The temporal character of timbre

Diploma Thesis

by

Miha Ciglar

miha.ciglar@irzu.org

Supervised by

Dr. Alois Sontacchi (IEM)

IEM - Institute of Electronic Music and Acoustics

University of Music and Performing Arts Graz

April 2009

Contents

.

Kurzfassung

Das Thema der Diplomarbeit bezieht sich auf die Klangfarbenanalyse. Im Speziellen wird auf eine ganz bestimmte Klangkategorie eingegangen, nämlich auf kurze, polyphone Audiosamples bzw. aus einem größerem Kontext genommene Signalbursts, die zwischen zwei aufeinander folgenden Onsets zu finden sind und meistens in der Größenordnung eines einzigen Taktschlages vorliegen.

Die Arbeit beschäftigt sich auch mit Fragen der generellen Definition von Klangfarbe und vertritt die These, dass die allgemeine Auffassung von Klangfarbe nicht durch eine statische Spektralkomposition sondern erst durch eine zeitliche Änderung dieser d.h. mittels einer zeitlichen Abfolge von Veränderungen in der Spektralkomposition konstituiert wird.

Zuerst wird ein Überblick verschiedener, bestehender Analysemethoden gegeben. Diese unterscheiden sich durch das zu modellierende Klangmaterial, welches von einfachen, einstimmigen harmonischen Klängen, über mehrstimmige Klangfarben-Mixturen bis hin zu geräuschartigen Klangtexturen reicht. Jede dieser Klangkategorien verlangt also nach einer speziellen Methode um die wahrnehmungsrelevanten Features der zu beschreibenden Klänge, in möglichst kompakter Form zusammenfassen zu können.

Weiters wird untersucht, inwiefern bekannte Modellierungsansätze aus der Teiltonverlaufanalyse auf die zeitliche Organisation der durch Analysemethoden polyphoner Klangfarbenmixturen generierten Features angewendet werden können. Konkret wird versucht die einzelnen zeitlichen Trajektorien der Mel Frequency Cepstral Coefficients (MFCC) innerhalb eines isolierten, mehrstimmigen Signalbursts zu beschreiben, mit der Absicht der Klangfarbenklassifikation bzw. der späteren Identifikation.

Abstract

This work is about timbre analysis and is aimed at the identification of timbral structures in music progression, which are characteristic for a particular class of sounds, namely, short polyphonic signal bursts. It gives an overview of timbre definitions and different timbre analysis approaches developed over the last 40 years. The general structural diversity of a musical audio signal is spanning a range from simple, monophonic harmonic sounds, over polyphonic timbre mixtures, to noise-like textures, all of which demand a particular analysis method in order to sufficiently describe its perceptually relevant features. The purpose of those features on one hand is automatic discrimination from different sounds found in the same category, that is, classification, as well as sound similarity matching in different fields of music information retrieval (e.g. search for similar sounding music, acoustic fingerprinting, concatenative sound synthesis, score following, etc.). The focus of this work however is on the temporal character of timbre. It points out the importance of a particular sequence of timbral features as being a crucial information carrier for the purpose of sound classification and identification. The temporal sequence of change, taking place in the timbral structure – or later, in its extracted features – is an important recognition cue when identifying sounds. The location, strength and the inter-relation of individual harmonic partials, especially their evolution in the attack and the release segments, plays an important role for the discrimination of different sounds or sound sources. Comparing the degree of fine-structure exhibited by monophonic-signal analysis approaches (e.g. sinusoidal plus residual models) and the tools for polyphonic timbre analysis (e.g. Mel Frequency Cepstral Coefficient (MFCC) representations), it is evident that the later are generally aimed at identifying larger song segments like choruses, verses, etc. i.e. a general timbral character, rather than a fine-structure within a shorter homogenous segment e.g. one beat or note. This work on the other hand investigates the possibilities of applying formal techniques like partial envelope modeling from the field of monophonic timbre analysis for re-organizing the content gained by methods of polyphonic timbre analysis like MFCC representations, etc.

Thesis Outline

The thesis is divided into three major parts, whereof the first is exploring methods of monophonic timbre analysis techniques. Starting with a short overview of different analysis approaches, the focus continues to move towards a *sinusoidal model for monophonic harmonic sounds* and its precise derivation. This model is applied to isolated harmonic partials exclusively and is further simplified following the work of Serra [Serra *et al.* 1997] and Jensen [Jensen 1999] by introducing the *High level Attributes (HLA), Minimum Description Attributes (MDA)* and *Instrumental Definition Attributes (IDA)*. In the *HLA* section, different *harmonic-spectral descriptors* are derived and further extended by the introduction of pure *spectral descriptors* applied to the complete spectrum, where the harmonicity of the signal is not given and where it is not possible to isolate individual partials.

The second major part is exploring methods for polyphonic timbre analysis. It starts with the introduction and deduction of *MFCC features*. Further, different concepts of their organization, ranging from static e.g. *Gaussian Mixture Models* (*GMM)* to dynamic, like *first and second order derivatives*, *self similarity representations* and *Hidden Markov Models (HMM),* are introduced. Next, some of the polyphonic timbre analysis methods specified by the *MPEG-7* standard are examined, namely, the *AudioSpectrumBasis* and *AudioSpectrumProjection* concepts. The section concludes with the *comparison* of those with the previously introduced MFCC features.

In the third section, a new method for the analysis of short samples, i.e. isolated beats or notes with polyphonic timbral character, is presented. An alternative organization of MFCC features is proposed, that is based on the formal methods presented in the first section, in particular, the attack and release times and their MDA approximations, which were there applied to isolated harmonic partials.

1. Introduction

This work points out the different interpretations and purposes of identifying timbral structures, found in a variety of audio signal categories and contexts. More precisely, the focus is on exhibiting different approaches to analyzing the timbre of an unprocessed audio signal, which are conditioned by its harmonic complexity as well as by the length of the sound sequence that is to be described. Furthermore, this work aims to emphasize the importance of the temporal character of timbre as it appears to be a crucial information carrier for the classification and recognition of sound. A particular goal however, is to introduce a concept of temporal organization of MFCC features in polyphonic timbre mixtures, namely to propose a model, describing the evolution of each coefficient during short sound bursts e.g. beats or single notes (chords), isolated from a larger context e.g. song or phrase.

A first subdivision of timbre analysis concepts may be performed with respect to the degree of complexity they are trying to deal with in the sound's vertical dimension or its spectral domain. On one hand, a lot of work has been focusing on monophonic sound samples, where the sound may be abstracted and modeled by the structure and the temporal progression / development of its individual harmonic partials. This kind of research is aimed at playing stile recognition, and further, at instrument identification [Herrera-Boyer *et al.* 2003], i.e. identifying if a note is being played on a saxophone or on a violin. Other work is focusing on analyzing real world timbre mixtures or polyphonic timbres, such as the ones found in music performed by music ensembles, where several instruments play at the same time [Aucouturier 2006].

Time is another important timbral dimension, and for the next - parallel to the prior - subdivision of work in this field, different analysis approaches, which are conditioned by the extent of the observed sound in time, shall be considered.

The temporal evolution and the relational dynamics of individual frequency components seem to be as important as a general, "global" constellation of those for a classification of particular sounds. The author's speculative assumption is that it would

only make sense to talk about “timbre” when the temporal extent of an analyzed sound exceeds a minimum duration (a condition of perceptual relevance) so that a structural alteration of its spectral components becomes audible. Thus, if only a snapshot i.e. an instant or perhaps even a longer, but completely rigid sound or signal is to be analyzed, it might be more appropriate to describe it as a mixture of individual frequency components with individually weighted amplitudes, rather than “timbre”.

The timbre analysis methods are also varying with respect to the length of the time frame to be analyzed. From this point of view, a research field emerges that is focused on the analyses of short sound samples on one hand. Generally these are individual notes played by a monophonic instrument, again, with the goal of instrument identification and further of expression or playing style classification (legato, staccato, etc.). On the other hand - as the length of the observed audio segment increases - the focus and purpose of classification is shifting towards compositional analysis, that is, identification of chord sequences and further, larger song segments like: choruses, verses, etc. Yet another step further, the term “Global timbre” - also referred to as audio fingerprint - comes up, which is an attribute aimed at describing a timbral quality that applies to the whole sequence of music or perhaps to a whole song, as opposed to a particular temporal segment or instrument. One of the main purposes of this analysis approach is genre classification [Haitsma *et al.* 2002] and it will not be a subject of research in this thesis.

The research contributions from the field of audio analysis and music information retrieval are usually a mix of the above described classes, however, we can recognize a general tendency of analyzing short segments like single notes in the context of monophonic music, whereas longer analysis frames, spanning complete songs or musical sequences on the other hand, are rather being considered when describing polyphonic timbre mixtures, where several instruments are playing together. Timbre descriptions using MFCC’s for example, are usually deployed to roughly describe homogenous sounding sequences, indicating changes in instrumentation, or general perceptual differences amongst song segments e.g. intro, chorus, verse, etc.

Following a description of existing timbre analysis approaches applied to sound in different contexts, a new method for the classification of short signal bursts like single beats and notes with a polyphonic timbral character – which are usually not encountered

self standing but have to be isolated from a larger context of polyphonic music – is proposed. This is achieved by interbreeding the structural concept of attack/release envelope times of isolated harmonic partials with the content generated by the MFCC analysis of polyphonic timbre mixtures. The idea is motivated by the concept of information reduction observed in the case of sinusoidal modeling in monophonic analysis approaches. Given the monophonic, together with the harmonic character of the sound to be analyzed, a sufficient description of the sounds perceptually relevant features can be achieved by reducing its spectral information to the location, strength and temporal evolution of individual partials, while disregarding the remaining spectral information. The idea behind polyphonic sound description by the lower end of the MFCC's exhibits a similar goal of dimension reduction, for it describes merely the general shape of the spectral envelope. Perhaps a similar effect of sound discrimination and identification could be achieved by describing a short polyphonic segment / signal burst, by the temporal evolution of its individual cepstral coefficients or perhaps already by their individual attack and release times. It is important to state however, that the two concepts of information reduction – partial and MFCC models – do not describe the same audible feature.

1.1 The definition of timbre

Indicating an unresolved terminological concern, here, a list of different historic timbre definitions found in literature. From today's point of view, some may be problematic and even false, however, the following statements will indicate how the timbral features have gradually been discovered throughout the years and how much confusion is still associated with this term.

Helmholtz's definition – a translation from the original text 1877 :
"...the amplitude of the vibration determines the force or loudness, and the period of vibration the pitch. Quality of tone can therefore depend upon neither of these. The only possible hypothesis, therefore; is that the quality of tone should depend upon the manner in which the motion is performed within the period of each single vibration." ..... "Certain characteristic peculiarities in the tones of several instruments depend on the mode in which

they begin and end. When we speak in what follows of musical quality of tone, we shall disregard these peculiarities of beginning and ending, and confine our attention to the peculiarities of the musical tone which continues uniformly." ..... "The quality of the musical portion of a compound tone depends solely on the number and relative strength of its partials simple tones, and in no respect on their differences of phase." [Helmholtz 1954]

"...timbre depends principally upon the overtone structure; but large changes in the intensity and the frequency also produce changes in the timbre." [Fletcher 1934.]

"… it can hardly be possible to say more about timbre than that it is a 'multidimensional' dimension." [Licklider 1951]

American Standards Association (ASA 1960)
"Timbre is that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar." [ASA 1960]

"In general, we may say that, aside from accessory noises and inharmonic elements, the timbre of a tone depends upon (1) the number of harmonic partials present, (2) the relative location or locations of these partials in the range from the lowest to the highest, and (3) the relative strength or dominance of each partial. …depends upon its harmonic structure as modified by absolute pitch and total intensity… we must also take phase relations into account." [Seashore 1967]

"Timbre depends primarily upon 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." [ANSI 1960,1970]

"In most textbooks timbre is defined as the overtone structure or the envelope of the spectrum of the physical sound. This definition is hopelessly insufficient, as I hope to prove by demonstrating that timbre can be expressed in terms of at least five major parameters. 1.The range between tonal and noise-like character, 2.The spectral envelope, 3.The time

envelope in terms of rise, duration and decay, 4.The change both of spectral envelope (formant glide) or fundamental frequency (micro-intonation), 5.The prefix, an onset of a sound quite dissimilar to the ensuing lasting vibration." [Schouten 1968]

"Timbre means tone quality-coarse or smooth, ringing or more subtly penetrating, "scarlet" like that of a trumpet, "rich brown" like that of a cello, or "silver" like that of the flute. These color analogies come naturally to every mind … The one and only factor in sound production which conditions timbre is the presence or absence, or relative strength or weakness, of overtones." [Scholes 1970]

" …timbre depends upon several parameters of the sound including the spectral envelope and its change in time, periodic fluctuations of the amplitude or the fundamental frequency, and whether the sound is a tone or noise."
"Clearly...timbre is determined by the absolute frequency position of the spectral envelope rather than by the position of the spectral envelope relative to the fundamental… Von Bismarck found that sharpness as the major attribute of timbre is primarily related to the position of the loudness centre on an absolute frequency scale rather than to a particular shape of the spectral envelope …Low frequency tones do indeed sound dull and high-frequency tones sharp..."
"… the spacing of the harmonics, determined by the fundamental frequency, is responsible for the timbre dissimilarity of sounds with different pitch but similar spectral envelopes." [Plomp 1976]

"Timbre is, after pitch and loudness, the third attribute of the subjective experience of musical tones… Especially important is the relative amplitude of the harmonics. …temporal characteristics of the tones may have a profound influence on timbre as well … Both onset effects (rise time, presence of noise or inharmonic partials during onset, unequal rise of partials, characteristic shape of rise curve, etc.) and steady state effects (vibrato, amplitude modulation, gradual swelling, pitch instability; etc) are important factors in the recognition and, therefore, in the timbre of tones." [Rasch et al. 1982]

Bregman comments on the ASA definition:
“This is, of course; no definition at all … it implies that there are some sounds for which we cannot decide whether they possess the quality of timbre or not. In order for the definition to apply; two sounds need to be able to be presented at the same pitch, but there are some sounds … that have no pitch at all … Either we must assert that only sounds with pitch can have timbre, meaning that we cannot discuss the timbre of a tambourine or of the musical sounds of many African cultures, or there is something terribly wrong with the definition.”....“Until such time as the dimensions of timbre are clarified perhaps it is better to drop the term timbre.” [Bregman 1990]

“Timbre is generally assumed to be multidimensional. It is the perceived quality of a sound, where some of the dimensions of the timbre, such as pitch, loudness and duration, are well understood, and others, including the spectral envelope, time envelope, etc., are still under debate.” [Jensen 1999]

“The word timbre is empty of scientific meaning and should be expunged from the vocabulary of hearing science” [Martin 1999]

As we can see from the above statements, the basic definition of timbre is still a rather problematic issue and may cause more confusion than clarity. Although timbre is a prominent word in a most common musical vocabulary as well as it is a crucial component in the terminology of hearing science, it seams, that a consensus on a precise scientific definition is still due to be reached. Although the terminological debate is beyond the scope of this work, an indirect contribution to it can still be recognized. Since this work inevitably operates with this particular terminology - in order to describe its primary subject discussing concrete technological content - the questions of timbre definition can not be completely avoided. The author’s mere opinions on the definition of timbre can thus be observed indirectly and by bringing this up, it should only be made clear that he is conscious of the loose definition of “timbre”, which also the reader should be made aware of.

1.2 An overview of sound analysis/synthesis approaches

The historic development of sound analysis methods went hand in hand with the sound synthesis methods/models. In the mid-1960s a number of approaches were introduced, dealing with analysis and synthesis of music and speech using computers. At Bell Labs, Jean-Claude Risset and Max Mathews studied the trumpet tone quality in order to determine the important physical correlates of trumpet timbre [Risset 1965]. They performed pitch-synchronous harmonic analysis of trumpet tones. This analysis technique assumed that the sound was quasi-harmonic and produced output data formatted in such a way that the trumpet tones could be accurately resynthesized using an additive synthesis technique. The output of their analysis stage was represented as a time series of amplitude and frequency values of a number of sinusoidal components. These values were then approximated with linear segments and formatted in such a way that the sequence of amplitude and time break points controlled linear ramp generators for amplitudes of the sinusoidal components. By systematically altering the tones' parameters, it was found that a few physical features were highly important for the perception and recognition of brasslike timbre: the attack time (which is shorter for the low-frequency harmonics than for the high-frequency ones), the fluctuation of the pitch frequency (which is of small amplitude, fast, and quasirandom), the harmonic content.

Analog vocoders were widely used for speech modeling in the 1960's, but the development of the digital phase vocoder [Portnoff 1976] was the key to high quality analysis/resynthesis. This development depended on the availability of faster digital computers and the rediscovery in 1965 of the FFT [Cooley *et al.* 1965]. Moorer was the first to adapt these ideas for use with music [Moorer 1987].

Linear predictive coding of speech was also developed at the end of the 1960's [Makhoul 1975] and later applied to the sung voice and other musical sounds. Sinusoidal representations of speech have widely influenced both the speech and computer music research communities.

Researchers have started to explore alternate time-frequency representations and wavelets [Gersem *et al.* 1979] for audio and music applications. In the next chapters a few

prominent tools for sound analysis/synthesis are presented as representatives of different concepts / approaches they are based on.

1.2.1 Sinusoidal Models

The sinusoidal model tries to approximate a sound merely by a sum of sinusoids.

1.2.1.1 SNDAN - programs for sound analysis, resynthesis, display and transformation

The SNDAN analysis/synthesis package [Beauchamp *et al.* 1997] from James Beauchamp's group at the University of Illinois Urbana/Champaign (UIUC) can perform either a pitch-synchronous short-time Fourier (Phase Vocoder) analysis or analysis based on the McAulay-Quartieri algorithm [McAulay *et al.* 1984]

SNDAN's phase-vocoder analysis produces a fixed number of harmonic partials of a fixed fundamental frequency. The fundamental is given as a parameter to the analysis; the frequencies of the analysis bins are integer multiples of this given fundamental. The system outputs are the initial phases (in radians) for all partials, followed by a series of frames containing the frequency deviation and amplitude for each partial.

SNDAN's McAulay-Quartieri-style analysis produces a time-varying number of sinusoidal tracks with frequencies of any relationship. The data consists of a series of frames, each containing a collection of partials with amplitude, frequency, and phase, and a "link" field to associate partials with each other across frames.

1.2.1.2 Lemur

Another group at UIUC under Lippold Haken has released a sinusoidal analysis/synthesis package called Lemur [Fitz *et al.* 1995]. Lemur is based on Maher's and Beauchamp's extension of the McAulay-Quatieri algorithm. Lemur analysis consists of a series of short-time Fourier spectra from which significant frequency components are selected. Similar components in successive spectra are linked to form time-varying partials, called tracks.

The number of significant frequency components and thus, the number of tracks may vary over the duration of a sound. Synthesis is performed by a bank of oscillators.

1.2.2 Source – Filter Models

In the source-filter analysis/synthesis methods, "source" refers to a vibrating object, such as a guitar string and "filter" represents the resonant structure of the rest of the instrument which colors the produced sound. A source-filter analysis is estimating the global spectral shape or the spectral envelope of a sound representing the "filter" part of the model. There are a number of possible techniques of estimating the spectral envelope:

- **The channel vocoder**: estimates the amplitude of the signal inside a few frequency bands.

- **Linear prediction coding (LPC)**: estimates the parameters of a filter that matches the spectrum of the sound.

- **Cepstrum analysis**: performs a Discrete Cosinus Transformation (DCT) on the logarithm of the spectrum (see equation 3.2), which yields an additive representation of the source and filter components.

1.2.3 Resonance Models

The work with resonance models grew mainly from singing voice research [Rodet *et all.* 1989], based initially on tracked formant analyses. A (frequency, amplitude, formant bandwidth) triplet was used to represent each formant. At IRCAM and CNMAT a multitude of variations on these triplets has grown from this work.

1.2.3.1 CHANT

The "CHANT" system for example uses five time-domain formant wave functions (FOFs) [Rodet 1984.], with a particular resonance characteristic used to model individual formants. Those functions would generate an approximation of the resonance spectrum of the first five formants of a female singer. Later work extended the use of FOFs to consonants and

unvoiced phonemes [Richard *et al.* 1992]. The CHANT system however was primarily designed to be a synthesis tool for composers.

1.2.4 Sinusoidal + Noise Models

1.2.4.1 SMS

The Spectral Modeling Synthesis (SMS) system, originally developed from Xavier Serra's dissertation work at Stanford [Serra 1989], later became a project of his group in Barcelona. The particular approach of SMS is based on modeling sounds as stable sinusoids (partials) plus noise (residual component), therefore analyzing sounds with this model and generating new sounds from the analyzed data. The analysis procedure detects partials by studying the time-varying spectral characteristics of a sound and represents them with time-varying sinusoids. These partials are then subtracted from the original sound and the remaining "residual" is represented as a time-varying filtered white noise component. The synthesis procedure is a combination of additive synthesis for the sinusoidal part, and subtractive synthesis for the noise part.

1.2.5 Sinusoidal + Noise + Transient Models

The fundamental assumption behind the sinusoids + noise model is that sound signals are composed of slowly-varying sinusoids and quasi-stationary broad-band noises. This view is quite schematic, as it neglects the most interesting part of sound events: transients. Again, the deficiency of an analysis approach is most clearly visible at the synthesis stage. The Spectral Modeling Synthesis presented above gives good results when applied to audio signals only composed of sinusoids and noise. Once transients occur in an audio signal, they will eventually appear in the noise part of the signal model. This will raise the spectral envelope of the noise during a residual approximation, yielding a synthesized signal with artifacts. Also, for sound modification purposes, better results can be achieved if first, the signal is taken apart and the transients are treated separately. The explicit handling of transients provides a more robust signal model and is essential for synthesizing realistic

attacks of many instruments. For example, when time stretching a signal, it is desirable for transients to move to their proper onset locations but remain localized, while the durations of harmonic partials – represented by sinusoids – as well as the noise parts of a signal stretch. For these reasons, a new sines + noise + transients (SNT) framework for sound analysis was established [Verma et al. 1997]. Until today, it remains a state of the art model for sound analysis and synthesis [Nsabimana *et al.* 2007]

2. Monophonic Timbre

The next chapters will focus on the derivation of features describing the perceptually relevant characteristics of monophonic sounds / timbre, mostly following the work of Serra [Serra *et al.* 1997] and Jensen [Jensen 1999]. The concepts that will be introduced are based on a sinusoidal model, which is certainly not a state of the art approach, compared to Sine+Noise+Residual model, for example. However, it is not the aim of this work to explore the details of a most comprehensive abstraction of a sound; instead, the following section should rather be understood as one component contributing to the punch line of this thesis, by its formal characteristics. Thus, the main purpose of theses chapters lies in examining the methods of timbre-feature organisation.

2.1 Deriving a timbral model for musical instruments using harmonic descriptors

Starting by deriving the Short-Time Fourier Transform (STFT), a sinusoidal model for monophonic sounds is introduced. Next, further concepts of describing the sine-model based timbral features are described, concluding with the following methods of their organisation, interpretation and further compression:

· *High Level Attributes (HLA)*:
Higher level information such as: pitch, spectral shape, vibrato, or attack/release characteristics, can be extracted from the sinusoidal or a sine+noise representation. [Serra *et al.* 1997]

· *Minimum Description Attributes (MDA)*:
MDA are a further information reduction. It extracts the smallest number of parameters necessary to define a sound from the HLA model by describing its parameters with the fundamental value and the evolution over the partial index. [Jensen 1999]

· *Instrument Definition Attributes (IDA)*:

The Instrument Definition Attribute model additionally models the evolution of all parameters that are characteristic for a particular instrument and are captured in an extended range of playing styles, tempi, velocities and tonal registers. The IDA model is therefore a collection of many MDA sets. [Jensen 1999]

The following chapters contain analysis methods for single note samples of pitched monophonic sounds with a quasi-harmonic character. The term quasi-harmonic denotes instruments whose partial frequencies are close to harmonic, excluding drums, cymbals, bells and other instruments with an inharmonic constellation of partials, or noise-like spectra.

The harmonic sounds can be decomposed into additive sinusoidal components called partials. Those partials have time-varying amplitude and frequency. In general, the sinusoidals would correspond to the fundamental frequency and the harmonic overtones of the sound being analyzed. Then the frequencies of the partials are ideally multiples of the fundamental frequency. The frequency of the harmonic partials is equidistant in the frequency domain.

Conceptually, sinusoidal models are rooted in basic Fourier theory, which states that any periodic sound $s(t)$ can be expressed mathematically as a sum of sinusoids:

$$s(t) = \sum_{k=1}^{K} A_k(t)\cos(\omega_k t + \phi_k) \qquad \textbf{(1.1)}$$

where t denotes time; $\omega_k = 2\pi k/T$ the kth harmonic radian frequency, where T is the sinusoidal period in seconds; $A_k(t)$, and φ_k are the amplitude and phase of the kth harmonic sinusoidal component, and K is the number of the highest audible harmonic.

For the sake of completeness it is important to mention that in real world, the sounds would generally not exhibit this kind of symmetry, especially in the higher spectral regions. It has been suggested that no partials higher than the 5th to 7th, regardless of the fundamental

frequency, are resolved individually. Studies have shown that the upper harmonics, rather than being perceived independently are heard as a group [Howard *et al.* 2001]. Sinusoidal plus noise signal model are a good way to simplify the representation of sounds.

$$s(t) = \sum_{k=1}^{K} A_k(t)\cos(\omega_k t) + r(t) \qquad \textbf{(2.2)}$$

where $r(t)$ is a noise residual, which is represented with a stochastic model. However, let us continue with the assumption of "harmonicity" and a sinusoidal model.

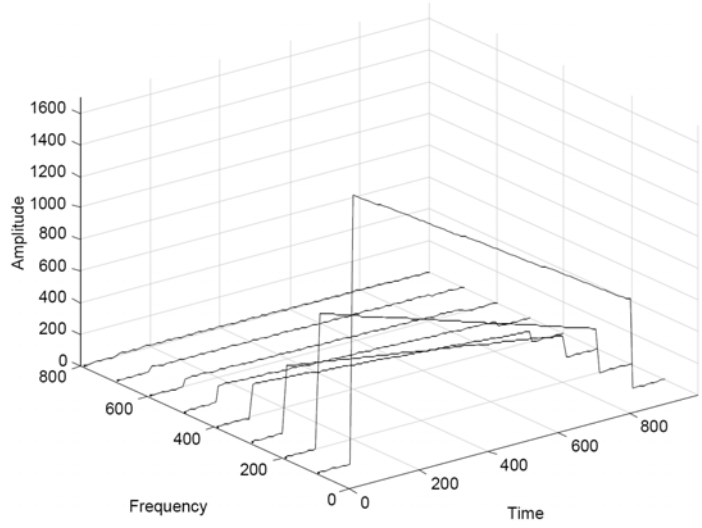


Figure 1. Plot of a harmonic signal with a fundamental frequency of 100 Hz, showing the evolution of the amplitude of each partial.[1] [Jensen 1999]

[1] time and frequency units used in all plots are milliseconds and Hertz respectively, unless otherwise noted.

2.1 FFT based analysis

Most of the research in the frequency domain analysis has been based around the Fourier transform. Specifically, the Discrete Fourier Transform (DFT) version and its efficient Fast Fourier Transform (FFT) version comprise the backbone of many studies in the frequency domain. The FFT transforms a time domain signal into the frequency domain. The input to the FFT is a frame of N time domain samples, where N is a power of two. The output of the FFT is used to compute the power density spectrum of the window represented as N/2 "frequency bins". The bins are evenly spaced and represent frequencies between zero and half the sample rate. A spectrogram can be generated by computing a series of FFTs. The output of each FFT represents the frequency amplitude levels over a narrow slice of time. The series of slices can be used to show how the frequency components of a signal change through time. It is also common practice for these FFT windows to be overlapping and averaged together to smooth out edge transitions. There is a trade-off between the resolution of frequency and time. Larger FFT windows can resolve more frequencies (more frequency bins), but are wider in time (more samples), and thus have lower resolution in time. Shorter FFT windows are shorter in time (fewer samples) and can therefore observe faster changes in time, but can't resolve as many frequencies.

In general, the time resolution should be at least as good as the fastest transient time under analysis, in the order of a few ms. The FFT-based analysis can further be optimized by a two-pass analysis, one with a good time-resolution, and one with a good frequency resolution (figure 2.). For the purpose of timbre analysis however, it would be advisable to take into account the time resolution efficiency of the human hearing system. In this case it is not so important to consider a detailed frequency analysis of fast transients, since these would in general not be perceived as pitched sounds. The data gained by very short analysis windows applied to special situations including very fast transients would represent a rather marginal contribution to the overall timbre description.

The FFT-based analysis is generally done on a sliding time-domain window. The FFT peaks are found by analyzing the FFT of a windowed time signal.

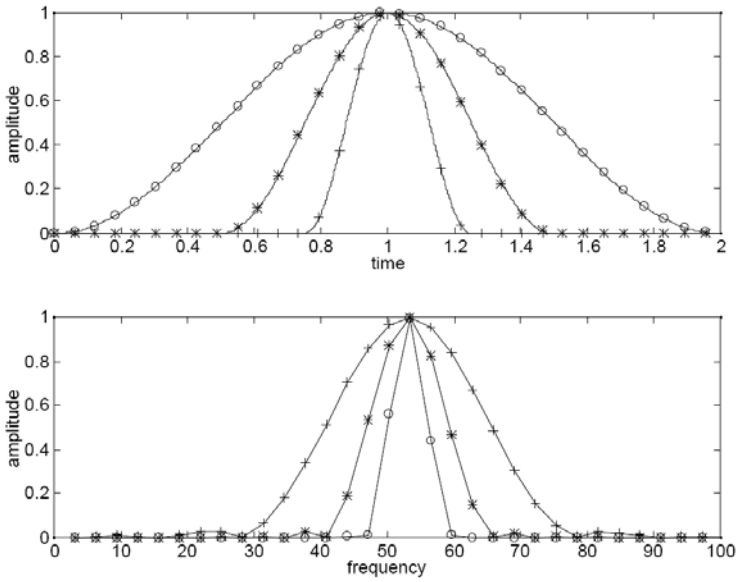


Figure 2. Illustration of the time / frequency window discrimination. A small time domain window yields a large frequency domain window, and vice verse. [Jensen 1999]

2.2.1 The fundamental frequency and higher harmonic content

The fundamental frequency of a musical sound is an important timbre attribute. Several algorithms for the estimation of fundamental frequency have been presented in the last few decades. The fundamental frequency estimation can be done in the time domain, the cepstral domain, or the frequency domain [Rabiner *et al.* 1976]. In the following chapters a rather primitive frequency domain method is presented which successfully estimates the pitch of most quasi-harmonic sounds.

The fundamental frequency is generally seen as the frequency of the first prominent spectral peak (the fundamental partial), or as the frequency difference between two adjoining harmonic overtones.

In order to find the fundamental frequency, the FFT should be performed on a certain segment of sound which is found right after the strongest (loudest) segment in the sound. The strongest is usually the “attack” segment, which is often containing too much transient behavior for a reliable estimation. After calculating the absolute of the complex valued FFT, the frequencies and amplitudes of the most pronounced peaks can be estimated.

2.2.1.1 General frequency and amplitude estimation

The discrete Fourier transform *X(k)* of a discrete time signal *x(n)* is computed as follows:

$$X(k) = DFT[x(n)] = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad \textbf{(2.3)}$$

where the frequency bin index *k* runs from 0 to *N-1*, with *N* being the window size in samples. The resulting N samples X(k) are complex-valued:

$$X(k) = X_R(k) + jX_I(k) \quad \textbf{(2.4)}$$

The resulting spectrum is composed of *N* equidistant frequency points with discrete frequency values from 0 to *(N-1)fs/N* Hz in steps of *fs/N*, where *fs* denotes the sampling frequency.

The inverse discrete Fourier Transform (IDFT) allows for the transformation of spectra in discrete frequency to signal in discrete time and is defined as:

$$x(n) = IDFT[X(k)] = \tfrac{1}{\mathbf{N}} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N} \quad \textbf{(2.5)}$$

In general, the time signal is multiplied by a window to avoid discontinuity effects – resulting from sharp changes at the signal boundaries – generating spectral leakage at the transformation. Leakage results in the signal energy smearing out over a wide frequency range in the FFT when it should be in a narrow frequency range. A Windowing function minimizes the effect of leakage to better represent the frequency spectrum of the data. Possible window forms include: Hamming Blackman, Hann, Triangular, Gaussian and Kaiser-Bessel windows. [Harris 1978]

2.2.1.2 Extraction of fundamental frequency and harmonic partials

There are many ways of detecting the fundamental frequency of a signal. However, it is not in the scope of this work to debate extensively on the state of the art approaches. Instead, a rather schematic approach shall be demonstrated in the next chapters with the aim to exposing some basic principles and the general relations between the fundamental and its harmonic partials.

In order to generate a sinusoidal representation of the signal, it needs to be reduced to the locations and strengths of individual peaks in the frequency domain. A general way of finding candidates for those harmonic partials from the frequency domain representation of a signal, is, by looking for peaks in the absolute value representation of $X(k)$. At this point it is important to state that not all bins that are higher than the two closest neighboring bins should be regarded as frequency domain peaks. When selecting relevant peaks it is important to consider a global relation of the identified peak to the complete signal and to pick only prominent and explicitly pronounced peaks. The frequency of a peak f_k found in the k-th bin can be defined as:

$$f_k = f_s * k / N \tag{2.6}$$

and its amplitude a_k - within the given representation - can roughly be described by:

$$a_k = |X(k)| \tag{2.7}$$

The frequency differences can now be calculated, by subtracting the frequency values of all pairs of neighboring peaks.

$$fd_1 = f_1 - 0, \qquad fd_2 = f_2 - f_1, \qquad \ldots\ldots \qquad fd_n = f_n - f_{n-1} \tag{2.8}$$

Then, all frequency differences that lie outside a certain percentage of the mean frequency (eq. 2.9) should be removed. The mean of the remaining frequencies can be defined as the fundamental frequency f_0.

$$F_0 = \text{mean}(fd) \qquad \textbf{(2.9)}$$

This estimation can be used to add possible missing harmonic frequencies and remove non-harmonic frequencies from the harmonic partial candidates. The resulting frequencies can now be defined as overtones of a harmonic sound.

Inharmonicity is an attribute to characterize pitched sounds with partial frequencies deviating harmonic frequencies. Those are also described as quasi-harmonic, which implies that the partial frequencies can be either stretched, or compressed. The frequency of the harmonic partial k can thus be a little higher or lower than k * fundamental. Inharmonicity is an attribute of stiff strings, for example, i.e. the piano sound. This is visualized best by dividing the overtones by their partial tone number / index resulting in a straight line for perfectly harmonic sounds and in a curve for quasi harmonic sounds.

Formula for the quasi-harmonic frequencies of a stiff piano string:

$$f_k = kf_0\sqrt{1+\beta k^2} \qquad \textbf{(2.10)}$$

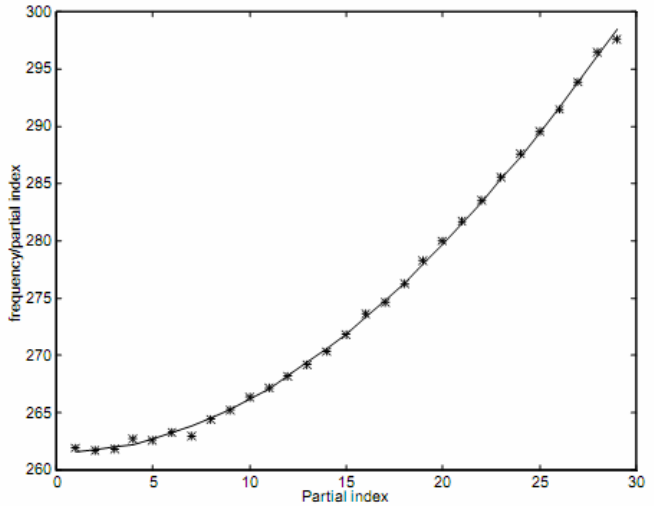


Fig. 3: Inharmonicity [Jensen 1999]

Here, f_0 stands for the fundamental frequency while β represents the strength of the inharmonicity.

2.2.1.3 Initial frequencies

Once the harmonic partials of a sound are extracted from the original FFT representation, a good initial estimation of the frequency content of a sound is provided. This data can then be used to further analyze and describe the sound.

The analyzed sounds are supposed to be harmonic or quasi-harmonic, but the extracted partials are often missing some harmonic partials, and can also contain strong non-harmonic partials, which are defined as spurious frequencies *or* "phantom partials" [Conklin 1997]. A spurious frequency is introduced, if it is sufficiently far away from the neighboring harmonic frequencies and if it is relatively strong compared with the neighboring frequencies and compared to the strongest partial. Those frequencies can also participate in the identification of an instrument. It is therefore necessary to consider them for inclusion in the sinusoidal model.

2.2.1.4 Partial track

In order to get a useful series of partials it is supposed that the frequencies and amplitudes can be connected in a series of connected lines, called tracks. The frequencies of these tracks can be harmonic, but they don't have to be, and there are often some shorter spurious partials in between the longer harmonic tracks. Several methods for tracking partials have been developed with locally optimized techniques [Serra 1989] or globally optimized techniques using hidden Markov modeling. [Depalle *et al.* 1993]. When the frequencies and the amplitudes are slowly varying, and the sounds are harmonic, the task of connecting the points is fairly easy, but noise and natural variations can often mask the partials. Supposing the partials up to time segment *k* have been connected. The *k-th* block has *N* partials and the *k*+1-*th* block has *M* partials. Then, the partials should connect if the difference in frequency, and perhaps also the difference in amplitude, is the smallest. All the close frequencies are analyzed and a matching value is calculated for each one of them. Here a rather simple locally optimized solution:

$$match(n,m) = k_a \mid a_{k+1}^n - a_k^m \mid + k_f \mid f_{k+1}^n - f_k^m \mid \qquad \textbf{(2.11)}$$

where partial n from block k+1 is connected to the partial m from block k with the best (lowest) match. The weights k_f and k_a are chosen experimentally. Good results can already be achieved if k_f is set to one, and k_a is set to zero. A more stable tracking is obtained if the slopes of the frequency and amplitude are used. Notably, partial crossing is then possible [Depalle *et al.* 1993].

Examples of partial tracks of different instruments can be seen in figure 4.

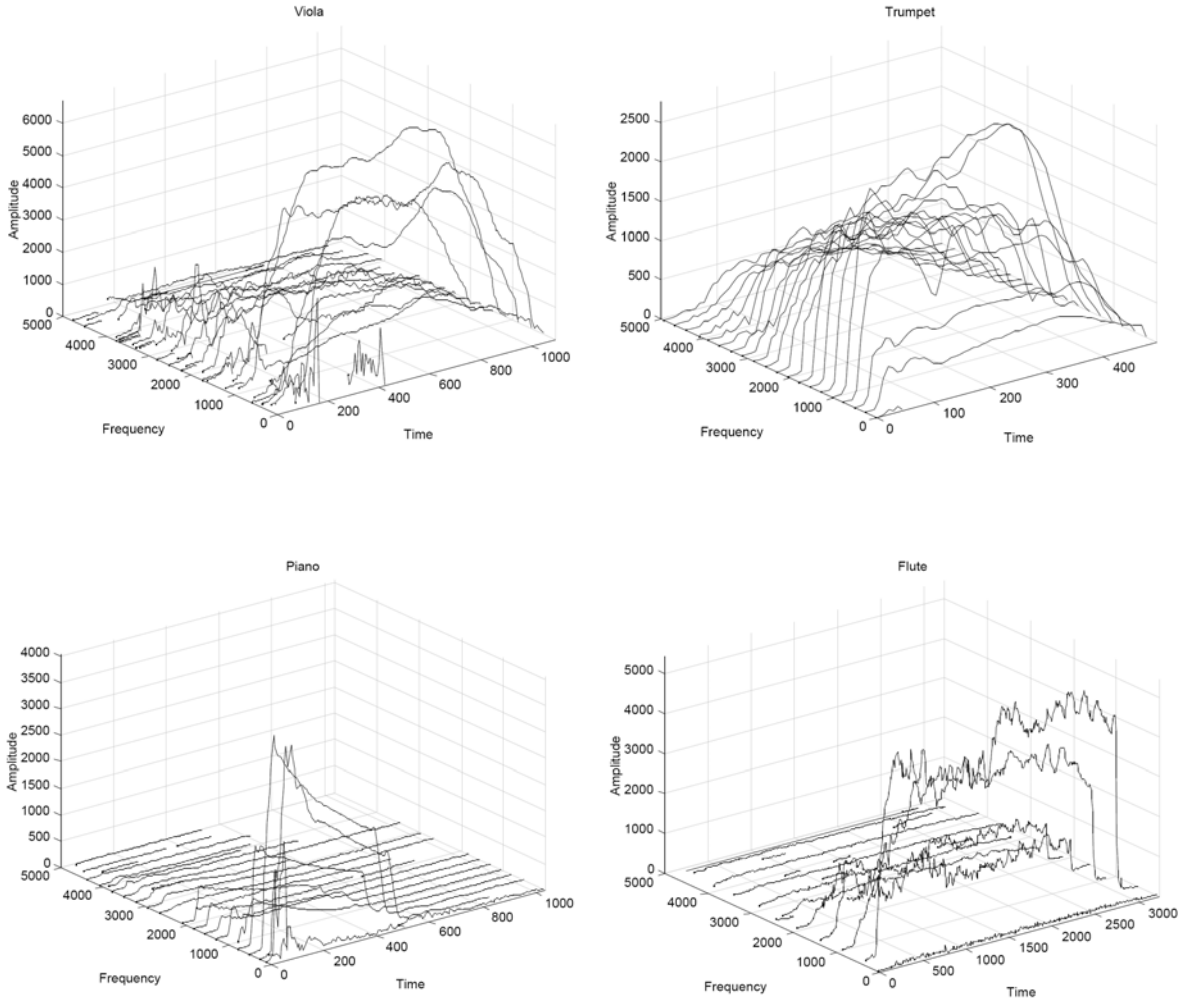


Figure 4. FFT analyzed additive parameters for viola, trumpet, piano and flute. [Jensen 1999]

2.2.2 Envelope modeling

The modeling of amplitude or other time-varying parameter in discrete time/value pairs is as old as electronic music. The ADSR envelope generator, which was introduced with the first analog synthesizers, divides the envelope in four steps, Attack, Decay, Sustain and Release, see figure 5. The ADSR approach operates with a reduced parameter set - defining amplitudes, durations and curve forms - which corresponds well with the perceptual quality of the amplitude.

Generally, the ADSR model is imposed on control parameters, such as amplitude, or filter frequency, and not on individual additive partials. The instrument model presented here will abstract its real-world instance by a sum of sinusoidals, also called partials, with time-varying amplitudes and frequencies.

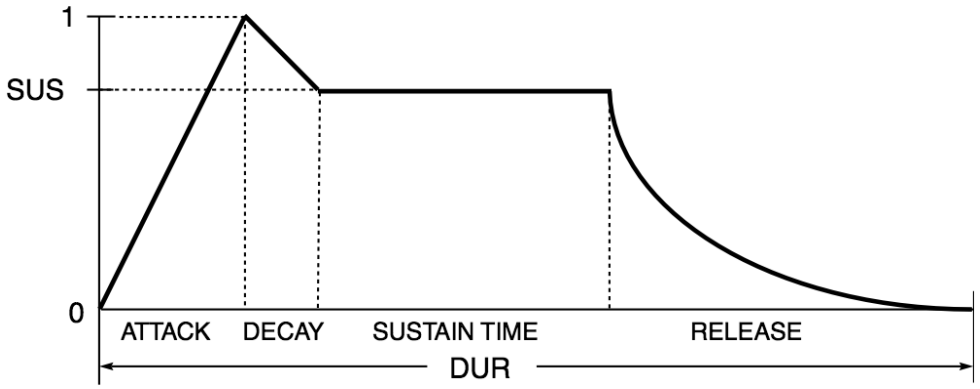


Figure 5. ADSR envelope.

The envelope is the evolution over time of the amplitude of a sound. It is one of the important timbre attributes. A faithful reproduction of a noiseless sound with no glissando or vibrato can be created using the individual amplitude envelopes of the additive frequency parameters. The analyzed amplitude envelopes often contain too much information to be easily manipulated; therefore, a model of the envelope is necessary. The envelope model presented in [Jensen 1999] is relatively simple, having only 4 split-points. The main characteristics of this model is the attack time, the sustain or decay as a homogenous segment of varying length, and the release time.

The envelope model can be seen as a data reduction of the additive frequency parameters. In [Horner *et al.* 1996], different envelope approximations are compared.

The model introduced by Jensen combines the intuitive simplicity of the ADSR model with the flexibility of the additive frequency model. His idea was to model each partial amplitude as four time/value pairs, here called start of attack (soa), end of attack (eoa), start of release (sor) and end of release (eor). Furthermore, the interval between each split point is modeled by a curve the quality of which (exponential/logarithmic) can be varied with one parameter. This model does not take into account tremolo or other effects. The sounds are supposed to be glissando-, vibrato- and tremolo-free, but these effects can be added to the additive parameters at any time.

2.2.2.1 Timing extraction

A method for the extraction of the attack and release times – also presented in [Jensen 1999] – finds the envelope times by analyzing the derivatives of the amplitude. The envelope times found are the start and end of the attack and release. The attack and release are found by searching for the maximum and minimum of the derivative of the envelope curve.

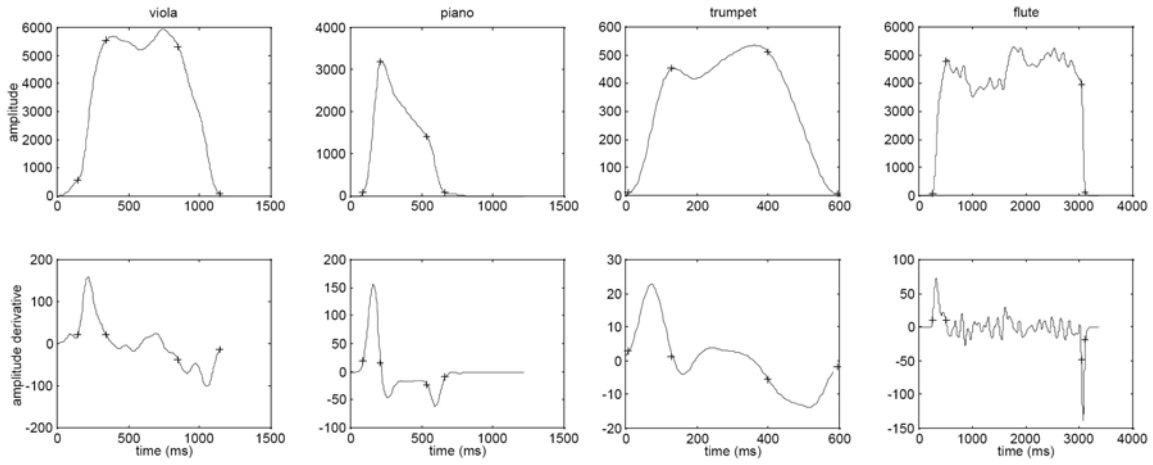


Figure 6. Amplitude and first derivative for the smoothed fundamental of four sounds with envelope times. [Jensen 1999]

The principle is illustrated in figure 6, where the smoothed envelope (top) and the first derivative (bottom) are shown for the viola, the piano, the trumpet and the flute. The start and end of the attack and release are indicated with '+'.

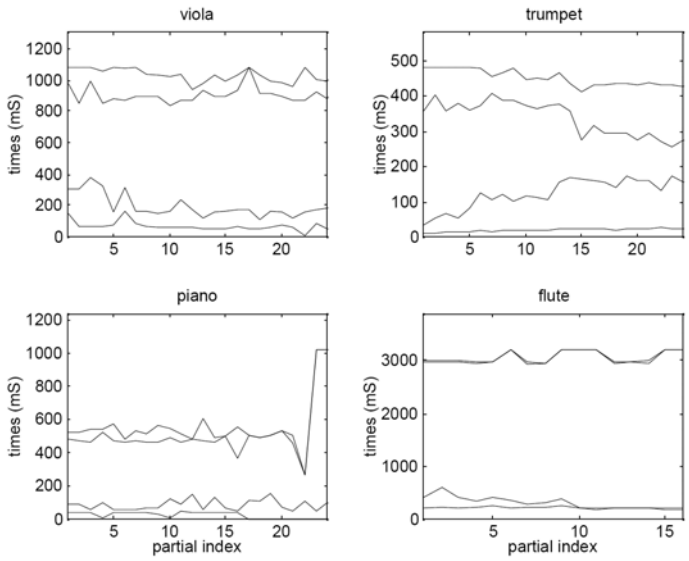


Figure 7. Slope times for the viola, the trumpet, the piano and the flute. [Jensen 1999]

2.2.2.2 Reconstruction of the envelope

An estimation of the envelope times is now available, but the curve between the envelope points is not known. The evolution between the envelope points can be modeled by a curve which has a parameter defined exponential/logarithmic slope. Obviously, no oscillation or irregularity is modeled, for these are assumed to be either tremolo or noise. There are five segments with a curve form for each partial; the start, attack, sustain, release and end segments. The recreated envelopes of the fundamental of the viola, the trumpet, the piano and the flute are shown in figure 8. The envelope split points are marked with plus signs in the plots. The detailed deduction of selecting and defining a best fit curve used for modeling one segment between 2 split points can be found in [Jensen 1999].

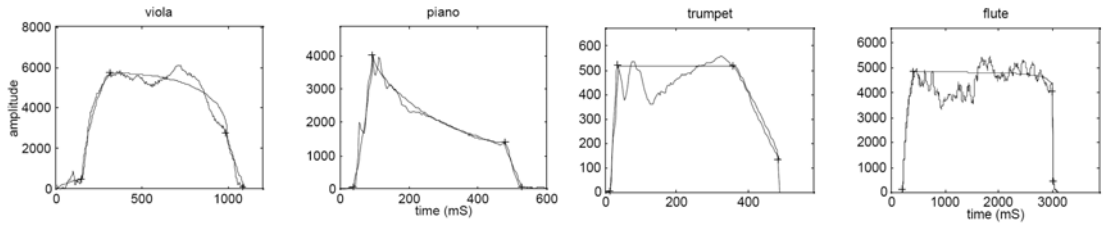


Figure 8. The original and the approximated envelope for four sounds. [Jensen 1999]

2.3 High Level Attributes (HLA)

The additive frequency parameters description is a good model of harmonic or quasi-harmonic instrument sounds, but it has a very large, non-intuitive parameter set. The High Level Attribute (HLA) term was introduced by [Serra *et al.* 1997] and can be seen as a data reduction of the additive frequency parameters. The HLA model is well suited for isolated sounds. It does not model vibrato, tremolo or glissando; however, its parameters help in the understanding of timbre and the perceived difference of sounds. Important timbre cues, such as the spectral envelope, the envelope timing, and the noise are easily extracted and visualized from this model. The parameters used for modeling each partial are:

- **amplitude envelope**
- **spectral envelope**
- **frequency**
- **noise**

Then the HLA model is then further re-organized, using the **Spectral envelope model.**

The interrelation of these four parameters is as follows. The amplitude envelope is based on an attack-sustain/decay-release model, where the maximum amplitude defines the second parameter, namely, the spectral envelope, the mean frequency further defines the frequency of each partial – the third parameter – and the irregularity of the partial amplitude and frequency models the noise of the sound (the fourth parameter of the HLA). These parameters will be discussed individually in the next chapters.

2.3.1 Amplitude envelope

The envelope of each partial is modeled in five segments, a start and end segment, supposedly close to silent, and an attack, sustain segment and release segment. Thus, there are 6 amplitude/time split points, where the first and the last amplitude values are zero, since all partials are supposed to start and end in silence. The amplitudes are defined as a percentage of the maximum of the amplitude, and the times are defined in ms. The perceptually most important envelope parameters seem to be the attack and release times. These are easily calculated from the difference between the absolute times. Furthermore, the curve form for each segment is modeled by either an exponential, logarithmic or linear curve, the choice of which would depend on a best-fit approximation of the original curve form.

2.3.2.1 Synchronicity

Synchronicity is an attribute, often accompanying the amplitude envelope. It is defined as the degree of time alignment of harmonic partials. Synchronicity in the onset part of a sound can be clearly observed in many acoustic instruments. An example for non-synchronous amplitude envelopes can be found, in woodwind instruments, where in general the starting time of the fundamental frequency occurs first, followed by the 2nd and 3rd harmonics.

2.3.2 Spectral envelope

The spectral envelope is defined in this work as the maximum amplitude of each partial. The spectral envelope is very important for the perceived effect of the sound; indeed, the spectral envelope alone is often enough to distinguish or recognize a sound. This is especially true for the recognition of vowels, which are entirely defined by the spectral envelope.

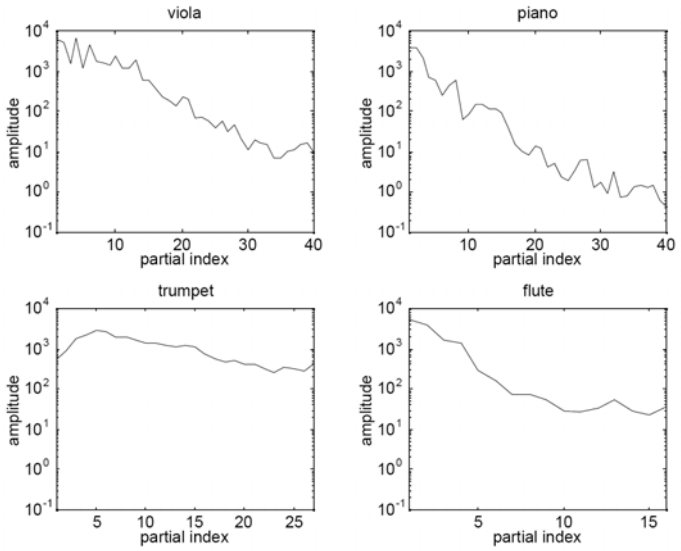


Figure 9. Spectral Envelope for the viola, the piano, the trumpet and the flute. [Jensen 1999]

2.3.3 Frequency

The frequency of each partial is modeled as the mean of the frequency for the sustain part. Assuming a stationary sound behavior, most sustained instruments are supposed to be perfectly harmonic. A particularly interesting representation of the parameter "frequency" is when the individual frequencies are divided by their partial index as seen in figure 10. The frequencies divided by the partial index will have a constant value for perfectly harmonic sound, i.e. for sounds exhibiting a constant frequency difference in all pairs of its neighboring partials. The degree of inharmonicity for the piano is easy to see. Notice the y-axis scale for the piano.

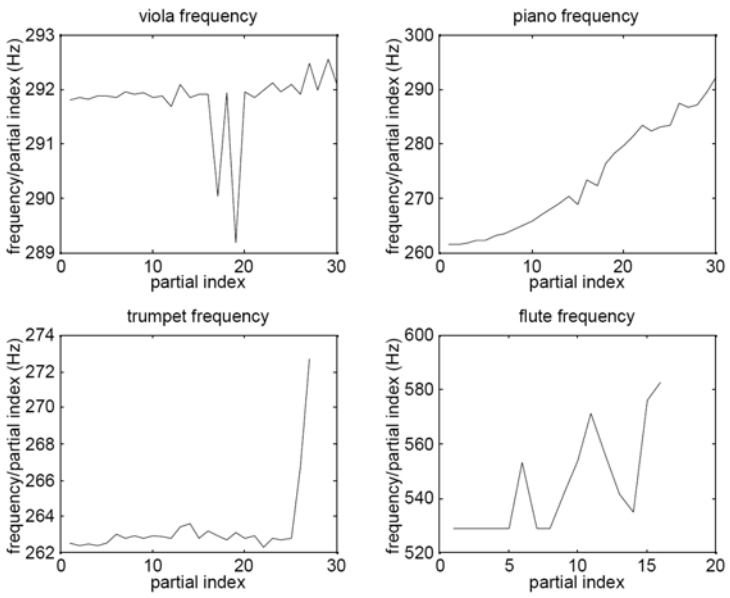


Figure 10. Frequency divided by the partial index for the viola, the piano, the trumpet and the flute. [Jensen 1999]

2.3.4 Noise

The simplified amplitude and frequency envelopes have the general shape of the original envelopes, however it is easy to see that there is a great deal of irregularity left, which is not modeled. The noise on the amplitude envelope is called shimmer, and the noise on the frequency is called jitter [Richard *et al.* 1996]. Shimmer is an additive component in the frequency domain, whereas jitter increases the bandwidth of the sinusoidal. Those two types of noise are modeled for the attack, sustain and release segments. The noise is supposed to have a Gaussian distribution; the amplitude of the noise is then characterized by the standard deviation. Shimmer is correlated with the maximum amplitude of the partial, whereas jitter is correlated with the mean of the frequency of the partial.

$$\sigma_{shimmer} = std(\frac{a_t - c_t}{c_t}) \quad \textbf{(2.12)}$$

$$\sigma_{jitter} = std(\frac{f_t - f}{f}) \quad \textbf{(2.13)}$$

Where a_t and f_t are the time-varying amplitudes and frequencies of the partial, f is the mean frequency and c_t is the curve found by the envelope model.

2.4 Spectral Envelope Model

Some perceptually meaningful attributes can be derived directly from the sound's spectral envelope, which is defined as the maximum amplitude of the harmonic partials of a sound. A model of the spectral envelope based on those attributes is presented in the following. This model, using perceptive attributes, is valid for non-formantic sounds.

The parameters of the spectral envelope model are**:**

- **Brightness (Spectral Centroid)**
- **Tristimulus**
- **Odd / Even relation**
- **Irregularity**

2.4.1 Brightness (Spectral Centroid)

The spectral centroid can be thought of as the center of gravity for the frequency components of a signal [Beauchamp 1982] and is correlated with the subjective quality of brightness [McAdams *et al.* 1995]. The Spectral Centroid, currently one of the MPEG-7 timbre descriptors, is defined as:

$$SC_{Hz} = \frac{\sum_{k=1}^{N-1} f[k]X[k]}{\sum_{k=1}^{N-1} X[k]} \qquad \textbf{(2.14)}$$

X[k] is the magnitude corresponding to frequency bin *k, f(k)* is the center frequency of that bin, *N* is the length of the DFT and *SC* is the spectral centroid in Hertz.

The brightness in the "partial domain" is calculated as:

$$brightness = \frac{\sum_{k=1}^{Np} ka_k}{\sum_{k=1}^{Np} a_k} \quad \textbf{(2.15)}$$

Here, *Np* is the number of extracted partials and *a* stands for their amplitude. If the partial multiplication *k* is replaced with the frequency of the partial, the brightness is expressed in Hertz. For harmonic sounds, this is equivalent to multiplying the brightness value expressed through the partial index with the fundamental frequency. Generally, it can be said that sounds with dark qualities tend to have more low frequency content and those with a brighter sound are dominated by higher frequencies.

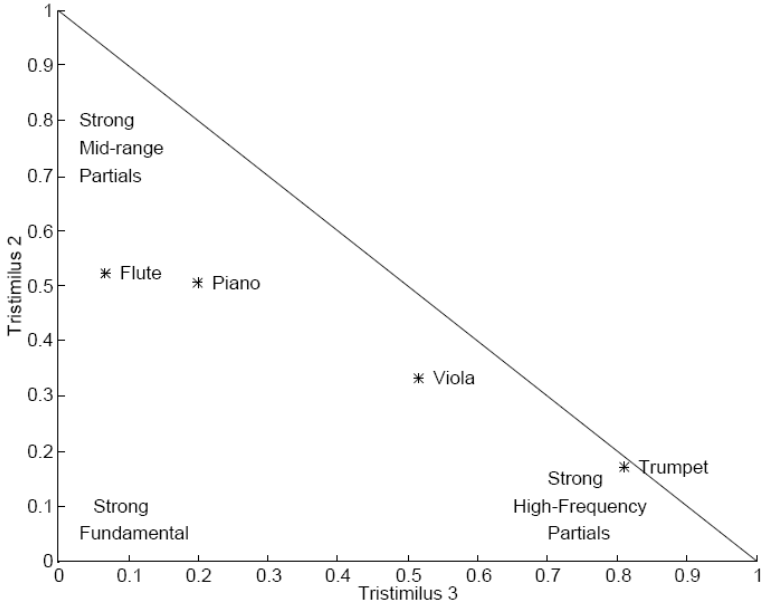
2.4.2 Tristimulus

The tristimulus is also a descriptor for the spectral energy distribution. It measures the energy in the fundamental-, the first three partials, and the higher partials in relation to the whole energy. Since the sum of Tristimulus one, -two and -three equals "1" only two values need to be calculated. The same accounts for the odd/even relation since Tristimulus 1+odd+even equals 1. The tristimulus values have been introduced in [Pollard *et al.* 1982] as a timbre equivalent to the color attributes in the vision. They used it for analyzing the transient behavior of musical sounds and for classification of musical timbre. Tristimulus is defined by the following three equations:

$$tristimulus1 = \frac{a_1}{\sum_{k=1}^{Np} a_k}$$

$$tristimulus2 = \frac{a_2 + a_3 + a_4}{\sum_{k=1}^{Np} a_k}$$

$$tristimulus3 = \frac{\sum_{k=5}^{Np} a_k}{\sum_{k=1}^{Np} a_k}$$



(2.16), (2.17), (2.18) **Fig. 11: Tristimulus 3 plotted against Tristimulus 2** **[Jensen, 1999]**

2.4.3 Odd / Even relation

This is a measure for the energy distribution on even and odd harmonics and is related to the subjective sensation of *fullness* of a sound. For instance the nasality and hollowness of the clarinet sound is caused by the dominance of odd harmonics [Benade *et al.* 1988].

$$odd = \frac{(\sum_{k=2}^{Np/2} a_{2k-1})}{\sum_{k=1}^{Np} a_k} \qquad \textbf{(2.19)}$$

$$odd = \frac{(\sum_{k=1}^{Np/2} a_{2k})}{\sum_{k=1}^{Np} a_k} \qquad \textbf{(2.20)}$$

To avoid too much correlation between the odd parameter and the tristimulus 1 parameter, the odd parameter is calculated from the third partial. Since tristimulus 1 + odd + even equals 1, it is necessary only to save one of the two relations. The odd parameter is saved.

2.4.4 Irregularity / Spectral Smoothness

Spectral irregularity also referred to as spectral smoothness (SSm) [McAdams 1999] basically shows the irregularity of a signal usually computed with the STFT where the average of the current, next, and previous amplitude values, i.e. the local mean, are compared with the current amplitude value. Bregman [Bregman 1990] remarks that the smoothness of a spectrum is an indicator for partials belonging to a same sound source and a single higher intensity partial is more likely to be perceived as an independent sound. It has also been found to be useful in revealing complex resonant structures of string instruments.

$$irregularity = \sum_{k=2}^{Np-1} | a_k - \frac{a_{k-1} + a_k + a_{k+1}}{3} | \qquad \textbf{(2.21)}$$

There is also an alternative version [Park 2004] of the conventional spectral irregularity algorithm, called Spectral Smoothness where the power of the spectrum is highlighted by the nonlinear square operator.

$$SSm = \frac{\sum_{k=1}^{Np-1} (a_k - a_{k-1})^2}{\sum_{k=1}^{Np-1} a_k{}^2} \qquad \textbf{(2.22)}$$

2.4.5 Other spectral descriptors applied to the complete spectrum (not to isolated partials)

In the previous chapter descriptors like *Brightness* and *Irregularity* were already introduced. Those find use in both, the spectral and the partial domain i.e., can be applied to isolated partials or to the full spectral representation. Choosing which context a descriptor shall be applied in, depends on the spectral complexity of the signal. In order to compare, analyze and classify sounds with respect their timbral quality, only monophonic signals with a strong harmonic content should be analyzed by their partial-structure / dynamics. Of

course, partial behavior analysis can be conducted on any kind of sound, or a polyphonic mixture of sound, however it might not be the most sophisticated way to generate meaningful or easily interpretable information. Whether harmonic descriptors are applicable to a sound or not may be expressed by the HarmonicEnergyRatio descriptor, described below.

2.4.5.1 Harmonic Energy Ratio

The harmonic energy ratio (HER) expresses the amount of signal energy resulting from harmonic partials over the total energy of the signal:

$$HER = \frac{\sum_{k=1}^{Np} a_k{}^2}{energy} \tag{2.23}$$

where the total energy of the signal is defined as:

$$energy = \frac{2}{N}\sum_{k=1}^{N/2} X[k]^2 \tag{2.24}$$

Here, N denotes the FFT frame length.

Next, a few other important spectral descriptors deployed in the analysis of the complete FFT magnitude spectrum shall be introduced.

2.4.5.2 Spectral Flux

The spectral flux (SF) defines the amount of frame-to-frame fluctuation in time. It is computed by the *2-norm* difference between consecutive STFT frames.

$$SF = \left\| X_t[f] - X_{t-1}[f] \right\| \tag{2.25}$$

where the general *q* norm is defined as:

$$\|X[f]\| = (\sum_{k=0}^{N-1} X[k]^q)^{1/q} \quad \textbf{(2.26)}$$

$X_t[f]$ denotes the magnitude components of frame *f* at time "t" and $X_{t-1}[f]$ at time "t-1". Both frame's magnitude components are of equal vector size. *SF* also known as the *delta magnitude spectrum* has also been used to discriminate speech and musical signals [Scheirer *et al.* 1997]. It exploits the fact that speech signals generally change faster than musical signals. In musical signals however, drastic changes tend to vary on a lesser degree.

2.4.5.3 Log Spectral Spread

The Log Spectrum Spread (SS) describes the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid, or else spread out over the spectrum. It allows differentiating between tone-like and noise-like sounds. The Spectrum Spread is defined as the RMS deviation of the log-frequency power spectrum with respect to its center of gravity, i.e the Spectral Centroid. The spread is somewhat similar to the spectral distribution found by Grey [Grey 1977] and is also compared to the *richness* of a sound, however no attempts have been made to quantify this factor. The spectral spread has also been specified as one of the MPEG-7 audio descriptors. It is calculated as:

$$SS = \sqrt{\frac{\sum_{k=0}^{N/2} (f_k - SC)^2 \left|X[k]\right|^2}{\sum_{k=0}^{N/2} \left|X[k]\right|^2}} \quad \textbf{(2.27)}$$

where $X[k]$ denotes the magnitude of the k-th frequency component (f_k). SC is the frequency of the spectral centroid.

2.4.5.4 Spectral Flatness Measure

The Spectral Flatness Measure (SFM) describes the flatness properties of the short-term power spectrum of an audio signal. This descriptor expresses the deviation of the signal's power spectrum over frequency from a flat shape (corresponding to a noise-like or an impulse-like signal). A high deviation from a flat shape may indicate the presence of tonal components. The spectral flatness analysis is calculated for a number of frequency bands. It is defined as the ratio between the *geometric mean* (Gm), and the *arithmetic mean* (Am). As SFM approaches "0" the signal becomes more sinusoidal and as SFM approaches "1" the signal becomes more flat and de-correlated.

$$SFM_{dB} = 10\log\left(\frac{Gm}{Am}\right) = 10\log_{10}\left(\frac{\left(\prod_{k=0}^{k=N-1}|X[k]|\right)^{1/N}}{\frac{1}{N}\sum_{k=0}^{N-1}|X[k]|}\right) \quad \textbf{(2.28)}$$

2.4.5.5 Roll-off

The *roll-off* point in Hertz is defined as the frequency boundary where 85% of the total power spectrum energy resides. It is commonly referred to *skew* of the spectral shape and is frequently used in differentiating percussive and highly transient sounds (which exhibit higher frequency components) from more constant sounds such as vowels [Park 2004].

$$\sum_{k=0}^{R} X[k] = 0.85\sum_{k=0}^{N-1} X[k] \quad \textbf{(2.29)}$$

where R is the frequency roll-off point with 85 % of the energy.

2.4.6 Time varying spectral envelope

Until now, the analysis of the spectral envelope was demonstrated on only one isolated frame, with the exception of the spectral flux descriptor, of course, which is conceptually conceived to operate with the deviation measure of two successive frames. In general, the envelope constructed with the maximum amplitudes of the quasi-harmonic partials that occurred on a sustained segment of a sound sample, was examined. Presented were also descriptor concepts that were not applied to the partial model but on the complete spectrum instead. In both cases only one frame (one instant of a sound) was analyzed, or, at best the difference in two consecutive frames (e.g. Spectral Flux). Since this work is about the temporal character of timbre, i.e. the evolution of the spectrum across time, a time axis needs to be introduced to the model. All the above introduced spectral envelope model parameters can of course be calculated for the time-varying spectrum.

The time varying spectral envelope model parameters for four test sounds can be seen in the following figures. The tristimulus is plotted only for the times where the amplitude is above 10 percent of the maximum amplitude. There is no time axis for the tristimulus, where tristimulus 2 is plotted as a function of tristimulus 3, but the time can be followed from the start '+' to the end 'o'

The spectral envelope parameters seems rather stable in the sustain part of the sound. The trumpet has much higher brightness in the middle of the sound than in the beginning and end of the sustain, even though the amplitude is rather stable throughout the sustain. The flute also has this behavior, although not as pronounced. The viola and the piano have falling brightness with time. These observations are made on the non-zero amplitude times. The viola has a lot of tristimulus variations, but most of this probably occurs in the attack. The trumpet has almost no tristimulus 1 and the flute has no tristimulus 3. The trumpet has a relatively high odd value, and the flute has a low odd value. The viola has a very high irregularity where the trumpet has a very low irregularity. [Jensen 1999].

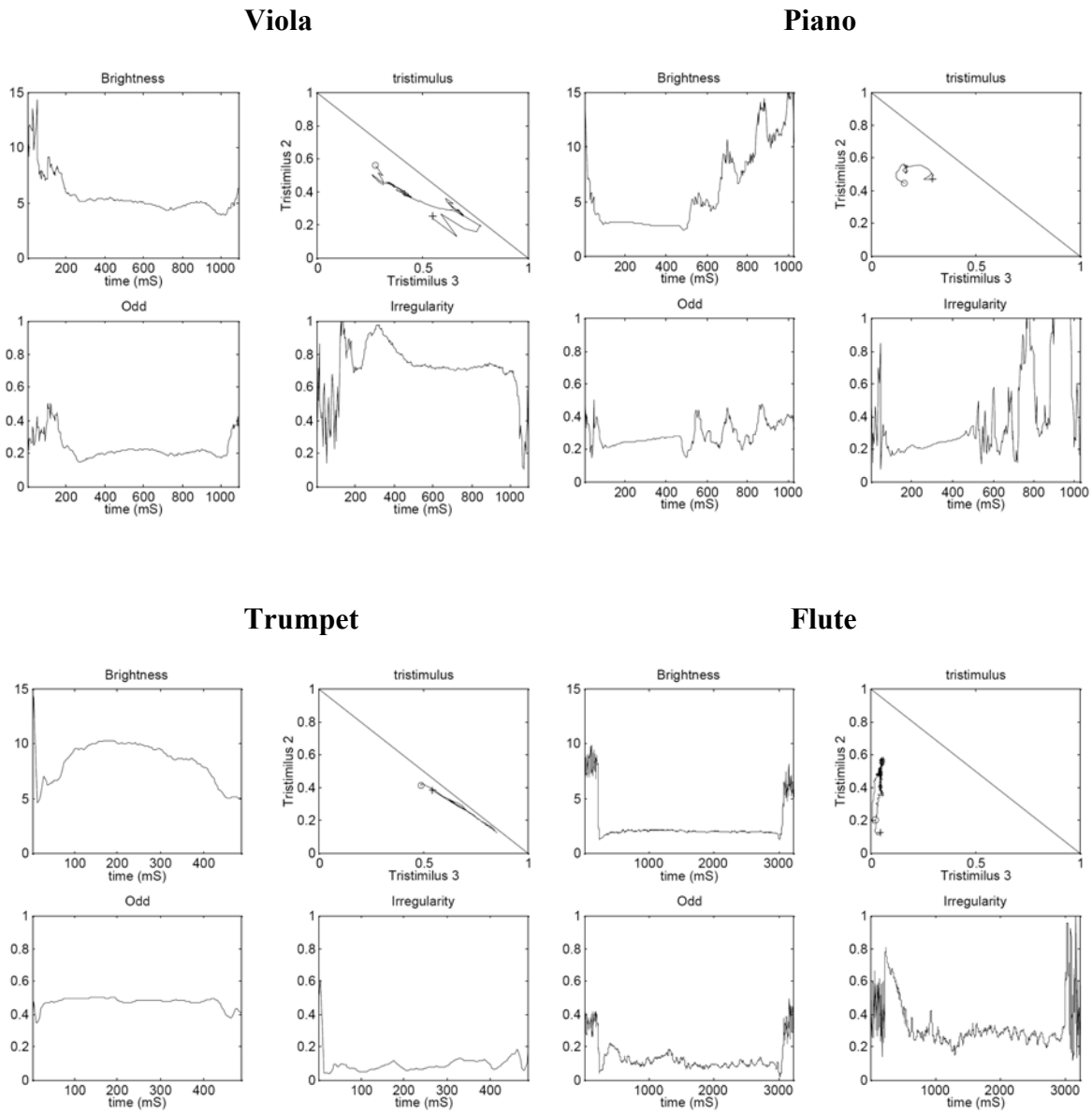


Fig. 11: Time-varying parameters of the spectral envelope model for the viola, the piano, the trumpet and the flute [Jensen 1999]

By analyzing a time-varying spectral envelope, well differentiated models of harmonic sounds can be made. It is important to state however, that the sounds analyzed here are monophonic instrument samples. Although most musical instruments can be modeled like this, it is not a universal method for characterizing sound samples.

2.5 Minimal Description Attributes

The Minimal Description Attributes (MDA) are calculated with HLA model parameters. The MDA model is an attempt to distil the minimum number of parameters necessary to characterize the identity and quality of an instrument. Instead of keeping one parameter for each HL-Attribute and partial, a model of the curve along the partial axis for each attribute is found and modeled using a few parameters. The MDA model is created by curve fitting [Lancaster *et al.* 1986] the data of the HLA model to a simple curve (usually exponential or 2^{nd} order polynomial). The MDA model generally generates two values (curves) for each attribute, a fundamental value, and a partial evolution value.

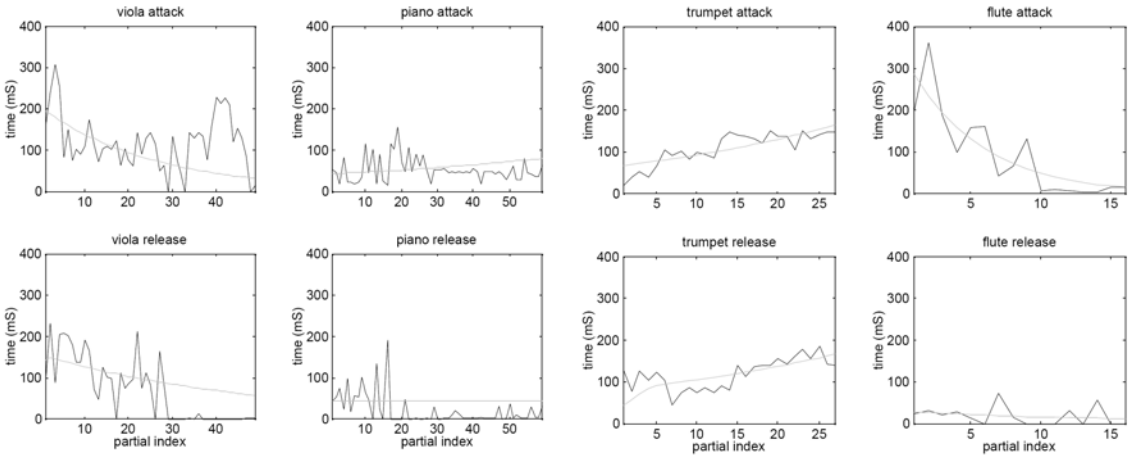


Figure 12. Attack and release times for the 4 sounds, with the MDA model envelope times (dotted). Attack (top) and release (bottom) [Jensen 1999]

Of course the resulting description doesn't meet the degree of detail contained in the HLA model but it reflects the general trend.

Figure 12. shows an example of the MDA approximation for the envelope model parameters. Details on all other parameters can be found in [Jensen 1999].

2.6 Instrument Definition Attributes

The instrument definition attributes (IDA) model has been introduced to collect the timbre attributes for many executions of the same instrument. The IDA model can visualize changes of timbre attributes as a function of *fundamental frequency*, different *playing styles* (e.g. *legato* and *staccato*, etc.) different *tempi* or different *intensities*. The IDA parameters are assumed to give a complete description of a musical instrument, ranging from the definition of the timbre of one sound, to the evolution of the timbre as a function of note or expression.

This model keeps the mean of every MDA parameter for each half octave band, for each playing style, intensity and tempo. The IDA parameters become more stable, the more sounds there are in each frequency band.

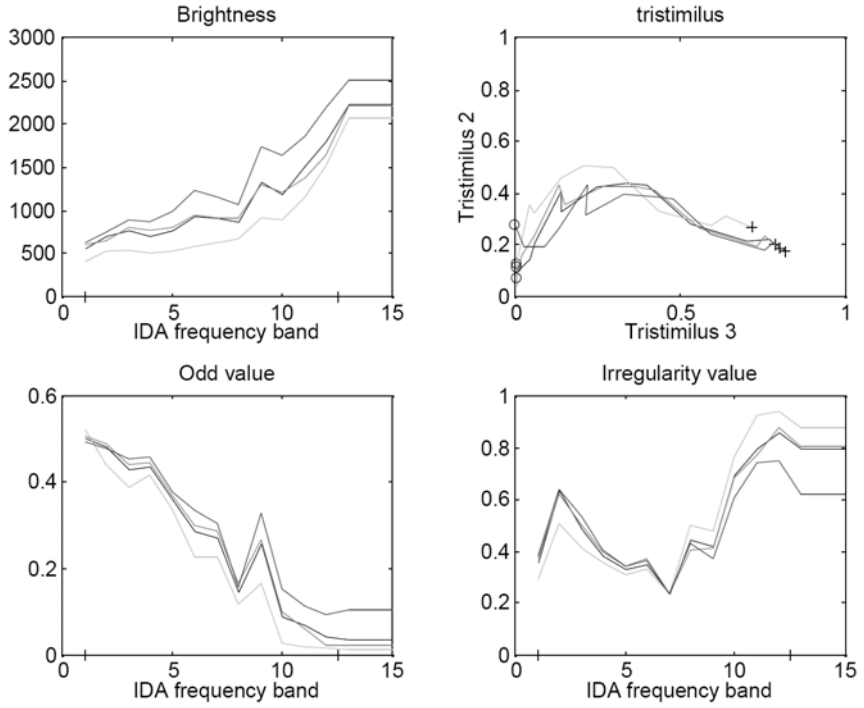


Figure 13. Spectral Envelope parameters for three different loudnesses for the piano. All loudnesses (solid), *piano* (dotted), *mezzo forte* (dashdotted) and *forte* (dashed). [Jensen 1999]

The IDA frequency range is divided into 15 bands, which range from 32 Hz to 4 kHz in half octave steps. All MDA parameters are then searched for each band, and the ones whose fundamental frequency is between the band ± 1/4 octave are used. Each parameter (spectral envelope, frequencies, envelope, noise, etc.), including the partial evolution, of the band is set to the mean of the corresponding parameter of the MDAs used.

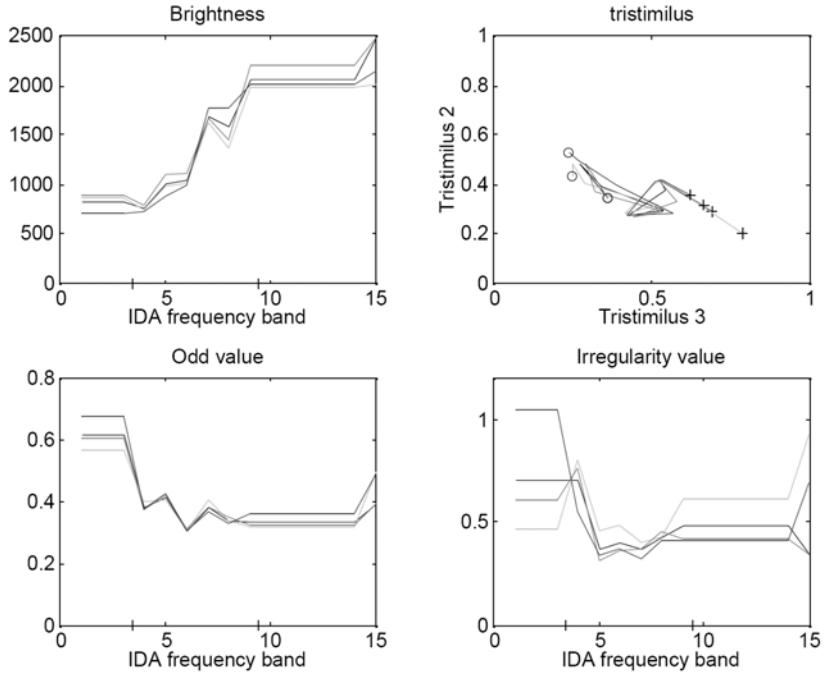


Figure 14. Spectral envelope parameters for the different styles of the cello. Complete cello set (solid), *staccato* (dotted), *spiccato* (dashdotted) and dashed (*legato*). [Jensen 1999]

In figures 13. and 14. the effects of changing velocity and playing stile are visualized for four of the spectral envelope attributes, respectively. A detailed discussion on all other attributes and parameters can be found in [Jensen 1999].

3. Polyphonic Timbre Mixtures

The next major section of this thesis should form a counterpoint to chapter 2 (Monophonic timbre) and will focus on the derivation of polyphonic timbre analysis methods applied at musical segments where several instruments play simultaneously. An interesting work documenting the perception of timbre mixtures of multiple instruments was conducted by [Kendall *et al.* 1991]. The authors have collected human dissimilarity judgments for pairs of instruments playing either single tones (at unison or in distance of a major third interval) or simple melodies (again, at unison and in harmony). The test listeners evaluated the timbral quality of different instruments with semantic descriptions / verbal attributes like "rich", "brilliant" and "nasal". They compared the vector-like timbre descriptions of the composed timbres with the timbre descriptions of each individual instrument. They found that, to a limited extent, a quasi-linear vector model could explain the perception of timbre combinations on the basis of the vector sum of the positions of individual timbre vectors. This suggests that attributes of timbre are perceived for linear combinations of sounds as a sum of the individual sound attributes. However, they also point that a linear mapping from the physical / acoustical dimension space onto a verbal was rather unlikely. Also, many features discussed for monophonic sounds in the second chapter are not linear functions of the signal. Especially temporal descriptors such as attack / release times, but also e.g. RMS-energy in frequency bands can not be added together this way. Therefore the computed value for mixtures of signals is usually not a linear combination of the individual timbre values. This suggests that, the characteristic timbral features computed for monophonic signals cannot directly extract musically meaningful data from polyphonic signals. Perhaps it would be possible to apply the monophonic analysis mechanisms if source separation of the individual components could be performed prior to analysis; however, this is still a difficult research problem of its own [Martin 1999], [Plumbley *et al.* 2002]. In the following chapters methods for the analysis of polyphonic timbre will demonstrated and discussed.

3.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are a popular feature used for speech analysis and recognition. The process of forming MFCCs executes two basic steps. First, the information contained in the FFT representation is reduced (compressed) by the warping the linear frequency scale in Hz, to the Mel frequency scale i.e. by grouping and weighting of the individual frequency channels according to that (Figure 17). Second, the calculation of the Discrete Cosine Transform (DCT) is performed in order to decorrelate these Mel-spectral vectors, and thus compress the spectral information into the lower coefficients. Different approaches involving MFCCs have been widely deployed by researchers to model music and audio sounds [Foote 1999], [Aucouturier *et al.* 2005], [Morchen et al. 2005], etc. The MFCCs – e.g. see [Rabiner *et al.* 1993] – are short-term spectral features. They are calculated as follows:

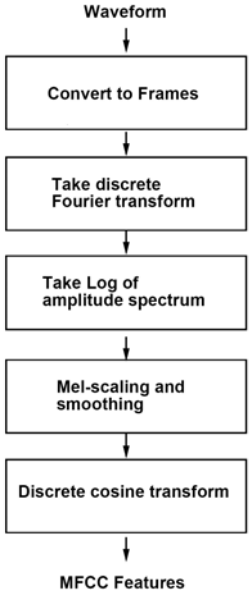


Figure 15. Process of generating MFCC features

The Mel-scale was proposed in 1940 by Stevens [Stevens *et al.* 1940] as the result of an experiment, where the difference between the real and the sensed pitch should be detected. The concept of the Mel scale is to organize pitch values which were judged by listeners to be equal in distance from one another. Apparently, the human auditory system does not perceive pitch in a linear manner. The Mell mapping is approximately linear below 1000

Hz and logarithmic above. The Mel scale is defined in the next equation, where *f* is the frequency in Hz.

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \qquad \textbf{(3.1)}$$

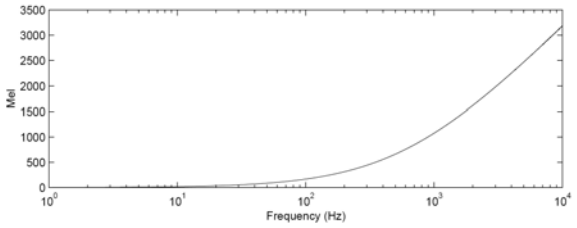


Figure 16. Mel scale vs. linear frequency scale

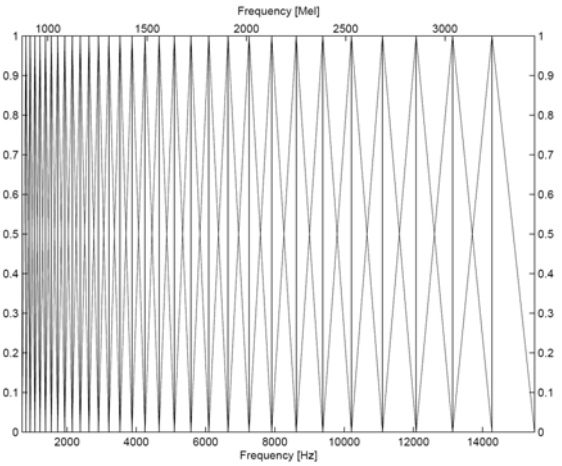


Figure 17. Mel filter-bank in frequency space

After the frequencies have been warped to the Mell scale, the signal is transformed to the Cepstral domain through the calculation of its Discrete Cosinus Transform (DCT). Generally, the cepstrum is defined as the inverse Fourier Transform of the log-spectrum [Oppenheim et al. 1968].

$$c(n) = IDFT(\log|X[k]|) = \tfrac{1}{\mathbf{N}} \sum_{k=0}^{N-1} \log|X[k]| \cdot e^{j2\pi nk/N} \qquad \textbf{(3.2)}$$

Calculation of cepstral coefficients with the IFFT. - *c*(*n*) is called the *n*th cepstral coefficient.

In practice however, the (DCT) is used instead of the inverse DFT (FFT). The DCT is defined as:

$$X_{DCT}(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi}{N}(n + \frac{1}{2})k\right] \qquad \textbf{(3.3)}$$

It is similar to a DFT but it only deploys a cosinus function to modulate the input signal. Like the DFT, it has the property that most of the energy of its output signal is concentrated in the first few coefficients (thus effectively compressing the spectral information).

The first few (low order) MFCC's account for the spectral envelope, which describes the general trend of the spectral representation. The higher order coefficients, describe the finer variations of the spectrum, i.e. the local details.

The fact that the logarithmic spectrum can be interpreted as the sum of two spectra, and therefore, the cepstrum is the sum of two components, has been exploited in the Source-Filter modeling approaches [Zölzer *et al.* 2002]. Cepstral analysis is based on the observation that a logarithmic spectrum can be decomposed as the sum of source and filter spectra. It is known that filtering in the frequency domain is achieved by a multiplication operation, which corresponds to adding logarithms. A filtered spectrum can thus be derived by adding logarithmic spectra.

The real cepstrum method will perform a spectral envelope estimation, based on the magnitude of the FFT alone. It can be written as:

$$\left|X[k]\right| = \left|S[k]\right| \cdot \left|E[k]\right| \qquad \textbf{(3.4)}$$

Taking the logarithm yields:

$$\log\left|(X[k])\right| = \log\left|(S[k])\right| + \log\left|(E[k])\right| \qquad \textbf{(3.5)}$$

The S term would represent an event sequence (e.g. a pulse sequence with a frequency of 100 Hz), which can be defined as "carrier" or the sound source. The term E corresponds to "envelope" in the frequency domain. In other words, S varies more quickly than F. Therefore, it is possible to apply some kind of filter to separate log(X(ω)) into "high-frequency" components from "low-frequency" components.

The two components can now easily be separated by executing another Fourier transformation on the given signal. Roughly, the lower coefficients – resulting from the modulation – would represent the spectral envelope (i.e. the formants, i.e. the filter) while

the higher ones would represent finer details of the spectrum, (i.e. the excitation source, i.e. the fundamental frequency) [Oppenheim *et al*. 2004].

Looking at speech or the singing voice from the source–filter point of view of, the voice's spectral envelope can be treated almost independently of its pitch. If we consider a concrete example, and try to transpose a vowel up by one octave, e.g. by resampling, the spectral envelope will be transposed as well. This effect would sound quite unnatural since formants are shifted up one octave, which would correspond to shrinking the vocal tract to a half of its length. Obviously, this is not the natural behavior of the vocal tract. Through a source–filter separation it is possible to pitch-shift only the source part, while the "filter" remains in an authentic shape. For music, source corresponds to vibrations (e.g. vibrating strings in plucked or bowed string instrument) and filter corresponds to the body of the instrument

Since MFCC parameterization discards pitch information, it can be objected that it might not be an appropriate method for music analysis (as opposed to speech). However, in the perception of timbre, the fundamental frequency is not a carrier of significant information. The lower coefficients of the MFCC representations will tend to describe general timbral shapes rather than exact pitches.

3.1.1 Organization of MFCC features

In recent years, different approaches analyzing polyphonic textures were developed and automatic systems were conceived, which are able to extract high-level descriptions (HLD) of music signals. A popular approach, to describe polyphonic timbre is the MFCC representation described above. In order to identify temporal structure of the obtained MFCC features that were computed for each signal frame, they need to be organized and analyzed.

3.1.2 Global models

There are a number of modeling approaches, which do not model a temporal structure, but rather try to describe an overall timbral character of an analyzed sequence. The most

popular way of aggregating a low level feature time series is the usage of mean and standard deviation [Aucouturier *et al.* 2004]. All feature vectors are fed to a classifier which models the global distributions of the features of signals using Gaussian Mixture Models (GMM). A GMM estimates a probability density as the weighted sum of *M* Gaussian densities, called components or states of the mixture.

$$p(x) = \sum_{m=1}^{m=M} w_m N(x, \mu_m, \Sigma_m) \tag{3.6}$$

where x is the feature vector observed at time t, N is a Gaussian probability density function (pdf) with mean μ_m, covariance matrix $\sum_m$, and w_m is a mixture coefficient (also called state prior probability).

$$N(x, \mu_m, \Sigma_m) = \frac{1}{2\pi^{\frac{d}{2}} \left|\Sigma_m\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1}(x-\mu_m)\right) \tag{3.7}$$

where d is the dimension of the feature vector x. The parameters of the GMM are initialized by random distribution and further trained with the classic Expectation Maximization (E-M) algorithm [Bishop 2006].

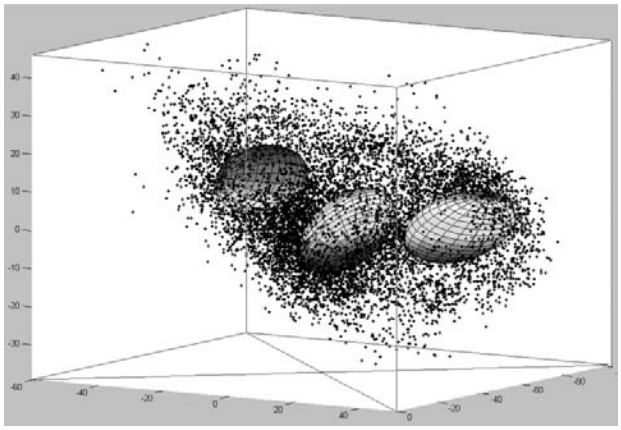


Figure 18: (GMM modeling an MFCC distribution (The Beatles—"Let It Be") – displaying three pronounced clusters of feature vectors. The axes correspond to the first three components from a set of 12-dim MFCC vectors [Aucouturier *et al.* 2005]

The dimension of the feature vector equals the number of MFCCs (N) extracted from each frame of data. The more MFCCs are kept, the more precise the approximation of the signal's spectrum, which also means more variability on the data. In this work, the interest lies in the spectral envelopes, not in the finer details, therefore a large number of coefficients (e.g. above 20) may not be appropriate. In general it can be said that the more Gaussian components (M) are defined to model the MFCCs, the better the precision of the model[2]. N and M are not independent and there is an optimal to be found between high dimensionality and high precision of the modeling. Figure 18 shows a three-dimensional (3-D) projection of a typical (higher dimensional) feature space. The dots represent MFCCs and the ellipsoids are the projection of the Gaussian distributions in the trained GMM.

Those models however do not take into account the dimension of time as one of the components contributing to the description of timbre. All frames are typically modeled as a whole, without any account of their time ordering.

3.2 Modeling the temporal dynamics of MFCC features

The next step towards temporal structure analysis is the computation of first and second order differences amongst successive MFCC frames [Berenzweig *et al.* 2003], [Morchen *et al.* 2005], etc. While this method still does not model a concrete sequence of the analyzed events, it may provide a general, idea about the global dynamics of timbral evolution in the observed sequence.

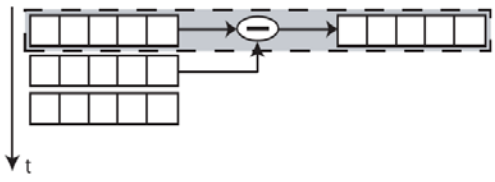


Figure 19: Computation of delta coefficients from a sequence of MFCC features. [Aucouturier 2006]

[2] It should be mentioned here, that it is not reasonable to increase M without any limit. The model will not be useful in an extreme case, where M would equal the number of available feature vectors

The same concept can be applied to the delta coefficients again to obtain the acceleration coefficients. The resulting modified feature set may contain, for each frame, the static feature values and their local delta values. The general idea behind these concepts however is to describe the rate of change with reference to the temporal dimension and the degree of deviation in successive feature frames.

3.2.1 Texture windows

A method to capturing the long term nature of sound, while still assuring that the features are computed on small stationary windows, is to average those local static features (typically extracted every 50 ms) over larger-scale windows (typically several seconds) [Tzanetakis *et al.* 2001]. Several statistics can be used on such so-called *texture windows*, e.g. mean, standard deviation, skewness, range, etc.

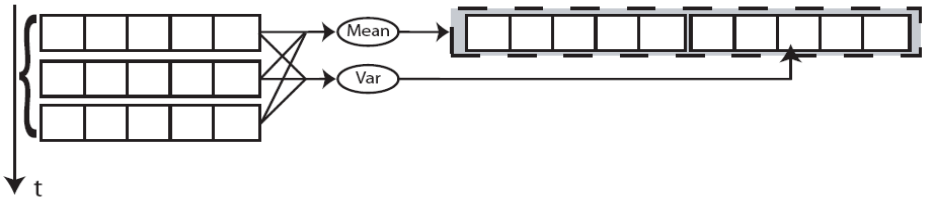


Figure 20: Texture windows [Aucouturier 2006]

3.2.2 Dynamic features

Another strategy to characterize the dynamics of the static features is to compute features on the signal constituted by the static feature sequence. For instance, an FFT can be taken on to analyze a sequence of MFCC feature frames, representing several seconds of audio. The FFT is executed individually, on each coefficient's temporal trajectory (Figure 21). Then, the low-frequency variations of the features (e.g. $[1 - 50Hz]$) may be taken as features instead of the original ones.

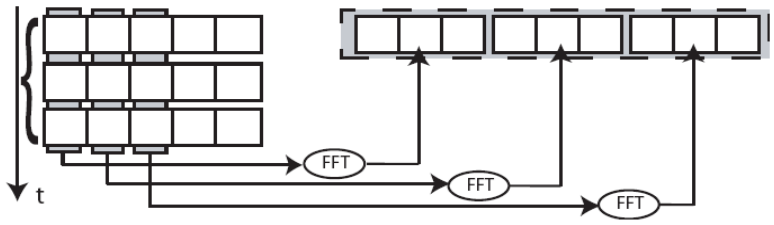


Figure. 21: Dynamic features [Aucouturier 2006]

3.3 Self similarity

The next step towards temporal structure analysis can be done by increasing the time lag in the MFCC–frame comparison process and further, calculating derivatives between more distant samples as well. The method involves the computation of autocorrelation- or self similarity matrices and was deployed in the work of Foote [Foote 1999]. The timbral similarity between any two instants of the analyzed sample can be displayed in a two-dimensional representation. Similar or repeating elements are visually distinct, allowing identification of structural and rhythmic characteristics.

Self similarity and repetition is a general feature of nearly all music. That is, the coda often resembles the introduction, the second chorus sounds like the first, and a theme is more or less similar to its variations. On a shorter time scale, successive bars are often repetitive, especially in popular music. Figure 22 shows a self similarity plot, where an audio file, or better to say, its structure - in terms of timbral development - is represented as a square. Each side of the square is proportional to the length of the piece, and time runs from left to right as well as from bottom to top.

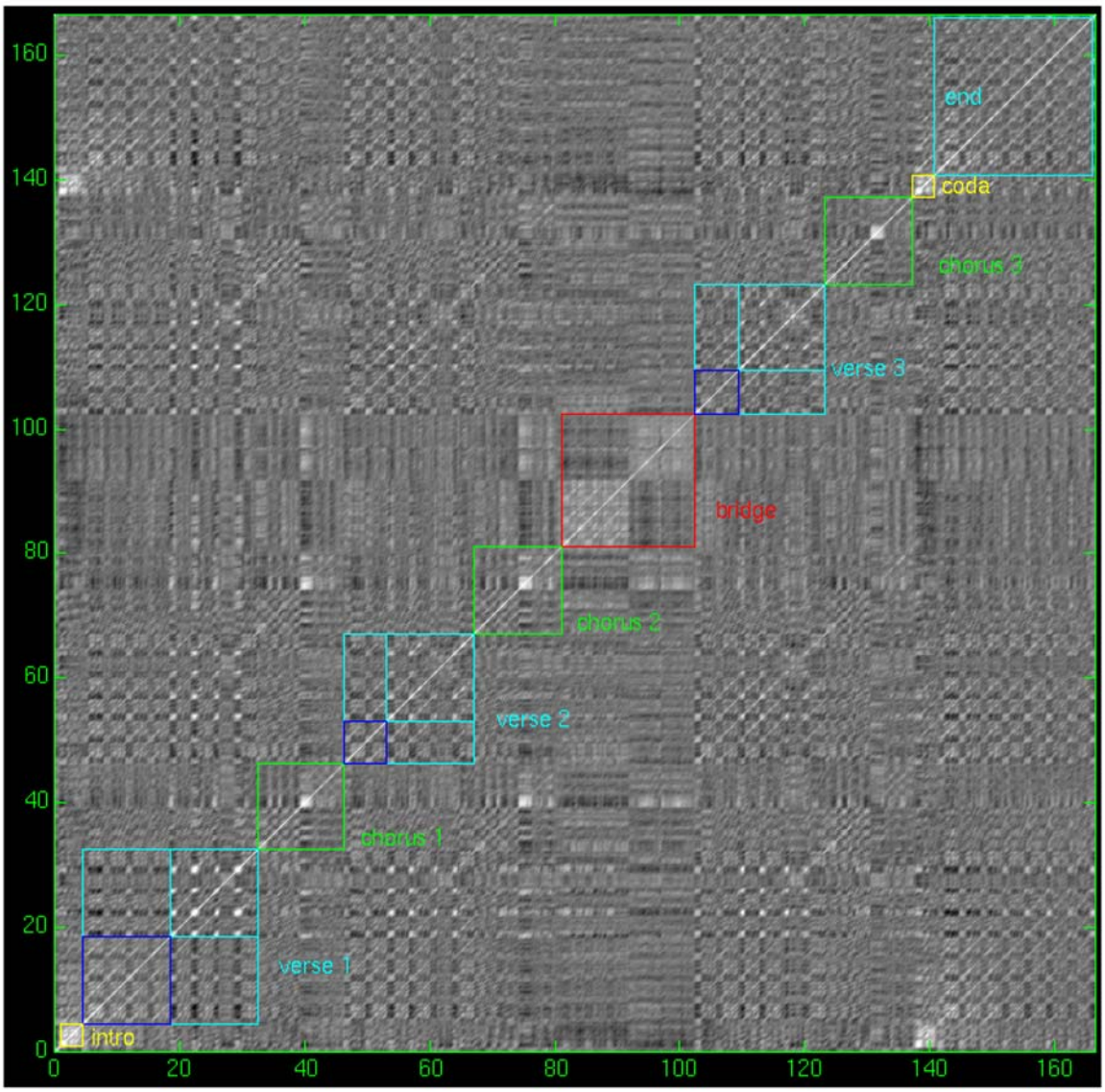


Figure 22. Self-similarity visualization - *Day Tripper* performed by the Beatles [Foote 1999]

The bottom left corner of the square corresponds to the beginning of the piece, while the top right corresponds to the end. In the square, the brightness of a point is proportional to the audio similarity at times *i* and *j*. Similar regions are bright while dissimilar regions are dark. Thus there is always a bright diagonal line running from bottom left to top right, since the similarity of two audio segments is always at its maximum when the segment is compared against itself. (Technically, the autocorrelation is always a maximum at a lag of zero.) In figure 22 repetitive similarity features such as repeating notes or motifs, show up

as a checkerboard pattern. Long repeated themes are visible as diagonal lines parallel to and separated from the main diagonal by the time difference between repetitions.

The feature vectors, calculated for each frame of the signal are used to calculate the similarity between two audio "instants" (frames). The similarity measure implemented by Foote is based on feature vector correlation. Given two MFCC feature vectors derived from audio frames *i* and *j*, a simple metric of vector similarity *s* is the scalar (dot) product of the vectors.

$$s(i,j) = v_i \cdot v_j \tag{3.8}$$

This product will be large if the vectors are both large and similarly oriented. Because the analysis frames, hence feature vectors, occur at a rate much faster than typical musical events, a better similarity measure *S* can be obtained by computing the vector correlation over a several frames. Thus:

$$S_w(i,j) = \frac{1}{w}\sum_{k=0}^{w-1}(v_{i+k} \cdot v_{j+k}) \tag{3.9}$$

To visualize an audio sample, a number of frames *w* is chosen, and the similarity measure *S*(*i*,*j*) is calculated for all pairs of frame combinations at start time indices *i* and *j*. Then an image is constructed so that each pixel at location *i*, *j* is given a grayscale value proportional to the similarity measure. The similarity values are scaled according to the maximum value, which is given the maximum brightness.

3.4 Other timbre segmentation models

The self similarity method proposed by Foote [Foote 1999] is retrieving merely structural information within a given sequence, based on the distance of each and every frame to all other frames. It operates with high temporal resolution i.e the results are displayed at the FFT frame level. While comparing every single frame or the average of several consecutive

frames, the aim is to generate a detailed representation of local and global musical structure.

It would be very interesting however, to be able to analyze a given song on a larger scale and to segment it into larger sections of homogeneous timbre (e.g., extracting the guitar solo in the middle). This issue is addressed in a later publication by Foote [Foote *et al.* 2003], as well as in [Aucouturier *et al.* 2005], where a method for representing the timbre of whole sequence of music as a GMM of MFCC features is presented. The model is extended with a temporal dimension, by breaking the GMM apart in order to model a sequence with segments of possibly very different Gaussian distributions. A polyphonic musical sequence can be represented by several different (local) timbre models or feature-cluster centers, preserving the information about particular locations of timbral quality changes. For instance, the song "Let it be" from the Beatles may be represented by one Gaussian for the texture "piano+voice" and another Gaussian for the "electric guitar solo" in the middle of the song.

Once the complete – GMM based – timbre model of a given sequence (song) is extracted the segmentation is simply achieved by labeling each frame with the component it is most probably generated by. The E-M algorithm used to fit several Gaussian components to the trajectory of MFCCs is computed as follows:

In his 2005 paper entitled "The way it sounds" Aucouturier explains the process of generating the segment models as follows. In the E-step, frames are labeled with their most probable segment (section) model – starting with a random distribution. In the M-step, in turn, the frames in each model are used to generate the segment models. After a certain number of iterations, all homogenous sounding segments are linked together – representing the different timbre models. The now generated timbre model can be used to decode further (unlabeled) data, i.e., to label each of its frames with its most probable component index.

Figure 23 shows the results of such an analysis on 20 s of music, a 1960's French song by Bourvil, modeled by a 3-state timbre model. Its instrumentation consists of a male singer accompanied by an accordion, and a discrete rhythmic section. The segmentation is very accurate: the background accompaniment comes at the end of every vocal phrase, sometimes even between the sung words. The accordion introduction appears very clearly, as well as the periodicities of the verse.

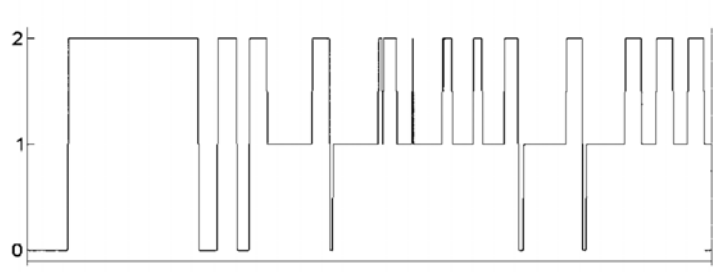


Figure 23. Segmentation of Bourvil's song. State 0 is {silence}, state 1 is {voice + accordion + accompaniment}, and state 2 is {accordion + accompaniment}, plotted against time. [Aucouturier *et al.* 2005]

Each of the three clusters represents a combination of instruments. One cluster has all frames with voice; the other has all frames of accordion, etc. On the other hand, we can also generate a model with more clusters, however, it is again important to consider and perhaps to predict how complex or diverse the timbral composition in an analyzed sequence really is, and thus how many clusters the model would needs to expose this diversity. For instance, a Vocal-Jazz song with a 50-component timbre model, would generate 17 clusters accounting for voice frames, 11 clusters for piano frames, ten clusters for percussive frames, three for double bass frames, and nine are clusters containing mixed frames. If we are only interested in the attributes describing instrumental mixtures, a 50 component analysis of a small Jazz combo, might not necessarily yield better results than a 5 component model.

3.5 MPEG-7 Higher-Level Descriptors

It is interesting to see that the MPEG-7 standard [ISO 2001] formalizes a very similar set of transforms that are closely related to the concept of Cepstrum, in order to describe the spectral shape and timbral qualities. By defining low level descriptors like the *AudioSpectrumEnvelopeD,* which form the basis for the *AudioSpectrumBasisD* and *AudioSpectrumProjectionD* descriptor, MPEG-7 provides an autonomous framework for timbre similarity classification and what is more, it offers a standardized tool for temporal

organization of timbral textures defined within the higher-level descriptor: *SoundModelStatePathD*.

The following chapters describe the general MPEG-7 (audio) framework and it's potential.

3.5.1 A brief introduction to MPEG-7 audio

MPEG-7's most important goal is to provide a set of methods and tools for the different classes of multimedia content description. It defines a series of elements that can be used to describe content, but it does not specify all the algorithms required to compute values for those descriptions. The building blocks of MPEG-7 description are descriptors, description schemes (complex structures made of aggregations of descriptors), and the Description Definition Language (DDL), which defines the syntax that an MPEG-7 compliant description has to follow. The DDL makes hence possible the creation of non-standard, but compatible, additional descriptors and description schemes. This is an important mechanism because different needs will call for different kinds of structures, and for different instantiations of them. In the audio section, music-specific descriptors for melody, rhythm or timbre can be found.

A low-level descriptor [Allamanche *et al.* 2001] can be computed from the time-series data in a direct or derived way (i.e. after signal transformations like Fourier or Wavelet transforms, after statistical processing like averaging, after value quantization like assignment of a discrete note name for a given series of pitch values, etc.). Most of low-level descriptors make little sense to the majority of users but, on the other hand, their exploitation by computing systems is usually easy. They can be also referred to as *"signal-centered descriptors"*

In addition to traditional timbre analysis methods that apply only to isolated musical instrument notes (as presented in the first part of the thesis), the MPEG-7 standardization is also designed to represent noise textures, environmental sounds, music recordings, melodic sequences, vocal utterances and sounds containing mixtures of sources.

Higher-level descriptors require an induction operation that goes from available data towards an inferred generalization about them. These descriptors usually pave the way for

labeling contents, as for example a Hidden Markov Model that makes it possible to segment a song according to timbre similarities. Machine learning and statistical modeling make higher-level descriptors possible, but in order to take advantage of those techniques and grant the validity of the models, large sets of observations need to be gathered. Those descriptors are also sometimes referred to as *"object-centered descriptors."*

In the following chapters, two components of MPEG-7 audio will be discussed. The first is the use of decorrelated spectral features for low-dimensional sound representation. The second component is the estimation of general sound similarity using finite-state probabilistic inference models [Casey 2001]

3.5.2 Decorrelated spectral features

Spectrum-based features are often considered as an elementary requisite for audio applications, but it is widely known that direct spectrum features are generally incompatible with classification applications due to their high dimensionality and their inconsistency. Each spectrum slice is an *n*-dimensional vector, with *n* being the number of spectral channels; therefore, typical values of a linearly-spaced spectrum are between 64 and 1024 dimensions.

The MPEG-7 standard defines the spectral envelope according to a logarithmic frequency scale – similar to the already presented Mel scale. The spectral envelope low level descriptor is called *AudioSpectrumEnvelopeD* and consists of one coefficient representing the power between 0 Hz and a "low edge" boundary at 62.5 Hz; a series of coefficients representing logarithmically spaced frequency channels in non-overlapping ¼-octave bands spanning between 62.5 Hz and 8 kHz; and finally and single coefficient representing the power above the "high edge" boundary of 8kHz. The output of the logarithmic frequency range is the weighted sum of the power spectrum in each logarithmic sub-band.

The logarithmic form of the spectral representation already yields a dimensionality reduction, i.e. a sub-summation of spectral information in ¼ octave bands. This action is justified by taking into account the logarithmic resolution of the human auditory perception system. From the point of view of computational efficiency however, probability classifiers

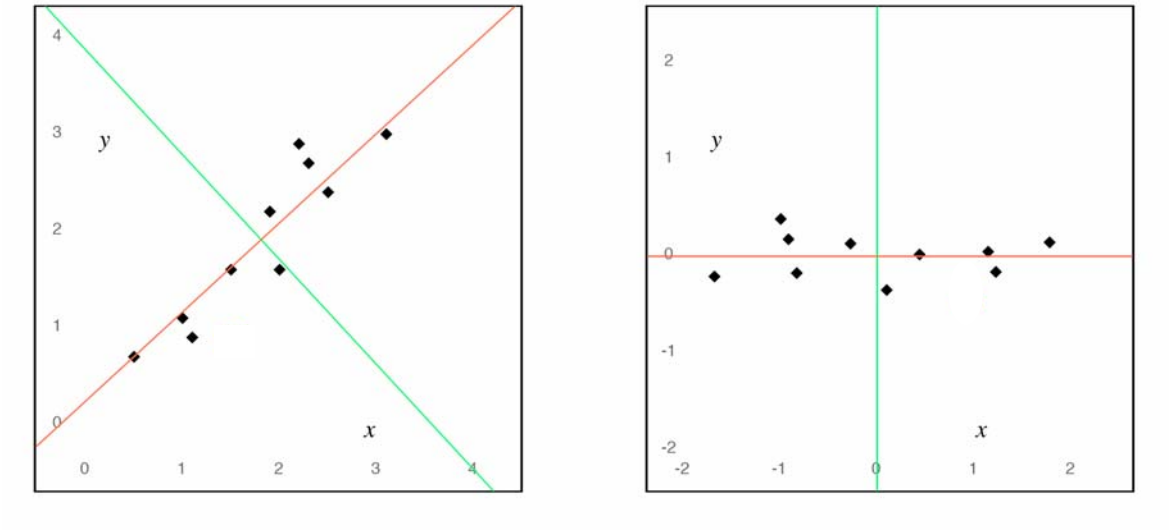
require even lower-dimensional data representations. What is required is a representation that makes a compromise between dimensionality reduction and information loss. A well-known technique for reducing the dimensionality of data whilst retaining maximum information is to use data-derived basis functions, such as computed by *principal component analysis* (PCA), *singular value decomposition* (SVD) or *Karhunen Löve Transform* (KLT).

3.5.3 Principal component analysis

Principal Component Analysis (PCA) [Smith 2002] is a widely adopted method for exploring multidimensional data sets. With PCA it is possible to identify unknown trends and cross-dimensional correlations in a collection of feature vectors. It finds a new coordinate system for the data set by reorganizing the data according to its decreasing covariance. Covariance tells whether changes in any two variables move together (are correlated). The data dimensions with the largest covariance are then projected to the first axis of the rotated coordinate system, and are now described by a single dimension – called, the first principal component. The data with the second greatest variance is projected to the second axis, and so on.

For the sake of convenience, a two dimensional sample data set will be examined distributing as in figure 24(a). Optimal lines can be found (indicated by the red and green lines) along the directions the sample points are distributed. PCA finds these optimal lines and rotates the space in such a way that the main distributing dimensions become the new axes (figure 24b).

After PCA executes a rotation of the data a dimension reduction can easily be performed by rebuilding the data set, using only the first (strongest) principal components. In this way, a reduction of the data dimensionality is achieved with minimum data loss. It can easily be observed, that the second dimension (in the given example) does not carry a lot of information, thus we can simply discard it.



(a) The original data points **(b) The data points rotated by PCA**

Figure 24. Basics of PCA [Shiraishi 2006]

Implementing PCA is the equivalent of applying Singular Value Decomposition (SVD) on the covariance matrix of a data set, where it extracts its eigenvalues and eigenvectors. The eigenvectors represent the lines as we see in figure 24, namely the principal components. The variances along these components are contained in the eigenvalues.

Computation of SVD:

If for example, X is an $n \times k$ array of n observations – stored in n rows – occurring in k dimensions (columns), the covariance matrix $\sum$ is defined as:

$$\sum = \frac{A^T A}{(n-1)} \tag{3.10}$$

where A is a new data set with zero mean, obtained by subtraction of μ, (the k-dimensional mean of the n observations), from each individual observation (e.g. feature vector).

Next, the eigenvectors and eigenvalues are derived by Singular Value Decomposition (SVD), where the covariance matrix $\sum$ is decomposed as:

$$\sum = U \cdot S \cdot V^T \quad \textbf{(3.12)}$$

U and V are unitary matrices. *S* is diagonal, and its elements are ordered in decreasing values. The eigenvectors od the covariance matrix $\sum$ are stored in the colums of V and the eigenvalues are held in S. The eigenvector with the highest eigenvalue is the first principal component of the data set and the one with the second highest is the second principal component and so on. The rotation of the data to the new coordinate system in the lower-dimensional space, is simply done by matrix multiplication.

$$Y = X \cdot V^T{}_L \quad \textbf{(3.13)}$$

where *Y* is the rotated, and X, the original data. $V^T{}_L$ is a feature vector consisting of the first *L* principal components.

3.5.4 AudioSpectrumBasisD / AudioSpectrumProjectionD

With Principal Component Analysis it is possible to reconstruct a spectrogram by using a set of decorrelated frequency basis functions. Fewer functions are required to reconstruct a given spectrogram than the total number of frequency channels, hence the possibility for dimensionality reduction. Figure 25 shows a spectrum of five seconds of pop music reconstructed using only four basis functions. The functions on the left of the figure are the frequency basis functions, those above the figure are the reduced dimension features (*AudioSpectrumProjectionD*) used for classification. In this case, 70% of the original 32-dimensional data is represented by only the 4-dimensional functions.

The *AudioSpectrumBasisD* descriptor contains basis functions that are used to project high-dimensional spectrum descriptions into a low-dimensional representation contained by the *AudioSpectrumProjectionD* descriptor. The reduced bases consist of decorrelated features of the spectrum with the important information described much more efficiently than the direct spectrum representation. This reduced representation is well suited for use with probability model classifiers that typically perform best when the input features consist of fewer than 10 dimensions.

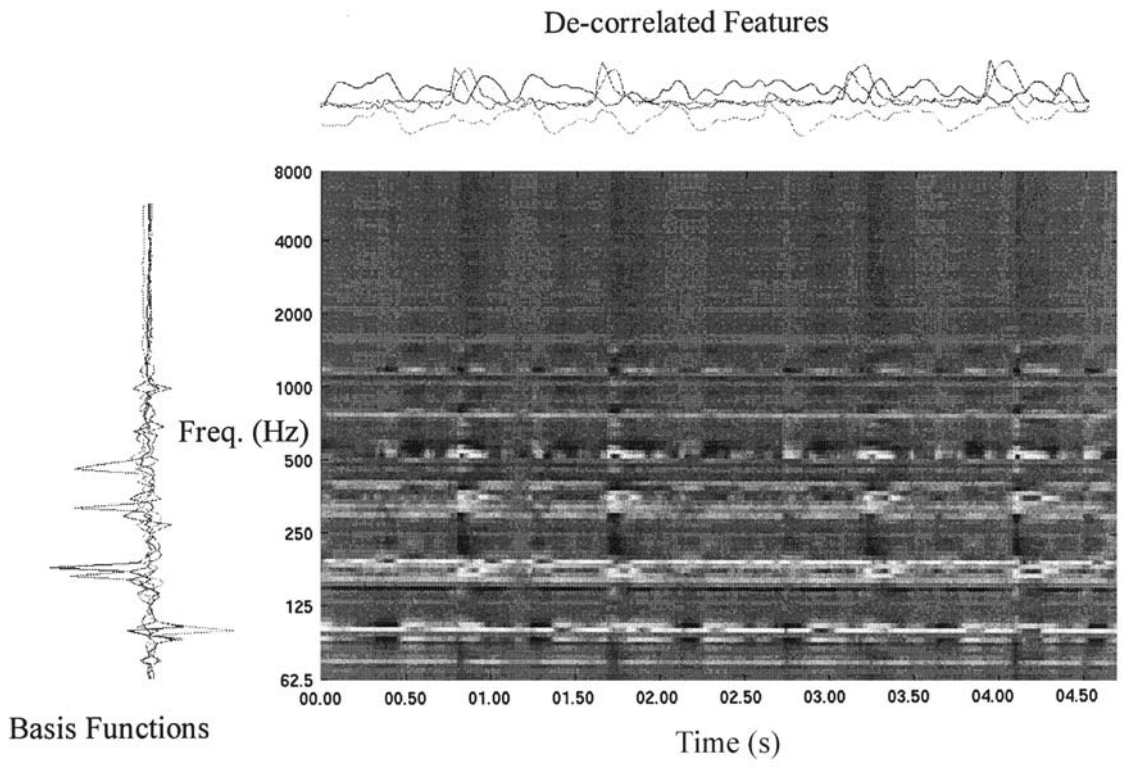


Figure 25. Reconstruction of a spectrogram with data-derived basis functions; in this case the functions were derived from the spectrum using singular value decomposition (SVD). [Casey 2001]

3.5.4.1 Spectral Basis function extraction method

The extraction method for *AudioSpectrumBasisD* and *AudioSpectrumProjectionD* is described in detail within the MPEG-7 standard [ISO 2001]. Within each step however, there is opportunity for alternate implementations. The following chapters outline the standardized extraction method for basis functions as described in [Casey 2001]:

(1) *Power spectrum*:

First, an *AudioSpectrumEnvelopeD* descriptor is instantiated using the extraction method defined in *AudioSpectrumEnvelopeD* descriptor. The resulting data will be a series of spectral vectors representing the selected signal frames.

(2) *Log-scale norming*:

For each spectral vector, **x**, in *AudioSpectrumEnvelopeD*, the power spectrum needs to be converted to a decibel scale:

$$z = 10\log_{10}(x) \tag{3.14}$$

Next, the *L2-norm* of the vector elements is computed:

$$r = \sqrt{\sum_{k=1}^{N} z_k^2} \tag{3.15}$$

The new unit-norm spectral vector is calculated by:

$$\tilde{x} = \frac{z}{r} \tag{3.16}$$

(3) *Observation matrix*

Each observation vector $\tilde{x}$ should then be placed *row-wise* into a matrix. The size of the resulting matrix is $M \times N$, where M is the number of time frames and N is the number of frequency bins. The matrix will have the following structure:

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_M^T \end{bmatrix} \tag{3.17}$$

(4) *Basis extraction*

The spectral basis is extracted using singular value decomposition (SVD). The SVD algorithm can be found implemented as a built-in function in many software packages. In Matlab for example it can be done using the command [U,S,V] = SVD(X). The SVD factors the matrix from step (3) in the following way:

$$\widetilde{X} = USV^T \tag{3.18}$$

where $\widetilde{X}$ is factored into the matrix product of three matrices; the row basis **U**, the diagonal singular value matrix **S**, and the transposed column basis functions **V**. The basis should then be reduced by retaining only the first *K* basis functions, i.e. the first *K* columns of ***V***:

$$V_k = \begin{bmatrix} v_1 & v_2 & \dots & v_k \end{bmatrix} \tag{3.19}$$

The SVD basis functions are stored in the columns of a matrix within the *AudioSpectrumBasisD* descriptor.

3.5.4.2 Spectrum Projection extraction

The *AudioSpectrumProjectionD* is the compliment to the *AudioSpectrumBasisD* and is used to represent low-dimensional features of a spectrum after projection against a reduced rank basis. These two types are always used together. The low-dimensional features of the *AudioSpectrumProjectionD* consist of a vector series, one vector for each frame, of the normalized input spectrogram $\widetilde{\mathbf{x}}_t$. Each spectral frame from steps (1)-(3) above yields a corresponding projected vector: $\widetilde{\mathbf{Y}}_k$.

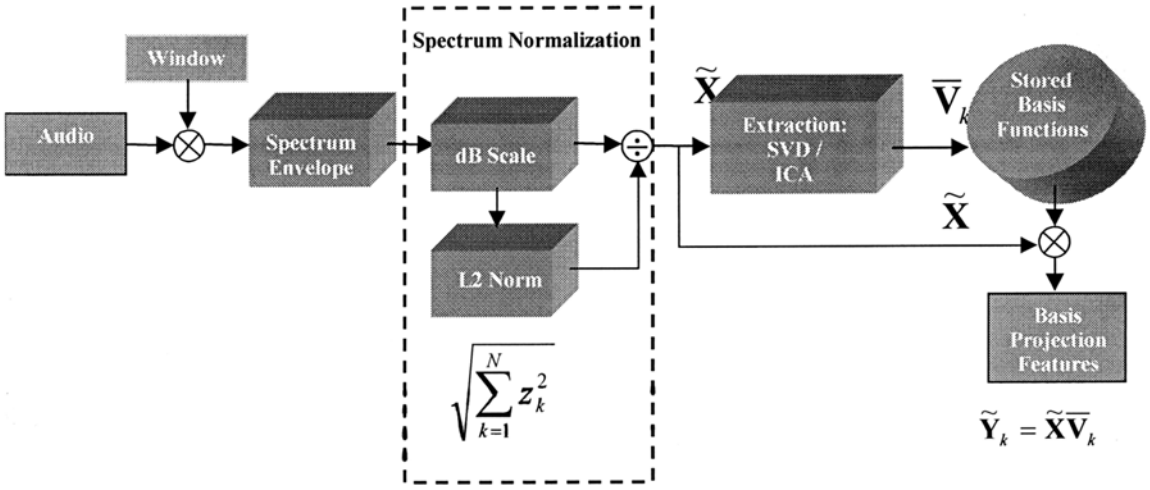


Figure 26. Extraction method for *AudioSpectrumBasisD* and *AudioSpectrumProjectionD*. [Casey 2001]

An approximation of the spectrogram can now be constructed the with the reduced dimension features. The individual vector pairs corresponding to the K_{th} vector in *AudioSpectrumBasisD* and *AudioSpectrumProjectionD*, are deployed in the reconstruction equation:

$$X_k = \widetilde{Y}_k \overline{V}_k^T \tag{3.20}$$

3.5.5 Automatic sound classification

The AudioSpectrumBasisD and AudioSpectrumProjectionD are used for automatic classification of audio segments using probabilistic models. In the following application, basis functions are computed for the set of training examples and are stored along with a probabilistic model of the training sounds. The method involves training statistical models to learn to recognize the classes of sound defined in a taxonomy.

3.5.5.1 Finite state models

Sound phenomena are dynamic. The spectral features vary in time and it is this variation that gives a sound its characteristic fingerprint for recognition. MPEG-7 sound recognition models partition a sound class into a finite number of states based on the spectral features; individual sounds are described by their trajectories through this state space. Each state is modeled by a continuous probability distribution such as a Gaussian. The dynamic behavior of a sound class through the state space is modeled by a $k \times k$ transition matrix that describes the probability of transition to each of the k states in a model, given a current state. For a transition matrix, **T**, the i_{th} row and j_{th} column entry is the probability of transitioning to state j at time t given state i at time $t-1$.An initial state distribution, which is a $k \times 1$ vector of probabilities, is also required for a finite-state model. The k_{th} element in the vector is the probability of being in state k in the first observation frame.

3.5.5.2 Continuous Hidden Markov Models

A continuous Hidden Markov model is a finite state model with Gaussian distributions approximating each state's probability distribution. The states are *hidden* since the states are not given along with the data. Rather, the observable data must be used to deduce the hidden states. The states are clusters in the feature space of the sound data, namely, the *SpectrumBasisProjectionD* audio descriptor discussed earlier. Each row of the projected feature matrix, defined above, is a point in an *n*-dimensional vector space. The cloud of points is divided into multiple states (Gaussian clusters) defined by multidimensional mean values and a covariance matrix. Figure 27 shows a representation of four Gaussian-distributed states (vector point clouds) in two dimensions.

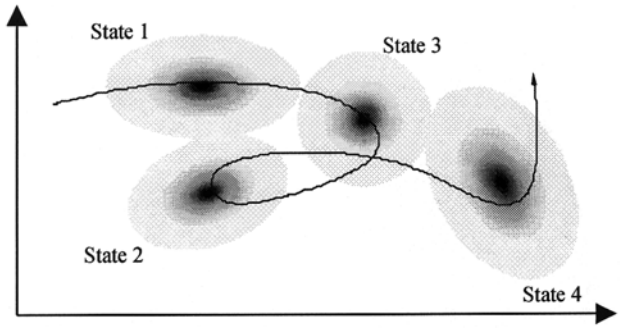


Figure 27. Four estimated Gaussian states depicted in a two dimensional vector space. Darker regions have higher probabilities. Sounds are represented as trajectories in such a vector space, the states are chosen to maximize the probability of the model given the observable evidence, i.e. the training data. The line shows a possible trajectory of a sound vector through the space. [Casey 2001]

3.5.5.3 Training the Hidden Markov Models

A statistical model is trained on the spectrum projections for a sound sequence. During training, the parameters for a sound model are estimated from the feature vectors of the training set. The HMM can then be used as a classifier since it represents the temporal evolution of important features extracted from audio data. The forward-backward, i.e. Baum-Welch algorithm [Rabiner *et al.* 1993] is used for the training of the HMMs. The procedure starts with random initial values for all of the parameters and optimizes them by an iterative

re-estimation. Each iteration runs through the entire set of training data in a process that is repeated until the model converges to satisfactory values.

3.5.5.4 Sound Model State Path

MPEG-7 defines the *SoundModelStatePath* descriptor, which contains the dynamic state path of a sound through a HMM model. Sounds are indexed by segmentation into model states or by sampling of the state path at regular intervals. Figure 28 shows a spectrogram of a dog bark sound with the state path through the 'DogBark' HMM shown below. The state path is an important method of description since it describes the evolution of a sound with respect to physical states. The state path shown in the figure indicates physical states for the dog bark; there are clearly delimited onset, sustain and termination/silent states. This is true of most sound classes; the individual states within the class can be inspected via the state path representation and a useful semantic interpretation can often be inferred.

The character of the sound as well as its temporal evolution can thus be represented by a set of basis functions and the temporal evolution of the basis projections.

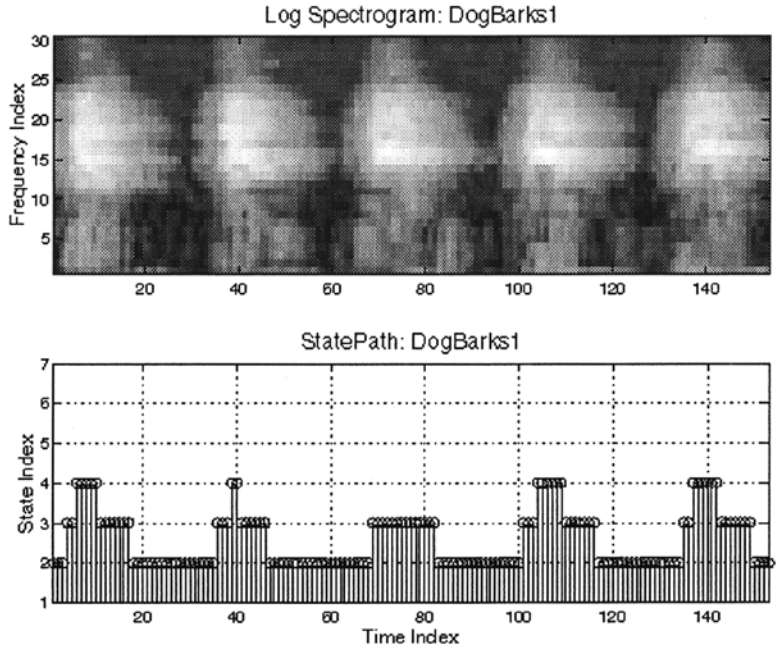


Figure 28. Dog bark spectrogram and the state path through the dog bark continuous hidden Markov model. [Casey 2001]

3.5.6 Comparison of the *MFCC*- and the *spectral basis / projection* concept

Two methods for the description and analysis of the temporal evolution of polyphonic timbre have now been introduced and it is interesting to see that both approaches are based on a similar concept.

One of the characteristics of the MFCC method is that it transforms the linear frequency scale to the Mel scale. As already discussed in chapter 3.1, the Mel scale is a perceptual scale and is derived from the nonlinear frequency resolution of the human auditory system i.e. on the basis of perceived relative pitch proportions. A similar frequency warping is deployed in MPEG-7, within the above described and standardized AudioSpectrumEnvelope descriptor, which introduces ¼ octave, logarithmically spaced coefficients.

The main difference however is in the transformation that follows: Computing MFCC, the nonlinear spectral representation is multiplied with a series of cosine functions (DCT) in order to achieve a compact representation consisting of decorrelated principal components. Similar than in the Fourier transform the multiplication (i.e., a modulation) of a signal with a periodic function, yields the exposure of its periodic events of the same frequency. The purpose of the DCT is to deconstruct the spectral representation into individual components shaped like the modulating cosinus functions. Thereby, the first, low frequent components would already exhibit the general shape of the spectrum, so all the higher components, would contribute a further, though proportionally less informative detail of the spectrum.

On the other hand, MPEG-7 approaches the same task from a different point of view. Through PCA, it first has to compute, the decorrelated features (the basis functions), which are then used to (de)-modulate the spectrum, yielding the basis projections. In [Logan 2000] it was found out that PCA derived basis functions calculated for either speech or music signals, were very similar to cosine functions. Therefore, also the concept of basis projections would correspond to the above described result of the DCT.

Figure 29. shows a sequence of 15 eigenvalues and the corresponding eigenvectors (basis functions), obtained by PCA on approx. 3 hours of recorded speech. In figure 30, shows the first 15 cosine functions used to multiply a signal in the DCT. Figure 31. shows a

sequence of 15 eigenvalues and the corresponding eigenvectors, calculated by PCA on approx. 300 minutes of Beatles songs.

The order of PCA basis functions is slightly different than their corresponding cosine functions. The cosine functions are ordered with respect to their increasing frequency argument, whereas the PCA functions are ordered by their corresponding eigenvalues, which does not necessarily give exactly the same order. In particular, it can be observed that the 3-th and 4-th eigenvector of the speech sequence appear to be reversed, compared to the corresponding cosine functions. However, their eigenvalues have almost the same size, which may explain the situation.

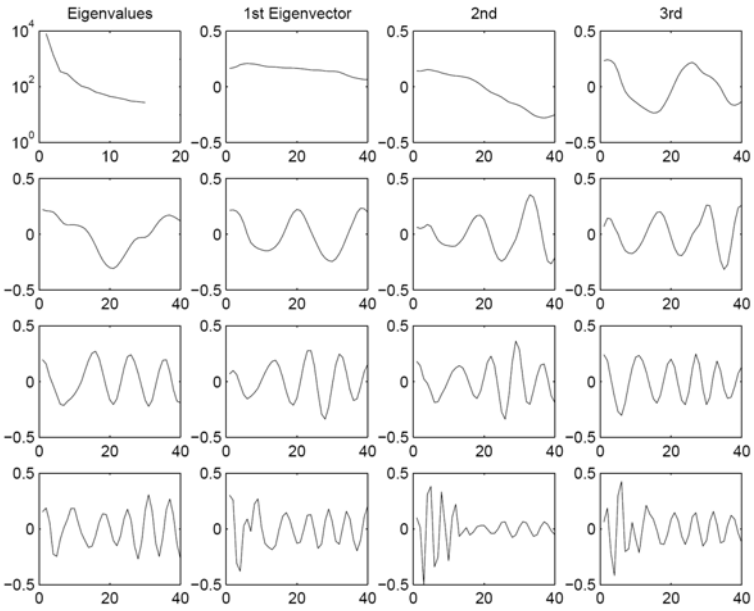


Figure 29. The first 15 eigenvalues and the corresponding eigenvectors for the covariance matrix of Mel log spectral vectors for 3 hours of clean speech. [Logan 2000]

It is interesting to observe that the eigenvector- derived basis functions are cosinus like, in particular, for the more important first few functions, with the largest corresponding eigenvalues.

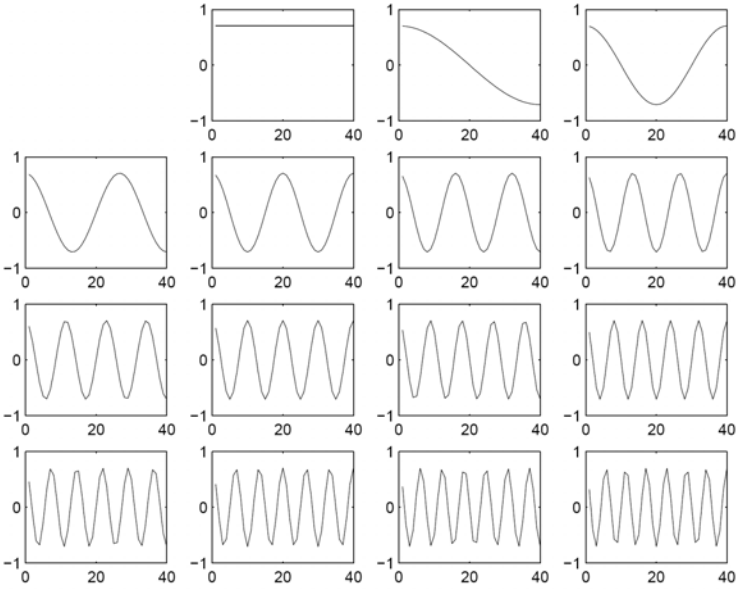


Figure 30. The first 15 cosine basis functions [Logan 2000]

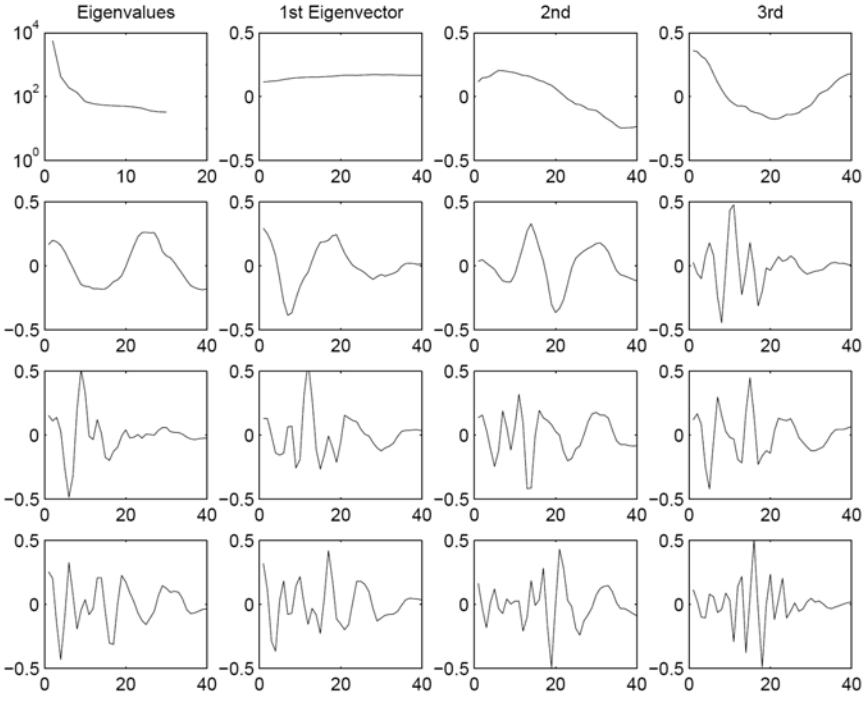


Figure 31. The first 15 eigenvalues and the corresponding eigenvectors for the covariance matrix of Mel log spectral vectors for 300 minutes of Beatles songs. [Logan 2000]

To conclude these observations one could say that the PCA method applied in MPEG-7 can give a more accurate description of the analyzed signal within its first principal spectrum components, compared to the DCT. If, for instance several different sounds would need to be compared and classified and the basis functions were derived for each individual sound in a corpus, the first basis functions could perhaps be used as an additional classifier of timbral characteristics. Opposed to the cosinus functions deployed at the DCT, they would exhibit an individual form for each analyzed sound. Here, both the first basis functions and the progression of the basis projections, provide a general timbre description, as well as they define the evolution of timbre over time. In terms of computational efficiency however the DCT would outperform the MPEG-7 methods, since it does not need to perform the PCA in order to calculate its basis functions, rather, it assumes they are Cosinus-like.

4. Modeling short polyphonic signal bursts

In the previous chapters a few basic concepts and approaches to timbre analysis were introduced, which were organized in two major research fields, namely polyphonic and monophonic sound analysis. The main aim of this thesis however, is to pick out and combine some concepts from different MIR research fields discussed in previous chapters and to study the claim of a possible temporal character of timbre. It was tried to find a model for the temporal evolution of MFCC features while reducing the amount of information in the extracted features as much as possible. In order to do so, several experiments were undertaken in which plausible quantization limits for each feature dimension were estimated. In terms of data reduction, three dimensions were considered, namely: the number of MFCCs, along with the temporal and amplitude resolution of their trajectories, needed for a sufficient description of a sounds timbre.

The sound material to be modeled by this approach were short signal bursts like single beats and notes with a polyphonic timbral character, which first of all had to be isolated from a larger context of polyphonic music.

4.1 Segmentation

A note onset detection algorithm that is needed to perform the task of isolating single beats / notes is itself an important factor in the later comparison, recognition and verification process. By undertaking a prior segmentation based on note onsets, a first step towards extraction of musically meaningful information is already done. The polyphonic timbre descriptions in this work are partly based on the interpretation of corpus based query results, which were obtained through direct comparison of the selected timbral features. Thus, by aligning the start-points of all the analyzed sound snippets in the given corpus according to the found onset a first criterion for the later comparison / verification process is defined. The comparison algorithm does not blindly evaluate the distances amongst all available frames or their statistical average in the defined segments. Accordingly, timbre

similarity judgments are based the cross-comparison of selected frames with respect to their position in the segments, e.g. the attack segment and the particular sequence of consecutive segments.

The program that was used to find the note onsets was the "Sonic Visualizer"[3], which is being developed at the Centre for Digital Music, at Queen Mary University of London. Initially, only one song – "Careless Whisper" from George Michael – was taken for analysis. For a pop song it exposed a rather diverse instrumentation as well as a diverse dynamic range within itself. At this point it is important to note, that when attempting to model an isolated sound segment, it is crucial to confine the selection of sound material perhaps to a single musical genre or style. Already when choosing an onset detection algorithm, along with its parameter constellations, a specific onset behavior is tracked. The onset behavior of musical sequences may be significantly different when comparing for example mainstream pop music with classical vocal-polyphonic pieces. In order to find beats and note onsets in all kinds of musical styles, a variety of onset detection systems were developed.

Onset detection in the time domain focuses on the change in the amplitude or energy of the signal or on the signal change in relation to the signal level. In the frequency domain, some authors analyze the change between the energy of successive short time spectra; others correlate short-time power spectra or evaluate changes in the complex frequency domain. Approaches using dyadic wavelet decomposition and Transient Modeling Synthesis have also been used. Onset events can also be found through a probabilistic approach where the conformity of the audio signal to a signal model is evaluated. The different approaches are described in detail in [Mikula 2008].

It is not in the scope of this work to explore the pros, cons and application areas of different onset detection approaches. How the material used for further analysis was segmented was not important, thus an appropriate onset detection algorithm was selected experimentally, by testing several different algorithms offered in the "Sonic Visualizer" application.

[3] http://www.sonicvisualiser.org/

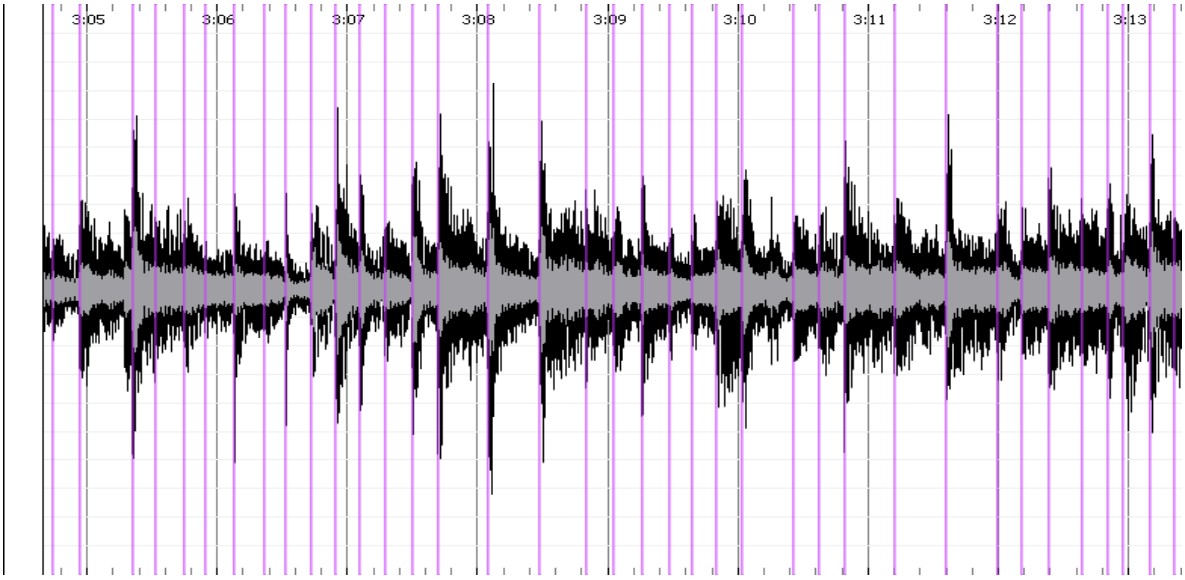


Figure 32. found onsets on a 8 second audio excerpt from the middle of the song "Careless whisper"

Through empiric verification – in form of careful listening and observations of the proposed onsets – an algorithm based on Spectral difference was determined as the best choice for segmenting pop music like in our case the "Careless whisper". Like stated above, it was not crucial to explore the reasons why a particular segmentation algorithm would be a best choice for a particular musical style or genre. A segmentation result showing an arbitrarily selected 8 second excerpt (from minute 3.05 to 3.17 – e.g. from the middle of the song) is displayed in figure 32. It can easily be observed that the peaks in the audio file in general correspond to the found onsets, which are marked by pink stripes and one can conclude that the results are actually representing musically meaningful segments. This was exactly the targeted material that is supposed to represent the subject of research in this thesis.

At a sampling rate of 44100 Hz, the average duration of such a segment turned out to be around 10.000 Samples or ¼ of a second (also visible in figure 32), thus, at a manually estimated average tempo of around 70 to 75 BPM, this duration would roughly represent a quarter note, which again is a reasonable result.

In order to move on to the main task of finding a timbre model, it was decided to accept this segmentation method without further, time consuming evaluations and manually segmented ground truth references.

4.2 Timbral similarity and MFCC trajectory deviation of two arbitrary samples

In the following chapters three different approaches will be documented, which continuously led to the concluding results about the sequential dependency and the required resolutions of timbral features for the purpose of classification and recognition. Common to all approaches is an algorithm for MFCC calculation, which was taken from the "RASTA/PLP/MFCC - feature calculation and inversion" toolkit[4] developed at LabROSA – the Laboratory for the Recognition and Organization of Speech and Audio - at Columbia University, New York, USA. All audio snippets generated by the above mentioned segmentation method were transformed / analyzed with the "melfcc.m" routine. The following parameters were set:

PARAMETER	VALUE
sample rate	44100 Hz
frame length	2048 (samples)
hop size	512 (samples)
number of Mel bands	40
Center frequency of lowest Mel filter	10 Hz
Center frequency of highest Mel filter	16000 Hz

Table 4.1 parameter / value pairs used for the calculation of MFCCs

4.2.1 Exploring the timbral difference by an analysis-resynthesis model

A first attempt to verify the hypothesis that timbral quality of sound may – to a great extent – be imposed by a particular temporal sequence of features was undertaken by deploying the analysis–resynthesis model shown in figure 35. The purpose of this approach was to find out, weather it was possible to employ a filter-bank with a time varying output gain section in order to alter the timbre of sample A, in a way, that it would approximate, or perhaps even match the timbre of sample B.

[4] http://labrosa.ee.columbia.edu/

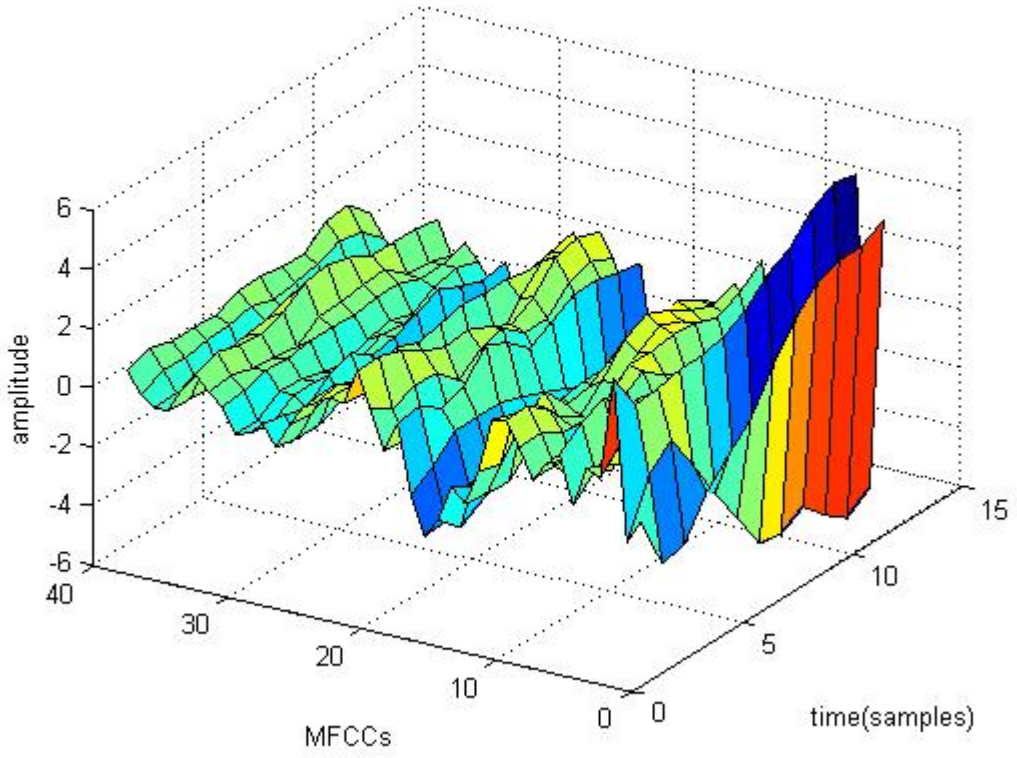


Figure 33. an MFCC representation of an arbitrary short polyphonic sound segment (without the first coefficient) – showing the evolution of 39 MFCCs over fourteen, 2048 sample long windows[5]

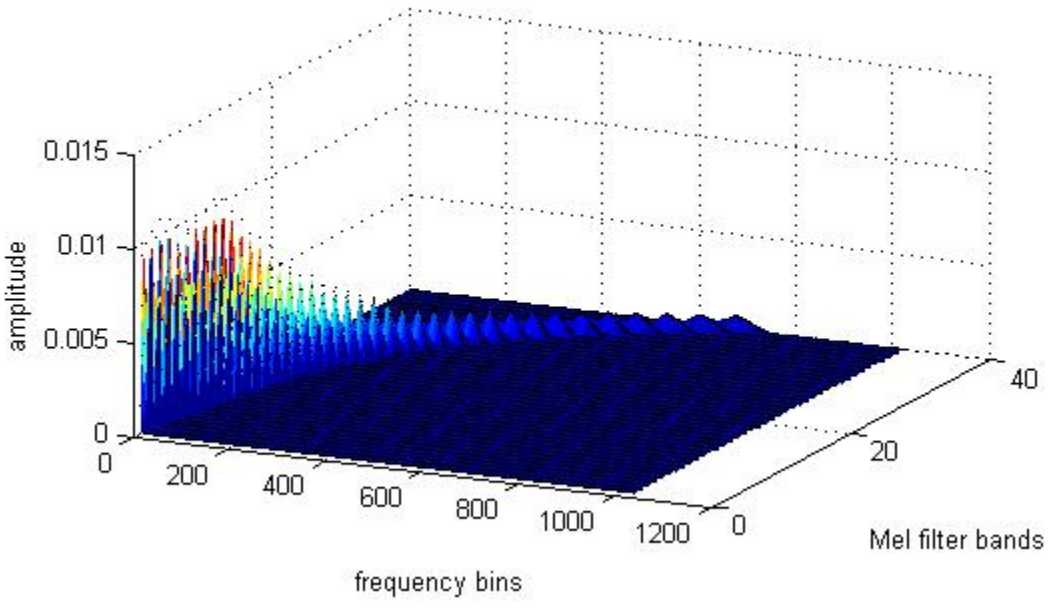
First, two arbitrary sounds x[t] and y[t] were selected out of the corpus of 1472 sound snippets, previously generated by the note onset detection - segmentation algorithm, which was executed on the song "Careless whisper". Both original sound segments were first cut to equal length and then played back one after another in order to get an audible impression and to observe the two different timbres. Next, each sample was transformed by an FFT algorithm using the above mentioned parameters (frame-length: 2048, hop-size: 512). Further, the absolute value of the short time energy spectrum was summed up and weighted by 40 Mel-spaced filters illustrated in figure 34. This operation, which was executed in the frequency domain, by multiplication of each signal frame with the filter impulse responses, yields a new - 40 chanel - signal representation, thus, compressing the 1024 bin information

[5] the labelling of the „time(samples)" axis in this and all following plots showing the evolution of MFCCs refers to the „frame length" value. Thus, one unit in the displayed MFCC domain represents the amount of time domain samples defined by „frame length".

from the spectral representation of the signal. The filter responses were calculated for a frequency range between 10 and 16000 Hz.

In the following step a logarithm of the summed and weighted amplitudes was calculated and the results carried on to the next section where this set of 40 parallel signals was modulated by a set of cosinus functions, within the Discrete Cosinus Transformation. The output of this modulation finally generates the Mel Frequency Cepstrum Coefficients. Given the two signals x[t] and y[t] in their cepstral representation ($x_c[t]$ and $y_c[t]$) a subtraction "$x_c[t] - y_c[t]$" was performed in order to estimate their cepstral deviation: $e_c[t]$.

Parallel to the now described analysis and subtraction process, a re-synthesis engine was prepared, that would make use of the cepstral deviation $e_c[t]$ in order to generate time-variant filter parameters modulating the time domain signal y[t] in such a way, that it would approximate the timbre of signal x[t]. The process began by generating ideal – 1024 samples large - frequency domain Mel-filter shapes shown in figure 34. The filter-banks are already implemented within the RASTA/PLP/MFCC package, in the "melfcc.m" script. Those were then used to generate the corresponding time domain filter impulse responses, which were obtained via an Inverse Fourier Transformation of the frequency domain filter representations.



.

Figure 34. frequency domain representation of the Mel filter-bank

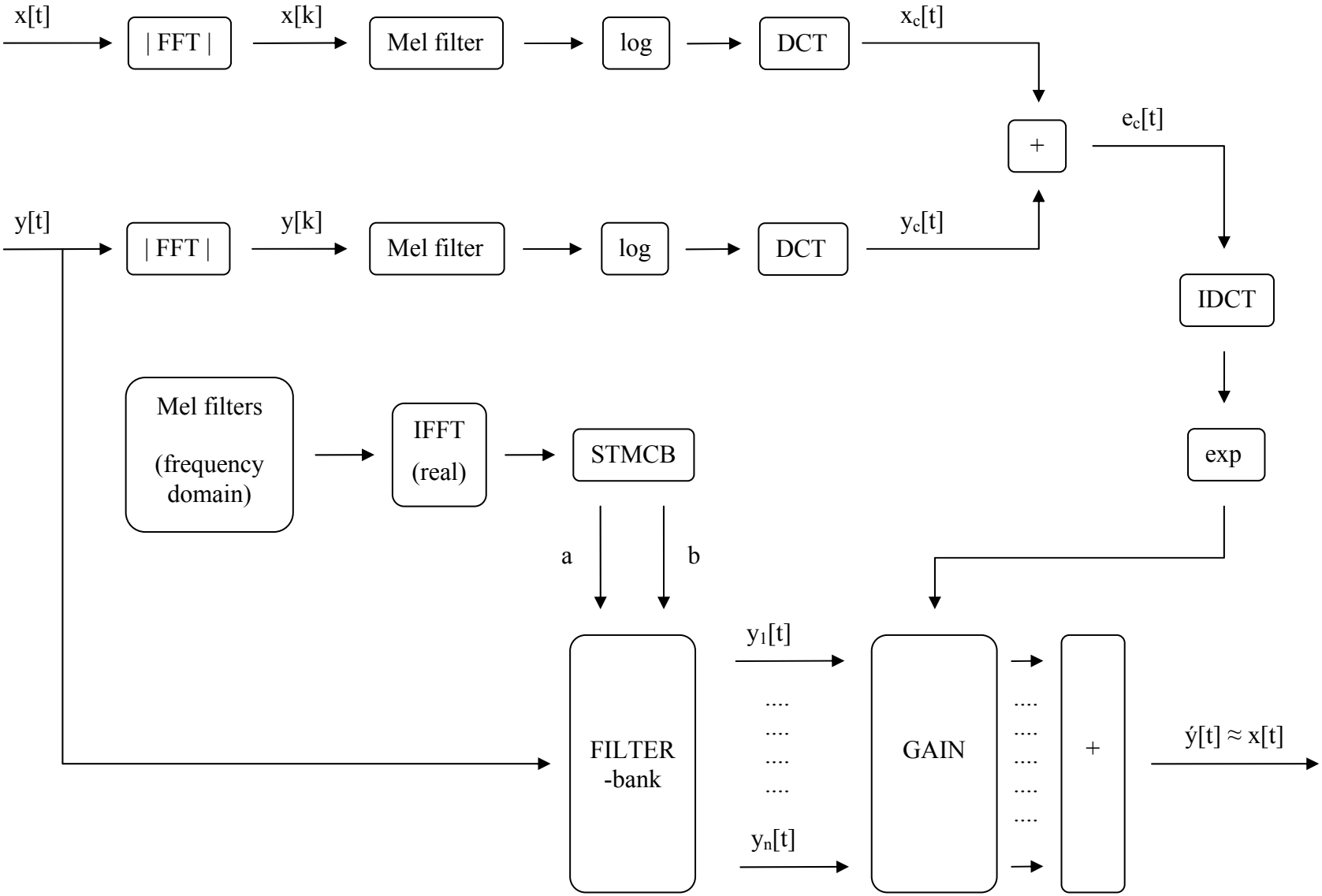


Figure 35. block diagram of the analysis-resynthesis algorithm

Next, the built-in Matlab function “stmcb.m” was deployed to perform a Steiglitz-McBride iteration, which, given a time domain impulse response, estimates the coefficients B(z)/A(z) of a linear system. The number of pole/zeros was set to 4. A filter-bank of 40 parallel filters, each defined by the 4 A and B coefficients was now ready to split the time domain signal y[t] into 40 band-passed signals.

The Steiglitz-Mcbride iteration was chosen with the aim of reducing the computational demand required for the filtering process. Given a time domain filter impulse response, a Finite Impulse Response (FIR) filtering operation may as well be performed by convolving the input signal with the given filters impulse response. However, since the impulse responses were 1024 samples long, performing 40 parallel convolution processes, would turn out to be computationally rather inefficient. With the Steiglitz-Mcbride iteration, it is possible to estimate an Infinite Impulse Response (IIR) filter system, with merely 4 “B” - (the feed-forward) and 4 “A” - (the feed-backward) filter coefficients. These would represent an adequate approximation of the FIR system, while drastically reducing the computation capacity required for the filtering process.

After performing an IFFT and taking the exponential of the signal e[c], the resulting data was used to drive a 40 channel – time varying – amplification stage, taking the 40 band-passed signals ($y_1[t]$ to $y_{40[}t]$) as its input, and outputting their sum ý[t], which should now give a similar audible impression as the first of the original samples: x[t].

4.2.1.1 Observations

More than 200 arbitrarily selected pairs of sound snippets – from acoustically distant to similar – were compared and tried to process with this method in order to achieve a convergence in timbre. Different amplitude parameter weightings and ranges – from linear to exponential – were deployed to transfer the cepstral difference $e_c[t]$ of signals $x_c[t]$ and $y_c[t]$ to the amplification engine gain parameters. At this gain stage a volume adaptation of individual filter output channels of the selected source signal y[t] was performed, with the aim of achieving a timbral alteration towards the sound of signal x[t]. After several attempts to refine the filter parameters by increasing the number of filter coefficients, and further, trying to set all possible smoothing parameter values for the temporal sequence

adaptation at the final amplification stage, the bottom line was clear. It was almost impossible to get any satisfying results in terms of timbre alignment of source- (x[t]) and target sound (y[t]), with the initially chosen, high resolution parameters (frame-length: 2048 / hop-size 512 / Mel frequency channels: 40), therefore, a research of possibilities regarding data reduction – by quantization to a lower-dimensional model – was not a realistic option. The best sounding results that have been observed would give a rough approximation of the dynamic progression of the target sound, whereas the timbral quality could never been successfully altered and has always adhered to the acoustic identity of the original source sound.

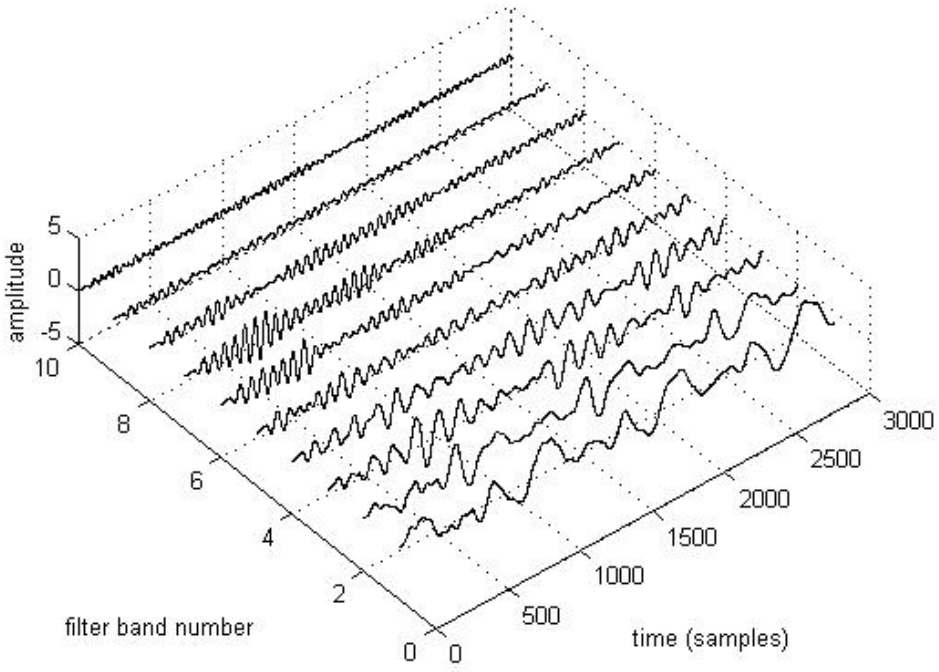


Figure 36. the first 10 Mel filter-bank output channels, showing the band-passed time domain components of a sample sound[6]

Figure 36 shows the first 10 band-passed time domain components of a sample sound. The reason for displaying only 10 channels is to avoid an unclear image, which would be the case if displaying all 40 channels at once. Also, the Mel-filter bandwidths in the lower frequency regions are a lot narrower than those in the upper regions, summing only very

[6] The axis label "time (samples)" refers to time domain samples. Sampling frequency = 44100 Hz

few, thus, not to distant frequency components to a single amplitude value. Therefore, the problem that will be pointed out next would grow even bigger in the upper frequency bands where the Mel-filter bandwidths increase. In figure 36, it can easily be observed that summing up a 1024 component spectral representation within 40 bands, results in a rather diverse frequency mix summed up in the individual channels, which can be inferred on the basis of the non-sine-like appearance of the time domain signal representation (figure 36). It is also an obvious conclusion that this frequency mix may occur in a multitude of different weighting proportions. Summing up all possible frequency combinations to a single overall amplitude value can thus have its source in as many different sonic situations as there are combinations of weighted frequency components. On a first glance, this insight could undermine the general credibility of the MFCC analysis as a sufficiently accurate tool for modeling musical timbre. However, at this point, the holistic nature of the MFCC representation should be emphasized, since the last step in the MFCC computation algorithm (the DCT) would consider the interrelations and the shape of all present – in our case: 40 – Mel-frequency bands for computing each coefficient. In other words, the DCT is bonding the influences of the energy in all frequency bands to the resulting values of individual coefficients. The most obvious reason however for this approach to fail is the fact that there is simply a lack in presence of specific frequency components or even worse, the lack of energy in complete frequency bands at certain times of the arbitrarily selected source sample (figure 36), which, in general could be interpreted as the main cause for it, having a different acoustic appearance than then the target sound.

One could argue that there is still one open possibility, to prove the given hypothesis – that timbral quality is connected with a particular sequence of short-time timbral features – by the now discussed analysis-resynthesis model. If the amount of Mel filter channels together with the amount of MFC Coefficients would further be increased, a yet more refined and differentiated filter-bank could be constructed, in order to split and modulate the given signal even more precisely. However, since the aim of this work - amongst others - is to find a reduced dimension timbre model, the decision to further boost the dimensions, which are already expected to deliver a high resolution timbral representation, would practically not make any sense.

For this reason, it was decided to temporarily drop any further verification attempts following this direction. An alternative approach, which will be presented in the next section was conceived, however, the idea of modulating the amplitude of individual filter output channels in order to alter the timbre of a source sample to approximate a different target-samples timbre was re-contextualized and re-applied again in the final sections of this work.

4.2.2 A corpus based approach for determining timbre similarity

After failing to take any successful steps towards the verification of a sequential inter-dependency of timbral features, let alone a dimension-reduced model of timbre representation with the analysis-resynthesis model, an alternative - corpus based - method was implemented, aimed at the verification of the temporal character of timbre in the first place. The general idea is visualized in figure 37.

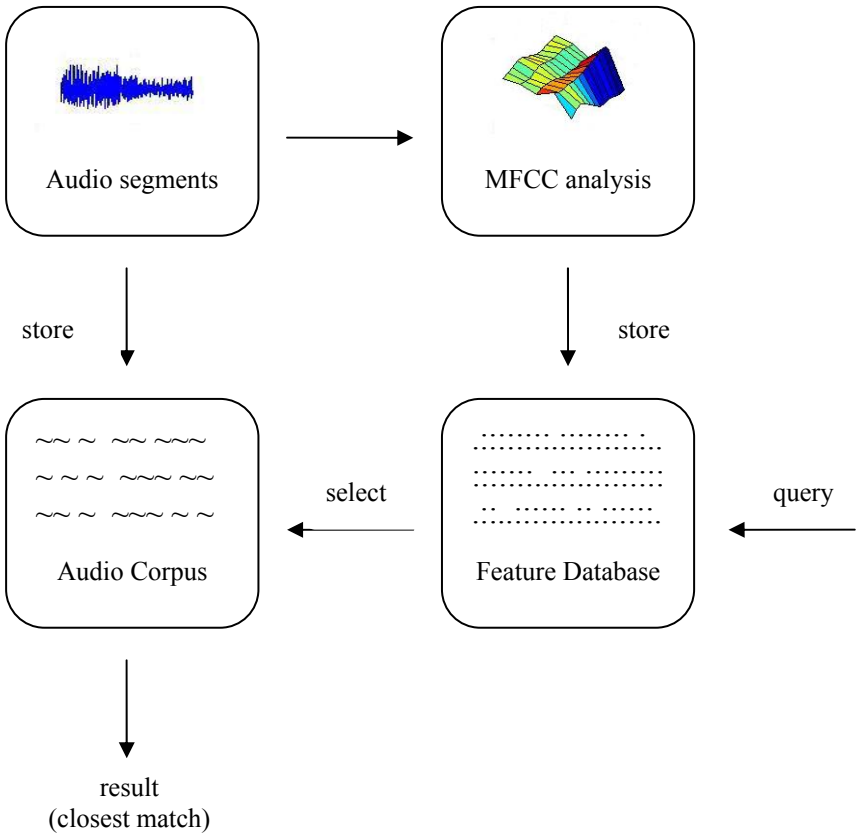


Figure 37. Schematic display of the query-by MFCC feature similarity algorithm

The corpus contained all sound snippets gained by the segmentation of the song “Careless whisper”. The goal was to first select an arbitrary target sample from this corpus and to find the most similar sounding segment out of the same corpus in the next step. Therefore, all sound segments were first transformed to the cepstral domain - with the high resolution parameters defined in table 4.1 - and stored in a database. At this resolution, a sound segment with the duration of 10240 samples or 0,232 seconds would be represented by a 663 dimensional vector, or a 17 timepoints * 39 cepstral coefficient[7] large plane (matrix), with the first cepstral coefficient excluded from evaluation and the comparison process.

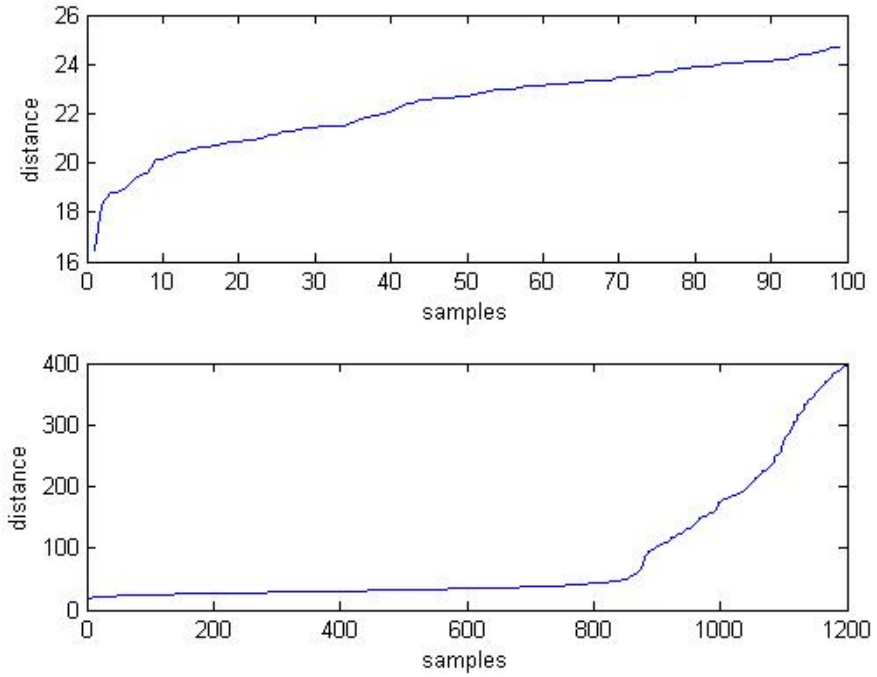


Figure 38. Sorted distances – upper image: first 100 closest sounding samples – lower image: all 1200 compared samples

This data structure can also be interpreted as a time-feature vector, representing one point in the multidimensional feature space, which is describing the temporal evolution of the 39

[7] This precise declaration of sample length was chosen in order to generate a comprehensible MFCC conversion result, regarding the dimensionality of the feature space. 10240 is an integer multiple of the deployed frame length. The Transformation of signals with the sample length between 10240 and 12288 would thus yield a MFCC time-feature vector of equal dimensionality (663), since the MFCC conversion algorithm would cut off the remaining – backmost – samples resulting from the modulo operation: “sample length % frame length”

dimensional short-time MFCC feature vector at a high resolution. In the next step, this arbitrarily selected sample would be compared to the all other samples in the database, by calculating the Euclidean distance to all vectors describing the corpus samples.

Figure 38 displays the results, i.e. the distances between a selected and all other corpus samples, sorted by increasing value. There were 1471 samples in the corpus. Before comparing 2 samples, the duration of both is adapted – cut to equal, and the resulting distance proportionally weighted, however, under a condition of a maximum length deviation of 10%. If the discrepancy in duration of two samples exceeds this tolerance limit, the corresponding source sample is not included in the distance evaluation process. The weighting of source samples exhibiting a different length than the target sample was performed as follows. If, lets say the source sample length would deviate from the length of the target sample for "+" or "– " 10%, the score gained through the comparison process has also been considered as a 90 % portion of the actual 100% score which was proportionally estimated in the next step. In an extreme case for example, where the selected source samples MFCC description would fully match the MFCC description of the target sample, but at the same time, both would exhibit a 50% discrepancy in length in either way, it would not be appropriate to regard this particular source sample as the "best fit", since 50% of it, or of the target sample was not even considered in the comparison process.

In the above stated example only 1200 out of 1471 samples were considered due to the 10% maximum deviation limitation. In a case, where the selected target sample is extremely long or short, the amount of eligible source samples meeting the condition of a similar duration would turn out to be even much lower.

The results of this approach were very promising. It was observed that it was actually possible to retrieve acoustically similar sounds with the comparison of the MFCC-trajectory based feature vector and the evaluation of the shortest measured distance. It was quite easy to evaluate the results, since the first suggested corpus sample would always have matched the instrumentation of the target sample, and what is more, the second, third, and so on closest matches would exhibit a timbre which was slowly drifting away from the target sample, very often keeping the same instrumentation in the first 6 to 7 closest matches.

In the next step, another test aimed at the verification of the sequential dependency of timbral features was performed. Here, the trajectories of the first five MFCC coefficients were reversed, which was an action that would preserve the statistical appearance of the feature vector time series, in terms of mean value and variance[8]. By reversing 5 out of 39 trajectories of the target sample and performing the above discussed comparison process again, an interesting result was observed. As it was expected, the now found closest matches form the corpus samples had a rather different acoustic appearance than the target sample, but again they exhibited a slowly fading acoustic similarity amongst themselves.

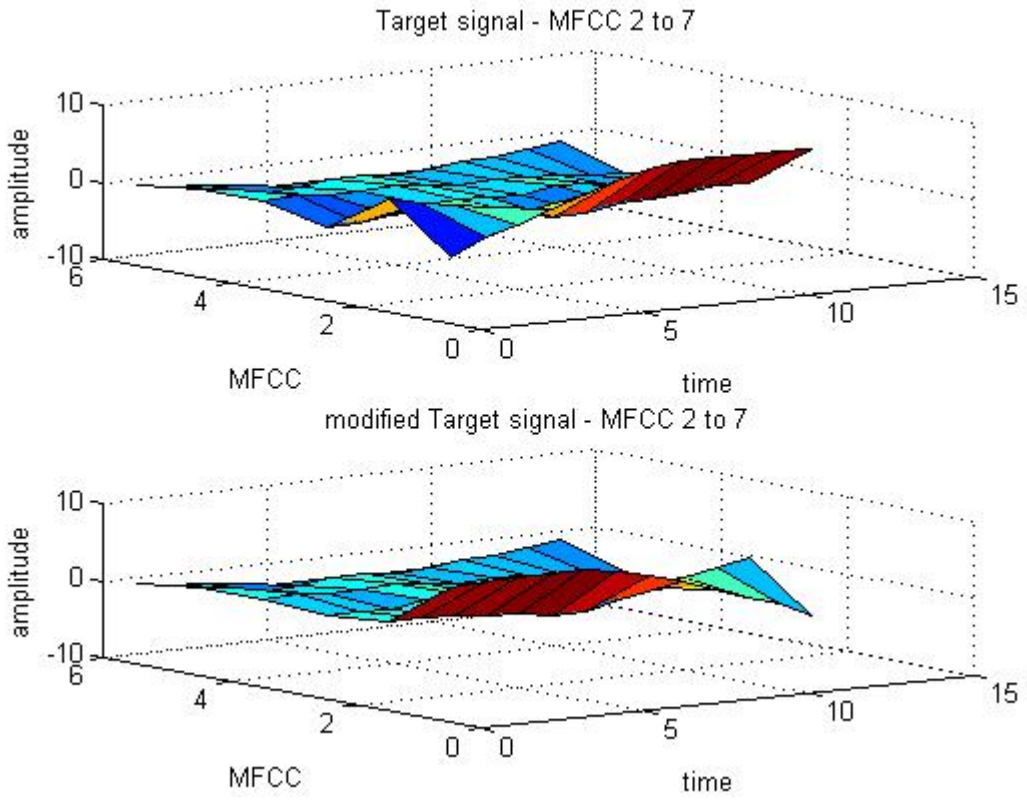


Figure 39. First 6 MFCC trajectories: 1. Original and 2. First 3 trajectories reversed

The next experiment involved a query based on the information contained in the trajectories of first 6 MFCCs (coefficients 2 to 7) alone. Amazingly, the results have matched the query

[8] In the MIR literature, the mean and variance are amongst the most popular classifiers when describing sound segments represented by MFCC features.

containing all 39 MFCC trajectories to a great extent, often introducing merely a change in the order of the first 10 closest matches. Again, the query was repeated, now with the first three trajectories time reversed (Figure 39.), yielding a very similar result as in the experiment with 39 MFCCs and the first 5 reversed, discussed above.

The fact that merely changing (reversing) the sequence of short-time timbral features has a significant impact on the perception of timbre was perhaps the most remarkable discovery in this research work. The results of this simple test might perhaps raise the demand for reconsideration of the most common methodologies in statistical timbre analysis and classification. This work does not include a deeper research of statistical timbre modeling concepts aimed at genre classification, music- / artist- / inter-song similarity research and therefore it could perhaps be inappropriate to comment on those in this context. However, it is still a bit peculiar to find contemporary research work from these categories, where it is still being tried to compute music similarity measures while not considering a prior – musically meaningful – segmentation, let alone the possibility of a sequential interdependency of timbral features and timbre perception. In some of the latest publications at the ISMIR conference: [Gasser *et al.* 2008], [Flexer *et al.* 2008], a statistical representation of a song is generated by training a Gaussian model with the first 20 MFCC values, collected throughout the song. If considering a sequential timbre model, perhaps the dimensionality in the MFCC domain could be reduced, while some more relevant timbral information could be captured by directing the attention towards the temporal dimension.

4.2.2.1 Data reduction

The aim of this work was not only to demonstrate the sequential dependency of timbral features but also to find a reduced dimension model exploiting the sequential dependency for this purpose. The testing began with a high resolution representation of the sound (40 MFCCs / 2048 sample frame length / 75 % overlapped).

First, the number of MFCCs was reduced, by gradually excluding the higher order coefficients. The comparison process was executed for each reduced set of coefficients, and it was observed that the results (the closest matches) obtained with only the first 10

coefficients, would be highly similar to the results generated with the full-sized set, operating with all 39 coefficients. Even with only the first 6 coefficients it was possible to get very reasonable sounding suggestions in the first 5 to 6 closest matches.

Despite those interesting observations, the problem with this corpus based approach is that it is not really possible to draw precise conclusions from the results, since one is always confined to the finite size of the available corpus. It is not possible to judge on the audible consequences of fine changes of feature values or to gradually approach and define the limit of perceptually relevant deviations in those values or in their resolution. The finite corpus size would necessarily introduce discontinuities with the limited available data, which would definitively account for false results. This is also the reason why the search for a dimension reduced timbre model in the temporal and amplitude domain was not pursued further with this – the corpus based – method.

4.2.3 Re-synthesis by band-passed noise + estimating a model

Having made one step further by roughly demonstrating a sequential dependence of MFCC features within a timbral character, the task of reducing the dimensionality of the sequential model and to finding a limit of perceptual relevance supporting that model was still open. A third approach was chosen, which would estimate the reduced dimension timbral model for purely synthesized sounds. The sounds used in the previous examples were re-utilized again here and taken as a model for the synthetic sounds.

First a full resolution MFCC analysis was performed on the real world sounds, with the high-resolution parameters already defined above. Next, the function "invmelfcc.m"[9] was deployed in order to re-synthesize the original sound back from its MFCC representation to a time domain signal. The re-synthesis is performed, by band-pass filtering white noise frame by frame, by a 40 channel Mel-spaced filter-bank.

The resulting time domain signal would sound like a noisy version of the original sound, yet preserving an impression of its initial timbral quality. The benefit of such a signal however is, that it would allow any kind of manipulation of its parameters in the cepstral domain while making the changes audible after an inverse transformation. Thus, a

[9] Found in "RASTA/PLP/MFCC - feature calculation and inversion" from LabROSA

search for quantization limits is a slightly more realistic task as if working with a finite corpus which wouldn't allow a precise manipulation range and would therefore produce unreliable results.

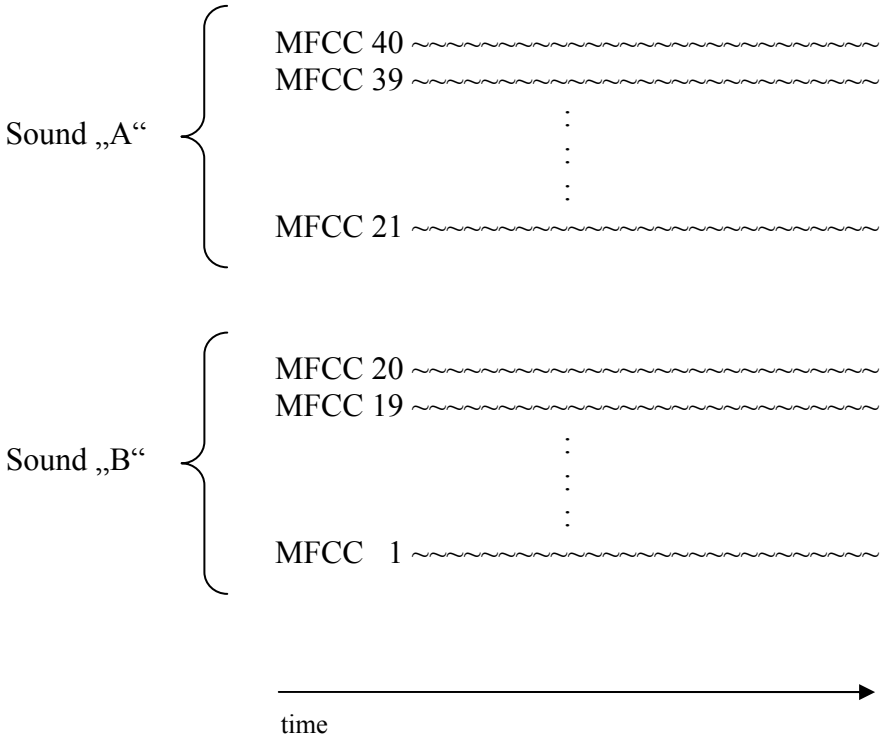


Figure 40. Morphing between the sounds: "A" and "B", by replacing individual trajectories of MFCC coefficients. Both sounds are synthetic (composed by 40 channel band-passed noise) which enables an authentic time domain reconstruction of the MFCC morphing.

Similar to the first approach where it was tried to transform one sound into another by manipulating the gain stages of individual filter outputs, an idea of morphing sound "A" into sound "B" was utilized again here. Two sounds with a distinct timbre were selected and resynthesized with "invmelfcc.m". In the next steps the individual MFCC trajectories of sound "A" were successively replaced with those of sound "B", starting with the 40th - the highest order - coefficient and continuing towards the first. Figure 40 shows the concept of replacing the individual MFCC trajectories. The example displays a temporary state from the middle of the morphing process, representing a sound segment, composed from the upper 20 MFCC trajectories of sound "A" and the lower 20 of sound "B". After each coefficient trajectory was replaced the sound was played back and the changes in timbre

subjectively compared to the original sounds “A” and “B”. Similar to the second experiment, with the corpus based retrieval of most similar sounds, it was observed that the perceptually relevant acoustic effects would slowly start to appear after the replacement of the first 30 highest order coefficient trajectories. So, the sound would not effectively start to morph its timbral character until the 10^{th} and the next lower coefficient trajectories have been replaced.

A second empirical test regarding the effects on the perception of timbre, similar to the prior, was performed with the MFCC trajectory replacement order reversed. Starting by replacing coefficient number 1, 2, 3 and so on, a drastic change of morphing speed was observed, making sound “B” appear very clearly already after replacing the first 5 coefficients. What is more, the timbre character of sound “A” would immediately alter towards that of sound B, already at replacement of the first coefficient. That was indeed an interesting insight, since in the previous (the corpus based) experiment the first coefficient was always excluded from the comparison and retrieval process, for it was assumed that it merely carried the information about the short time energy content of the signal, or its overall loudness, which was not expected to be of relevance for the timbral identity of a sound.

4.2.3.1 Verification

Until now, only tests regarding the data reduction in the first dimension – the amount of MFCCs – were carried out, while the possibilities of temporal and amplitude quantization – let’s call them dimension 2 and 3 – still remain unexamined. Another disadvantage of the experiments carried out thus far, is the selected mode of verification. All judgments on the contribution of the MFCC features to the timbral identity of the sounds were made on a subjective basis of acoustic evaluation carried out solely by the author himself.

Striving to avoid listening tests involving a multitude of selected individuals, an alternative verification method was needed, which could impose a measurable indication confirming the subjective acoustic observations already made by the author.

A method of estimating the variance in the de-trended version of the temporal MFCC-trajectory data set was implemented with the aim of evaluating the contribution of

each coefficient according to the presence of its variance in relation to the variance of the remaining coefficients. The degree of variance in MFC coefficients was assumed to be an indicator for the importance and the informative value of a coefficient in describing a timbral feature. In order to obtain a reliable estimation on the variance of each coefficient, a de-trended reference is needed, for the isolated beats and notes from pop music always exhibit a peak at the onset and fading amplitude character along with the further progression of the note. The native Matlab function "detrend.m" was deployed here, since it removes the best straight-line fit linear trend from the analyzed data.

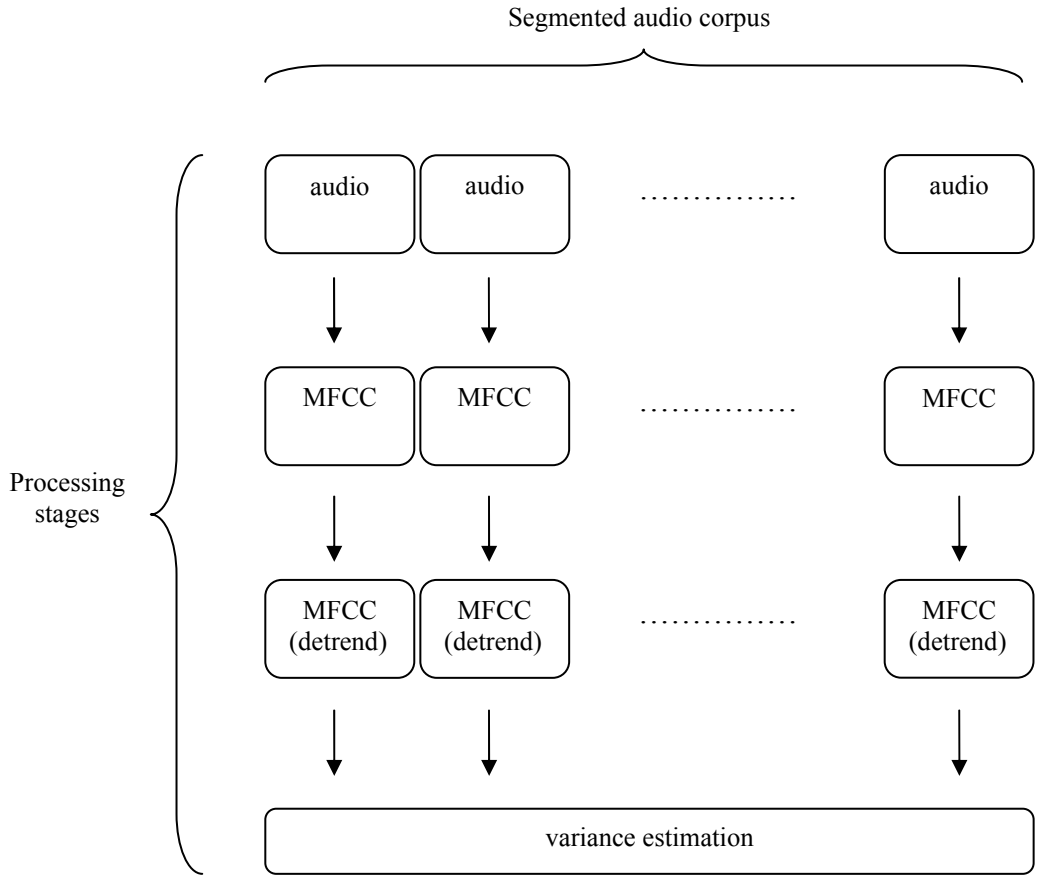


Figure 41. Schematic display of the variance estimation algorithm. The boxes indicate segment boundaries determined by onset segmentation. The linear trend in the temporal MFCC trajectories is eliminated from each audio segment individually before the variance is calculated.

Figure 42 shows an example of the de-trend principle. In figure 41, the audio processing stages in the MFCC-trajectory variance estimation procedure are displayed. The results

represent the averaged variance across the whole corpus of 1471 sound snippets and were not to difficult to interpret (figure 43). Figure 43 shows a steep decrease of the variance in the first 5 coefficients. From coefficient number 6 to 40, however a slowly fading linear trend can be recognized with very similar, almost constant variance values. This and the following variance plots display the mean variance value, averaged over all segments (left image), while the right images show the variance of this variance.

Going back to the listening examples, perhaps no one would argue the claim that the contribution of the highest order coefficient (no. 40) is rather marginal and thus it could be excluded from the sounds timbral model without any further concern. Now, if taking into account its variance, which could be an indicator of the coefficients contribution or better to say, an indicator of the relative importance of the timbral feature this coefficient is describing, it could well be asserted that other timbral features, modeled by other coefficients exhibiting similar variance values, could also carry only relatively marginal information on the sounds timbral quality.

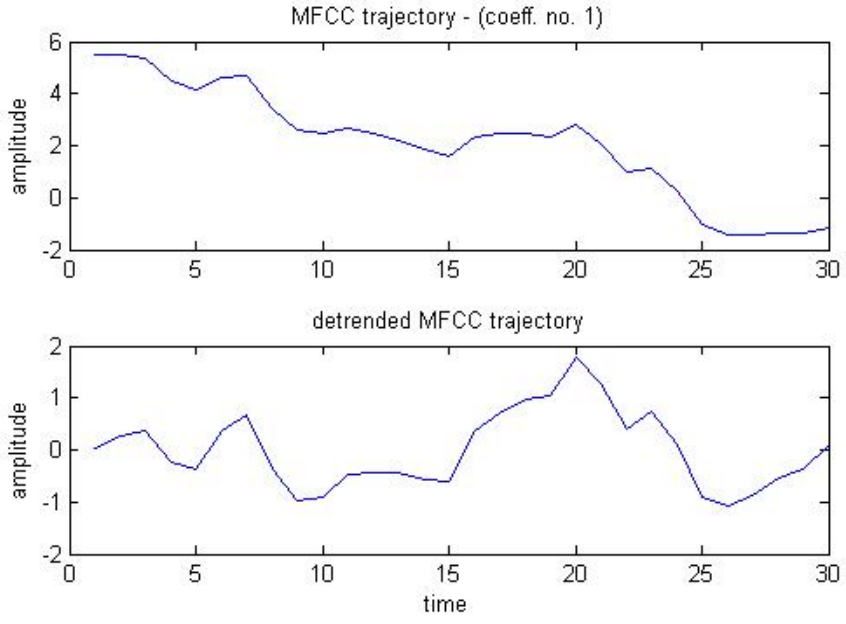


Figure 42. Original and the de-trended version of an MFCC trajectory

To conclude with, the acoustic observations where the most important carriers of timbral information were found to be the first 5 to 6 coefficients could be confirmed by the results of the MFCC-trajectory variance analysis.

4.2.3.2 Amplitude and time quantization

Having analyzed the possibilities of data reduction in the first dimension (The amount of MFCCs), there was still an open potential to reduce the data by quantizing the values in the time and amplitude dimensions of the MFCC representation.

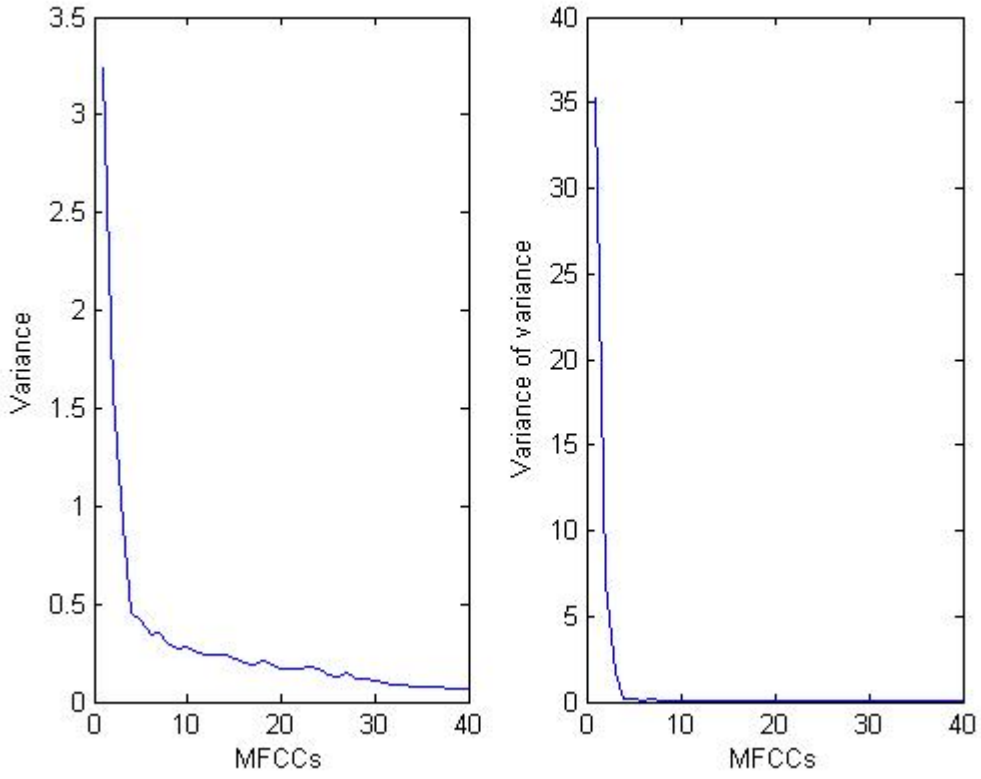


Figure 43. The variance plots of individual MFCCs at the following resolution: (FFT frame length: 2048 samples / hop-time: 512 samples / no quantization of the amplitude values)

Amplitude quantization:

Following plots show how the variance of individual MFCCs changes if the MFCC amplitude is quantized to integer values. The displayed quantization step-sizes are “1”, ”2”, “3” and “7”. Parallel to the variance estimation also listening tests were conducted with arbitrarily selected, amplitude quantized sound samples. Through the author’s subjective evaluation, it was found out that the acoustic examples would preserve an authentic acoustic impression up to a minimum step-size 2.

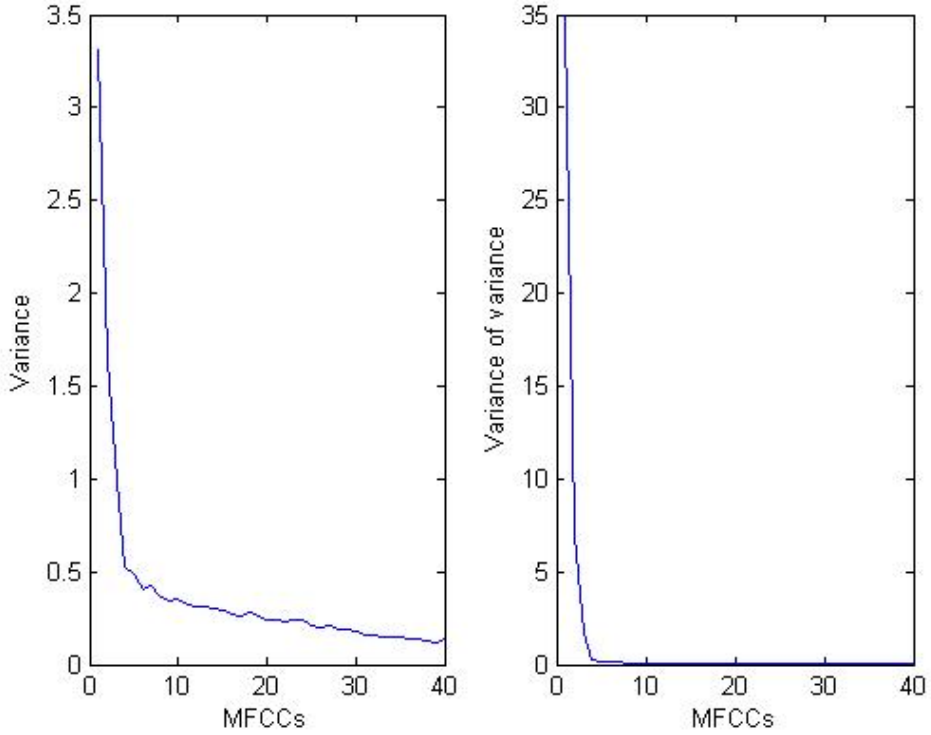


Figure 44. Variances of individual MFCCs amplitude-quantized with smallest step-size "1".

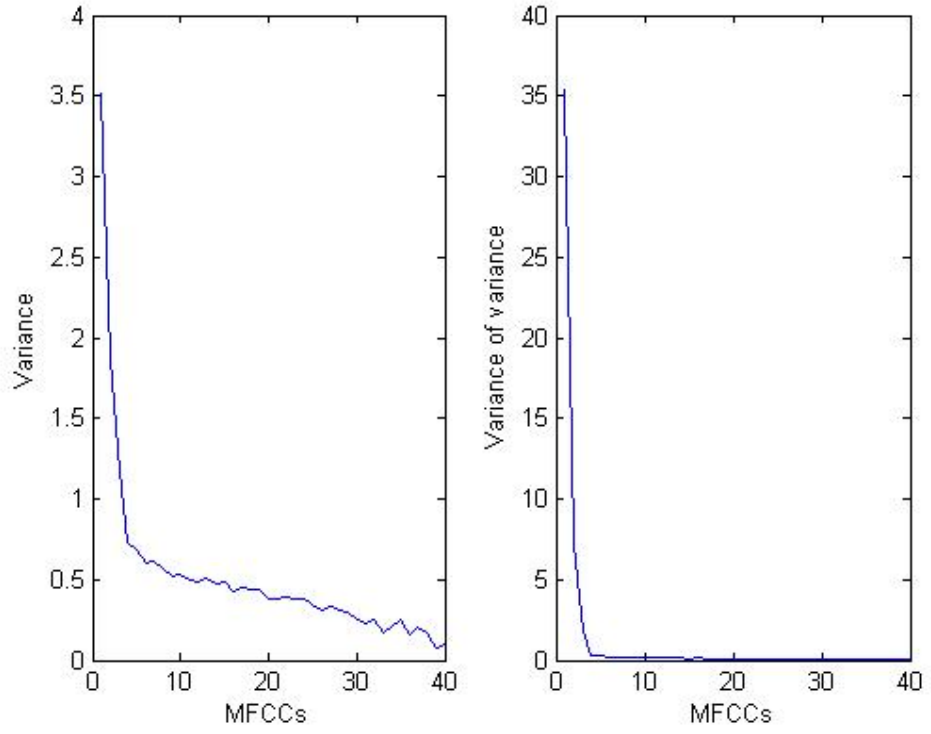


Figure 45. The variances of individual MFCCs amplitude-quantized with smallest step-size "2".

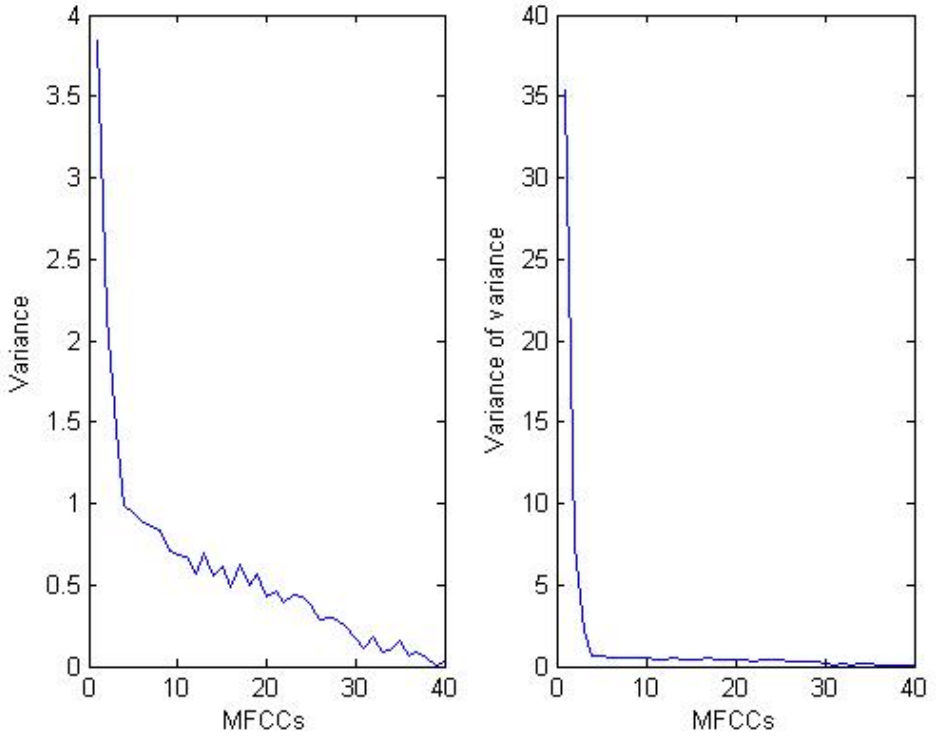


Figure 46. The variance of individual MFCCs amplitude-quantized with smallest step-size "3"

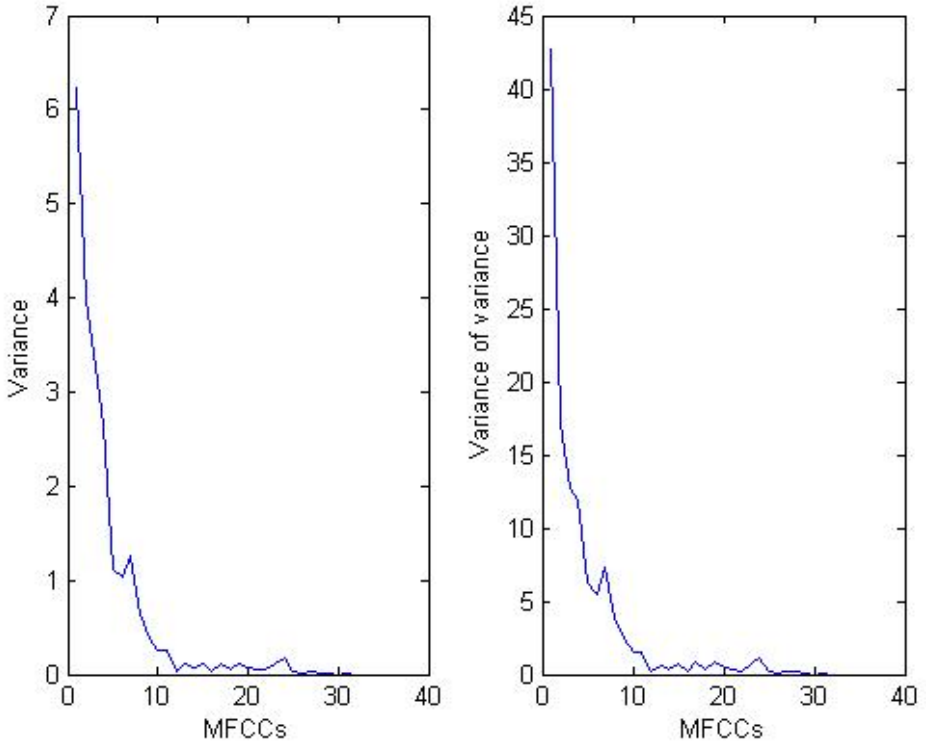


Figure 47. The variance of individual MFCCs amplitude-quantized with smallest step-size "7"

In the variance plots, it can be observed that there is a tendency of the variance to increase in the first coefficients, while the variance in the higher order coefficients is pushed towards zero (figure 47). In general, the higher order coefficients have very low amplitudes, therefore it was well expected that their variance values would run towards zero after quantization. More relevant however for determining the acceptable quantization limits are the lower order coefficients, which is why it is important for their variance values not to grow to large. Perhaps, in this case it would be important to observe and to limit the variance of the first 5 coefficients. If the original version – without amplitude quantization – (figure 43) is compared with the quantized versions in figures 44 and 45, it can be seen that the value of the variance in the first coefficient is held between 3 and 3,5, while the overall shape remains identical. In figure 46, the value of coefficient 1 is already 3.8 and coefficient 4 is twice as large as in the original version, which might be an argument for reaching a limit in the acceptable quantization step-size.

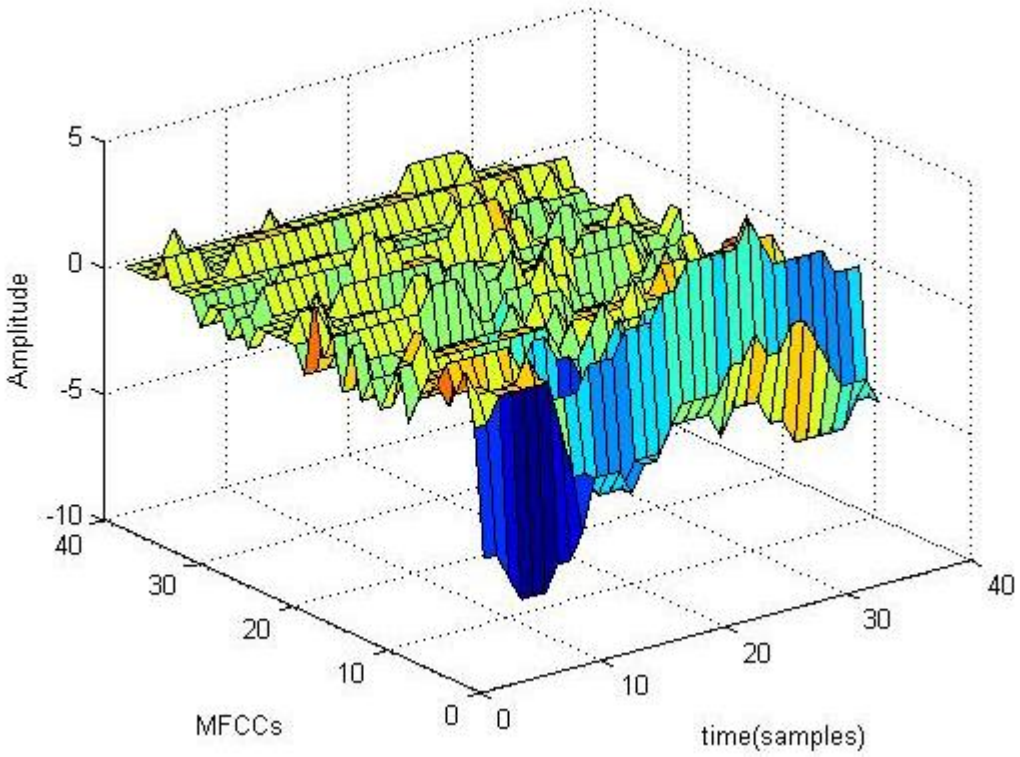


Figure 48. MFCC-trajectories, amplitude-quantized with smallest step-size "1"

Figure 47, is displayed for tendency-demonstration purposes only, since the quantization step-size "7" would already yield rather mutilated acoustic results. Figures 48 and 49 show

the amplitude-quantized MFCC trajectories of the same, arbitrarily selected sound sample, with quantization step-size “1” and “3” respectively. The image in figure 49 could be another apparent argument for the rather bizarrely sounding inversion (sonic reconstruction) of the amplitude-quantized MFCC trajectories.

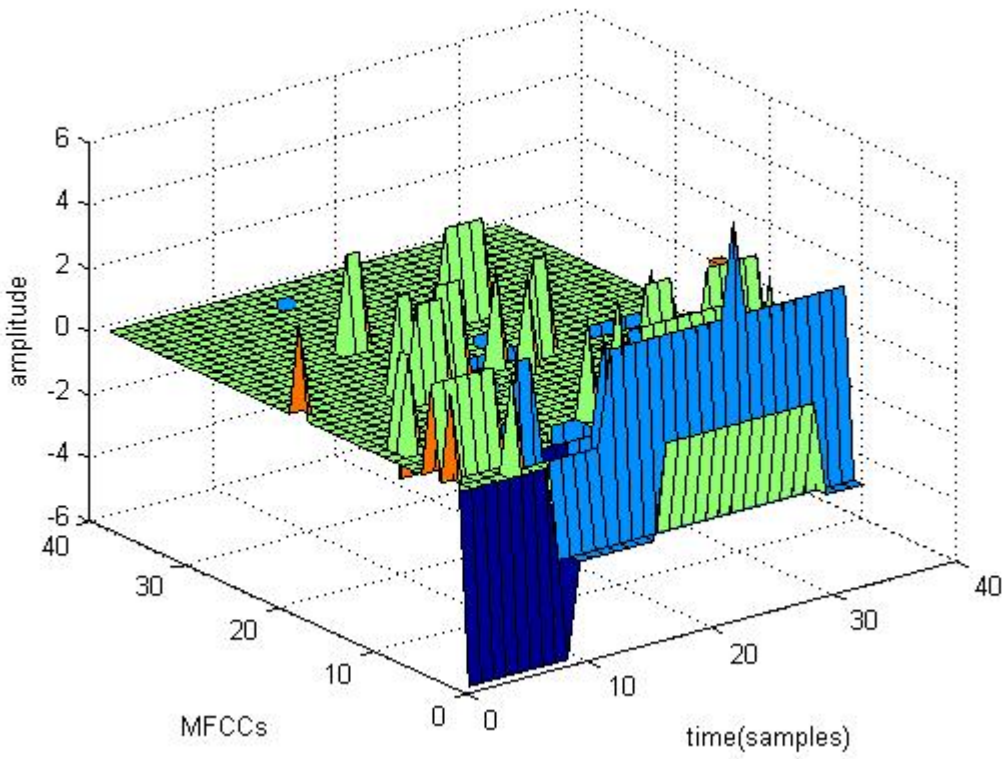


Figure 49. MFCC-trajectories, amplitude-quantized with smallest step-size “3”

Time quantization:

Until now, all transformations were done with the same time resolution (frame-length: 2048 samples / hop-time 512 samples), at a sampling rate of 44100 Hz), which also is in conformance with standard resolutions applied in related work like: [Foote 1999], [Aucouturier *et al.* 2005], etc. According to the author’s personal acoustic judgments, any transformations done with higher resolutions would not yield any better acoustic results, so those parameters were assumed to represent a high resolution standard.

In the next step, the possibilities of augmenting the hop-size and replacing the missing data using linear interpolation were studied. Figure 51 shows a MFCC trajectory

plot, with a hop-size of 2048 samples, i.e. with no overlap of analysis frames, while reconstructing the missing values with linear interpolation in order to preserve the original sample density. The corresponding MFCC variance values are shown in figure 53 and if compared with the variance values of the full resolution analysis (figure 43) a decline of the values at all coefficient can be observed, which is coherent with the theory and thus, a well expected consequence. Again, most of the attention shall be directed towards the changes taking place in the variances of the first coefficients. An indicator for reaching a quantization limit is – like in the amplitude quantization above – the variance deviation magnitude – with reference to the optimal resolution. It is rather difficult however to set a boundary condition for this case, so the results of realizations deploying different quantization parameters can merely be compared amongst themselves i.e. judged on a relational basis. Listening tests on the other hand have still shown an acceptable acoustic reconstruction of the original sample at this particular resolution, combined with linear interpolation between sample points.

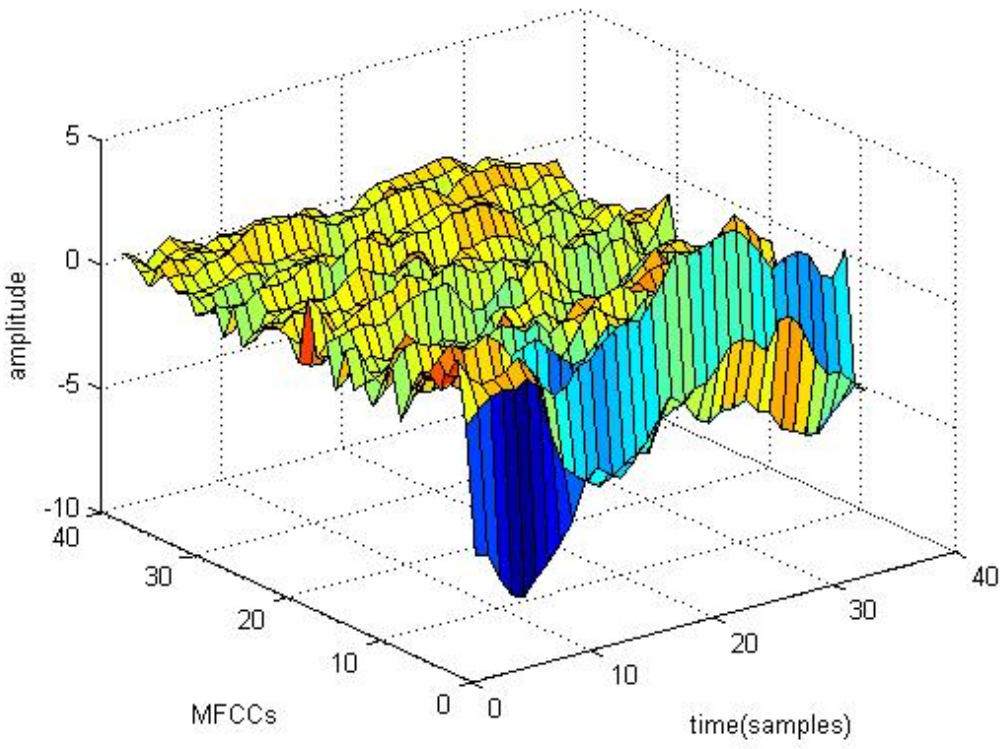


Figure 50. MFCC-trajectories, - same sample as above - no quantization

Working with an average sound-sample length of 226 milliseconds or 10.000 samples, there are not many more options for a larger time-frame quantization. A test with a hop-size of 4096 samples was conducted, yielding unacceptable acoustic results, also confirmed by a drastic decrease of variance values (figure 54). The problem with the larger hop-sizes (e.g. 4096 samples) is also the following. If, for instance, sounds with a duration under 10240 samples – which is the case with about 700 out of 1471 samples in the here analyzed sound corpus – are taken for analysis, the output result would be represented by two temporal values (for each coefficient), resulting in an unrealistic variance value. Perhaps, a reasonable minimal number of sample points for describing 1/4 of a second long audio snippets can be defined to "4", making a hop-size of 2048 samples a reasonable choice. The frequency rate, at which temporal changes in the amplitude values of the first MFCC coefficients take place seem to be low enough for minimizing the data loss at an approximation via linear interpolation.

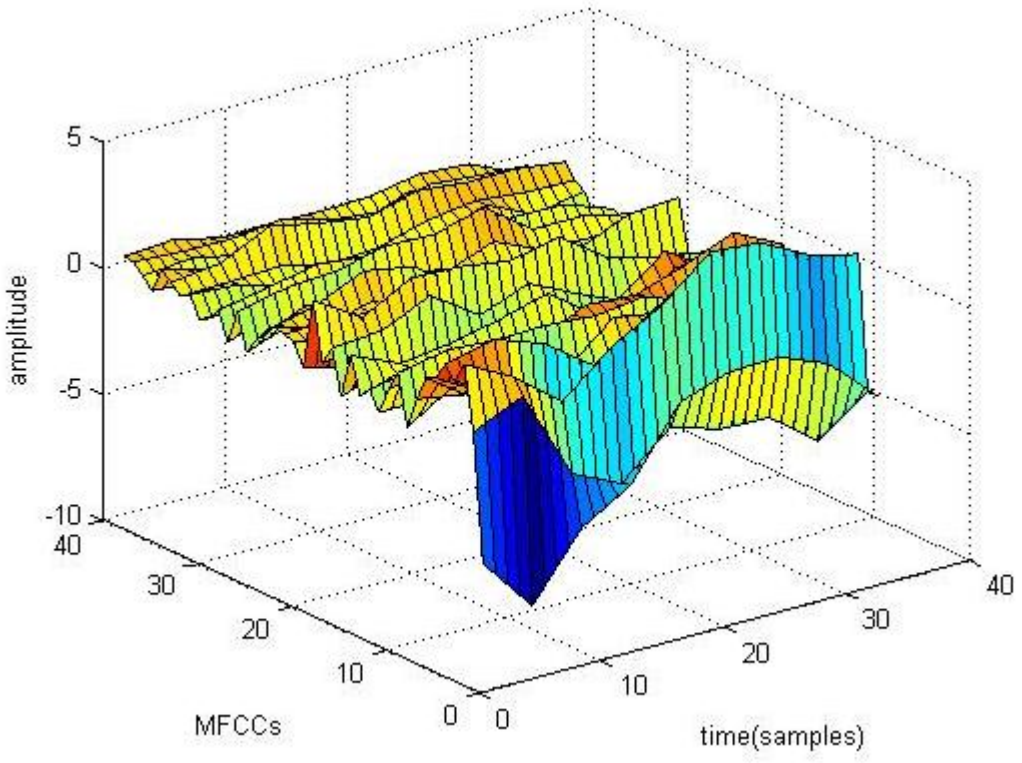


Figure 51. MFCC-trajectories, - FFT frame-length and hop-size = 2048 - with linear interpolation

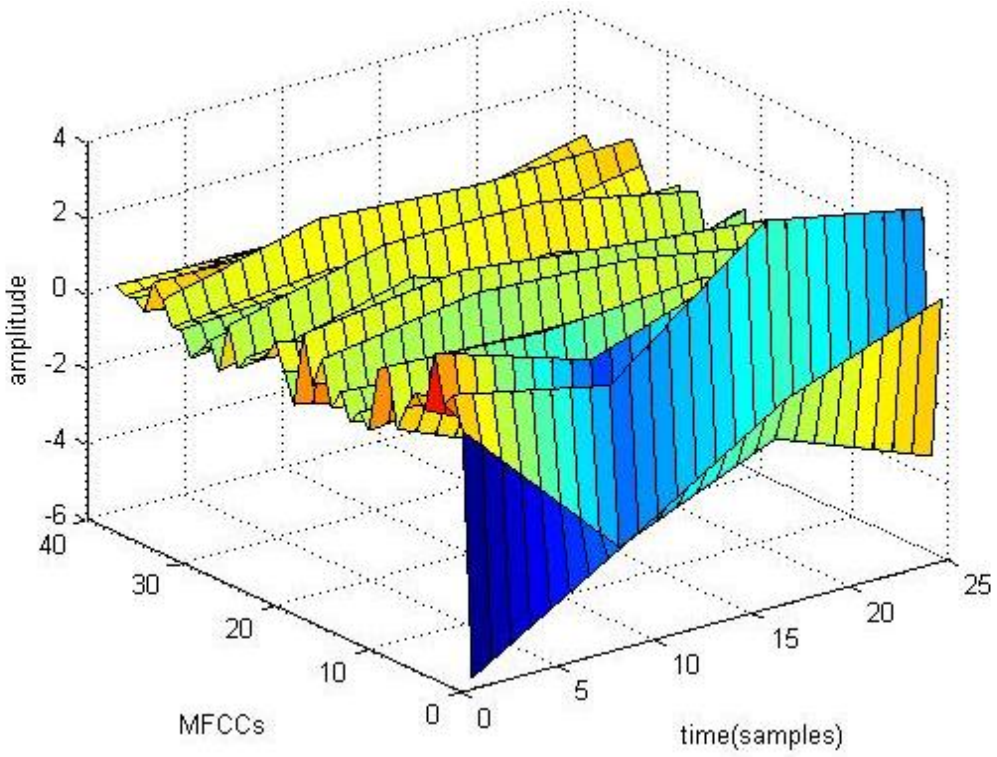


Figure 52. MFCC-trajectories, - FFT frame-length 2048 and hop-size =4096 - with linear interpolation

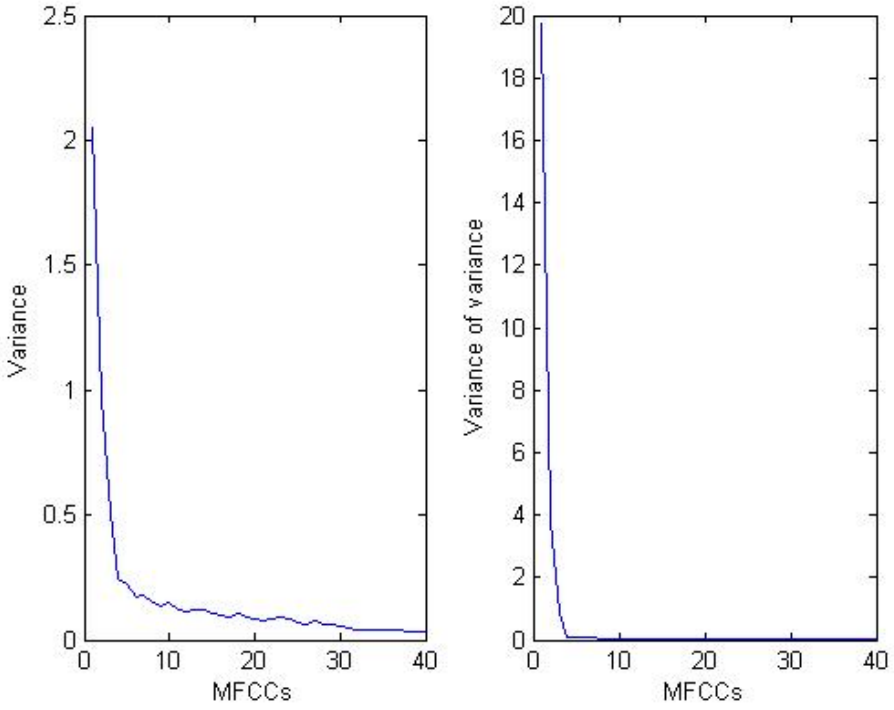


Figure 53. Variances of individual MFCCs: frame-length and hop-size = 2048 (linear interpolation)

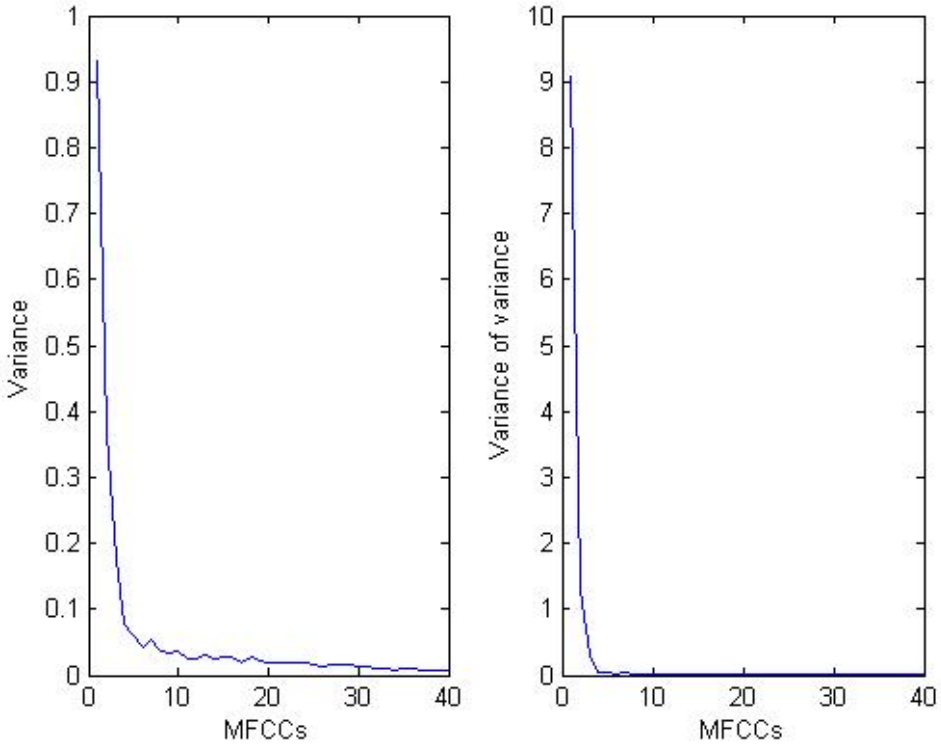


Figure 54. Variances of individual MFCCs with FFT frame-length 2048 and hop-size = 4096 and linear interpolation

4.3 A possible practical application

The research work presented in this thesis was motivated by an idea of realizing a concrete musical application based on sound sample similarity. Unfortunately it is not in the scope of this work to debate extensively on this particular application, in fact the thesis results would firstly represent a basis on which further tests and steps towards the realization of the application are yet to be conducted.

The conceived application is based on the idea of sound modulation (transformation), however the aim is not to actually modulate a given sound but rather to define the modulation parameters in the first step and then to replace the given, target sound with an existing real-world version taken out of a giant corpus, implying an extensive user network as an ideal scenario. From this point of view, the task of defining and interfacing the modulation parameters was the primary goal of this thesis. First tests were already

conducted within the context of the corpus based algorithm presented in section 4.2.2. A conceptual design of a "difference plane" interface was implemented, where the user could manually model a plane (figure 55), which would represent the difference between the temporally arranged values of the first 5 to 6 MFCCs, modeling two distinct sound samples i.e. the timbral distance between the target and the desired source sample replacing it. The length of the difference plane would depend on the choice of the target sample, where each value on the time axis would represent a 2048 sample long signal block. The amplitude values were confined to a minimum quantization step-size with value "1". In the next step, this difference plane could be imposed on a whole sequence of previously segmented sound snippets and could thus alter, or better to say, replace a target sequence of originally connected samples with an alternative sounding sequence of originally discontiguous samples, sharing a common character. The search for of the closest match would look as follows.

$$\mathbf{T[c] - S[c] = d[c]} \qquad (4.1)$$

The cepstral difference plane: **d[c]** is the result of subtracting the temporal progression of the first 5 MFCC values representing the source sample **S[c],** from the same representation of the selected target sample **T[c]**. The user however is supposed to select a target sample and further, to model the desired difference plane. The MFCC model of the closest sounding source sample is determined by subtracting the difference plane from the target sample MFCC representation. The search for the actual sample is continued by comparing the distances between the ideal (the computed) source sample – respectively its MFCC representation – and the MFCC representations of all samples in the available corpus. The closest match is selected by calculating and evaluating the Euclidean distance between the desired and all other vectors describing the corpus samples. Whether the Euclidian distance measure is the best choice for this purpose is not clear and could be subject of further research. Perhaps the search results could be improved by weighting the contributions of individual coefficients in accordance with their priority in describing the timbral quality.

For a proper functionality of this application however, an extensive corpus of sound material would be required, ideally realized as a networked audio plug-in.

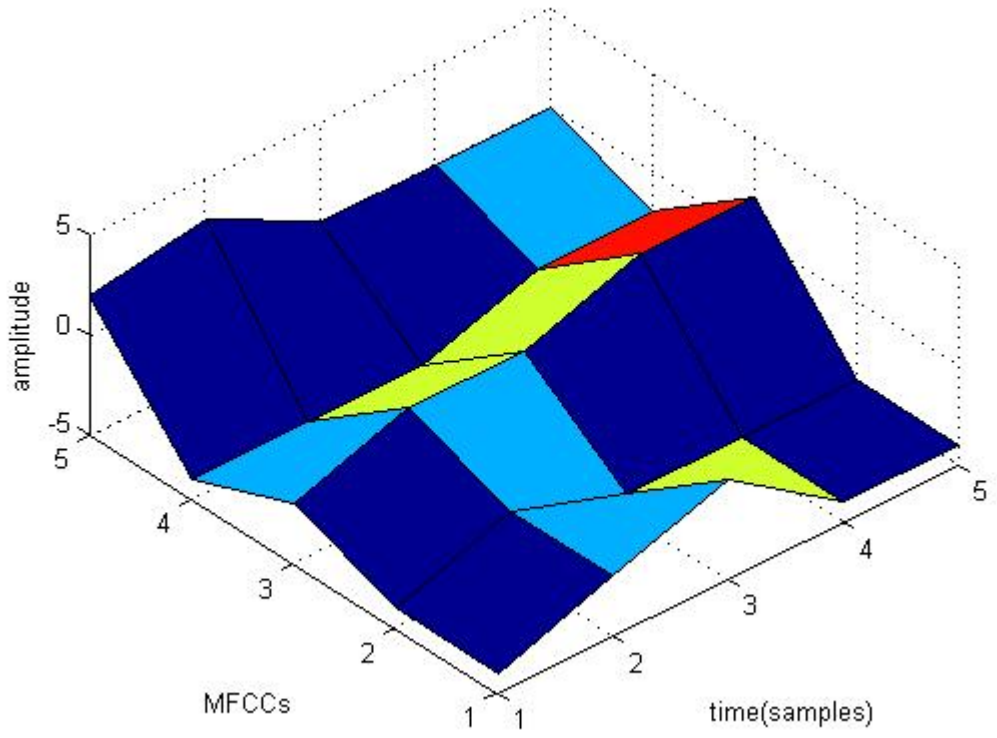


Figure 55. MFCC difference plane – a possible interface for sound "modulation"

4.4 Conclusion

In the first place, this thesis was motivated by an idea for a practical application based on timbre similarity, which however was in the background throughout the work.

In order to gain insight into existing timbre modeling approaches, a broad amount of research work was studied and the crucial steps and conclusions summarized in the chapters 2. and 4. This was done very systematically with the aim of exposing and legitimatizing some methods, which would get recombined in the main idea of the thesis, which was introduced, analyzed and documented in chapter 4. Starting with the examination of monophonic timbre modeling techniques on one hand and continuing with the research of polyphonic timbre analysis techniques, on the other, a concept of describing very particular sonic material - namely short polyphonic beats and notes, isolated from a larger context of a musical sequence - was introduced. In the proposed method, formal concepts of arranging features calculated with monophonic music analysis tools were

applied for modeling the content, or better to say the features generated with polyphonic signal analysis methods. The main idea behind this strategy was the assumption that a particular temporal sequence of timbral features would represent a highly important factor for timbre classification. This concept could represent an alternative to the mainstream methods from the field of polyphonic music analysis research, which do not pay attention to the temporal character of timbre, instead they rather operate with statistical – "bag of frames" – timbre models, applying to longer sequences of polyphonic music.

A number of approaches aiming at the verification of the timbre's temporal character were introduced and documented. In chapter 4.2.3 time, amplitude and MFCC quantization parameters were deduced, proposing a temporal model at a reasonable resolution, so that an acoustic reconstruction would preserve an identical and recognizable timbral image compared with the original sound sample. This model can be used for timbre classification and identification of previously segmented sound material exhibiting a polyphonic character.

Regarding the complicated question of timbre definition however, no concrete conclusions can be drawn from this work. Perhaps if the work would be continued and the application proposed in chapter 4.3 actually implemented at its full extent, some conclusions regarding the timbre definition could be drawn based on the user experience and feedback. Until then, perhaps one would best keep out harm's way by sticking to the definition of Keith D. Martin from [Martin 1999].

5. References

[Allamanche *et al.* 2001] Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., Kastner, & T., Cremer, M. Content-based Identification of Audio Material Using MPEG-7 Low-level Description. Proceedings of the International Conference on Music Information Retrieval 2001.

[ANSI 1960,1970] ANSI (American National Standards Institute) 1960, 1970.

[Aucouturier *et al.* 2004] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences, 1(1), 2004.

[Aucouturier *et al.* 2005] The Way It Sounds": Timbre Models for Analysis and Retrieval of Music Signals - Jean-Julien Aucouturier, François Pachet, and Mark Sandler, IEEE Transactions on Multimedia, Vol. 7, No. 6, December 2005

[Aucouturier 2006] Julien Aucouturier – PhD 2006 10 experiences about Modelling of Polyphonic Timbre – Sony Computer Science Lab, Paris in collaboration with the Laboratory of informatics, Paris

[Beauchamp 1982] J. Beauchamp, Synthesis by spectral amplitude and "Brightness" matching of analyzed musical instrument tones. J. Acoust. Eng. Soc., Vol. 30, No. 6. 1982.

[Beauchamp *et al.* 1997] J. W. Beauchamp and A. Horner, "Spectral modelling and timbre hybridisation programs for computer music," *Organised Sound*, vol. 2, num. 3, pp. 253-8, 1997.

[Benade *et al.* 1988] A. H. Benade, S. N. Kouzoupis, The clarinet spectrum: Theory and experiment. J. Acoust. Soc. Am. 83(1), January 1988.

[Bishop 2006] Bishop C. M. Pattern Recognition and Machine Learning (ISBN 0387310738) - Springer 2006

[Bregman 1990] Bregman, A. 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press: Cambridge.

[Berenzweig *et al.* 2003] A. Berenzweig, D. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In Proceedings ICME-03, 2003

[Casey 2001] Casey, M. A.: General sound classification and Similarity in MPEG-7, *Organised Sound* 6(2): 153–164, 2001 Cambridge University Press, United Kingdom.

[Conklin 1997] H. A. Conklin, Piano strings and 'phantom' partials, J. Acoust. Soc. Am., Vol. 102, No. 1, 1997.

[Cooley et al. 1965] J. W. Cooley, J. W. Tukey, "An algorithm for machine calculation of complex Fourier series" (1965), In Math. Comp.

[Depalle *et al.* 1993] P. Depalle, G. Garcia, X. Rodet, Tracking of partials for additive sound synthesis using hidden markov models. Proc. of the IEEE, 1993.

[Fitz *et al.* 1995] K. Fitz, L. Haken, and B. Holloway, "Lemur - A Tool for Timbre Manipulation," International Computer Music Conference, Banff, Canada, 1995.

[Fletcher 1934.] Fletcher, H. 1934. Loudness, Pitch and Timber of Musical Tones and their Relations to the Intensity, the Frequency and the Overtone Structure, in JASA, Vol. 6. No. 2, pp. 59 - 69.

[Flexer *et al.* 2008] Flexer A., Schnitzer D., Gasser M., Widmer G.: Playlist Generation using Start and End Songs, to appear in: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08), Philiadelphia, USA, 2008

[Foote 1997] Foote, Jonathan. “Content-Based Retrieval of Music and Audio,” in C.-C. J. Kuo et al., editor, Multimedia Storage and Archiving Systems II, Proc. of SPIE, Vol. 3229, pp. 138-147, 1997.

[Foote 1999] J. Foote, **“**Visualizing music and audio using self-similarity**,”**. in Proc. of the 7th ACM Intl. Conf. on Multimedia (Part 1),. New York, NY, USA, 1999

[Foote *et al.* 2003] J. Foote and M. Cooper, “Media segmentation using self-similarity decomposition,” in *Proc. SPIE Storage and Retrieval for Multimedia Databases*, vol. 5021, J. 2003.

[Gasser *et al.* 2008] Gasser M., Flexer A., Widmer G.: Streamcatcher: Integrated visualization of music clips and online audio streams: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08), Philiadelphia, USA, 2008

[Gersem *et al.* 1979] P. De Gersem, B. De Moor, and M. Moonen, “Applications of the continuous wavelet transform in the processing of musical signals,” - 13th International Conference on Digital Signal Processing, Santorini, Greece, 1997.

[Grey 1977] Grey J. M. 1977. Multidimensional Perceptual Scaling of Musical Timbre. Journal of the Acoustical Society of America Vol. 61, pp. 1270 - 1277.

[Haitsma *et al.* 2002] Haitsma, Jaap, and T. Kalker. 2002. A highly robust audio fingerprinting system. International Symposium on Musical Information Retrieval (ISMIR2002), pp.144-8.

[Harris 1978] F. J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform. Proc IEEE, Vol. 66, No. 1, January 1978.

[Helmholtz 1954] Helmholtz, H. L 1954. On the Sensation of Tone as a Physiological Basis for the Theory of Music (translation of original text 1877), New York: Dover Publications.

[Herrera-Boyer *et al.* 2003] D. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *J. New Music Res.*, vol. 32, no. 1, pp. 3–21, 2003.

[Horner *et al.* 1996] A. Horner, J. Beauchamp, Piecewise-linear approximation of additive synthesis envelopes: A comparison of various methods, Computer Music Journal, Vol. 20, No. 2, summer 1996.

[Howard *et al.* 2001] Howard, D. M., Angus, J. 2001. "Acoustics and Psychoacoustics", Oxford, Boston: Focal Press.

[ISO 2001] ISO/IEC FDIS 15938 4:2001(E) Information Technology — Multimedia Content Description Interface — Part 4: Audio

[Jensen 1999] Jensen K. "Timbre models of musical sounds" - Kristoffer Jensen - PHD 1999, University of Copenhagen.

[Kendall *et al.*1991] Kendall, R.A. and Carterette, E.C.(1991). "Perceptual scaling of simultaneous wind instrument timbres" in Music Perception, 8:369–404.

[Lancaster *et al.* 1986] P. Lancaster, K. Salkauskas, Curve and surface fitting: An introduction, Academic Press, 1986.

[Licklider 1951] Licklider, J. C. R. 1951. Basic Correlates of the Auditory Stimulus. (In Handbook of Experimental Psychology, S.S. Stevens ed.) New York: Wiley.

[Logan 2000] B. Logan. Mel frequency cepstral coefficients for music modeling. In Proceedings of International Symposium on Music Information Retrieval (ISMIR), 2000.

[Logan *et al*. 2000] B. Logan and S. Chu, "Music summarization using key phrases," in International Conference on Acoustics, Speech and Signal Processing pp. II–749–752 (2000).

[Makhoul 1975] J. Makhoul, “Linear prediction: a tutorial review,” *Proceedings of the IEEE*, vol. 63, num. 4, pp. 561-80, 1975.

[Martin 1999] Martin, K. 1999. Sound-Source Recognition: A Theory and Computational Model. Ph.D. Dissertation, MIT.

[McAdams *et al.* 1995] S. McAdams, S. Winsberg, S. Donnadieu, G. de Soete, J. Krimphoff, Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological Research, 58, pp. 177-192. 1992.

[McAdams *et al.* 1999] McAdams, S., J. W. Beauchamp, S. Meneguzzi 1999. Discrimination of Musical Instruments Sounds Resynthesized with Simplified Spectrotemporal Parameters, JASA 104(2).

[McAulay *et al.* 1984] McAulay, R.J. and T.F. Quatieri. 1984 “Magnitude-Only Reconstruction Using a Sinusoidal Speech Model” - IEEE 1984, pp. 27.6.1-27.6.4.

[Mikula 2008] Luka Mikula “Concatinative music composition based on recontextualisation utilizing rhythm-synchronous feature extraction” - Diploma Thesis - Institute of Electronic Music and Acoustics - University of Music and Dramatic Arts Graz, Austria

[Morchen *et al.* 2005] Morchen, F., Ultsch, A., Thies, M., Lohken, I., Nocker, M., Stamm, C., Efthymiou, N. & Kummerer, M. (2005). MusicMiner: Visualizing Timbre Distance of Music as Topographical Maps. *Tech Report*. Department of Mathematics and Computer Science, University of Marburg, Germany.

[Moorer 1987] J. A. Moorer, “The use of the phase vocoder in computer music applications,” Journal of the Audio Engineering Society, vol. 26, num. 1-2, pp. 42-5, 1978.

[Nsabimana *et al.* 2007] F.X. Nsabimana and U. Zolzer. Transient encoding of audio signals. using dyadic approximations. In Proc. 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France, September 10-15, 2007

[Oppenheim *et al.* 1968] Oppenheim, A.V., R.W. Schafer, and T.G. Stockham, Jr., Non-Linear Filtering of Multiplied and Convolved Signals. Proc. IEEE, 1968. 56(8): p. 1264-1291

[Oppenheim *et al*. 2004] A. Oppenheim and R. Schafer. From frequency to quefrency: A history of the cepstrum. IEEE Signal Processing Magazine, 21(5):95–106, 2004.

[Park 2004] Tae Hong Park "Towards Automatic Musical Instrument Timbre Recognition" Ph.D. Dissertation, Princeton University, NJ, USA (November, 2004)

[Plomp 1976] Plomp, R. 1976. Aspects of Tone Sensation. A Psychophysical study. New York: Academic Press.

[Plumbley *et al.* 2002] Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti,G., and Sandler, M.B.(2002). Automatic music transcription and audio source separation. Cybernetics and Systems , 33(6): 603–627.

[Pollard *et al.* 1982] H. F. Pollard, E. V. Jansson, A tristimulus method for the specification of musical timbre. Acustica, vol. 51. 1982.

[Portnoff 1976] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-24, num. 3, pp. 243-8, 1976.

[Rabiner *et al.* 1976] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, C. A. McGonegal, A comparative performance study of several pitch detection algorithms. IEEE Trans. ASSP, Vol. ASSP - 24, No. 5, October 1976.

[Rabiner *et al.* 1993] Rabiner, L. and Juang, B., 1993, Fundamentals of Speech Recognition, Prentice-Hall.

[Rasch *et al.* 1982] Rasch, R. and Plomp, R. 1982. The Perception of Musical Tones. Psychology of music. Academic Press: New York.

[Richard *et al.* 1992] G. Richard, C. d'Alessandro, and S. Grau. Unvoiced speech synthesis using poissonian random formant wave functions. In Signal Processing VI: European Signal Processing Conference, pages 347–350, 1992.

[Richard *et al.* 1996] G. Richard, C. d'Allesandro, Analysis, Synthesis and modification of the speech aperiodic component, Speech Communication 19, 1996.

[Risset 1965] J. C. Risset, "Computer Study of Trumpet Tones," Journal of the Acoustical Society of America, vol. 33, pp. 912, 1965.

[Rodet et al. 1984.] X. Rodet, Y. Potard, and J.-B. Barri`ere. The CHANT project: From the synthesis of the singing voice to synthesis in general. Computer Music Journal, 8(3):15 31,1984.

[Rodet 1984.] Rodet, X., "Time-Domain Formant-Wave-Function Synthesis." Computer Music Journal 8(3):9-14, 1984.

[Rodet *et al.* 1989] X. Rodet and G. Bennett, Synthesis of the singing voice. Cambridge, MA, USA: MIT Press, 1989.

[Scheirer *et al.* 1997] E. Scheirer and M. Slaney, Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, ICASSP'97, Munich, Vol. II, pp. 1331-1334.

[Shiraishi 2006] Satoshi Shiraishi, masters thesis: "A Real-Time Timbre Tracking Model Based on Similarity" Institute of Sonology Royal Conservatory, The Hague, June 2006

[Scholes 1970] Scholes, P. A. 1970. The Oxford Companion to Music. London: Oxford University Press.

[Schouten 1968] Schouten, J.F. 1968. The Perception of Timbre. In Reports of 6th International Congress on Acoustics, Tokyo, Japan.

[Seashore 1967] Seashore, C.E. 1967. Psychology of Music. (Originally published by McGraw-Hill in 1938, and reprinted) Dover Publications.

[Serra 1989] X. Serra, "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," PhD dissertation, STAN-M-58, Music, CCRMA, Stanford University, 1989.

[Serra 1996] Serra, X. 1996. "Musical Sound Modeling with Sinusoids plus Noise", in G. D. Poli, A. Picialli, S. T. Pope, and C. Roads, editors, Musical Signal Processing. Swets & Zeitlinger Publishers.

[Serra *et al.* 1997] X. Serra, J. Bonada, P. Herrera, R. Loureiro, Integrating complementary spectral models in the design of a musical synthesizer. Proc. of the Int. Comp. Music Conf. 1997.

[Smith 2002] Smith, L. (2002). A tutorial on Principal Components Analysis. online PDF: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[Stevens *et al.* 1940] S. Stevens and J. Volkmann. The relation of pitch to frequency. American Journal of Psychology, 53:329, 1940.

[Tzanetakis *et al.* 2001] Tzanetakis, G., Essl, G., and Cook, P. (2001). Automatic musical genre classification of audio signals. In proceedings ISMIR.

[Verma *et al.* 1997] T. S. Verma, S. N. Levine, and T. H. Y. Meng. "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals". In Proc. International Computer Music Conference, pages 164-167, Thessaloniki, Greece, Sept. 1997. ICMA.

[Yantorno 2000] Yantorno, R. E. 2000. A Study of Spectra Autocorrelation Peak Valley Ratio (SAPVR) as a Method for Identification of Usable Speech and Detection of Cochannel Speech. Final Report for: Summer research faculty program. Speech Processing Lab, EE&CE, Temple University.

[Zölzer *et al.* 2002] Zölzer, U. (Ed). "DAFX: Digital Audio Effects". John Wiley and Sons (2002) - Chapter 8: Arfib, D., Keiler, F. and Zölzer, U., "Source-filter Processing".