysis (see section 1.2).

An efficient two-stage implementation of the Harmonic Matching Pursuit (HMP) algorithm for fixed atom duration is proposed. Its application in a musical context is illustrated by two examples of pitch determination for musical phrases and piano notes separation respectively [DCSD04].

B.1 Harmonic signals

A harmonic signal is defined as a signal composed of several sinusoidal components having their frequencies as integer multiples of a fundamental frequency f_0. In figure B.1 are shown the waveform, magnitude and phase representations of a harmonic signal.

There is an infinite number of acoustic signals having harmonically or locally harmonically related components: most musical tones than can be produced by instruments or the voiced sounds produced by the human voice box exhibit harmonic properties.

The fundamental or first partial is defined as the lowest frequency of the harmonic series. The second partial is twice the frequency of the fundamental, which makes it an octave higher. The third harmonic partial, at three times the frequency of

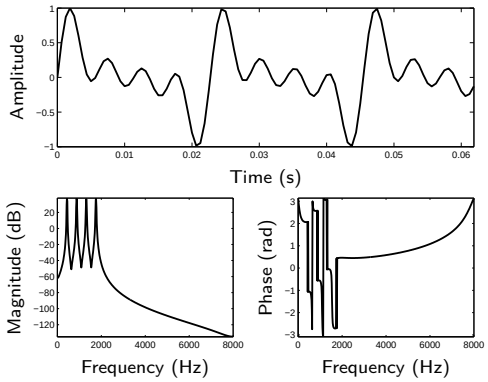


Figure B.1: Harmonic signal time-domain waveform (top) and frequency representation using the STFT. Magnitude (bottom left) and phase (bottom right) after Hanning windowing.

the fundamental, is a perfect fifth above the second harmonic. Similarly, the fourth harmonic partial is four times the frequency of the fundamental; it is a perfect fourth above the third partial (two octaves above the fundamental).

However, not all musical instruments have partials that exactly match the harmonic series described above. In [Bla65], it is quoted that

> *Most musical instruments produce tones whose partial tones, or overtones, are harmonic: their frequencies are whole-number multiples of a fundamental frequency. The piano is an exception.*

In music, the concept of inharmonicity refers to the degree to which the frequencies of the overtones of a fundamental differ from whole number multiples of the fundamental frequency. These inharmonic overtones are often distinguished from the harmonic ones: since the harmonics contribute to the sense of sounds as pitched or unpitched, the more inharmonic a sound, the less definite it becomes in pitch. Many percussion instruments such as cymbals, toms, chimes or the piano create complex and inharmonic sounds.

The decomposition of audio signals into harmonic atoms is interesting for the following reasons:

- first in terms of high-level musical interpretation. As mentioned in section 1.1.5, the harmonic relation between partials characterises among other factors the pitch and the timbre of musical signals. The concept of pitch used in musical annotation to classify notes and chords on a quantised frequency scale is related to the partials frequency distribution. Techniques for harmonic signal decom-

position have found numerous applications in automatic music transcription [Kla03] or more generally for signal analysis and modelling purposes.

- second for coding and compression purposes. The use of harmonic or harmonic plus noise models (e.g. Harmonic and Individual Line plus Noise models (HILN) in the MPEG-4 standard) allows the bit rate to be significantly reduced compared to conventional algorithms using sinusoidal components. The principle is to decompose the input signal into harmonic audio objects corresponding to more appropriate models regarding the signal structure. For instance, objects such as sinusoids, harmonic tones, and noise are used in the HILN coder. This approach allows the introduction of more advanced source model than merely assuming a stationary signal for the duration of a frame.

B.2 Principle

Harmonic Matching Pursuit (HMP) algorithms are a direct extension of the Matching Pursuits Decomposition (MPD) introduced by Mallat and Zang in [MZ93]. This iterative procedure consists of finding a sub-optimal signal representation in a highly redundant dictionary of Gabor atoms of the form:

$$g_{s,u,\xi}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-u}{s}\right) e^{j2\pi\xi(t-u)} \tag{B.1}$$

where u, s and ξ are the amplitude scaling, time and frequency shift factors respectively of a mother window $w(t)$. Due to the harmonic nature of audio and musical signals, it is interesting to consider harmonic atoms for the decomposition. The HMP theoretical background has been exposed in [GB03]. In essence, an harmonic atom is of the form:

$$h(t) = \sum_{k=1}^{K} c_k g_{s,u,\xi_k}(t) \tag{B.2}$$

where the factors c_k weight the contribution of each Gabor function in the atom $h(t)$. It is further assumed that $\xi_k \approx k\xi_0$, $k = 1, \ldots, K$, models the harmonic relationship between fundamental and overtones frequencies. By considering real-valued atoms, the problem of extracting, at each iteration, the harmonic series removing the most energy from the signal can be tackled in the frequency domain using the analysis/synthesis techniques described in the preceding sections.

Although having been independently carried out, the work presented in this section and the one reported in [GB03] share the same basic principles. The main difference

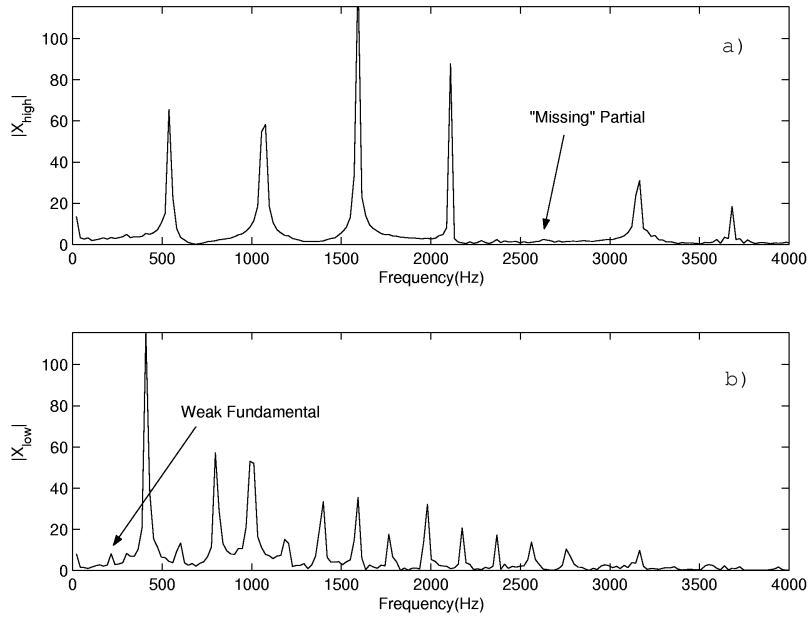


Figure B.2: Spectral plots of both (a) low and (b) high violin notes showing respectively a missing partial and a weak fundamental frequency.

between the two techniques is concerned with the resolution used. In [GB03], the notion of duration or scale of the harmonic grain is introduced. In essence, the technique requires the calculation of multiple FFT (one for each scale) at each iteration. Although being more crude, this approach deals with fixed resolution atoms and aims at decomposing locally stationary signals that can be obtained, for example, after a transient/sinusoidal signal separation.

In basic matching pursuit, atoms from a continuous and overcomplete dictionary have to be considered at each iteration. As a consequence, this exhaustive search often leads to large computational requirements. A modified two-stage implementation for fixed-scale atoms is presented in the following. More specifically, it consists of a dual resolution spectral approach that significantly reduces the computational requirements while still maximising the energy extracted at each stage.

It has been mentioned in section 1.1.5 that the fundamental frequency may not be present in the signal for some musical tones without having an effect on the perceived pitch. Likewise, musical sounds might have missing partials. As an illustration, figure B.2(a) shows a missing partial, and figure B.2(b) shows a weak fundamental frequency component for two violin tones. These two sounds are however perceived as if the latter frequency components were present.

Because of these issues, simply choosing the position of the lowest frequency high

energy sinusoidal component as a possible fundamental frequency may not be relevant in terms of meaningful grain extraction. Therefore, we propose to use a harmonic energy criteria for the harmonic grain selection. More generally, the algorithm consists of an initial low-resolution pitch analysis followed by a high resolution harmonic grain extraction based on local complex interpolation within the spectral domain. It is composed of:

- Low resolution harmonic energy analysis: during this stage, the harmonic energy is calculated for each potential fundamental frequency within a FFT frame. The harmonic series corresponding to the maximum harmonic energy is chosen. A rough value of the fundamental frequency is then estimated from the bin location of the fundamental and its overtones.
- High resolution harmonic grain extraction: using the selected value of the fundamental frequency from the previous stage, local interpolation of each partial within the FFT frame is performed in order to determine more accurate values of the frequencies together with their corresponding amplitudes and phases. The resulting harmonic grain is then synthesised and subtracted from the original grain.

B.3 Low-resolution harmonic energy analysis

Within an FFT frame, a frequency domain component contributes to the harmonic energy of the series if it is the maximum component within the corresponding *harmonic window*. This window increases in width linearly with the partial number.

To be more specific, let us consider an FFT frame with a resolution of 20 Hz and a harmonic series corresponding to the fundamental frequency bin 100–120 Hz. The first partial can appear anywhere between 200 and 240 Hz, corresponding to a *harmonic window* twice the width of the resolution value. Similarly, the next partial can appear anywhere between 300 and 360 Hz, i.e. three times the initial window width. The *harmonic window* width (in bins) as a function of the partial number p is given by:

$$\nu(p) = p, \quad p = 1, ..., \rho$$

where p is the partial number; $p = 1$ represents the fundamental and ρ the total number of overtones. The use of such harmonic windows has the other advantage of being able to capture a certain degree of inharmonicity that the signal could exhibit.

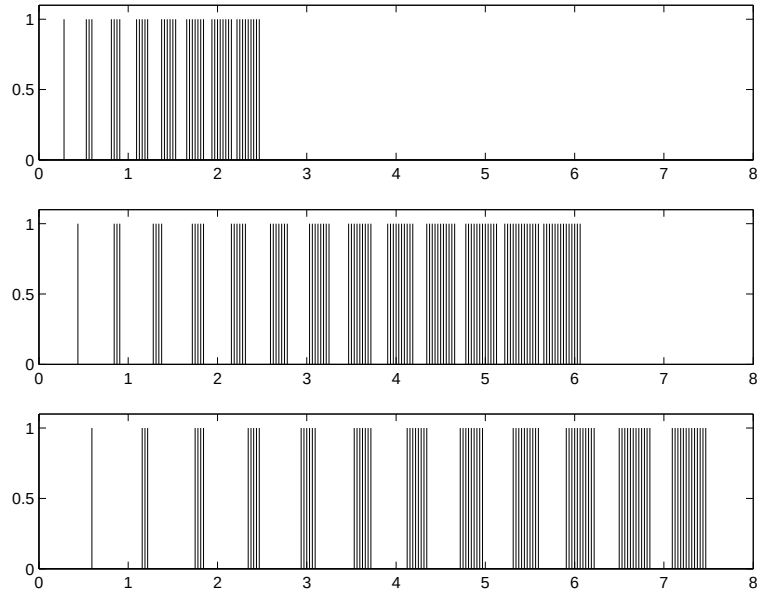


Figure B.3: Three masking functions used for the calculation of the spectral harmonic energy. The fundamental frequencies are (from top to bottom): 281.25 Hz, 437.50 Hz and 593.75 Hz. They correspond to integer multiples of fs/N.

In figure B.3, three masking functions corresponding to three fundamental frequencies used to weight the short-term spectra before the harmonic energy calculation are depicted. The maximum number ρ of retained overtones is respectively 7, 12 and 11 for the three fundamentals frequencies 281.25 Hz, 437.50 Hz and 593.75 Hz as shown in figure B.3, for a 16 kHz sampling rate.

The harmonic energy $\Lambda(k)$ is calculated for each potential fundamental index k as:

$$\Lambda(k) = |X(k)|^2 + \sum_{p=2}^{\rho} \max_{\nu} |X(kp + \nu(p) - 1)|^2$$

where $X(k)$ is the complex FFT of the input frame. The harmonic series with the highest energy is then retained.

The selected series is used to calculate an approximation of the fundamental frequency:

$$\tilde{f_0} = \frac{1}{\rho} \sum_{p=1}^{\rho} \frac{f_{k_0,p}}{p}$$

where $f_{k_0,p}$ is the frequency corresponding to the maximum amplitude over the con-

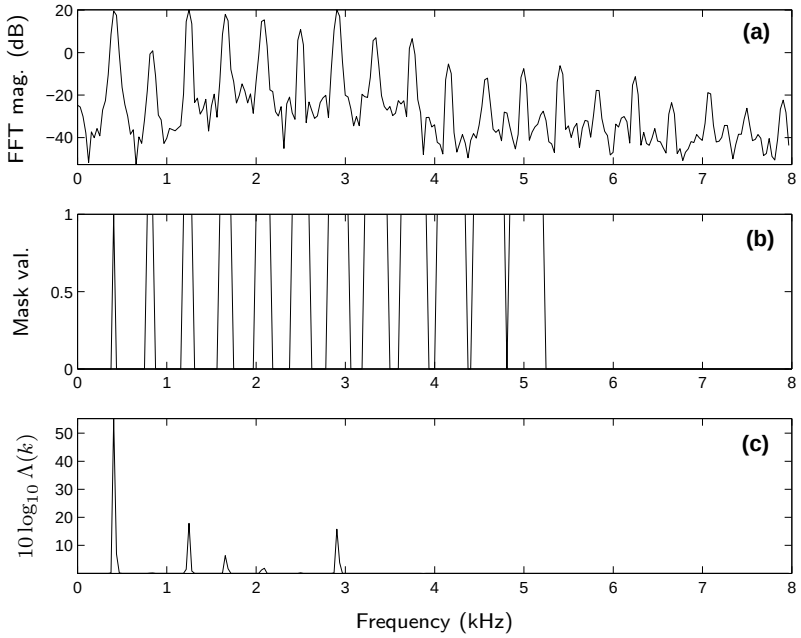


Figure B.4: Harmonic energy calculation. The input frame (N=512) have been extracted from a clarinet solo phrase resampled at $f_s = 16$ kHz. The FFT magnitude is shown in (a). In (c) are plotted the logarithm of the harmonic energy calculated for each possible fundamental frequency. The retained fundamental frequency value is determined from the maximum of the function and its corresponding harmonic comb is shown in plot (b).

sidered window $\nu(p)$. k_0 is the bin index of the retained fundamental:

$$f_{k_0,p} = f_{\max_\nu |X(k_0 p+\nu(p))|}$$

This new measure of the fundamental frequency is used to select the most relevant partial bins $\tilde{f}_p$, $p = 1,...,\rho$, by rounding the expected partial position

$$\tilde{f}_p = p\tilde{f}_0, \quad p = 1, ..., \rho$$

to the nearest bin value. These bins values are used in the high resolution harmonic grain extraction stage.

B.4 High-resolution grain extraction

Prior to the grain extraction, frequencies and corresponding phases of the selected fundamental and partials are interpolated in order to counterbalance the FFT finite resolution and therefore to maximise the energy extracted at each iteration.

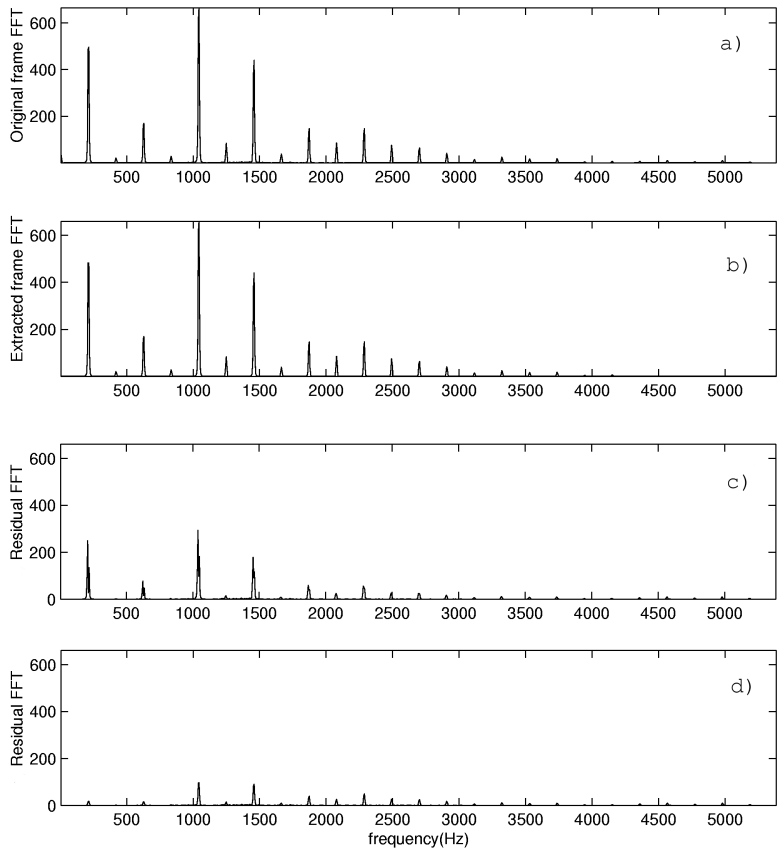


Figure B.5: Illustration of a harmonic grain extraction within a FFT frame. (a) original FFT frame. (b) FFT of the extracted grain after interpolation. (c) FFT of the residual without interpolation. (d) FFT of the residual after interpolation ($\xi = 20$).

The interpolation is done using a complex local interpolation scheme similar to the chirp Fourier Transform. This technique is identical to a zero-padding, but is applied locally to a region of the FFT around the considered peak (typically the width of the harmonic window as defined in section B.3, extended by 2, 3 or 4 bins). It has the advantage of interpolating both phase and amplitude together. Note also that quadratic interpolation techniques such as the one presented in section 3.5.2.2 can also be used.

1. The interpolation function is calculated using the FFT of a zero padded rectangular window (by a factor ξ). These values are calculated once and saved in a table.

2. For each input frame, the frequency bins required for the interpolation (the fundamental and its related partials) are extracted as vectors and upsampled

by the interpolation factor ξ.

3. The vectors are then convolved with the interpolation function values, achieving a local, complex-domain interpolation.

Figures B.5(c) and B.5(d) show the advantages of using the interpolation in terms of the extracted energy per FFT frame.

The synthesis is performed using the IFFT/OLA approach described in the previous section and illustrated in figure 3.15. The synthesised time-domain grain is then subtracted from the original grain and the process is eventually repeated.

B.5 Examples of application

Two examples of applications of our implementation of HMP are presented in this section. The first one is concerned with the determination of the pitch of a monophonic musical phrase while the second one consists of separating a signal mixture of two piano notes.

Tests audio signals (trumpet and jazz guitar pieces) were sampled at 44.1 kHz. The analysis stage was performed on successive frames of 2048 samples weighted by a Hanning window. Using such a window clearly sacrifices temporal localisation. A hop size of 25% (i.e. 512 samples) was therefore used to retain some signal timing information. The parameter ξ was set to 20 and the local interpolation is applied on a 7 bins vector (the selected peak plus 3 bins on each side). A final point regarding robust implementation of this algorithm is to ignore the first three bins corresponding to the frequencies 0–65 Hz, as issues related to the lack of orthogonality become more preponderant in that frequency range. However, this should not induce severe artifacts as it is quite uncommon for musical signals to contain notes at such low frequency. Interpolated amplitudes, frequencies and phases are then used to synthesise the time-domain grain using the method presented in section 3.5.2.3.

Figure B.6 is an example of pitch extraction for a monophonic trumpet signal. The algorithm accurately captures the evolution of the fundamental frequency as a function of time. Figure B.7 is an illustration of the complete application of the harmonic matching pursuits on a monophonic jazz guitar piece after having removed the transients [DDS01]. Figure B.7 (e) represents the time waveform residual after subtraction of the synthesised signal from the original. One can notice that the transients are still slightly present, but nevertheless much more attenuated than if the whole original signal was used during the decomposition.

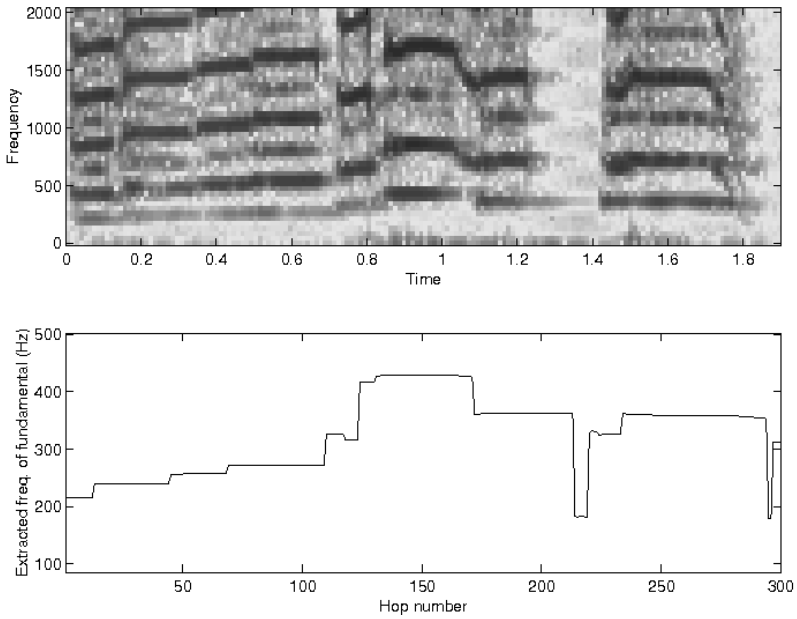


Figure B.6: Extracted pitches (fundamental frequencies) as a function of time for a harmonic monophonic trumpet piece.

Finally, an example of two notes extraction from a mixture is shown in figure B.8. An artificial mixture of two piano notes without overlapping harmonics (B–250 Hz and F–350 Hz) has been synthesised. Two successive iterations of the algorithm are needed to decompose the input frame in two harmonic grains corresponding to the two individual notes. Residual signals (plotted in figures B.8(c) and B.8(f)) mainly contain noise and informal listening tests did not show any perceivable differences between the original and extracted notes.

B.6 Conclusion

In this chapter, we introduced an efficient two-stage implementation of the matching pursuit algorithm based on a harmonic grain extraction within the spectral domain. At each iteration, a complete series of sinusoidal components is extracted, as opposed to standard sinusoidal matching pursuit, where only one peak is picked at a time. Thus, far fewer iterations are required for the decomposition. Not only dramatically reducing the computational requirements, this frequency domain method offers a meaningful approach to the musical sounds modelling problem by improving spectral components grouping.

This implementation performs well for stationary audio signals. The quality of the

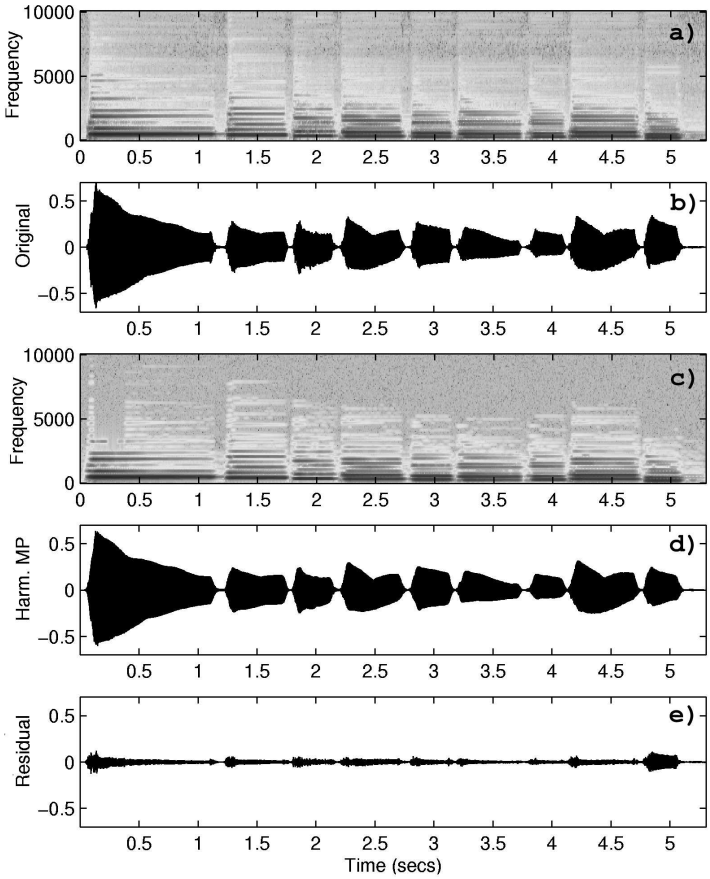


Figure B.7: Complete HMP analysis/synthesis decomposition of a monophonic jazz guitar signal. (a) spectrogram of the original signal. (b) original waveform. (c) spectrogram of the re-synthesised using a single harmonic grain per frame. (d) re-synthesised waveform. (e) time-domain residual.

extraction is good even though we make no assumptions about the type of instrument that is played.

Improvements, however, are needed to make the algorithm more robust and applicable to a wider range of signals. Firstly, problems arise with polyphonic audio mixtures containing overlapping harmonics. This is a common drawback of all the spectral analysis techniques. In such a case, harmonics corresponding to a given pitch may be assigned to another harmonic series, thus introducing some false notes in the considered grain. This can be overcome for example by making assumptions about the harmonic distributions [Kla01]. Secondly, if the signal is not purely steady-state, the residual is shaped into harmonics which may introduce artifacts in the extracted signal. Particular attention should therefore be paid regarding the quality of the transient/sinusoidal signals decomposition.

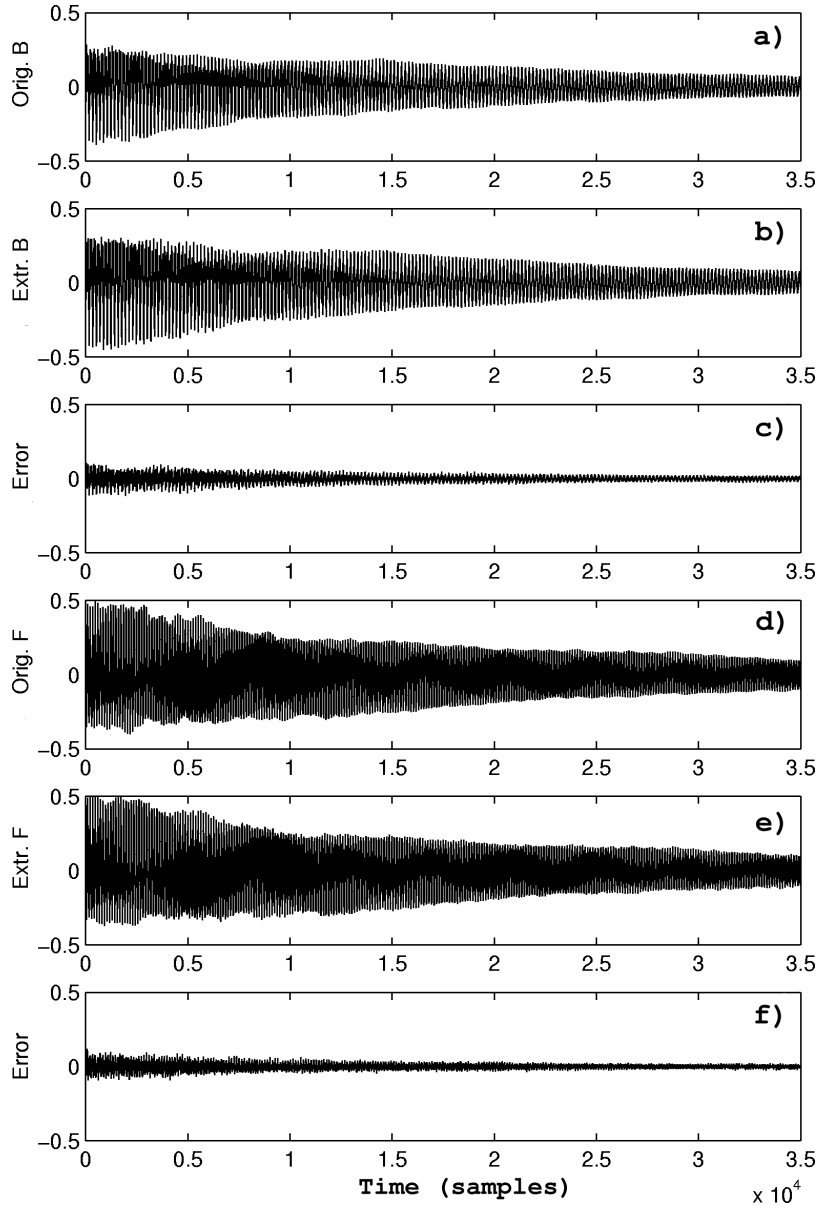


Figure B.8: Example of two piano notes separation. (a) and (d) are the original waveforms (respectively B-250Hz and F-350Hz), (b) and (e) the re-synthesised signals, (c) and (f) the corresponding residual errors.

Bibliography

[ABL02] X. Amatriain, J. Bonada, and A. Loscos. *DAFx - Digital Audio Effects*, chapter Spectral Processing, pages 373–438. Udo Zölzer, 2002.

[ALP01] G. Agostini, M. Longari, and E. Pollastri. Content-based classification of musical instrument timbres. *International Workshop on Content-Based Multimedia Indexing*, 2001.

[BDA+05] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Trans. Speech and Audio Process.*, 13(6):1035–1047, 2005.

[Ben76] A. H. Benade. *Fundamentals of Musical Acoustics*. Oxford University Press, 1976.

[Ber63] K. W. Berger. Some factors in the recognition of timbre. *J. Acoust. Soc. Am.*, 36:1888–1891, 1963.

[BHM01] J. C. Brown, O. Houix, and S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *J. Acoust. Soc. Am.*, 109(3):1064–1072, 2001.

[Bil97] J. A. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models. *Technical Report, University of Berkeley, ICSI-TR-97-021*, 1997.

[Bla65] E. D. Blackham. The physics of the piano. In Scientific American, editor, *The Physics of Music*, chapter 3, pages 24–33. W. H. Freeman and Company, San Fransisco, 1965.

[Blu80] A. Blumenthal. *Wilhelm Wundt and the Making of a Scientific Psychology*. R.W. Rieber, New York: Columbia University, 1980.

[Bre90] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[Bro99] J. C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.*, 105(3):1933–1941, 1999.

[Bur98] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):955–974, 1998.

[Cal00] Calliope. *La parole et son traitement automatique*. Presse Universitaire de France, 2000. Second Edition.

[Cas02] M. A. Casey. Generalized sound classification and similarity in MPEG-7. *Organized Sound*, 6(2), 2002.

[CDS05a] N. Chétry, M. Davies, and M. Sandler. Identification of monophonic instrument recordings using K-means and Support Vector Machines. In *Proc. Digital Music Research Network (DMRN) conference*, 2005.

[CDS05b] N. Chétry, M. Davies, and M. Sandler. Musical instrument identification using LSF and K-means. In *Proc. AES 118th Convention*, 2005.

[Che57] E. C. Cherry. *On human communication: a review, survey and a criticism*. MIT Press, Cambridge, MA, 1957.

[CHM97] G. Charbonneau, C. Hourdin, and T. Moussa. A multidimensionnal scaling analysis of musical instruments time-varying spectra. *Computer Music Journal*, 21:40–55, 1997.

[CL01] C. C. Chang and C. J. Lin. *LibSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[CS06] N. Chétry and M. Sandler. Linear predictive models for musical instrument identification. *accepted for publication in Proc. ICASSP*, 2006.

[DC95] C. J. Darwin and R. P. Carlyon. *Auditory grouping*, chapter Handbook of Perception and Cognition: Hearing. Edited by Brian Moore. New York, NY: Academic Press, 1995.

[dCK02] A. de Cheveigne and H. Kawahara. YIN, a fundamental pitch estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

[DCSD04] C. Duxbury, N. Chétry, M. Sandler, and M. Davies. An efficient two-stage implementation of harmonic matching pursuit. In *Proc. EUSIPCO*, 2004.

[DDS01] C. Duxbury, M. Davies, and M. Sandler. Extraction of transient content in musical audio using multiresolution analysis techniques. In *Proc. DAFx*, 2001.

[Deu82] D. Deutsch. Grouping mechanisms in music. In D. Deutsch, editor, *The Psychology of Music*, pages 99–130. Academic Press, 1982.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39:1–38, 1977.

[Dol86] M. B. Dolson. The phase vocoder: a tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

[dPLY04] H. B. de Paula, M. A. Loureiro, and H. C. Yehia. Representation and classification of the timbre space of a single musical instrument. *Proc. ISCA Workshop on Statistical and Perceptual Audio Process.*, 2004.

[DR02] S. Dubnov and X. Rodet. Investigation of phase coupling phenomena in sustained portion of musical instruments sound. *J. Acoust. Soc. Am.*, 112(6), 2002.

[EB03] J. Eggink and G. J. Brown. A missing feature approach to instrument identification in polyphonic music. In *Proc. ICASSP*, 2003.

[EB04] J. Eggink and G. J. Brown. Instrument recognition in accompanied sonatas and concertos. In *Proc. ICASSP*, 2004.

[EE47] H. V. Eagleson and O. W. Eagleson. Identification of musical instruments when heard directly and over a public-address system. *J. Acoust. Soc. Am.*, 19:338–342, 1947.

[Ell96] D. P. W. Ellis. Prediction-driven computational auditory scene analysis. *Massachussets Institute of Technology*, Ph.D. dissertation, 1996.

[ELR+05] S. Essid, P. Leveau, G. Richard, L. Daudet, and B. David. On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments. In *Proc. AES 118th Convention*, 2005.

[ERDa] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *submitted for publication in IEEE Trans. on Speech and Audio Process.*

[ERDb] S. Essid, G. Richard, and B. David. Musical instrument recognition by pairwise classification strategies. *To appear in IEEE Trans. on Speech and Audio Process.*

[ERD04] S. Essid, G. Richard, and B. David. Musical instrument recognition based on class pairwise feature selection. In *Proc. ISMIR*, 2004.

[Ero01] A. Eronen. Comparison of features for musical instrument recognition. In *Proc. WASPAA*, 2001.

[ESS97] R. Eichert, L. Schmidt, and U. Seifert. Logic, Gestalt theory, and neural computation in research on auditory perceptual organization. In Marc Leman, editor, *Music, gestalt and computing: studies in cognitive and systematic musicology, Lecture Notes in Artificial Intelligence*, pages 70–88. Springer, 1997.

[Eur00] European Telecommunications Standards Institute – ETSI. GSM 06.10 version 7.1.0 - Full rate speech. TS-100 961, 2000.

[FG66] J. L. Flanagan and R. M. Golden. Phase vocoder. *The Bell System Technical Journal*, 45:1493–1509, 1966.

[Fle34] H. Fletcher. Loudness, pitch, and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *J. Acoust. Soc. Am.*, 6:59–69, 1934.

[Fle40] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12:47–65, 1940.

[FM33] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *J. Acoust. Soc. Am.*, 5:82–108, 1933.

[Fuj98] I. Fujinaga. Machine recognition of timbre using steady-state tone of acoustic musical instruments. In *Proc. ICMC*, 1998.

[Fur81] S. Furui. Cepstrum analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Process.*, 29:254–272, 1981.

[GB03] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Trans. Signal Process.*, 51(1):101–111, 2003.

[GG99] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic publishers, 1999.

[GR04] O. Gillet and G. Richard. Automatic transcription of drum loops. In *Proc. ICASSP*, 2004.

[Gre77] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, 61(5):1270–1277, 1977.

[HDG03] P. Herrera, A. Dehamel, and F. Gouyon. Automatic labelling of unpitched percussion sounds. In *Proc. AES 114th Convention*, 2003.

[Hel54] H. L. F. Helmholtz. *On the sensations of tone as a physiological basis for the theory of music*. Dover Publications, Inc., New York, 1954.

[HL01] A. Harma and U. K. Laine. A comparison of warped and conventional linear predictive coding. *IEEE Trans. on Speech and Audio Process.*, 9(5):579–588, 2001.

[HT98] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[ISO] ISO/IEC IS 11172-13 – Information Technology. Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbits/s – Part 3: Audio.

[Ita75] F. Itakura. Line spectrum representation of linear predictive coefficients of speech signals. *J. Acoust. Soc. Am.*, 57:S35, 1975.

[Jen02] K. Jensen. Perceptual and physical aspects of musical sounds. In *Journal of Sangeet Research Academy, India*, 2002.

[Kac66] M. Kac. Can one hear the shape of a drum? *Amer. Math. Monthly*, 73, 1966.

[Ken86] R. A. Kendall. The role of acoustic signal partitions in listener categorization of musical phrases. *Music Perception*, 4:185–214, 1986.

[KGO03] T. Kitahara, M. Goto, and H. G. Okuno. Musical instrument identification based on f0-dependent multivariate normal distribution. In *Proc. ICASSP*, 2003.

[Kla01] A. P. Klapuri. Mutipitch estimation and sound separation by the spectral smoothness principle. In *Proc. ICASSP*, 2001.

[Kla03] A. Klapuri. Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Process.*, 11(6):804–816, 2003.

[KR86] P. Kabal and R. P. Ramachandran. The computation of Line Spectral Frequencies using Chebyshev polynomials. *IEEE Trans. Acoust., Speech, Signal Process.*, 34(6):1419–1426, 1986.

[KS04] A. G. Krishna and T. V. Sreenivas. Music instrument recognition: from isolated notes to solo phrases. In *Proc. ICASSP*, 2004.

[KW02] Y. R. Kim and B. Whitman. Singer identification in popular music recordings using voice coding features. In *Proc. ISMIR*, 2002.

[LAM] The LAME project. `http://lame.sourceforge.net/`.

[Lar99] J. Laroche. Improved phase vocoder time-scale modification of audio. *IEEE Trans. Speech and Audio Process.*, 7:323–332, 1999.

[LBG80] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28:702–710, 1980.

[LR03] A. Livshin and X. Rodet. The importance of cross-database evaluation in sound classification. In *Proc. ISMIR*, 2003.

[Mar99] K. D. Martin. Sound-source recognition: a theory and computational model. *Massachussets Institute of Technology*, Ph.D. dissertation, 1999.

[McL92] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley and Sons Inc., 1992.

[Mey56] L. B. Meyer. *Emotion and Meaning in Music*. University of Chicago Press, Chicago, 1956.

[MG76] J. D. Markel and A. H. Gray. *Linear Prediction of speech*. Springer–Verlag, 1976.

[MK98] K. D. Martin and Y. E. Kim. Musical instrument identification: a pattern-recognition approach. In *Proc. 136th meeting of the Acoustical Society of America*, 1998.

[MM99] J. Marques and P. J. Moreno. A study of musical instrument classification using gaussian mixtures models and support vector machines. *Compaq Cambridge Research Laboratory*, Tech. Report 99–4, 1999.

[Moo97] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 1997.

[MQ86] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoustic, Speech and Signal Process.*, 34(4):744–754, 1986.

[Mus] Music-DSP. Source Code Archive. `http://www.musicdsp.org/`.

[MZ93] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41:3397–3415, 1993.

[OPGB05] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *Proc. WASPAA*, 2005.

[OW87] F. Opolko and J. Wapnick. Mcgill university master samples (CD). 1987.

[PA93] K. K. Paliwal and B. S. Atal. Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Trans. Speech, Audio Process.*, 1:3–14, 1993.

[Pal88] K. K. Paliwal. A perception-based LSP distance measure for speech recognition. *J. Acoust. Soc. Am.*, 84:S14–S15, 1988.

[Pan95] D. Pan. A tutorial on MPEG/Audio compression. *IEEE multimedia*, 2(2):60–74, 1995.

[Pee03] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Proc. AES 115th Convention*, 2003.

[Plo70] R. Plomb. Timbre as a multidimensional attribute of complex tones. *Frequency analysis and periodicity detection in hearing*, pages 397–414, 1970.

[Por96] Boaz Porat. *A Course in Digital Signal Processing*. John Wiley and Sons, 1996. Second Edition.

[PR03] G. Peeters and X. Rodet. Hierarchical Gaussian tree with inertia ratio maximization for the classification of large musical instrument databases. In *Proc. DAFx*, 2003.

[Reu97] C. Reuter. Karl Erich Schumann's principles of timbre as a helpful tool in stream segregation research. In Marc Leman, editor, *Music, gestalt and computing: studies in cognitive and systematic musicology, Lecture Notes in Artificial Intelligence*, pages 362–372. Springer, 1997.

[Rit67] R. J. Ritsma. Frequencies dominant in the perception of the pitch of complex sounds. *J. Acoust. Soc. Am.*, 42:191–198, 1967.

[Ros89] T. D. Rossing. *The science of sound*. Addison-Wesley Publishing Company, Inc., 2nd edition, 1989.

[RR95] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech, Audio Process.*, 3:72–83, 1995.

[RS86] A. E. Rosenberg and F. K. Soong. Evaluation of a vector-quantization talker recognition system in text independent and text dependent modes. In *Proc. ICASSP*, 1986.

[RW82] J. C. Risset and D. L. Wessel. Exploration of timbre by analysis and synthesis. In D. Deutsch, editor, *The Psychology of Music*, pages 26–58. Academic Press, 1982.

[RWC] RWC Music Database. Music genre database and musical instrument sound database. `http://staff.aist.go.jp/m.goto/RWC-MDB`.

[SC64] E. L. Saldanha and J. F. Corso. Timbre cues and the identification of musical instruments. *J. Acoust. Soc. Am.*, 36:2021–2026, 1964.

[SC67] W. Strong and M. Clark. Synthesis of wood-instrument tones. *J. Acoust. Soc. Am.*, 41:39–52, 1967.

[Sch29] K. E. Schumann. *Physik der Klangfarben*. Berlin: Habilschr, 1929.

[Sch40] J. F. Schouten. The residue, a new component in subjective sound analysis. *Proc. Kon. Ned. Akad. Wetensch*, 43:356–365, 1940.

[Ser97] Xavier Serra. *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pages 91–123. Swets and Zeitlinger Publishers, 1997.

[SIM] SIMAC. Semantic interaction with music content. `http://www.semanticaudio.org/`.

[Sla98] M. Slaney. Auditory toolbox. *Apple Technical Report*, (1998-010), 1998.

[Smi88] B. Smith. *Foundations of Gestalt Theory*. Barry Smith, 1988.

[SRRJ87] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang. A vector quantization approach to speaker recognition. *AT&T Technical Journal*, 66:14–26, 1987.

[SS87] J. O. Smith and X. Serra. PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. In *Proc. ICMC*, 1987.

[SSF02] A. Srinivasan, D. Sullivan, and I. Fujinaga. Recognition of isolated instruments tones by conservatory students. In *Proc. ICMPC*, 2002.

[SV40] S. S. Stevens and J. Volkman. The relation of pitch to frequency. *American Journal of Psychology*, 53:329–353, 1940.

[The] The University of Iowa. Musical instrument database. `http://theremin.music.uiowa.edu`.

[Tre82] T. E. Tremain. The Government standard linear predictive coding algorithm: LPC-10. *Speech Technology*, pages 40–49, 1982.

[TSS82] E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J. Acoust. Soc. Am.*, 71:679–688, 1982.

[Vap95] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[vB74] G. von Bismarck. Timbre and steady sounds: a factorial investigation of its verbal attributes. *Acustica*, 30:146–158, 1974.

[VOI] VOICEBOX. Speech Processing Toolbox for MATLAB. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[VR04] E. Vincent and X. Rodet. Instrument identification in solo and ensemble music using independent subspace analysis. In *Proc. ISMIR*, 2004.

[War] WarpTB. Matlab Toolbox for Warped DSP. http://www.acoustics.hut.fi/software/warp/.

[Wes78] D. L. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2), 1978.

[YR04] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Process.*, 52(7):1830–1847, 2004.

[ZF99] E. Zwicker and H. Fastl. *Psycho-acoustics, Facts and Models*. Springer-Verlag, 2nd updated edition, 1999.