Features for Content-Based Audio Retrieval

DALIBOR MITROVIĆ,
MATTHIAS ZEPPELZAUER,
and CHRISTIAN BREITENEDER
Vienna University of Technology,
dalibor.mitrovic@computer.org,
{zeppelzauer | breiteneder}@ims.tuwien.ac.at

Abstract

Today, a large number of audio features exists in audio retrieval for different purposes, such as automatic speech recognition, music information retrieval, audio segmentation, and environmental sound retrieval. The goal of this paper is to review latest research in the context of audio feature extraction and to give an application-independent overview of the most important existing techniques. We survey state-of-the-art features from various domains and propose a novel taxonomy for the organization of audio features. Additionally, we identify the building blocks of audio features and propose a scheme that allows for the description of arbitrary features. We present an extensive literature survey and provide more than 200 references to relevant high quality publications.

Contents

To appear in: Advances in Computers Vol. 78, pp. 71-150, 2010.

1 Introduction

The increasing amounts of publicly available audio data demand for efficient indexing and annotation to enable access to the media. Consequently, content-based audio retrieval has been a growing field of research for several decades. Today, content-based audio retrieval systems are employed in manifold application domains and scenarios such as music retrieval, speech recognition, and acoustic surveillance.

A major challenge during the development of an audio retrieval system is the identification of appropriate content-based features for the representation of the audio signals under consideration. The number of published content-based audio features is too large for quickly getting an overview of the relevant ones. This paper tries to facilitate feature selection by organizing the large set of available features into a novel structure.

Audio feature extraction addresses the analysis and extraction of meaningful information from audio signals in order to obtain a compact and expressive description that is machine-processable. Audio features are usually developed in the context of a specific task and domain. Popular audio domains include audio segmentation, automatic speech recognition, music information retrieval, and environmental/general purpose sound recognition, see Section 6.1. We observe that features originally developed for a particular task and domain are later often employed for other tasks in other domains. A good example are cepstral coefficients, such as Mel-frequency cepstral coefficients (MFCCs, see Section 5.5.1). MFCCs have originally been employed for automatic speech recognition and were later used in other domains such as music information retrieval and environmental sound retrieval as well. Based on these observations, we conclude that audio features may be considered independently from their original application domain.

This paper provides a comprehensive survey on content-based audio features. It differs from other surveys in audio retrieval in the fact that it does not restrict itself to a particular application domain. We bring together state-of-the-art and traditional features from various domains and analyze and compare their properties.

It is nearly impossible to give a complete overview of audio features since they are widely distributed across the scientific literature of several decades. We survey publications in high quality audio and multimedia related journals and conference proceedings. The resulting literature survey covers more than 200 relevant publications. From these publications we select a manifold set of state-of-the-art features. Additionally, we include traditional features that are still competitive. The major criterion for selection is the maximization of heterogeneity between the features in relation to what information they carry and how they are computed. The result is a selection of more than 70 audio features together with references to the relevant literature. We direct the paper towards researchers in all domains of audio retrieval and developers of retrieval systems.

The presented set of audio features is heterogeneous and has no well-defined structure. We develop a taxonomy in order to structure the set of audio features into meaningful groups. The taxonomy groups the audio features by properties, such as the domain they live in, perceptual properties, and computational similarities. It organizes the entire set of selected features into a single structure that is independent of any application domain. This novel organization groups features with similar characteristics from different application domains. The taxonomy represents a toolkit that facilitates the selection of features for a particular task. It further enables the comparison of features by formal and semantic properties.

This paper is organized as follows. We give background information on audio retrieval in Section 2. Characteristics of audio features and the challenges in feature design are discussed in Section 3. Section 4 introduces a novel taxonomy for audio features. We summarize the features in Section 5. Section 6 is devoted to related literature. Finally, we summarize the paper and draw conclusions in Section 7.

2 Background

This section covers different aspects, that may allow for better understanding of the authors' view on content-based audio retrieval and its challenges.

2.1 A Brief Overview on Content-Based Audio Retrieval

There are different fields of research in content-based audio retrieval, such as segmentation, automatic speech recognition, music information retrieval, and environmental sound retrieval which we list in the following. *Segmentation* covers the distinction of different types of sound such as speech, music, silence, and environmental sounds. Segmentation is an important preprocessing step used to identify homogeneous parts in an audio stream. Based on segmentation the different audio types are further analyzed by appropriate techniques.

Traditionally, *automatic speech recognition* focuses on the recognition of the spoken word on the syntactical level [151]. Additionally, research addresses the recognition of the spoken language, the speaker, and the extraction of emotions.

In the last decade *music information retrieval* became a popular domain [40]. It deals with retrieval of similar pieces of music, instruments, artists, musical genres, and the analysis of musical structures. Another focus is music transcription which aims at extracting pitch, attack, duration, and signal source of each sound in a piece of music [86].

Environmental sound retrieval comprises all types of sound that are neither speech nor music. Since this domain is arbitrary in size, most investigations are restricted to a limited domain of sounds. A survey of techniques for feature extraction and classification in the context of environmental sounds is given in [36].

One major goal of content-based audio retrieval is the identification of perceptually similar audio content. This task is often trivial for humans due to powerful mechanisms in our brain. The human brain has the ability to distinguish between a wide range of sounds and to correctly assign them to semantic categories and previously heard sounds. This is much more difficult for computer systems, where an audio signal is simply represented by a numeric series of samples without any semantic meaning.

Content-based audio retrieval is an ill-posed problem (also known as inverse problem). In general, an ill-posed problem is concerned with the estimation of model parameters by the manipulation of observed data. In case of a retrieval task, model parameters are terms, properties and concepts that may represent class labels (e.g. terms like "car" and "cat," properties like "male" and "female," and concepts like "outdoor" and "indoor").

The ill-posed nature of content-based retrieval introduces a *semantic gap*. The semantic gap refers to the mismatch between high-level concepts and low-level descriptions. In content-based retrieval the semantic gap is positioned between the audio signals and the semantics of their contents. It refers to the fact that the same media object may represent several concepts. For example, a recording of Beethoven's Symphony No. 9 is a series of numeric values (samples) for a computer system. On a higher semantic level the symphony is a sequence of notes with specific durations. A human may perceive high-level semantic concepts like musical entities (motifs, themes, movements) and emotions (excitement, euphoria).

Humans bridge the semantic gap based on prior knowledge and (cultural) context. Machines are usually not able to complete this task. Today, the goal of the research community is to narrow the semantic gap as far as possible.

2.2 Architecture of a typical Audio Retrieval System

A content-based (audio) retrieval system consists of multiple parts, illustrated in Figure 1. There are three modules: the input module, the query module, and the retrieval module. The task of the input module is to extract features from audio objects stored in an *audio database* (e.g. a music database). *Feature extraction* aims at reducing the amount of data and extracting meaningful information from the signal for a particular retrieval task. Note that the amount of raw data would be much too big for direct processing. For example, an audio signal in standard CD quality consists of 44100 samples per second for each channel. Furthermore, a lot of information (e.g. harmonics and timbre) is not apparent in the waveform of a signal. Consequently, the raw waveform is often not adequate for retrieval.

The result of feature extraction are parametric numerical descriptions (features) that characterize meaningful information of the input signals. Features may capture audio properties, such as the fundamental frequency and the loudness of a signal. We discuss fundamental audio attributes in Section 2.4. Feature extraction usually reduces the amount of data by several orders of magnitude. The features are extracted once from all objects in the database and stored in a *feature database*.

The user communicates with the retrieval system by formulating queries. There are different types of queries. Usually, the user provides the system with a query that contains one or more audio objects of interest (query by example). Other possibilities are query by humming and query by whistling which are often applied in music retrieval systems. In these approaches, the user has to hum or whistle a melody which is then used as a query object. In both cases the user asks the system to find objects with similar content as that of the query object(s).

After formulation of a query, features are extracted from the query object(s). This is the same procedure as in the input module. The resulting features have

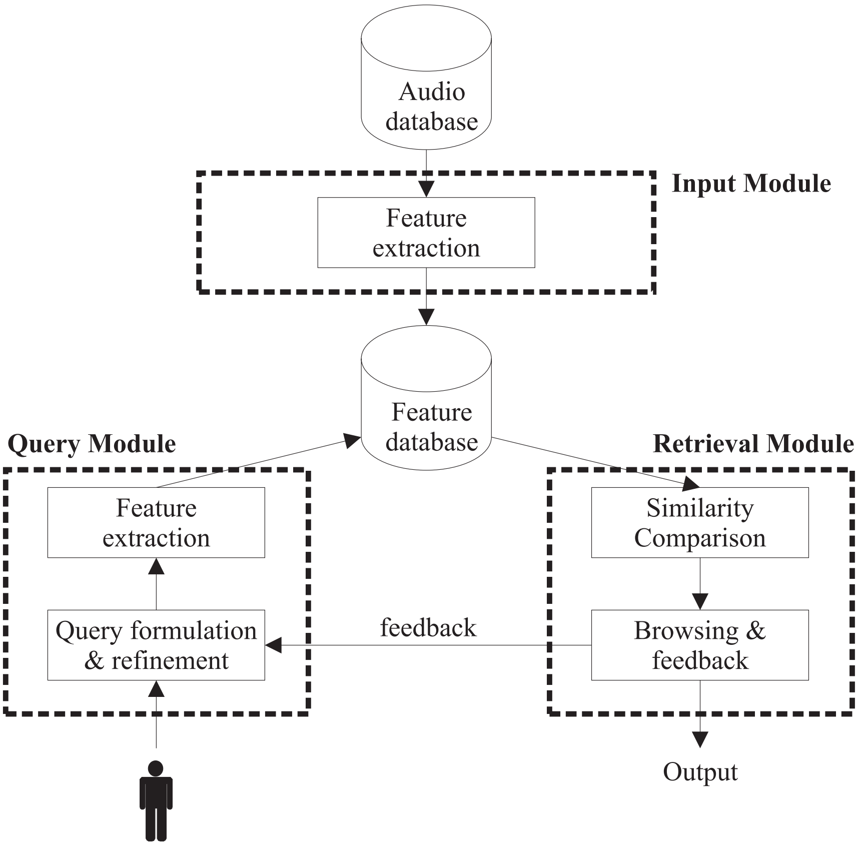


Figure 1: The components of a typical content-based audio retrieval system and their relations.

to be compared to the features stored in the feature database in order to find objects with similar properties. This is the task of the retrieval module.

The crucial step in the retrieval module is *similarity comparison* which estimates the similarity of different feature-based media descriptions. Similarity judgments usually base on distance measurements. The most popular approach in this context is the *vector space model* [158]. The basic assumption of this model is that the numeric values of a feature may be regarded as a vector in a high-dimensional space. Consequently, each feature vector denotes one position in this vector space. Distances between feature vectors may be measured by metrics (e.g. Euclidean metric). Similarity measurement is performed by mapping distances in the vector space to similarities. We expect that similar content is represented by feature vectors that are spatially close in the vector space while dissimilar content will be spatially separated.

Similarity measures derived from distance metrics are only appropriate to a certain degree, since mathematical metrics usually do not fully match the human perception of similarity. The mismatch between perceived similarity and computed similarity often leads to unexpected retrieval results.

After similarity comparison the audio objects that are most similar to the query object(s) are returned to the user. In general, not all returned media objects satisfy the query. Additionally, the query may be imperfect, for example in a query by humming application. Consequently, most retrieval systems offer the user the opportunity to give feedback based on the output of the retrieval process. The user may specify which of the returned objects meet her expectations and which do not (relevance feedback) [92]. This information may be used to iteratively refine the original query. Iterative refinement enables the system to improve the quality of retrieval by incorporating the user's knowledge.

In the following, we mainly focus on the process of feature extraction. Feature extraction is a crucial step in retrieval since the quality of retrieval heavily relies on the quality of the features. The features determine which audio properties are available during processing. Information not captured by the features is unavailable to the system.

For successful retrieval it is necessary that those audio properties are extracted from the input signals that are significant for the particular task. In general, features should capture audio properties that show high variation across the available (classes of) audio objects. It is not reasonable to extract features that capture invariant properties of the audio objects, since they do not produce discriminatory information. Furthermore, in some applications, e.g. automatic speech recognition the features should reflect perceptually meaningful information. This enables similarity comparisons that imitate human perception. In most applications, the features should be robust against signal distortions and interfering noise and should filter components of the signal that are not perceivable by the human auditory system.

In the following, we present three example sound clips together with different (feature) representations in order to show how different features capture different aspects of the signals and how features influence similarity measurements. The three example sounds are all one second long and originate from three different sound sources all playing the musical note A4 (440 Hz). The sources are a tuning fork, a flute, and a violin. Figures 2(a), 2(b), and 2(c) show plots of the sounds' amplitudes over time (also called waveforms). The sound produced by the tuning fork has higher amplitude at the beginning and lower amplitude at the end because it dies out slowly after striking the tuning fork. The flute's sound (hereafter flute) exhibits higher variation of the amplitude because it contains tremolo. The amplitude of the violin's sound (hereafter violin) slowly increases towards the end. Except for the similar range of values the waveforms are not similar at all. Signal properties and similarities can hardly be derived from the waveforms. A much more expressive visualization of sounds is the spectrogram which reveals the distribution of frequencies over time. The spectrogram of the fork sound in Figure 2(d) contains only one strong frequency component at 440 Hz. The spectrograms of flute (Figure 2(e)) and violin (Figure 2(f)) are similar to each other. They exhibit strong frequency components at 440 Hz and contain a large number of harmonics (multiples of the fundamental frequency). In the spectrogram of flute we further observe that the periodic change in amplitude is accompanied by a change in the frequency distribution.

We present two different feature representations of the example sounds and the similarities they reveal in Figure 3. Figures 3(a), 3(b), and 3(c) depict the content-based feature *pitch* which is an estimate of the fundamental frequency of a sound (see Sections 2.4 and 5.4.4). The values of the pitch feature are almost identical for all sounds (approximately at 440 Hz). Considering pitch, the three sounds are extremely similar and cannot be discriminated. However, the three sounds have significantly differing acoustic colors (timbre, see Section 2.4). Consequently, a feature that captures timbral information may be better-suited to discriminate between the different sound sources. Figures 3(d), 3(e), and 3(f) show visualizations of the first 13 *Mel-Frequency Cepstral Coefficients* (MFCCs) which coarsely represent the spectral envelope of the signals for each frame, see Section 5.5.1. We observe, that the three plots vary considerably. For example, the violin's sound has much higher values in the third and fifth MFCC than the fork and the flute. Under consideration of this feature all three sounds are different from each other.

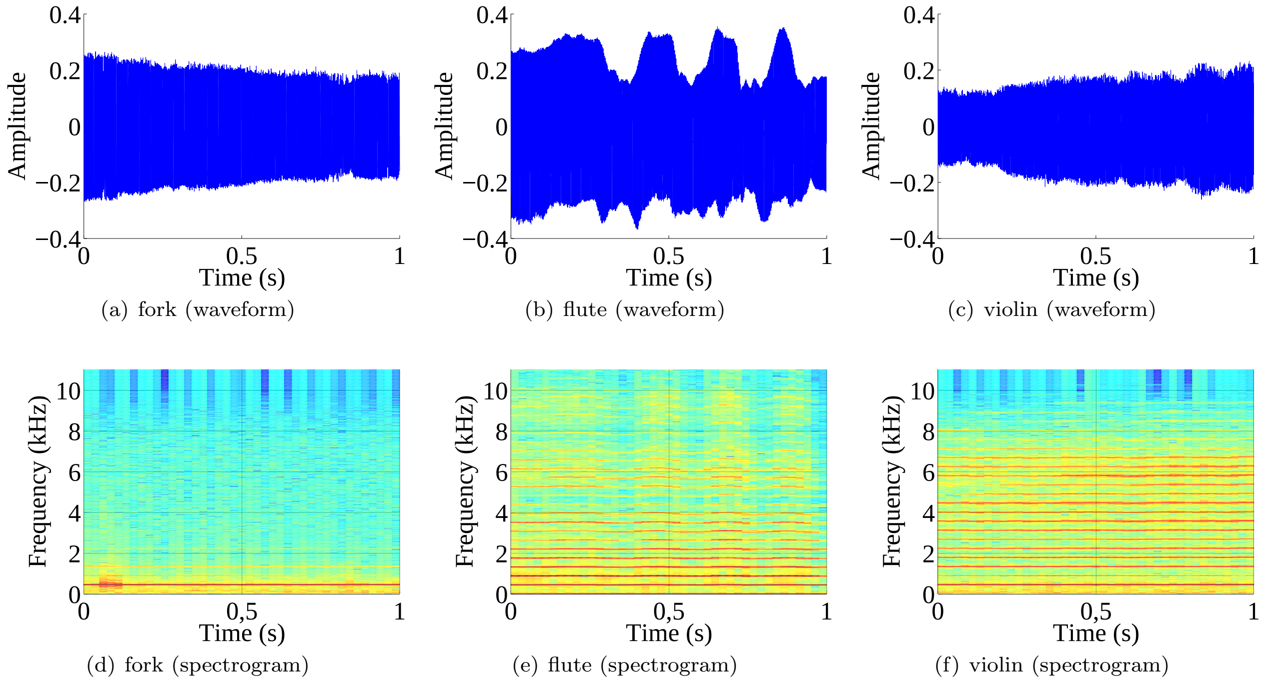
This example demonstrates that different content-based features represent different information and that the retrieval task determines which information is necessary for measuring similarities. For example, pitch is suitable to determine the musical note from a given audio signal (e.g. for automatic music transcription). Classification of sound sources (e.g. different instruments) requires a feature that captures timbral characteristics such as MFCCs.

We conclude that the selection and design of features is a non-trivial task that has to take several aspects into account, such as the particular retrieval task, available data, and physical and psychoacoustic properties. We summarize aspects of feature design in Section 3.

2.3 Objective Evaluation of Audio Retrieval Techniques

An open issue is the evaluation of content-based audio retrieval systems. The results of a retrieval system depend heavily on the input data. Hence, it may happen that a retrieval system is optimized for a specific data set. This may degrade the objectivity of the retrieval results.

The lack of readily available ground truths is an underestimated challenge. There is a need for standardized ground truths in order to objectively evaluate the performance of different retrieval systems. Currently, ground truths are mostly available in the domains of music information retrieval and automatic speech recognition. Due to legal and economic reasons they frequently are not for free. For speech data, high costs are introduced through the necessary transcription by humans. In the domain of music, copyrights constrain the availability of free data. The situation for environmental sounds is even worse. Due to the infinite range of environmental sounds it is difficult to build a representative ground truth. Furthermore, the partition of environmental sounds



(a) fork (waveform)

(b) flute (waveform)

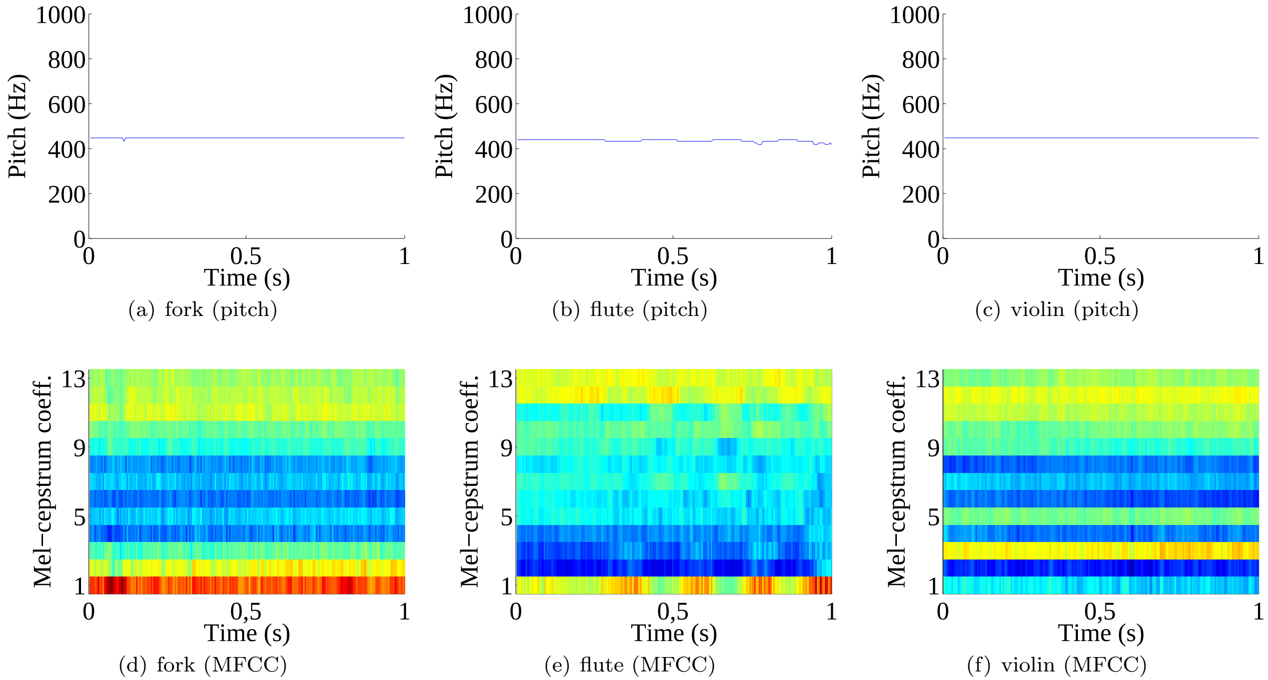
(c) violin (waveform)

(d) fork (spectrogram)

(e) flute (spectrogram)

(f) violin (spectrogram)

Figure 2: Three example sounds from different sound sources: tuning fork, flute, and violin. The first row (a-c) shows their waveforms and the second row (d-f) shows their spectrograms.



(a) fork (pitch)

(b) flute (pitch)

(c) violin (pitch)

(d) fork (MFCC)

(e) flute (MFCC)

(f) violin (MFCC)

Figure 3: Two features (pitch in the first row (a-c) and MFCCs in the second row(d-f)) for the tuning fork, flute, and violin. While all three sounds have similar pitch, their representations in terms of MFCCs differ considerably.

into distinct classes is much more demanding than in the domains of speech and music due to the vast amount of possible sound sources.

Recently, there have been attempts to standardize data and evaluation metrics for music retrieval, for example the audio description contest at the International Conference on Music Information Retrieval in 2004 [71] and the Music Information Retrieval Evaluation eXchange [120]. These contests provide ground truths for free to the participants. According to the authors' knowledge there are no such efforts in the context of environmental sound recognition.

We believe, that a set of freely available benchmarking databases and well-defined performance metrics would promote the entire field of audio retrieval. Additionally, independent domain experts should be employed in the process of building ground truths due to their unbiased view. Even though this leads to a decrease of performance, the objectivity and comparability of the results would improve. Although there are efforts in this direction, more attention has to be turned to standardized and easily available ground truths.

2.4 Attributes of Audio

Audio features represent specific properties of audio signals. Hence, we should briefly discuss the different types of audio signals and the general attributes of audio prior to studying audio features.

Generally, we distinguish between tones and noise. Tones are characterized by the fact that they are "capable of exciting an auditory sensation having pitch" [7] while noise not necessarily has a pitch (see below). Tones may be *pure tones* or *complex tones*. A pure tone is a sound wave where "the instantaneous sound pressure of which is a simple sinusoidal function in time" while a complex tone contains "sinusoidal components of different frequencies" [7].

Complex tones may be further distinguished into *harmonic complex tones* and *inharmonic complex tones*. Harmonic complex tones comprise of partials with frequencies at integer multiples of the fundamental frequency (so called harmonics). Inharmonic complex tones consist of partials whose frequencies significantly differ from integer multiples of the fundamental frequency.

There are different types of noise, distinguished by their temporal and spectral characteristics. Noise may be stationary or non-stationary in time. Stationary noise is defined as "noise with negligibly small fluctuations of level within the period of observation" while non-stationary noise is "noise with or without audible tones, for which the level varies substantially during the period of observation" [7].

The spectral composition of noise is important for its characterization. We distinguish between *broad-band noise* and *narrow-band noise*. Broad-band noise usually has no pitch while narrow-band noise may stimulate pitch perception. Special types of noise are for example *white noise*, which equally contains all frequencies within a band, and *colored noise* where the spectral power distribution is a function of frequency (e.g. pink ($1/f$) noise).

From a psychoacoustic point of view, all types of audio signals may be described in terms of the following attributes: duration, loudness, pitch, and timbre.

Duration is the time between the start and the end of the audio signal of interest. The temporal extent of a sound may be divided into attack, decay, sustain, and release depending on the envelope of the sound. Not all sounds necessarily have all four phases. Note that in certain cases silence (absence of audio signals) may be of interest as well.

Loudness is an auditory sensation mainly related to sound pressure level changes induced by the producing signal. Loudness is commonly defined as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from soft to loud" with the unit *sone* [7].

The American Standards Association defines (spectral) *pitch* as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high" with the unit *mel* [7]. However, pitch has several meanings in literature. It is often used synonymously with the fundamental frequency. In speech processing pitch is linked to the glottis, the source in the source and filter model of speech production. In psychoacoustics, pitch mainly relates to the frequency of a sound but also depends on duration, loudness, and timbre. In the context of this paper, we refer to the psychoacoustic definition.

Additionally, to spectral pitch, there is the phenomenon of *virtual pitch*. The model of virtual pitch has been introduced by Terhardt [175]. It refers to the ability of auditory perception to reproduce a missing fundamental of a complex tone by its harmonics.

An attribute related to pitch is *pitch strength*. Pitch strength is the "subjective magnitude of the auditory sensation related to pitch" [7]. For example, a pure tone produces a stronger pitch sensation than high-pass noise [204]. Generally, the spectral shape determines the pitch strength. Sounds with line spectra and narrow-band noise evoke larger pitch strength than signals with broader spectral distributions.

The most complex attribute of sounds is *timbre*. According to the ANSI standard timbre is "that attribute of auditory sensation which enables a listener to judge that two non-identical sounds, similarly presented and having the same loudness and pitch, are dissimilar." [7]. For example, timbre reflects the difference between hearing sensations evoked by different musical instruments playing the same musical note (e.g. piano and violin).

In contrast to the above mentioned attributes, it has no single determining physical counterpart [3]. Due to the multidimensionality of timbre, objective measurements are difficult. Terasawa et al. propose a method to compare model representations of timbre with human perception [174].

Timbre is a high-dimensional audio attribute and is influenced by both stationary and non-stationary patterns. It takes the distribution of energy in the critical bands into account (e.g. the tonal or noise-like character of sound and its harmonics structure). Furthermore, timbre perception involves any aspect of sound that changes over time (changes of the spectral envelope and temporal characteristics, such as attack, decay, sustain, and release). Preceding and following sounds influence timbre as well.

Each of the attributes duration, loudness, pitch, and pitch strength generally allow for ordering on a unidimensional scale. From a physical point of view, one may be tempted to consider them as independent. Unfortunately, the sensations of these attributes are not independent. In the following, we summarize some relations in order to illustrate the complexity of auditory perception.

Pitch perception is not only affected by the frequency content of a sound, but also by the sound pressure and the waveform [7, 169]. For example, the perceived pitch of sounds with frequencies above approximately 2 kHz increases with rising amplitudes, while sounds below 2 kHz are perceived to have lower pitch when the amplitude increases. Pitch is usually measured using models of the human perception. Evaluation is performed by comparison of the automatic measurements with human assessments.

There are only few sounds that do not have a pitch at all, such as broad-band noise. Non-pitched sounds are for example produced by percussive instruments. Byrd and Crawford list non-pitched sounds as one of the current real world problems in music information retrieval [21].

Pitch strength is related to duration, amplitude, and frequency of a signal. For example, in case of pure tones the pitch strength increases both with the amplitude and the duration. Additionally, it reaches a maximum in the frequency range between 1 and 3 kHz for pure sounds [204].

Loudness is a subjective sensation that does not only relate to the sound pressure but also to the frequency content and the waveform of a signal as well as its duration [7]. Sounds with durations below 100 ms appear less loud than the same sounds with longer durations [204]. Furthermore, loudness sensation varies with the frequency. This relation is described by equal-loudness contours (see Section 3.3.1).

Generally, audio features describe aspects of the above mentioned audio attributes. For example there is a variety of features that aim at representing pitch and loudness. Other features capture particular aspects of timbre, such as sharpness, tonality and frequency modulations. We present the overview of audio features in Section 5.

3 Audio Feature Design

Feature design is an early conceptual phase in the process of feature development. During this process, we first determine what aspects of the audio signal the feature should capture. This is performed in the context of the application domain in question and the specific retrieval task. The next step is the development of a technical solution that fulfills the specified requirements and the implementation of the feature.

In this section, we investigate properties of content-based audio features. Additionally, we analyze the fundamental building-blocks of features from a math-

Property	Values
Signal representation	linear coded, lossily compressed
Domain	temporal, frequency, correlation, cepstral, modulation frequency, reconstructed phase space, eigendomain
Temporal scale	intraframe, interframe, global
Semantic meaning	perceptual, physical
Underlying model	psychoacoustic, non-psychoacoustic

Table 1: The formal properties of audio features and their possible values.

ematically motivated point of view. Finally, we summarize important challenges and problems in feature design.

3.1 Properties of Audio Features

Content-based audio features share several structural and semantical properties that help in classifying the features. In Table 1, we summarize properties of audio features that are most frequently used in literature.

A basic property of a feature is the *audio representation* it is specified for. We distinguish between two groups of features: features based on linear coded signals and features that operate on lossily compressed (subband-coded) audio signals. Most feature extraction methods operate on linear coded signals. However, there has been some research on lossily compressed domain audio features, especially for MPEG audio encoded signals due to their wide distribution. Lossy audio compression transforms the signal into a frequency representation by employing psychoacoustic models which remove information from the signal that is not perceptible to human listeners (e.g. due to masking effects). Although lossy compression has different goals than feature extraction, features may benefit from the psychoacoustically preprocessed signal representation, especially for tasks in which the human perception is modeled. Furthermore, compressed domain features may reduce computation time significantly if the source material is already compressed. Wang et al. provide a survey of compressed domain audio features in [188]. We focus on features for linear-coded audio signals, since they are most popular and form the basis for most lossily compressed domain audio features.

Another property is the *domain* of an audio feature. This is the representation a feature resides in after feature extraction. The domain allows for the interpretation of the feature data and provides information about the extraction process and the computational complexity. For example, a feature in *temporal* domain directly describes the waveform while a feature in *frequency* domain represents spectral characteristics of the signal. It is important to note that we only consider the final domain of a feature and not the intermediate representations during feature extraction. For example, MFCCs are a feature in *cepstral* domain, regardless of the fact that the computation of MFCCs first takes place in frequency domain. We summarize the different domains in Section 3.2.

Another property is the *temporal scale* of a feature. In general, audio is a non-stationary time-dependent signal. Hence, various feature extraction methods operate on short frames of audio where the signal is considered to be locally stationary (usually in the range of milliseconds). Each frame is processed separately (eventually by taking a small number of neighboring frames into account, such as spectral flux) which results in one feature vector for each frame. We call such features *intraframe* features because they operate on independent frames. Intraframe features are sometimes called frame-level, short-time, and steady features [192]. A well known example for an intraframe feature are MFCCs which are frequently extracted for frames of 10-30 ms length.

In contrast, *interframe* features describe the temporal change of an audio signal. They operate on a larger temporal scale than intraframe features in order to capture the dynamics of a signal. In practice, interframe features are often computed from intraframe representations. Examples for interframe features are features that represent rhythm and modulation information (see Section 5.6). Interframe features are often called long-time features, global features, dynamic features, clip-level features and contour features [179, 192].

In addition to interframe and intraframe features, there are *global* features. According to Peeters a global feature is computed for the entire audio signal. An example is the attack duration of a sound. However, a global feature does not necessarily take the entire signal into account [101].

The *semantic interpretation* of a feature indicates whether or not the feature represents aspects of human perception. *Perceptual* features approximate semantic properties known by human listeners, e.g. pitch, loudness, rhythm, and harmonicity [201]. Additionally to perceptual features, there are *physical* features. Physical features describe audio signals in terms of mathematical, statistical, and physical properties without emphasizing human perception in the first place (e.g. Fourier transform coefficients and the signal energy).

We may further distinguish features by the type of the *underlying model*. In recent years, researchers incorporated psychoacoustic models into the feature extraction process in order to improve the information content of the features and to approximate human similarity matching [156]. Psychoacoustic models for example incorporate filter banks that simulate the frequency resolution of the human auditory system. Furthermore, these models consider psychoacoustic properties, such as masking, specific loudness sensation, and equal-loudness contours, see Section 3.3.1. Investigations show that retrieval results often benefit from features that model psychoacoustical properties [51, 63, 156, 173]. In the context of this work, we distinguish between *psychoacoustic* and *non-psychoacoustic* features.

Each audio feature can be characterized in terms of the above mentioned properties. We employ several of these properties in the design of the taxonomy in Section 4.

3.2 Building Blocks of Features

In this section, we analyze the mathematical structure of selected features and identify common components (building blocks). This approach offers a novel perspective on content-based audio features that reveals their structural similarities.

We decompose audio features into a sequence of basic mathematical operations similarly to Mierswa and Morik in [118]. We distinguish between three basic groups of functions: transformations, filters, and aggregations. *Transformations* are functions that map data (numeric values) from one domain into another domain. An example for a transformation is the discrete Fourier transform that maps data from temporal domain into frequency domain and reveals the frequency distribution of the signal. It is important that the transformation from one domain into the other changes the interpretation of the data. The following domains are frequently used in audio feature extraction.

Temporal domain. The temporal domain represents the signal changes over time (the waveform). The abscissa of a temporal representation is the sampled time domain and the ordinate corresponds to the amplitude of the sampled signal. While this domain is the basis for feature extraction algorithms the signals are often transformed into more expressive domains that are better suited for audio analysis.

Frequency domain. The frequency domain reveals the spectral distribution of a signal and allows for example the analysis of harmonic structures, bandwidth, and tonality. For each frequency (or frequency band) the domain provides the corresponding magnitude and phase. Popular transformations from time to frequency domain are Fourier- (DFT), Cosine- (DCT), and Wavelet transform. Another widely-used way to transform a signal from temporal to frequency domain is the application of banks of band-pass filters with e.g. Mel- and Bark-scaled filters to the time domain signal. Note that Fourier-, Cosine-, and Wavelet transforms may also be considered as filter banks.

Correlation domain. The correlation domain represents temporal relationships between signals. For audio features especially the *auto*correlation domain is of interest. The autocorrelation domain represents the correlation of a signal with a time-shifted version of the same signal for different time lags. It reveals repeating patterns and their periodicities in a signal and may be employed, for example for the estimation of the fundamental frequency of a signal.

Cepstral domain. The concept of cepstrum has been introduced by Bogert et al. in [16]. A representation in cepstral domain is obtained by taking the Fourier transform of the logarithm of the magnitude of the spectrum . The second Fourier transform may be replaced by the inverse DFT, DCT, and inverse DCT. The Cosine transform better decorrelates the data than the Fourier transform and thus is often preferred. A cepstral representation is one way to compute an approximation of the shape (envelope) of the spectrum. Hence, cepstral features usually capture timbral information [174]. They are frequently applied in automatic speech recognition and audio fingerprinting.

Modulation frequency domain. The modulation frequency domain reveals information about the temporal modulations contained in a signal. A typical representation is the joint acoustic and modulation frequency graph which represents the temporal structure of a signal in terms of low-frequency amplitude modulations [173]. The abscissa represents modulation frequencies and the ordinate corresponds to acoustic frequencies. Another representation is the modulation spectrogram introduced by Greenberg and Kingsbury in [55] which displays the distribution of slow modulations across time and frequency. Modulation information may be employed for the analysis of rhythmic structures in music [139] and noise-robust speech recognition [55, 84].

Reconstructed phase space. Audio signals such as speech and singing may show non-linear (chaotic) phenomena that are hardly represented by the domains mentioned so far. The non-linear dynamics of a system may be reconstructed by embedding the signal into a phase space. The reconstructed phase space is a high-dimensional space (usually $d > 3$), where every point corresponds to a specific state of the system. The reconstructed phase space reveals the attractor of the system under the condition that the embedding dimension d has been chosen adequately. Features derived from the reconstructed phase space may estimate the degree of chaos in a dynamic system and are often applied in automatic speech recognition for the description of phonemes [1, 100].

Eigendomain. We consider a representation to be in eigendomain if it is spanned by eigen- or singular vectors. There are different transformations and decompositions that generate eigendomains in this sense, such as Principal Components Analysis (PCA) and Singular Value Decomposition (SVD). The (statistical) methods have in common that they decompose a mixture of variables into some canonical form, for example uncorrelated principal components in case of the PCA. Features in eigendomain have decorrelated or even statistically independent feature components. These representations enable easy and efficient reduction of data (e.g. by removing principal components with low eigenvalues).

Additionally to transformations, we define *filters* as the second group of operators. In the context of this paper, we define a filter as a mapping of a set of numeric values into another set of numeric values residing in the *same* domain. In general, a filter changes the values of a given numeric series but not their number. Note that this definition of the term filter is broader than the definition usually employed in signal processing.

Simple filters are for example scaling, normalization, magnitude, square, exponential function, logarithm, and derivative of a set of numeric values. Other filters are quantization and thresholding. These operations have in common that they reduce the range of possible values of the original series.

We further consider the process of windowing (framing) as a filter. Windowing is simply the multiplication of a series of values with a weighting (window) function where all values inside the window are weighted according to the function and the values outside the window are set to zero. Windowing may be applied for (non-)uniform scaling and for the extraction of frames from a signal (e.g. by repeated application of hamming windows).

Similarly, there are low-pass, high-pass, and band-pass filters. Filters in the domain of audio feature extraction are often based on Bark- [203], ERB- [126], and Mel-scale [171]. We consider the application of a filter (or a bank of filters) as a filter according to our definition, if the output of each filter is again a series of values (the subband signal). Note that a filter bank may also represent a transformation. In this case the power of each subband is aggregated over time, which results in a spectrum of a signal. Consequently, a filter bank may be considered as both, a filter and a transformation, depending on its output.

The third category of operations are *aggregations*. An aggregation is a mapping of a series of values into a single scalar. The purpose of aggregations is the reduction of data, e.g. the summarization of information from multiple subbands. Typical aggregations are mean, variance, median, sum, minimum, and maximum. A more comprehensive aggregation is a histogram. In this case each bin of the histogram corresponds to one aggregation. Similarly, binning of frequencies (e.g. spectral binning into Bark- and Mel bands) is an aggregation.

A subgroup of aggregations are *detectors*. A detector reduces data by locating distinct points of interest in a value series, e.g. peaks, zero crossings, and roots.

We assign each mathematical operation that occurs during feature extraction to one of the three proposed categories (see Section 5.1). These operations form the building blocks of features. We are able to describe the process of computation of a feature in a very compact way, by referring to these building blocks. As we will see, the number of different transformations, filters, and aggregations employed in audio feature extraction is relatively low, since most audio features share similar operations.

3.3 Challenges in Features Design

The task of feature design is the development of a feature for a specific task under consideration of all interfering influences from the environment and constraints defined by the task. Environmental influences are interfering noise, concurrent sounds, distortions in the transmission channel, and characteristics of the signal source. Typical constraints are for example the computational complexity, dimension, and statistical properties and the information carried by the feature. Feature design poses various challenges to the developer. We distinguish between psychoacoustic, technical, and numeric challenges.

3.3.1 Psychoacoustic Challenges

Psychoacoustics focuses on the mechanisms that process an audio signal in a way that sensations in our brain are caused. Even if the human auditory system has been extensively investigated in recent years, we still do not fully understand all aspects of auditory perception.

Models of psychoacoustic functions play an important role in feature design. Audio features incorporate psychoacoustic properties in order to simulate human perception. Psychoacoustically enriched features enable similarity measurements that correspond to some degree to the human concepts of similarity.

We briefly describe the function of the human ear, before we present some aspects of psychoacoustics. The human ear comprises of three sections: the outer ear, the middle ear, and the inner ear. The audio signal enters the outer ear at the pinna, travels down the auditory canal, and causes the ear drum to vibrate. The vibrations of the ear drum are transmitted to the tree bones of the middle ear (Malleus, Incus, and Stapes) which in turn transmit the vibrations to the cochlea. The cochlea in the inner ear performs a frequency-to-place conversion. A specific point on the basilar membrane inside the cochlear is excited, depending on the frequency of the incoming signal. The movement of the basilar membrane stimulates the hair cells which are connected to the auditory nerve fibers. The inner hair cells transform the hydromechanical vibration into action potentials while the outer hair cells actively influence the vibrations of the basilar membrane. The outer hair cells receive efferent activity from the higher centers of the auditory system. This feedback mechanism increases the sensitivity and frequency resolution of the basilar membrane [124]. In the following, we summarize important aspects of auditory perception that are often integrated into audio features.

Frequency selectivity. The frequency resolution of the basilar membrane is higher at low frequencies than at high frequencies. Each point on the basilar membrane may be considered as a band-pass filter (auditory filter) with a particular bandwidth (critical bandwidth) and center frequency. We refer the reader to [204] and [124] for a comprehensive introduction to the frequency selectivity of the human auditory system.

In practice, a critical band spectrum is obtained by the application of logarithmically scaled band-pass filters where the bandwidth increases with center frequency. Psychoacoustical scales, such as Bark- and ERB-scale are employed to approximate the frequency resolution of the basilar membrane [125, 203].

Auditory masking. Masking is "the process by which the threshold of hearing for one sound is raised by the presence of another (masking) sound" [7]. The amount of masking is expressed in decibels. We distinguish between simultaneous masking and temporal masking. Simultaneous masking is related to the frequency selectivity of the human auditory system. One effect is that when two spectral components of similar frequency occur simultaneously in the same critical band, the louder sound may mask the softer sound [193]. Spectral masking effects are implemented for the computation of loudness for example in [139].

In temporal masking, the signal and the masker occur consecutively in time. This means for example that a loud (masking) sound may decrease the perceived loudness of a preceding sound. We distinguish between forward masking (also post-stimulus masking) which refers to a "condition in which the signal appears after the masking sound" and backward masking (also pre-stimulus masking) where the signal appears before the masking sound [7].

Loudness levels. The loudness of sinusoids is not constant over all frequencies. The loudness of two tones of same sound pressure level but different frequency varies [47]. Standardized equal-loudness contours relate tones of different frequencies and sound pressure levels to loudness levels (measured in phon) [72]. Figure 4 shows equal-loudness contours for different loudness levels. Pfeiffer presents a method to approximate loudness by incorporating equal-loudness contours in [144].

Psychophysical power law. According to Stevens, the loudness is a power function of the physical intensity [170]. A tenfold change in intensity (interval of 10 phons) approximately results in a twofold change in loudness. The unit of loudness is sone, where 1 sone is defined as the loudness of a pure 1000 Hz tone at 40 dB sound pressure level (40 phon). Figure 4 shows the phon and corresponding sone values of several equal-loudness contours.

In many domains psychoacoustically motivated features have shown to be superior to features that do not simulate auditory perception, for example in automatic speech recognition [51], fingerprinting [173], and audio classification [156].

3.3.2 Technical Challenges

An audio signal is usually exposed to distortions, such as interfering noise and channel distortions. Techniques robust to a wide range of distortions have been proposed for example in [5, 20]. Important factors are:

Noise. Noise is present in each audio signal and is usually an unwanted component that interferes with the signal. Thermal noise is always introduced during capturing and processing of signals by analog devices (microphones, amplifiers, recorders) due to thermal motion of charge carriers. In digital systems additional noise may be introduced through sampling and quantization. These types of noise are often neglected in audio retrieval.

More disturbing are background noise and channel distortions. Some characteristics of noise have already been summarized in Section 2.4. Additionally, noise may be characterized by the way it is embedded into the signal. The simplest case, is additive noise. A more complicated case is convolutional noise, usually induced by the transmission channel. Generally, noise is considered to be independent from the signal of interest, however, this is not true in all

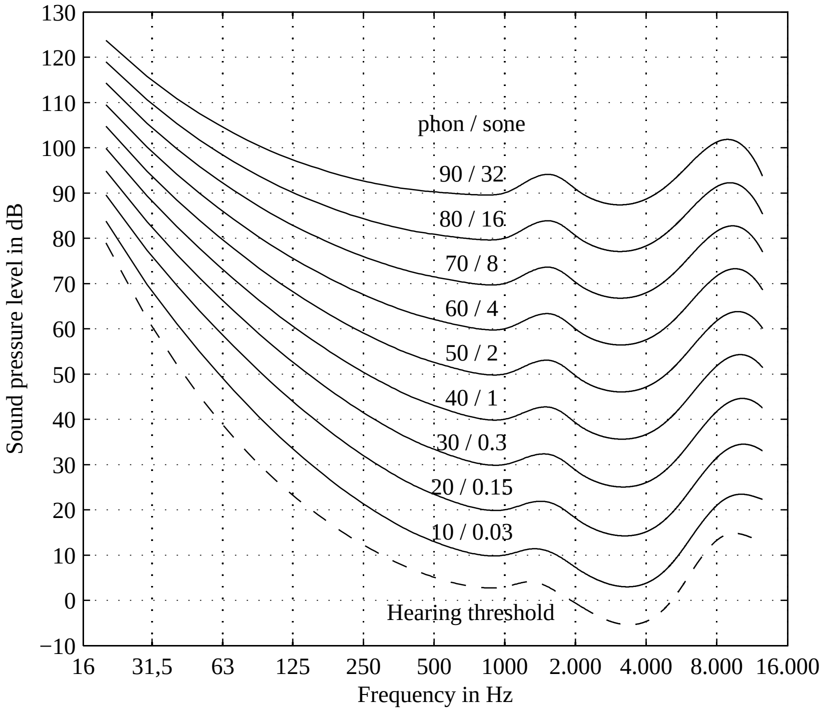


Figure 4: The solid lines are the equal-loudness contours for 10 to 90 phons as specified by the ISO 226 standard. Additionally, the corresponding sone values are given. The dashed line is the threshold of hearing. We are most sensitive to frequencies around 2 kHz and 5 kHz.

situations. Noise robustness is one of the main challenges in audio feature design [156, 164, 199].

Sound pressure level (SPL) variations. For many retrieval tasks it is desired that an audio feature is invariant to the SPL of the input signal (except for features that are explicitly designed to measure loudness, see Section 5.4.3). For example, in automatic speech recognition, an utterance at different SPLs should ideally yield the same feature-based representation.

Tempo variations. In most application domains uncontrolled tempo variations decrease retrieval performance. For example, in music similarity retrieval one is interested in finding all interpretations of a piece of music independent of their respective tempos. A challenge in feature design is to create audio descriptions that are invariant against temporal shifts and distortions. Therefore, it is important to maintain the original frequency characteristics [130, 173].

Concurrency. Concurrent audio signals (background noise and reverberation) pose problems to feature extraction. In many situations the audio signal contains components of more than one signal source, e.g. multiple instruments or a mixture of environmental sounds. It is difficult (and generally impossible) to filter all unwanted portions from the composite signal.

Available resources. Finally, the computational complexity of an audio feature is a critical factor especially in real-time applications. While feature extraction on standard PCs is often possible in real-time, applications on mobile devices, such as PDAs and mobile phones pose novel challenges to efficient feature extraction.

3.3.3 Numeric Challenges

The result of feature extraction is a numeric feature vector that represents particular aspects of the underlying signal. The feature vector should fulfill a number of statistical and numeric requirements depending on the employed classifier and similarity/distance measure. In the following, we summarize the most important statistical and numeric properties.

Compactness. This property refers to the dimensionality of the feature vector. A compact representation is desired in order to decrease the computational complexity of subsequent calculations.

Numeric range. The components of a feature vector should be in the same numeric range in order to allow for comparisons of the components. Different numeric ranges of components in the same vector may lead to unwanted bias in following similarity judgements (depending on the employed classifier and distance metric). Therefore, normalization may be applied after feature extraction.

Completeness. A feature should be able to completely cover the range of values of the property it describes. For example, a feature that describes the pitch of an audio signal, should cover the entire range of possible pitches.

Redundancy. The correlation between components of a feature vector is an indicator for its quality. The components of a feature vector should be decorrelated in order to maximize the expressive power. We find features with decorrelated components especially in the cepstral- and eigendomain (see Sections 5.5 and 5.7).

Discriminant power. For different audio signals, a feature should provide different values. A measure for the discriminant power of a feature is the variance of the resulting feature vectors for a set of input signals. Given different classes of similar signals, a discriminatory feature should have low variance inside each class and high variance over different classes.

Sensitivity. An indicator for the robustness of a feature is the sensitivity to minor changes in the underlying signal. Usually, low sensitivity is desired in order to remain robust against noise and other sources of irritation.

In general, it is not possible to optimize all mentioned properties simultaneously, because they are not independent from each other. For example, with increasing discriminant power of a feature, its sensitivity to the content increases as well which in turn may reduce noise robustness. Usually, tradeoffs have to be found in the context of the particular retrieval task.

4 A novel Taxonomy for Audio Features

Audio features describe various aspects and properties of sound and form a versatile set of techniques that has no inherent structure. One goal of this paper is to introduce some structure into this field and to provide a novel, holistic perspective. Therefore, we introduce a taxonomy that is applicable to general purpose audio features independent from their application domain.

A taxonomy is an organization of entities according to different principles. The proposed taxonomy organizes the audio features into hierarchical groups with similar characteristics. There is no single, unambiguous and generally applicable taxonomy of audio features, due to their manifold nature. A number of valid and consistent taxonomies exist. Usually, they are defined with particular research fields in mind. Hence, most of them are tailored to the needs of these particular fields which diminishes their general applicability.

We want to point out some issues related to the design of a taxonomy by discussing related approaches. Tzanetakis proposes a categorization for audio features in the domain of music information retrieval in [179]. The author employs two organizing principles. The first principle corresponds to computational issues of a feature, e.g. *Wavelet transform features*, *short-time Fourier transform-based features*. The second principle relates to qualities like texture, timbre, rhythm, and pitch. This results in groups of features that either are computed similarly or describe similar audio qualities.

Two groups in this categorization are remarkable. There is a group called *other features* that incorporates features that do not fit into any other group. This reflects the difficulties associated with the definition of a complete and clear taxonomy. The other remarkable group is the one named *musical content features*. This group contains combinations of features from the other groups and cannot be regarded to be on the same structural level as the other groups. Tzanetakis' categorization is appropriate for music information retrieval [98]. However, it is too coarse for a general application in audio retrieval.

Peeters promotes four organizing principles for the categorization of audio features in [142]. The first one relates to the *steadiness or dynamicity* of a feature. The second principle takes the *time extent* of a feature into account. The third principle is the *abstractness* of the representation resulting from feature extraction. The last organizing principle is the *extraction process* of the feature. Peeters describes an organization that is better suited for general use, though we believe a more systematic approach is needed.

We have identified several principles that allow for classification of audio features inspired by existing organizations and the literature survey presented in Section 6. Generally, these principles relate to feature properties, such as the domain, the carried information (semantic meaning), and the extraction process. The selection of organizing principles is crucial to the worth of a taxonomy. There is no broad consensus on the allocation of features to particular groups, e.g. Lu et al. [109] regard zero crossing rate (ZCR) as a perceptual feature, whereas Essid et al. [43] assign ZCR to the group of temporal features. This lack of consensus may stem from the different viewpoints of the authors.

Despite the aforementioned difficulties, we propose a novel taxonomy, that aims at being generally applicable. The taxonomy follows a method-oriented approach that reveals the internal structure of different features and their similarities. Additionally, it facilitates the selection of features for a particular task. In practice, the selection of features is driven by factors such as computational constraints (e.g. feature extraction on (mobile) devices with limited capabilities) or semantic issues (e.g. features describing rhythm). The proposed taxonomy is directed towards these requirements.

We believe that a taxonomy of features has to be as fine-grained as possible in order to maximize the degree of introduced structure. However, at the same time the taxonomy should maintain an abstract view in order to provide groups with semantic meaning. We aim at providing a tradeoff between these conflicting goals in the proposed taxonomy.

We assign features to groups in a way that avoids ambiguities. However, we are aware that even with the proposed organizing principles, certain ambiguities will remain. Generally, the number of computationally and conceptually valid views of features, renders the elimination of ambiguities impossible.

The proposed taxonomy has several levels. On the highest level, we distinguish features by their *domain* as specified in Section 3.1. This organizing principle is well-suited for the taxonomy, since each feature resides in one distinct domain. The domains employed for the taxonomy are presented in Section 3.2.

Figure 5 depicts the groups of the first level of the taxonomy. Note that we group features from frequency domain and from autocorrelation domain into the same group of the taxonomy (named frequency domain) since both domains represent similar information. The frequency domain represents the frequency distribution of a signal while the autocorrelation domain reveals the same frequencies (periodicities) in terms of time lags.

The domain a feature resides in reveals the basic meaning of the data represented by that feature e.g. whether or not it represents frequency content. Additionally, it allows the user to coarsely estimate the computational com-

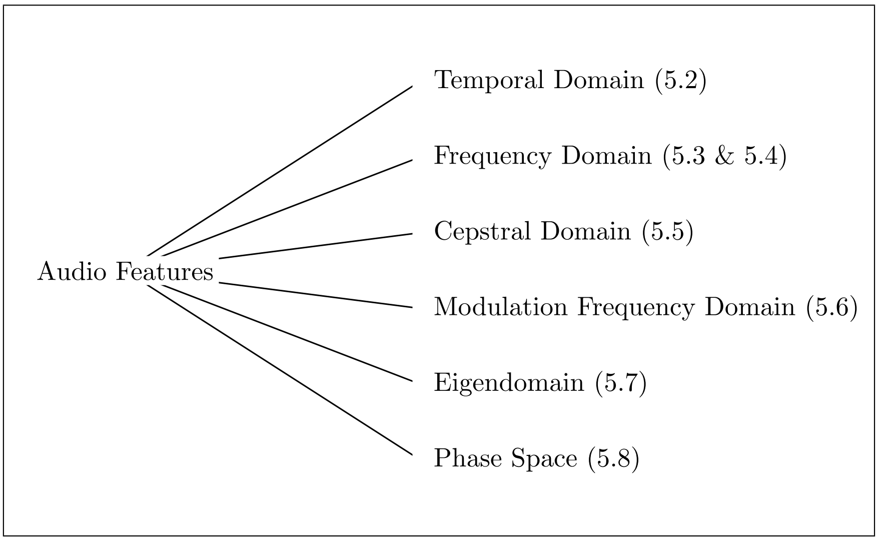


Figure 5: The first level of the proposed taxonomy. The organizing principle is the domain the features reside in. In brackets a reference to the section containing the corresponding features is given.

plexity of a feature. It further provides information on the data quality, such as statistical independence of the feature components.

On the next level, we apply organizing principles based on computational and semantic concepts. Inside one domain we consistently categorize features according to the property that structures them best. The structure of the *temporal domain* bases on what aspect of the signal the feature represents. In the temporal domain, depicted in Figure 6, we distinguish between 3 groups of features: amplitude-based, power-based, and zero crossing-based features. Each group contains features related to a particular physical property of the waveform.

For the frequency domain we propose a deeper hierarchy due to the diversity of the features that live in it. We introduce a semantic layer that divides the set of features into two distinct groups. One group are *perceptual* features and the other group are *physical* features. Perceptual features represent information that has a semantic meaning to a human listener, while physical features describe audio signals in terms of mathematical, statistical, and physical properties of the audio signal (see Section 3.1). We believe that this layer of the taxonomy supports clarity and practicability.

We organize the perceptual features according to semantically meaningful aspects of sound. These aspects are: brightness, chroma, harmonicity, loudness, pitch, and tonality. Each of these properties forms one subgroup of the perceptual frequency features (see Figure 7). This structure facilitates the selection of audio features for particular retrieval tasks. For example, if the user needs to ex-

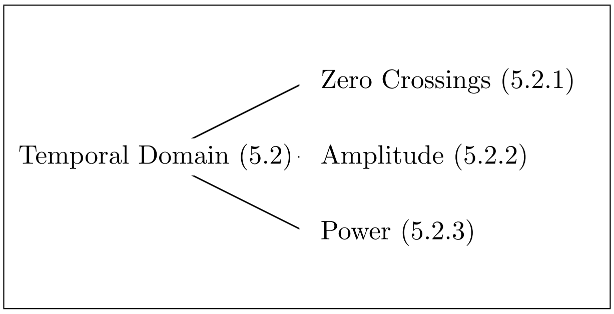


Figure 6: The organization of features in the temporal domain relates to physical properties of the signal. In brackets a reference to the section containing the corresponding features is given.

tract harmonic content, the taxonomy makes identification of relevant features an easy task.

Note that we do not employ timbre as a semantic category in the taxonomy because of its versatile nature. Its many facets would lead to an agglomeration of diverse features into this group. Many audio features represent one or more facets of timbre. In this taxonomy features that describe timbral properties are distributed over several groups.

A semantic organization of the physical features in the frequency domain is not reasonable, since physical features do not explicitly describe semantically meaningful aspects of audio. We employ a mathematically motivated organizing principle for physical features. The features are grouped according to their extraction process. We distinguish between features that are based on autoregression, adaptive time-frequency decomposition (e.g. Wavelet transform), and short-time Fourier transform. Features that base on short-time Fourier transform may be further separated into features that take the complex part into account (phase) and features that operate on the real part (envelope) of the spectrum.

Similarly to the physical features in the frequency domain, we organize the features in the cepstral domain. Cepstral features have in common that they approximate the spectral envelope. We distinguish between cepstral features by differences in their extraction process.

Figure 8 illustrates the structure of the cepstral domain. The first group of cepstral features employs critical band filters, features in the second group incorporate advanced psychoacoustic models during feature extraction and the third group applies autoregression.

Modulation frequency features carry information on long-term frequency modulations. All features in this domain employ similar long-term spectral analyses. A group of features we want to emphasize are rhythm-related features, since they represent semantically meaningful information. Consequently, these features form a subgroup in this domain.

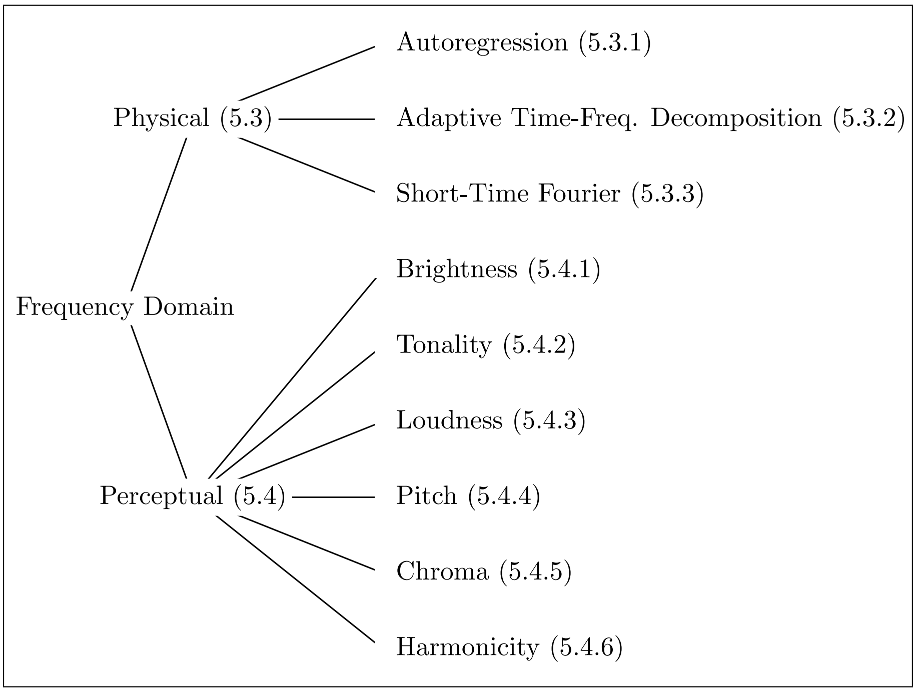


Figure 7: The organization of features in the frequency domain relates to physical and semantic properties of the signal. In brackets a reference to the section containing the corresponding features is given.

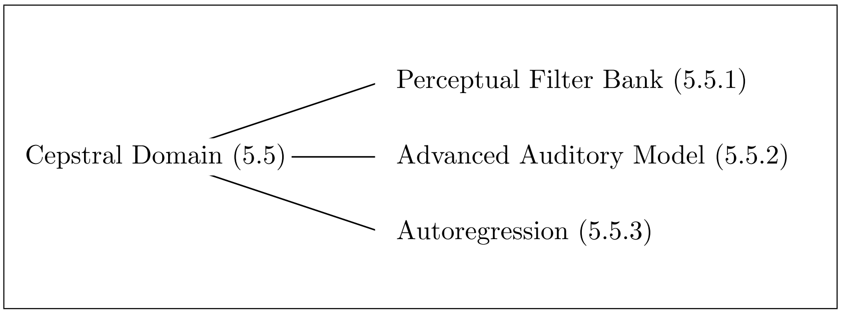

Perceptual Filter Bank (5.5.1)
Cepstral Domain (5.5)
Advanced Auditory Model (5.5.2)
Autoregression (5.5.3)

Figure 8: The organization of features in the cepstral domain relates to the computational properties of the features. In brackets a reference to the section containing the corresponding features is given.

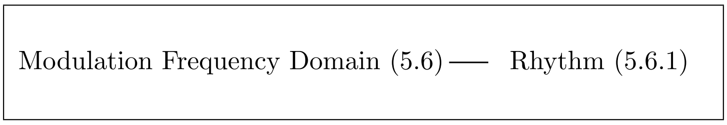


Figure 9: The organization of features in the modulation frequency domain. We group features that relate to rhythmic content into a separate semantic group. In brackets a reference to the section containing the corresponding features is given.

The remaining domains of the first level of the taxonomy are *eigendomain* and *phase space*. We do not further subdivide these domains, since the taxonomy does not profit from further subdivision. A further partition of the domains would decrease the general applicability of the taxonomy.

The taxonomy allows for the selection of features by the information the features carry (e.g. harmonic and rhythm-related features) as well as by computational criteria (e.g. temporal features). We believe that the taxonomy groups features in a way that makes it generally applicable to all areas of audio retrieval and demands only a small number of tradeoffs.

5 Audio Features

In the previous section, we have introduced a taxonomy that represents a hierarchy of feature groups that share similar characteristics. We investigate more than 70 state-of-the-art and traditional audio features from an extensive literature survey. In the following, we briefly present each audio feature in the context of the taxonomy. The sections and subsections reflect the structure of the taxonomy. We describe important characteristics of the features and point out similarities and differences. Before we describe the features in more detail, we give an overview of all covered features and introduce a compact notation for describing the feature extraction process. We compactly present properties of the features, such as the extraction process, domain, temporal structure, application domain, complexity etc. A tabular representation gives the reader the opportunity to structurally compare and survey all features.

5.1 Overview

Before we present the tables containing the properties of the features, we introduce a notation, that allows for the compact representation of the extraction process of a feature. In Section 3.2 we have introduced three groups of mathematical operations that are usually employed in audio feature extraction: transformations, filters, and aggregations. We identify the most important operators belonging to these categories by analyzing the features covered in this paper. The resulting sets of transformations, filters, and aggregations are listed

transformations	
A	Autocorrelation
R	Cross-Correlation
B	Band-pass Filter Bank
F	Discrete Fourier Transform (DFT)
C	(Inverse) Discrete Cosine Transform (DCT/IDCT)
Q	Constant Q Transform (CQT)
M	Modulated Complex Lapped Transform (MCLT)
V	Adaptive Time Frequency Transform (ATFT)
W	Discrete Wavelet (Packet) Transform (DW(P)T)
E	Phase Space Embedding
I	Independent Component Analysis (ICA)
P	(Oriented) Principal Component Analysis ((O)PCA)
S	Singular Value Decomposition (SVD)

Table 2: Frequent transformations employed in audio features and their symbols (upper-case letters, left).

in Tables 2, 3, and 4. We arrange similar operations into groups by horizontal bars in order to improve understanding and readability.

In the tables, we assign a character to each operation as an abbreviation. Transformations are abbreviated by upper-case Latin characters and filters by lower-case Latin characters. We assign Greek characters (lower- and upper-case) to aggregations. We observe that the number of identified operations (building blocks) is relatively small, considering, that they originate from the analysis of more than 70 different audio features.

The process of computation of a feature may be described as a sequence of the identified operations. We introduce a *signature* as a compact representation that summarizes the computational steps of the extraction process of a feature. A signature is a sequence of transformations, filters, and aggregations represented by the previously assigned symbols in Tables 2, 3, and 4. The characters are arranged from left to right in the order the corresponding operations are performed during feature extraction.

We demonstrate the composition of a signature by means of the well-known MFCC feature [18]. MFCCs are usually computed as follows. At first the Fourier transform of the windowed input signal is computed (a short-time Fourier transform). Then a Mel-filter bank, consisting of logarithmically positioned triangular band-pass filters is applied. After taking the logarithm of the magnitude of the band-pass filtered amplitudes, the Cosine transform is taken in order to obtain MFCCs.

We can easily construct the corresponding signature for MFCCs by selecting the necessary building blocks from Tables 2, 3, and 4. First, a single frame ("f") of the input signal is extracted and a Fourier transform ("F") is performed. Then spectral binning of the Fourier coefficients is performed to obtain the responses of the Mel-filters ("β"). Taking the logarithm corresponds to "l" and

Filters

b	Band-pass Filter (Bank)
c	Comb Filter (Bank)
o	Low-pass Filter
f	Framing / Windowing
w	(Non-) Linear Weighting Function
d	Derivation, Difference
e	Energy Spectral Density
g	Group Delay Function
l	Logarithm
x	Exponential Function
n	Normalization
a	Autoregression (Linear Prediction Analysis)
r	Cepstral Recursion Formula

Table 3: Frequent filters employed in audio features and their symbols (lower-case letters, left).

the completing Cosine transform matches "C". The resulting sequence for the MFCC feature is "f F β l C".

Additionally to transformations, filters, and aggregations, the signatures may contain two structural elements: Parenthesis and Brackets. Parenthesis indicate optional operations. We apply parenthesis in cases where different definitions of a feature exist in order to express that more than one computation is possible. Brackets label operations that are repeated for several (two or more) audio frames. For example, in the signature of MPEG-7 temporal centroid "[f ϖ] μ" the brackets indicate that the mean operator is applied to several root-mean-squared frames.

We construct signatures for all features in order to enable a *structural* comparison of the features and present them together with other properties in Tables 5, 6, and 7. The tables organize the features according to the taxonomy. The first column presents the domain of the features (which is the first level of the taxonomy). The second column contains references to the sections where the corresponding features are presented (each section covers a sub group of the taxonomy).

For each feature we specify its temporal scale: "I," "X," and "G" denote intraframe, interframe, and global features, respectively (see Section 3.1). "Y" and "N" in column "perceptual" indicate whether or not a feature is perceptual. The same is done in the column "psychoacoustic model." Furthermore, we rate the computational complexity of each feature ("L," "M," and "H" denote low, medium, and high). The next column lists the proposed dimension of the feature vectors. The character "V" indicates that the dimension of a features is parameterized (variable). Additionally, we list the "application domain" where the feature is mostly used. The abbreviation "ASR" stands for automatic speech recognition, "ESR" is environmental sound recognition, "MIR" is music

	Aggregations and Detectors
χ	Maximum
ι	Minimum
μ	Mean (weighted, arithmetic, geometric)
ϕ	Median
Σ	Sum, Weighted Sum
σ	Deviation, Sum of Differences
ϖ	Root Mean Square
ω	Power (Mean Square)
H	Entropy
π	Percentile
ρ	Regression
Λ	Histogram
β	Spectral binning
κ	Peak Detection
ψ	Harmonic Peak Detection
θ	Polynomial Root Finding
ζ	Zero-/Level Crossing Detector

Table 4: Frequent aggregations employed in audio features and their symbols (Greek letters, left). The subgroup of detectors are summarized at the bottom of the table.

information retrieval, "AS" is audio segmentation, "FP" is fingerprinting and "VAR" indicates that the feature is applied across several application domains.

The benefit of the signatures in Tables 5, 6, and 7 is not only the compact representation of the extraction process. More important is the ability to identify structurally similar features by comparing rows in the tables. Note that this may be done very quickly without decoding the signatures. Additionally to structural similarities, we may identify preferred operations for particular tasks (e.g. time-to-frequency transformation, analysis of harmonic structures), typical combinations of building blocks and coarsely estimate the complexity of a feature.

In the following, we summarize some observations from the signatures in Tables 5, 6, and 7. We observe that framing ("f") is part of almost every audio feature independent from the temporal scale. Most of the features are intraframe features, which means that the feature generates one vector for every frame (see Section 3.1). Features that contain brackets in their signature are most often interframe features, for example modulation frequency domain features. These features incorporate information from several frames and represent long-term properties, such as rhythm and tempo.

The signatures reveal the usage and distribution of mathematical transformations among the audio features. Most features employ the (short-time) Fourier transform ("f F") in order to obtain a time-frequency representation. We observe that the Cosine transform ("C") is mainly employed for the con-

Domain	Section	Feature Name	Temporal Scale	Perceptual	Psychoac. Model	Complexity	Dimension	Appl. Domain	Signature
Temporal	5.2.1	Zero Crossing Rate (ZCR)	I	N	N	L	1	VAR	f ζ
		Linear Prediction ZCR	I	N	N	L	1	ASR	f ζ a ζ
		Zero Crossing Peak Amplitudes (ZCPA)	I	N	Y	M	V	ASR	f b ζ κ l Λ Σ
		Pitch Synchronous ZCPA	I	N	Y	M	V	ASR	f b A χ ζ κ l Λ Σ
	5.2.2	MPEG-7 Audio Waveform	I	N	N	L	2	-	f χ ι
		Amplitude Descriptor	I	N	N	L	9	ESR	f μ σ ζ μ σ
	5.2.3	Short-Time Energy, MPEG-7 Audio Power	I	N	N	L	1	VAR	f ω
		Volume	I	N	N	L	1	VAR	f ϖ
		MPEG-7 Temporal Centroid	X	N	N	L	1	MIR	[f ϖ] μ
		MPEG-7 Log Attack Time	G	N	N	L	1	MIR	[f ϖ] κ l
Frequency - Physical	5.3.1	Linear Predictive Coding	I	N	N	L	V	ASR	f (b)a (F)
		Line Spectral Frequencies	I	N	N	M	V	VAR	f a θ
	5.3.2	Daubechies Wavelet Coef. Histogr.	I	N	N	M	28	MIR	f W Λ
		Adaptive Time-Frequency Transform	G	N	N	M	42	MIR	V Λ
	5.3.3	Subband Energy Ratio	I	N	N	L	V	VAR	f F β e n
		Spectral Flux	I	N	N	L	1	VAR	[f F] d Σ
		Spectral Slope	I	N	N	L	4	VAR	f F ρ
		Spectral Peaks	X	N	N	L	V	MIR	[f F χ] d
		(Modified) Group Delay	I	N	N	M	V	ASR	f F (o)g (C)
Freq. - Perc.	5.4.1	MPEG-7 Spectral Centroid	G	Y	N	L	1	MIR	F μ
		MPEG-7 Audio Spectrum Centroid	I	Y	Y	M	1	VAR	f F β l μ
		Spectral Centroid	I	Y	N	L	1	VAR	f F (β)(l) μ
		Sharpness	I	Y	Y	M	1	VAR	f F β w w μ
		Spectral Center	I	Y	N	L	1	MIR	f F e ϕ

Table 5: This table gives an overview of temporal and frequency domain features. For each feature, we list the domain, a reference to the describing section, temporal scale, whether or not the feature is perceptual and employs psychoacoustic models, the complexity, dimension, application domain, and signature.

Domain	Section	Feature Name	Temporal Scale	Perceptual	Psych. Model	Complexity	Dimension	Appl. Domain	Signature
Frequency - Perceptual	5.4.2	Bandwidth	I	Y	N	L	1	VAR	f F β (l) σ
		MPEG-7 Audio Spectrum Spread	I	Y	Y	M	1	VAR	f F β l σ
		Spectral Dispersion	I	Y	N	L	1	MIR	f F e ϕ σ
		Spectral Rolloff	I	Y	N	L	1	VAR	f F π
		Spectral Crest	I	Y	N	L	V	FP	f F β χ μ (l)
		Spectral Flatness	I	Y	N	M	V	FP	f F β μ (l)
		Subband Spectral Flux	I	Y	N	M	8	ESR	f F l n β d μ
		(Multi-resolution) Entropy	I	Y	N	M	V	ASR	f F n β H
	5.4.3	Sone	I	Y	Y	H	V	MIR	f F β o l w
		Integral Loudness	I	Y	Y	H	1	MIR	f F l Σ w x Σ
	5.4.4	Pitch (dominant frequency)	I	Y	N	L	1	VAR	f A χ
		MPEG-7 Audio Fundamental Freq.	I	Y	N	L	2	VAR	f A χ
		Pitch Histogram	X	Y	N	M	V	MIR	[f A κ] Λ (Σ)
		Psychoacoustical Pitch	I	Y	Y	H	V	VAR	b b w A Σ
	5.4.5	Chromagram	I	Y	N	M	12	MIR	f F l Σ
		Chroma CENS Features	I	Y	N	M	12	MIR	f B Σ n o
		Pitch Profile	I	Y	N	H	12	MIR	f Q κ Σ χ Λ χ χ Σ
	5.4.6	MPEG-7 Audio Harmonicity	I	Y	N	M	2	VAR	f A χ
		Harmonic Coefficient	I	Y	N	L	1	AS	f A χ
		Harmonic Prominence	I	Y	N	M	1	ESR	f A ψ
		Inharmonicity	I	Y	N	M	1	MIR	f A ψ σ
		MPEG-7 Harmonic Spectral Centroid	I	Y	N	M	1	MIR	f F ψ μ
		MPEG-7 Harmonic Spectral Deviation	I	Y	N	M	1	MIR	f F ψ μ l σ
		MPEG-7 Harmonic Spectral Spread	I	Y	N	M	1	MIR	f F ψ σ
		MPEG-7 Harmonic Spectral Variation	I	Y	N	M	1	MIR	[f F ψ] R
		Harmonic Energy Entropy	I	Y	N	M	1	MIR	f F ψ H
		Harmonic Concentration	I	Y	N	M	1	MIR	f F ψ e Σ
		Spectral Peak Structure	I	Y	N	M	1	MIR	f F ψ d Λ H
		Harmonic Derivate	I	Y	N	M	V	MIR	f F l d

Table 6: This table gives an overview of frequency domain perceptual features. For each feature, we list the domain, a reference to the describing section, temporal scale, whether or not the feature is perceptual and employs psychoacoustic models, the complexity, dimension, application domain, and signature.

Table 7: This table gives an overview of features in cepstral domain, modulation frequency domain, eigendomain, and phase space. The provided data is organized as in Table 5 and 6.

Domain	Section	Feature Name	Temporal Scale	Perceptual	Psychoac. Model	Complexity	Dimension	Appl. Domain	Signature
Cepstral	5.5.1	Mel-scale Frequency Cepstral Coef.	I	N	Y	H	V	VAR	f F β l C
		Bark-scale Frequency Cepstral Coef.	I	N	Y	H	V	VAR	f F β l C
		Autocorrelation MFCCs	I	N	Y	H	V	ASR	f A o F β l C
	5.5.2	Noise-Robust Auditory Feature	I	N	Y	H	256	ESR	f B w d o l C
	5.5.3	Perceptual Linear Prediction (PLP)	I	N	Y	H	V	ASR	f F β w w C a r
		Relative Spectral PLP	I	N	Y	H	V	ASR	f F β l b w w x C a r
		Linear Prediction Cepstral Coef.	I	N	N	M	V	ASR	f (b)a r
Modulation Frequency	5.6	Auditory Filter Bank Temp. Envelopes	I	N	Y	M	62	MIR	f b b e Σ
		Joint Acoustic and Modul. Freq. Feat.	X	N	Y	H	V	VAR	[f F β o] W Σ
		4 Hz Modulation Harmonic Coef.	X	N	N	M	1	AS	[f A χ] C b
		4 Hz Modulation Energy	X	N	Y	M	1	AS	[f F β] b e n Σ
	5.6.1	Band Periodicity	X	Y	N	M	4	AS	[f b A χ] Σ
		Pulse Metric	I	Y	N	M	1	AS	f b κ A κ
		Beat Spectrum (Beat Spectrogram)	X	Y	N	H	V	MIR	[f F l o] R A
		Cyclic Beat Spectrum	X	Y	N	H	V	MIR	o [f F d Σ] c o Σ κ
		Beat Tracker	X	Y	N	H	1	MIR	[f b o d c Σ] κ]
		Beat Histogram	X	Y	N	M	6	MIR	[f W o Σ A κ] Λ
		DWPT-based Rhythm Feature	X	Y	N	M	V	MIR	[f W A κ] Λ
		Rhythm Patterns	X	N	Y	H	80	MIR	[[f F β o l w] F w o] ϕ
Eigend.	5.7	Rate-scale-frequency Features	X	N	Y	H	256	ESR	[f B w d o] W Σ P
		MPEG-7 Audio Spectrum Basis	X	N	N	H	V	ESR	[f F β l n] S (I)
		Distortion Discriminant Analysis	X	N	N	H	64	FP	[f M l P] P
	5.8	Phase Space Features	I	N	N	H	V	ASR	f E

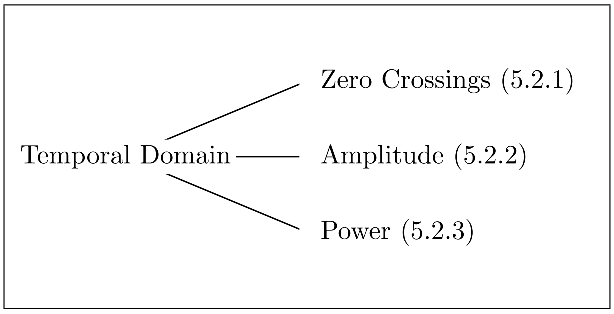


Figure 10: The organization of features in the temporal domain relates to the captured physical properties of the signal. In brackets a reference to the section containing the corresponding features is given.

version from frequency to cepstral domain (due to its ability to decorrelate the data). In the set of investigated features, the Wavelet transform ("W") appears rarely compared to the other transformations, although it has better time-frequency resolution than the short-time Fourier transform.

As already mentioned, the features in Tables 5, 6, and 7 are arranged according to the taxonomy (see Section 4). Usually, features from the same group of the taxonomy share similar properties. For example, most harmonicity features share the same building blocks (DFT "F" or Autocorrelation "A" followed by a peak detection "h"). Another observation is that Pitch and Rhythm features make extensive use of autocorrelation.

The identification of building blocks and signatures provides a novel perspective on audio features. Signatures give a compact overview of the computation of a feature and reveal basic properties (e.g. domain, temporal scale, and complexity). Additionally, they enable the comparison of features based on a unified vocabulary of mathematical operations that is independent of any application domain. The literature concerning each feature is listed separately in Section 6.2.

5.2 Temporal Features

The temporal domain is the native domain for audio signals. All temporal features have in common that they are extracted directly from the raw audio signal, without any preceding transformation. Consequently, the computational complexity of temporal features tends to be low.

We partition the group of temporal features into three groups, depending on what the feature describes. First, we investigate features that are based on zero crossings, then we survey features that describe the amplitude and the energy of a signal, respectively. Figure 10 depicts the groups of the taxonomy.

5.2.1 Zero Crossing-Based Features

Zero crossings are a basic property of an audio signal that is often employed in audio classification. Zero crossings allow for a rough estimation of dominant frequency and the spectral centroid [41].

Zero crossing rate (ZCR). One of the cheapest and simplest features is the zero crossing rate, which is defined as the number of zero crossings in the temporal domain within one second. According to Kedem the ZCR is a measure for the dominant frequency in a signal [77]. ZCR is a popular feature for speech/music discrimination [140, 159] due to its simplicity. However, it is extensively used in a wide range of other audio application domains, such as musical genre classification [114], highlight detection [27], speech analysis [33], singing voice detection in music [200], and environmental sound recognition [22].

Linear prediction zero-crossing ratio (LP-ZCR). LP-ZCR is the ratio of the zero crossing count of the waveform and the zero crossing count of the output of a linear prediction analysis filter [41]. The feature quantifies the degree of correlation in a signal. It helps to distinguish between different types of audio, such as (higher correlated) voiced speech and (lower correlated) unvoiced speech.

Zero crossing peak amplitudes (ZCPA). The ZCPA feature has been proposed by Kim et al. in [80, 81] for automatic speech recognition in noisy environments. The ZCPA technique extracts frequency information and corresponding intensities in several psychoacoustically scaled subbands from time domain zero crossings. Information from all subbands is accumulated into a histogram where each bin represents a frequency. The ZCPA feature is an approximation of the spectrum that is directly computed from the signal in temporal domain and may be regarded as a descriptor of the spectral shape. Kim et al. show that ZCPA outperforms linear prediction cepstral coefficients (see Section 5.5.3) under noisy conditions for automatic speech recognition [81].

Pitch synchronous zero crossing peak amplitudes (PS-ZCPA). PS-ZCPA is an extension of ZCPA that additionally takes pitch information into account [52]. Small peak amplitudes, which are prone to noise are removed by synchronizing the ZCPA with the pitch. Ghulam et al. show that the resulting feature is more robust to noise than ZCPA [52]. They further increase the performance of PS-ZCPA by taking auditory masking effects into account in [53].

5.2.2 Amplitude-Based Features

Some features are directly computed from the amplitude (pressure variation) of a signal. Amplitude-based features are easy and fast to compute but limited in their expressiveness. They represent the temporal envelope of the audio signal.

MPEG-7 audio waveform (AW). The audio waveform descriptor gives a compact description of the shape of a waveform by computing the minimum and maximum samples within non-overlapping frames. The AW descriptor represents the (downsampled) waveform envelope over time. The purpose of the descriptor is the display and comparison of waveforms rather than retrieval [73].

Amplitude descriptor (AD). The amplitude descriptor has been developed for the recognition of animal sounds [123]. The descriptor separates the signal into segments with low and high amplitude by an adaptive threshold (a level-crossing operation). The duration, variation of duration, and energy of these segments make up the descriptor. AD characterizes the waveform envelope in terms of quiet and loud segments. It allows to distinguish sounds with characteristic waveform envelopes.

5.2.3 Power-Based Features

The energy of a signal is the square of the amplitude represented by the waveform. The power of a sound is the energy transmitted per unit time (second) [124]. Consequently, power is the mean-square of a signal. Sometimes the root of power (root-mean-square) is used in feature extraction. In the following, we summarize features that represent the power of a signal (short-time energy, volume) and its temporal distribution (temporal centroid, log attack time).

Short-time energy (STE). STE describes the envelope of a signal and is extensively used in various fields of audio retrieval (see Table 9 in Section 6.2 for a list of references). We define STE according to Zhang and Kuo as the mean energy per frame (which actually is a measure for power) [201]. The same definition is used for the *MPEG-7 audio power descriptor* [73]. Note that there are varying definitions for STE that take the *spectral* power into account [32, 109].

Volume. Volume is a popular feature in audio retrieval, for example in silence detection and speech/music segmentation [76, 140]. Volume is sometimes called loudness, as in [194]. We use the term loudness for features that model human sensation of loudness, see Section 5.4.3. Volume is usually approximated by the root-mean-square (RMS) of the signal magnitude within a frame [104]. Consequently, volume is the square root of STE. Both, volume and STE reveal the magnitude variation over time.

MPEG-7 temporal centroid. The temporal centroid is the time average over the envelope of a signal in seconds [73]. It is the point in time where most of the energy of the signal is located in average. Note that the computation of temporal centroid is equivalent to that of spectral centroid (Section 5.4.1) in the fr equency domain.

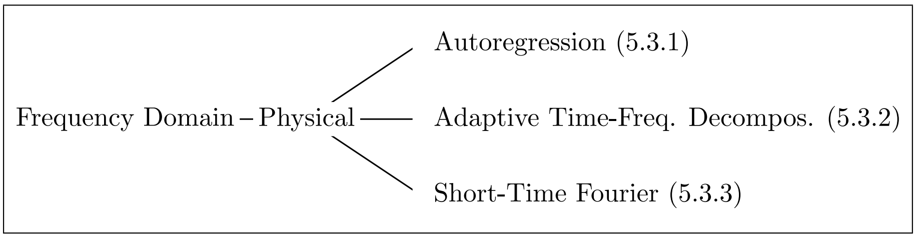

Frequency Domain – Physical
Autoregression (5.3.1)
Adaptive Time-Freq. Decompos. (5.3.2)
Short-Time Fourier (5.3.3)

Figure 11: The organization of physical features in the frequency domain. In brackets a reference to the section containing the corresponding features is given.

MPEG-7 log attack time (LAT). The log attack time characterizes the attack of a sound. LAT is the logarithm of the time it takes from the beginning of a sound signal to the point in time where the amplitude reaches a first significant maximum [73]. The attack characterizes the beginning of a sound, which can be either smooth or sudden. LAT may be employed for classification of musical instruments by their onsets.

5.3 Physical Frequency Features

The group of frequency domain features is the largest group of audio features. All features in this group have in common that they live in frequency or autocorrelation domain. From the signatures in Tables 5, 6, and 7 we observe that there are several ways to obtain a representation in these domains. The most popular methods are the Fourier transform and the autocorrelation. Other popular methods are the Cosine transform, Wavelet transform, and the constant Q transform. For some features the spectrogram is computed by directly applying a bank of band-pass filters to the temporal signal followed by framing of the subband signals.

We divide frequency features into two subsets: *physical* features and *perceptual* features. See Section 3.1 for more details on these two properties. In this section, we focus on physical frequency features. These features describe a signal in terms of its physical properties. Usually, we cannot assign a semantic meaning to these features. Figure 11 shows the corresponding groups of the taxonomy.

5.3.1 Autoregression-Based Features

Autoregression analysis is a standard technique in signal processing where a linear predictor estimates the value of each sample of a signal by a linear combination of previous values. Linear prediction analysis has a long tradition in audio retrieval and signal coding [152, 178].

Linear predictive coding (LPC). LPC is extensively used in automatic speech recognition since it takes into account the source-filter model of speech production (by employing an all-pole filter) [152]. The goal of LPC is to estimate basic parameters of a speech signal, such as formant frequencies and the vocal tract transfer function. LPC is applied in other domains as well, such as audio segmentation and general purpose audio retrieval where the LPC spectrum is used as an approximation of the spectral envelope [78, 79, 102].

In practice, the cepstral representation of LPC coefficients is mostly used due to their higher retrieval efficiency [195]. We address linear prediction cepstral coefficients (LPCC) in Section 5.5.3.

Line spectral frequencies (LSF). Line spectral frequencies (also called line spectral pairs) are an alternative representation of linear prediction coefficients. LSF are obtained by decomposing the linear prediction polynomial into two separate polynomials. The line spectral frequencies are at the roots of these two polynomials [24].

LSF characterize the resonances of the linear prediction polynomial together with their bandwidths [88]. While LSF describe equivalent information to LPC coefficients, they have statistical properties that make them better suited for pattern recognition applications [177]. LSF are employed in various application domains, such as in speech/music discrimination [41], instrument recognition [88], and speaker segmentation [108].

5.3.2 Adaptive Time-Frequency Decomposition-Based Features

The short-time Fourier transform (STFT) is widely used in audio feature extraction for time-frequency decomposition. This can be observed from the signatures in Tables 5, 6, and 7. However, STFT provides only a suboptimal tradeoff between time and frequency resolution since the frequency resolution of the STFT is the same for all locations in the spectrogram. The advantage of adaptive time-frequency decompositions, like the Wavelet transform is that they provide a frequency resolution that varies with the temporal resolution.

This group of the taxonomy comprises features that employ Wavelet transform and related transformations for time-frequency decomposition. Features in this group are based on the transform coefficients. For example, Khan and Al-Khatib successfully employ the variance of Haar Wavelet coefficients over several frames for speech/music discrimination in [79]. We consider such features as physical features since they do not have a semantic interpretation.

Daubechies Wavelet coefficient histogram features (DWCH). DWCHs have been proposed by Li et al. for music genre classification in [98]. The authors decompose the audio signal by Daubechies Wavelets and build histograms from the Wavelet coefficients for each subband. The subband histograms provide an approximation of the waveform variation in each subband. The first three statistical moments of each coefficient histogram together with the energy per subband make up the feature vector. Li et al. show that DWCHs improve efficiency in combination with traditional features for music genre classification [98]. Further studies on DWCHs in the fields of artist style identification, emotion detection, and similarity retrieval may be found in [95, 97].

Adaptive time frequency transform (ATFT) features. The ATFT investigated by Umapathy et al. in [185] is similar to the Wavelet transform. The signal is decomposed into a set of Gaussian basis functions of several scales, translations, and center frequencies. The scale parameter varies with the waveform envelope of the signal and represents for example rhythmic structures. It shows that the scale parameter contains discriminatory information for musical genres.

5.3.3 Short-Time Fourier Transform-Based Features

In this section, we group physical frequency features that employ the short-time Fourier transform (STFT) for computation of the spectrogram. The STFT yields real and complex values. The real values represent the distribution of the frequency components while the complex values carry information on the phase of the components. Consequently, we distinguish between features that rely on the frequency distribution (spectral envelope) and features that evaluate the phase information. First, we present features that capture basic properties of the spectral envelope: subband energy ratio, spectral flux, spectral slope, and spectral peaks. Then, we focus on phase-based features, such as the (modified) group delay function.

Subband energy ratio. The subband energy ratio gives a coarse approximation of the energy distribution of the spectrum. There are slightly different definitions concerning the selection of the subbands. Usually, four subbands are used as in [102]. However, Cai et al. divide the spectrum into eight Mel-scaled bands in [22]. The feature is extensively used in audio segmentation [76, 168] and music analysis [127]. See Table 9 in Section 6.2 for further references.

Spectral flux (SF). The SF is the 2-norm of the frame-to-frame spectral amplitude difference vector [162]. It quantifies (abrupt) changes in the shape of the spectrum over time. Signals with slowly varying (or nearly constant) spectral properties (e.g. noise) have low SF, while signals with abrupt spectral changes (e.g. note onsets) have high SF.

A slightly different definition is provided by Lu et al. in [106] where the authors compute SF based on the logarithm of the spectrum. Similarly to SF, the cepstrum flux is defined in [195]. SF is widely used in audio retrieval, e.g. in speech/music discrimination [76, 78, 79], music information retrieval [95, 180], and speech analysis [181].

Spectral slope. The spectral slope is a basic approximation of the spectrum shape by a linear regression line [127]. It represents the decrease of the spectral amplitudes from low to high frequencies (the spectral tilt) [142]. The slope, the y-intersection, the maximum- and median regression error may be used as features. Spectral slope/tilt may be employed for discrimination of voiced and unvoiced speech segments.

Spectral peaks. Wang introduces features that allow for a very compact and noise robust representation of an audio signal. The features are part of an audio search engine that is able to identify a piece of music by a short segment captured by a mobile phone [186, 187].

The author first computes the Fourier spectrogram and detects local peaks. The result is a sparse set of time-frequency pairs - the constellation map. From the constellation map, pairs of time-frequency points are formed. For each pair, the two frequency components, the time difference, and the time offset from the beginning of the audio signal are combined into a feature. Each piece of music is represented by a large number of such time-frequency pairs. An efficient and scalable search algorithm proposed by Wang allows for efficiently searching large databases built from these features. The search system is best described in [186].

The proposed feature represents a piece of music in terms of spatio-temporal combinations of dominant frequencies. The strength of the technique is that it solely relies on the salient frequencies (peaks) and rejects all other spectral content. This preserves the main characteristics of the spectrum and makes the representation highly robust to noise since the peak frequencies are usually less influenced by noise than the other frequencies.

Group delay function (GDF). The features mentioned above take the real part (magnitude) of the Fourier transform into account. Only a few features describe the phase information of the Fourier spectrum.

Usually, the phase is featureless and difficult to interpret due to polarity and wrapping artifacts. The group delay function is the negative derivative of the unwrapped Fourier transform phase [198]. The GDF reveals meaningful information from the phase, such as peaks of the spectral envelope.

The GDF is traditionally employed in speech analysis, for example for the determination of significant excitations [166]. A recent approach applies the GDF in music analysis for rhythm tracking [163]. Since the GDF is not robust against noise and windowing effects, the *modified* GDF is often employed instead [6].

Modified group delay function (MGDF). The MGDF algorithm applies a low-pass filter (cepstral smoothing) to the Fourier spectrum prior to computing the GDF [198]. Cepstral smoothing removes artifacts contributed by noise and windowing, which makes the MGDF more robust and better suited to speech analysis than the GDF [6]. The MGDF is employed in various subdomains of speech analysis, such as speaker identification, phoneme recognition, syllable detection, and language recognition [60, 61, 132, 134]. Murthy et al. show that the MGDF robustly estimates formant frequencies in [133].

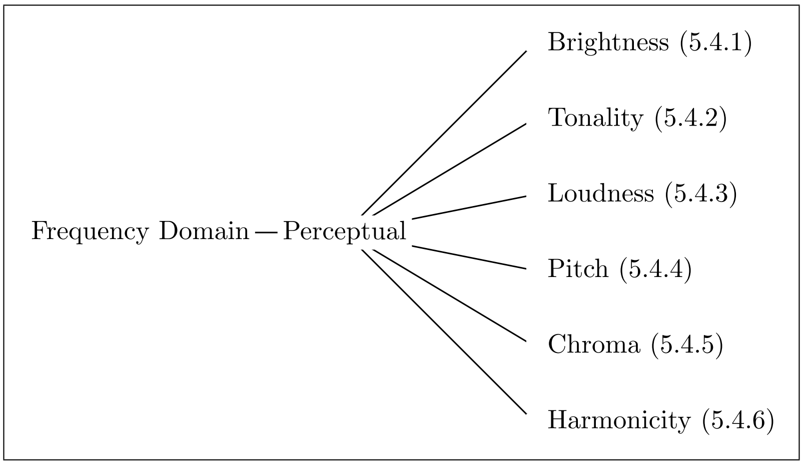


Figure 12: The organization of perceptual features in the frequency domain. In brackets a reference to the section containing the corresponding features is given.

5.4 Perceptual Frequency Features

So far we have focused on physical frequency features, that have no perceptual interpretation. In this section, we cover features that have a semantic meaning in the context of human auditory perception. In the following, we group the features according to the auditory quality that they describe (see Figure 12).

5.4.1 Brightness

Brightness characterizes the spectral distribution of frequencies and describes whether a signal is dominated by low or high frequencies, respectively. A sound becomes brighter as the high-frequency content becomes more dominant and the low-frequency content becomes less dominant. Brightness is often defined as the *balancing point* of the spectrum [102, 162]. Brightness is closely related to the sensation of sharpness [204].

Spectral centroid (SC). A common approximation of brightness is the SC (or frequency centroid). It is defined as the center of gravity of the magnitude spectrum (first moment) [99, 180]. The SC determines the point in the spectrum where most of the energy is concentrated and is correlated with the dominant frequency of the signal. A definition of spectral centroid in logarithmic frequency can be found in [167]. Furthermore, SC may be computed for several frequency bands as in [154].

The MPEG-7 standard provides further definitions of SC [73]. The *MPEG-7 audio spectrum centroid (ASC)* differs from the SC in that it employs a power spectrum in the octave-frequency scale. The ASC approximates the perceptual sharpness of a sound [83]. Another definition of SC is the *MPEG-7 spectral centroid*. The difference to SC is that MPEG-7 spectral centroid is defined for entire signals instead of single frames and that the power spectrum is used instead of the magnitude spectrum. The different definitions of spectral centroid are very similar, as shown by the signatures in Table 5.

Sharpness. Sharpness is closely related to brightness. Sharpness is a dimension of timbre that is influenced by the center frequency of narrow-band sounds. Sharpness grows with the strength of high-frequencies in the spectrum [204]. It may be computed similarly to the spectral centroid but based on the specific loudness instead of the magnitude spectrum. A mathematical model of sharpness is provided by Zwicker and Fastl [204]. Sharpness is employed in audio similarity analysis in [64, 142].

Spectral center. The spectral center is the frequency where half of the energy in the spectrum is below and half is above that frequency [163]. It describes the distribution of energy and is correlated with the spectral centroid and thus with the dominant frequency of a signal. Sethares et al. employ spectral center together with other features for rhythm tracking in [163].

5.4.2 Tonality

Tonality is the property of sound that distinguishes noise-like from tonal sounds [204]. Noise-like sounds have a continuous spectrum while tonal sounds typically have line spectra. For example, white noise has a flat spectrum and consequently a minimum of tonality while a pure sine wave results in high tonality. Tonality is related to the pitch strength that describes the strength of the perceived pitch of a sound (see Section 2.4). Sounds with distinct (sinusoidal) components tend to produce larger pitch strength than sounds with continuous spectra.

We distinguish between two classes of features that (partially) measure tonality: *flatness measures* and *bandwidth measures*. In the following, we first describe bandwidth measures (bandwidth, spectral dispersion, and spectral rolloff point) and then we focus on flatness measures (spectral crest, spectral flatness, subband spectral flux, and entropy).

Bandwidth. Bandwidth is usually defined as the magnitude-weighted average of the differences between the spectral components and the spectral centroid [194]. The bandwidth is the second-order statistic of the spectrum. Tonal sounds usually have a low bandwidth (single peak in the spectrum) while noise-like sounds have high bandwidth. However, this is not the case for more complex sounds. For example in music we find broadband signals with tonal characteristics. The same applies to complex tones with a large number of harmonics, that may have a broadband line spectrum. Consequently bandwidth may not be a sufficient indicator for tonality for particular tasks. Additional features (e.g. harmonicity features, see Section 5.4.6 and flatness features, see below) may be necessary to distinguish between tonal and noise-like signals.

Bandwidth may be defined in the logarithmized spectrum or the power spectrum [103, 109, 167]. Additionally, it may be computed within one or more subbands of the spectrum [4, 154].

In the MPEG-7 standard the measure for bandwidth is called *spectral spread* [73, 83]. Similarly to the bandwidth measures above, the MPEG-7 audio spectrum spread (ASS) is the root-mean-square deviation from the spectrum centroid (MPEG-7 ASC descriptor, see Section 5.4.1). Measures for bandwidth are often combined with that of spectral centroid in literature since they represent complementary information [4, 109, 154].

Spectral dispersion. The spectral dispersion is a measure for the spread of the spectrum around its spectral center [163]. See Section 5.4.1 for a description of spectral center. In contrast to bandwidth, the computation of spectral dispersion takes the spectral center into account instead of the spectral centroid.

Spectral rolloff point. The spectral rolloff point is the $N\%$ percentile of the power spectral distribution, where N is usually 85% or 95% [162]. The rolloff point is the frequency below which $N\%$ of the magnitude distribution is concentrated. It increases with the bandwidth of a signal. Spectral rolloff is extensively used in music information retrieval [96, 127] and speech/music segmentation [162].

Spectral flatness. Spectral flatness estimates to which degree the frequencies in a spectrum are uniformly distributed (noise-like) [74]. The spectral flatness is the ratio of the geometric and the arithmetic mean of a subband in the power spectrum [154]. The same definition is used by the MPEG-7 standard for the *audio spectrum flatness* descriptor [73]. Spectral flatness may be further computed in decibel scale as in [59, 90]. Noise-like sounds have a higher flatness value (flat spectrum) while tonal sounds have lower flatness values. Spectral flatness is often used (together with spectral crest factor) for audio fingerprinting [65, 90].

Spectral crest factor. The spectral crest factor is a measure for the "peakiness" of a spectrum and is inversely proportional to the flatness. It is used to distinguish noise-like and tone-like sounds due to their characteristic spectral shapes. Spectral crest factor is the ratio of the maximum spectrum power and the mean spectrum power of a subband. In [90] the spectral crest factor is additionally logarithmized. For noise-like sounds the spectral crest is lower than for tonal sounds. A traditional application of spectral crest factor is fingerprinting [65, 90, 154].

Subband spectral flux (SSF). The SSF has been introduced by Cai et al. in [22] for the recognition of environmental sounds. The feature is a measure for the portion of prominent partials ("peakiness") in different subbands. SSF is computed from the logarithmized short-time Fourier spectrum. For each subband the SSF is the accumulation of the differences between adjacent frequencies in that subband. SSF is low for flat subbands and high for subbands that contain distinct frequencies. Consequently, SSF is inversely proportional to spectral flatness.

Entropy. Another measure that correlates with the flatness of a spectrum is entropy. Usually, Shannon- and Renyi entropy are computed in several subbands [154]. The entropy represents the uniformity of the spectrum. A *multi-resolution entropy* feature is proposed by Misra et al. in [121, 122]. The authors split the spectrum into overlapping Mel-scaled subbands and compute the Shannon entropy for each subband. For a flat distribution in the spectrum the entropy is low while a spectrum with sharp peaks (e.g. formants in speech) has high entropy. The feature captures the "peakiness" of a subband and may be used for speech/silence detection and automatic speech recognition.

5.4.3 Loudness

Loudness features aim at simulating the human sensation of loudness. Loudness is "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud" [7]. The auditory system incorporates a number of physiological mechanisms that influence the transformation of the incoming physical sound intensity into the sensational loudness [204]. See Section 3.3 for a summary of important effects.

Specific Loudness Sensation (Sone). Pampalk et al. propose a feature that approximates the specific loudness sensation per critical band of the human auditory system [139]. The authors first compute a Bark-scaled spectrogram and then apply spectral masking and equal-loudness contours (expressed in phon). Finally, the spectrum is transformed to specific loudness sensation (in sone). The feature is the basis for rhythm patterns (see Section 5.6.1). The representation in sone may be applied to audio retrieval as in [127, 128].

Integral Loudness. The specific loudness sensation (sone) gives the loudness of a single sine tone. A spectral integration of loudness over several frequencies enables the estimation of the loudness of more complex tones [204]. Pfeiffer proposes an approach to compute the integral loudness by summing up the loudness in different frequency groups [144]. The author empirically shows that the proposed method closely approximates the human sensation of loudness. The integral loudness feature is applied to foreground/background segmentation in [147].

5.4.4 Pitch

Pitch is a basic dimension of sound, together with loudness, duration, and timbre. The hearing sensation of pitch is defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high" [7]. The term pitch is widely used in literature and may refer to both, a stimulus parameter (fundamental frequency or frequency of glottal oscillation) and an auditory sensation (the perceived frequency of a signal) depending on the application domain.

In this section, we first focus on features that capture the fundamental frequency and then present a technique that models the psychoacoustic pitch. Features that describe pitch are correlated to chroma and harmonicity features (see Sections 5.4.5 and 5.4.6).

Fundamental frequency. The fundamental frequency is the lowest frequency of a harmonic series and is a coarse approximation of the psychoacoustic pitch. Fundamental frequency estimation employs a wide range of techniques, such as temporal autocorrelation, spectral, and cepstral methods and combinations of these techniques. An overview of techniques is given in [66].

The MPEG-7 standard proposes a descriptor for the fundamental frequency (*MPEG-7 audio fundamental frequency*) which is defined as the first peak of the local normalized spectro-temporal autocorrelation function [29, 73]. Fundamental frequency is employed in various application domains [33, 180, 194].

Pitch Histogram. The pitch histogram describes the pitch content of a signal in a compact way and has been introduced for musical genre classification in [179, 180]. In musical analysis pitch usually corresponds to musical notes. The pitch histogram is a global representation that aggregates the pitch information of several short audio frames. Consequently, the pitch histogram represents the distribution of the musical notes in a piece of music. A similar histogram-based technique is the beat histogram that represents the rhythmic content of a signal (see Section 5.6.1).

Psychoacoustic Pitch. Meddis and O'Mard propose a method to model human pitch perception in [115]. First the authors apply a band-pass filter to the input signal to emphasize the frequencies relevant for pitch perception. Then the signal is decomposed with a gammatone filter bank that models the frequency selectivity of the cochlea. For each subband an inner hair-cell model transforms the instantaneous amplitudes into continuous firing probabilities. A running autocorrelation function is computed from the firing probabilities in each subband. The resulting autocorrelation functions are summed across the channels in order to obtain the final feature.

In contrast to other pitch detection techniques, the output of this algorithm is a series of values instead of one single pitch value. These values represent a range of frequencies relevant for pitch perception. Meddis and O'Mard point out that a single pitch frequency is not sufficient for approximation of the pitch perception of complex sounds. Consequently, they employ all values of the feature for matching pitches of different sounds.

5.4.5 Chroma

According to Shepard the sensation of musical pitch may be characterized by two dimensions: *tone height* and *chroma* [165]. The dimension of tone height is partitioned into the musical octaves. The range of chroma is usually divided into 12 pitch classes, where each pitch class corresponds to one note of the twelve-tone equal temperament. For example, the pitch class C contains the Cs of all possible octaves (C_0, C_1, C_2, ...). The pitches (musical notes) of the same pitch class share the same chroma and produce a similar auditory sensation. Chroma-based representations are mainly used in music information analysis and retrieval since they provide an octave invariant representation of the signal.

Chromagram. The chromagram is a spectrogram that represents the spectral energy of each of the 12 pitch classes [13]. It is based on a logarithmized short-time Fourier spectrum. The frequencies are mapped (quantized) to the 12 pitch classes by an aggregation function. The result is a 12 element vector for each audio frame. A similar algorithm for the extraction of chroma vectors is presented in [54].

The chromagram maps all frequencies into one octave. This results in a spectral compression that allows for a compact description of harmonic signals. Large harmonic series may be represented by only a few chroma values, since most harmonics fall within the same pitch class [13]. The chromagram represents an octave-invariant (compressed) spectrogram that takes properties of musical perception into account.

Chroma energy distribution normalized statistics (CENS). CENS features are another representation of chroma, introduced for music similarity matching by Müller et al. in [130] and by in Müller in [129]. The CENS features are robust against tempo variations and different timbres which makes them suitable for the matching of different interpretations of the same piece of music.

Pitch Profile. The pitch profile is a more accurate representation of the pitch content than the chroma features [202]. It takes pitch mistuning (introduced by mistuned instruments) into account and is robust against noisy percussive sounds (e.g. sounds of drums that do not have a pitch). Zhu and Kankanhalli apply the pitch profile in musical key detection and show that the pitch profile outperforms traditional chroma features [202].

5.4.6 Harmonicity

Harmonicity is a property that distinguishes periodic signals (harmonic sounds) from non-periodic signals (inharmonic and noise-like sounds). Harmonics are frequencies at integer multiples of the fundamental frequency. Figure 13 presents the spectra of a noise-like (inharmonic) and a harmonic sound. The harmonic spectrum shows peaks at the fundamental frequency and its integer multiples.

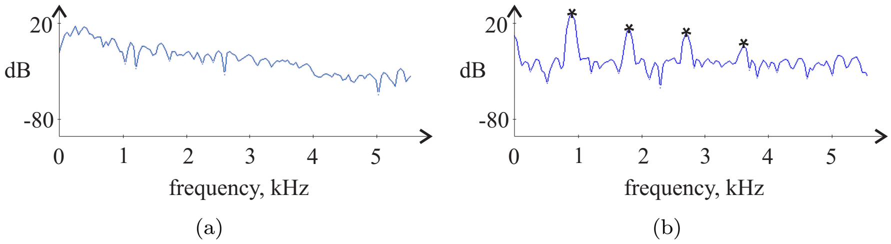


Figure 13: (a) The spectrum of a noise-like sound (thunder). (b) The spectrum of a harmonic sound (siren). The harmonic sound has peaks at multiples of the fundamental frequencies (the harmonic peaks are marked by asterisks), while the noise-like sound has a flat spectrum.

Harmonicity relates to the proportion of harmonic components in a signal. Harmonicity features may be employed to distinguish musical instruments. For example harmonic instrument sounds (e.g. violins) have stronger harmonic structure than percussive instrument sounds (e.g. drums). Furthermore, harmonicity may be useful in environmental sound recognition in order to distinguish between harmonic (e.g. bird song) and inharmonic (e.g. dog barks) sounds.

MPEG-7 audio harmonicity. The audio harmonicity descriptor of the MPEG-7 standard comprises two measures. The *harmonic ratio* is the ratio of the fundamental frequency's power to the total power in an audio frame [73, 83]. It is a measure for the degree of harmonicity contained in a signal. The computation of harmonic ratio is similar to that of MPEG-7 audio fundamental frequency, except for the used autocorrelation function.

The second measure in the audio harmonicity descriptor is the *upper limit of harmonicity.* The upper limit of harmonicity is the frequency beyond which the spectrum no longer has any significant harmonic structure. It may be regarded as the bandwidth of the harmonic components. The audio harmonicity descriptor is well-suited for the distinction of periodic (e.g. musical instruments, voiced speech) and non-periodic (e.g. noise, unvoiced speech) sounds.

A similar feature is the *harmonic coefficient* which is defined as the first maximum in the (spectro-temporal) autocorrelation function in [31]. Note that the definition is nearly equivalent to that of harmonic ratio, except for the employed autocorrelation function.

Inharmonicity measures. Most real world harmonic signals do not show a perfect harmonic structure. Inharmonicity features measure the difference between observed harmonics and their theoretical (predicted) values which are exactly at integer multiples of the fundamental frequency.

A straight-forward cumulative measure for the deviation of the harmonics from their predicted values is presented in [4] and [142]. A more enhanced and more accurate feature is *harmonicity prominence* which additionally takes the energy and the bandwidth of each harmonic component into account [22].

A related feature is *spectral peak structure* which is the entropy of the distances of adjacent peaks in the spectrum. For perfect harmonic sounds these distances are constant, while for non-harmonic sounds the distances may vary. Consequently, the entropy of the distances is a measure for inharmonicity.

MPEG-7 spectral timbral descriptors. The MPEG-7 standard defines a set of descriptors for the harmonic structure of sounds: *MPEG-7 harmonic spectral centroid (HSC)*, *MPEG-7 harmonic spectral deviation (HSD)*, *MPEG-7 harmonic spectral spread (HSS)*, and *MPEG-7 harmonic spectral variation (HSV)* [73, 143]. All descriptors are based on an estimate of the fundamental frequency and the detection of harmonic peaks in the spectrum (see the signatures in Table 6. The descriptors represent statistical properties (moments) of the harmonic frequencies and their amplitudes.

The HSC is the amplitude-weighted average of the harmonic frequencies. Similarly to spectral centroid (see Section 5.4.1) HSC is related to brightness and sharpness [83].

The HSS descriptor is the power-weighted root-mean-square deviation of the harmonic peaks from the HSC. It represents the bandwidth of the harmonic frequencies. HSC and HSS are first and second moment of the harmonic spectrum similarly to spectral centroid and bandwidth (spectral spread) which are first and second moment of the entire spectrum.

HSD measures the amplitude deviation of harmonic peaks from their neighboring harmonic peaks in the same frame. If all harmonic peaks have equal amplitude HSD reaches its minimum. While HSS represents the variation of harmonic frequencies, HSD reflects the variation of harmonics' amplitudes.

The HSV descriptor represents the correlation of harmonic peak amplitudes in two adjacent frames. It represents fast variations of harmonic structures over time. The MPEG-7 spectral timbral descriptors address musical instrument recognition, where the harmonic structure is an important discriminative property [143].

Further harmonicity features. Srinivasan and Kankanhalli introduce harmonicity features for classification of music genre and instrument family in [167]. *Harmonic concentration* measures the fraction of energy of the dominant harmonic component of the signal. *Harmonic energy entropy* describes the energy distribution of the harmonic components by computing the entropy of their energies. Finally, Srinivasan and Kankanhalli define the *harmonic derivate* as the difference of the energy of adjacent harmonic frequencies. The feature represents the decay of harmonic energy with increasing frequency.

There is a large number of features that capture harmonic properties in literature. Harmonicity features are related to pitch- and chroma features. Additionally, they are correlated to a high degree due to methodological similarities which may be observed from the signatures in Table 6.

5.5 Cepstral Features

The concept of the "cepstrum" has been originally introduced by Bogert et al. in [16] for the detection of echoes in seismic signals. In the domain of audio, cepstral features have first been employed for speech analysis [18, 37, 136]. Cepstral features are frequency smoothed representations of the log magnitude spectrum and capture timbral characteristics and pitch. Cepstral features allow for application of the Euclidean metric as distance measure due to their orthogonal basis which facilitates similarity comparisons [37]. Today, cepstral features are widely used in all fields of audio retrieval (speech-, music-, and environmental sound analysis), e.g. [101, 196].

We have identified three classes of cepstral features. The first group employs traditional filter banks, such as Mel- and Bark-filters. The second group bases on more elaborate auditory models. The third group are cepstral features that apply autoregression.

5.5.1 Perceptual Filter Bank-Based Features

Bogert et al. define the cepstrum as the Fourier Transform (FT) of the logarithm (log) of the magnitude (mag) of the spectrum of the original signal [16].

$$signal \rightarrow FT \rightarrow mag \rightarrow log \rightarrow FT \rightarrow cepstrum$$

This sequence is the basis for the cepstral features described in this section. However, in practice the computation slightly differs from this definition. For example, the second Fourier transform is often replaced by a DCT due to its ability to decorrelate output data.

Mel-frequency cepstral coefficients (MFCCs). MFCCs originate from automatic speech recognition but evolved into one of the standard techniques in most domains of audio retrieval. They represent timbral information (the spectral envelope) of a signal. MFCCs have been successfully applied to timbre measurements by Terasawa et al. in [174].

Computation of MFCCs includes a conversion of the Fourier coefficients to Mel-scale [171]. After conversion, the obtained vectors are logarithmized, and decorrelated by DCT in order to remove redundant information.

The components of MFCCs are the first few DCT coefficients that describe the coarse spectral shape. The first DCT coefficient represents the average power in the spectrum. The second coefficient approximates the broad shape of the spectrum and is related to the spectral centroid. The higher-order coefficients represent finer spectral details (e.g. pitch). In practice, the first 8-13 MFCC coefficients are used to represent the shape of the spectrum. However, some applications require more higher-order coefficients to capture pitch and tone information. For example in Chinese speech recognition up to 20 cepstral coefficients may be beneficial [190].

Variations of MFCCs. In the course of time several variations of MFCCs have been proposed. They mainly differ in the applied psychoacoustic scale. Instead of the Mel-scale, variations employ the Bark- [203], ERB- [126] and octave-scale [111]. A typical variation of MFCCs are Bark-frequency cepstral coefficients (BFCCs). However, cepstral coefficients based on the Mel-scale are the most popular variant used today, even if there is no theoretical reason that the Mel-scale is superior to the other scales.

Extensions of MFCCs. A noise-robust extension of MFCCs are *autocorrelation MFCCs* proposed by Shannon and Paliwal in [164]. The main difference is the computation of an unbiased autocorrelation from the raw signal. Particular autocorrelation coefficients are removed in order to filter noise. From this representation more noise-robust MFCCs are extracted.

Yuo et al. introduce two noise-robust extensions of MFCCs, namely RAS-MFCCs and CHNRAS-MFCCs in [199]. The features introduce a preprocessing step to the standard computation of MFCCs that filters additive and convolutional noise (cannel distortions) by cepstral mean substraction.

Another extension of MFCCs is introduced in [30]. Here, the outputs of the Mel-filters are weighted according to the amount of estimated noise in the bands. The feature improves accuracy of automatic speech recognition in noisy environments.

Li et al. propose a novel feature that may be regarded as an extension of Bark-frequency cepstral coefficients [93]. The feature incorporates additional filters that model the transfer function of the cochlea. This enhances the ability to simulate the human auditory system and improves performance in noisy environments.

5.5.2 Advanced Auditory Model-Based Features

Features in this group base on an auditory model that is designed to closely represent the physiological processes in human hearing.

Noise-robust audio features (NRAF). NRAF are introduced in [156] and are derived from a mathematical model of the early auditory system [197]. The auditory model yields a psychoacoustically motivated time-frequency representation which is called the auditory spectrum. A logarithmic compression of the auditory spectrum models the behavior of the outer hair cells. Finally, a DCT decorrelates the data. The temporal mean and variance of the resulting decorrelated spectrum make up the components of NRAF. The computation of NRAF is similar to that of MFCCs but it follows the process of hearing in a more precise way. A related audio feature of NRAF are rate-scale-frequency features addressed in Section 5.7.

5.5.3 Autoregression-Based Features

Features in this group are cepstral representations that base on linear predictive analysis (see Section 5.3.1).

Perceptual linear prediction (PLP). PLP was introduced by Hermansky in 1990 for speaker-independent speech recognition [62]. It bases on the concepts of hearing and employs linear predictive analysis for the approximation of the spectral shape. In the context of speech PLP represents speaker-independent information, such as vocal tract characteristics. It better represents the spectral shape than conventional linear prediction coding (LPC) by approximating several properties of human hearing. The feature employs Bark-scale as well as asymmetric critical-band masking curves in order to achieve a higher grade of consistency with human hearing.

Relative spectral - perceptual linear prediction (RASTA-PLP). RASTA-PLP is an extension of PLP introduced by Hermansky and Morgan in [63]. The objective of RASTA-PLP is to make PLP more robust to linear spectral distortions. The authors filter each frequency channel with a bandpass filter in order to alleviate fast variations (frame to frame variations introduced by the short-time analysis) and slow variations (convolutional noise introduced by the communication channel). RASTA PLP better approximates the human abilities to filter noise than PLP and yields a more robust representation of the spectral envelope under noisy conditions.

Linear prediction cepstrum coefficients (LPCCs). LPCCs are the inverse Fourier transform of the log magnitude frequency response of the autoregressive filter. They are an alternative representation for linear prediction coefficients and thus capture equivalent information. LPCCs may be directly derived from the LPC coefficients presented in Section 5.3.1 with a recursion formula [8].

In practice, LPCCs have shown to perform better than LPC coefficients, e.g. in automatic speech recognition, since they are a more compact and robust representation of the spectral envelope [2]. In contrast to LPC they allow for the application of the Euclidean distance metric. The traditional application domain of LPCCs is automatic speech recognition. However, LPCCs may be employed in other domains, such as music information retrieval as well [195].

5.6 Modulation Frequency Features

Modulation frequency features capture low-frequency modulation information in audio signals. A modulated signal contains at least two frequencies: a high carrier frequency and a comparatively low modulation frequency. Modulated sounds cause different hearing sensations in the human auditory system. Low modulation frequencies up to 20 Hz produce the hearing sensation of fluctuation strength [204]. Higher modulation frequencies create the hearing sensation of

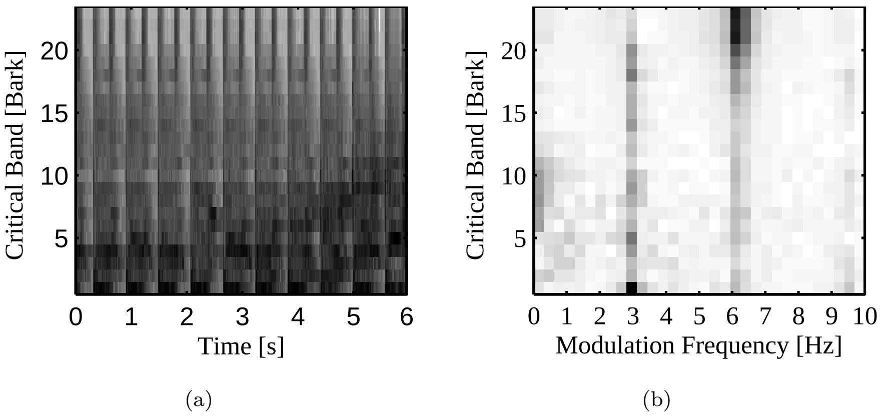


Figure 14: (a) The spectrogram of a 6 second excerpt of "Rock DJ" by Robbie Williams. (b) The modulation spectrogram reveals modulation frequencies at 3 Hz and 6 Hz. The modulation frequencies relate to the main beat and the sub beats of the song.

roughness. Modulation information is a long-term signal variation of amplitude or frequency that is usually captured by a temporal (interframe) analysis of the spectrogram.

Rhythm and tempo are aspects of sound (especially important in music) that are strongly related to long-time modulations. Rhythmic structures (e.g. sequences of equally spaced beats or pulses) may be revealed by analyzing low-frequency modulations over time. Figure 14 shows a short-time Fourier spectrogram together with the corresponding modulation spectrogram of a piece of music. The spectrogram represents the distribution of acoustic frequencies over time, while the modulation spectrogram shows the distribution of long-term modulation frequencies for each acoustic frequency. In Figure 14 we observe two strong modulation frequencies at 3 Hz and 6 Hz that are distributed over all critical bands. These frequencies relate to the main and sub beats of the song. We discuss features that represent rhythm and tempo-related information in Section 5.6.1.

4 Hz modulation energy. The hearing sensation of fluctuation strength has its peak at 4 Hz modulation frequency (for both, amplitude- and frequency modulated sounds) [46, 67]. This is the modulation frequency that is most often observed in fluent speech, where approximately four syllables per second are produced. Hence, the 4 Hz modulation energy may be employed for distinguishing speech from non-speech sounds.

Scheirer and Slaney extract the 4 Hz modulation energy by a spectral analysis of the signal [162]. They filter each subband by a 4 Hz band-pass filter along the temporal dimension. The filter outputs represent the 4 Hz modulation energy. A different definition that derives the 4 Hz modulation energy is given in [104].

Similarly to 4 Hz modulation frequency, Ghou and Gu define the *4 Hz modulation harmonic coefficient* which actually is an estimate of the 4 Hz modulation energy of the fundamental frequency of a signal [31]. The authors report that this feature better discriminates speech from singing than the 4 Hz modulation frequency.

Joint acoustic and modulation frequency features. Sukittanon and Atlas propose a feature for audio fingerprinting that represents the distribution of modulation frequencies in the critical bands [172]. The feature is a time-invariant representation and captures time-varying (non-stationary) behavior of an audio signal.

The authors first decompose the input signal into a Bark-scaled spectrogram. Then they demodulate the spectrogram by extracting frequencies of each subband envelope. A Wavelet transform produces one modulation frequency vector for each subband. The output of this procedure is a matrix (a modulation spectrogram) that contains the modulation frequencies for each acoustic frequency band. The modulation spectrogram is constant in size and time-invariant. Hence, it may be vectorized in order to build a feature vector (fingerprint) for retrieval.

Sukittanon et al. show that their modulation frequency feature outperforms MFCCs in presence of noise and time-frequency distortions [173]. A similar feature are rhythm patterns which have been developed for music similarity matching. We present rhythm patterns together with other rhythm-related features in Section 5.6.1 below.

A spectral representation that takes the temporal resolution of modulation information into account is the modulation spectrogram by Greenberg and Kingsbury [55]. In contrast to the features mentioned above, the modulation spectrogram shows the distribution of slow modulations across time *and* frequency. Experiments show that it is more robust to noise than the narrow-band spectrogram.

Auditory filter bank temporal envelopes. McKinney and Breebaart present another approach for the computation of modulation frequency features in [114]. They employ logarithmically spaced gamma tone filters for subband decomposition. The resulting subband envelopes are band-pass filtered in order to obtain modulation information. The feature represents modulation energy for particular acoustic frequency bands similarly to the joint acoustic and modulation frequency features (see above). The features have been successfully employed for musical genre classification and general purpose audio classification.

5.6.1 Rhythm

Rhythm is a property of an audio signal that represents a change pattern of timbre and energy over time [201]. According to Zwicker and Fastl, the hearing sensation of rhythm depends on the temporal variation of loudness [204]. Rhythm is an important element in speech and music. In speech, rhythm relates to stress and pitch and in music it relates to the tempo of a piece of music (in beats-per-minute). Rhythm may be important for the characterization of environmental sounds as well, for example for the description of footsteps [201].

Rhythm is a property that evolves over time. Consequently, the analysis windows of rhythm features are usually longer than that of other features. Typical analysis windows are in the range of a few seconds (≈ 3-5s) [180]. Rhythmic patterns are usually obtained by analyzing low-frequency amplitude modulations.

We first present two features that measure the strength of a rhythmic variation in a signal (pulse metric and band periodicity). Then we summarize features that estimate the main- and sub beats in a piece of music (beat spectrum representations, beat tracker) and finally we address features that globally represent the rhythmic structure of a piece of music (beat histograms and rhythm patterns).

Pulse metric. A measure for the "rhythmicness" of sound is proposed by Scheirer and Slaney in [162]. They detect rhythmic modulations by identifying peaks in the autocorrelation function of several subbands. The pulse metric is high when the autocorrelations in all subbands show peaks at similar positions. This indicates a strong rhythmic structure in the sound.

Band periodicity. The band periodicity also measures the strength of rhythmic structures and is similar to pulse metric [106]. The signal is split into subbands and the maximum peak of the subband correlation function is estimated for each analysis frame. The band periodicity for a subband is the mean of the peaks in all frames. It correlates with the rhythm content of a signal, since it captures the strength of repetitive structures over time.

Beat spectrum (beat spectrogram). The beat spectrum represents the self-similarity of a signal for different time lags (similarly to autocorrelation) [49, 50]. The peaks in the beat spectrum indicate strong beats with a specific repetition rate. Hence, this representation allows a description of the rhythm content of a signal. The peaks correspond to note onsets with high periodicity.

The beat spectrum is computed for several audio frames in order to obtain the beat spectrogram. Each column of the beat spectrogram is the beat spectrum of a single frame. The beat spectrogram shows the rhythmic variation of a signal over time. It is a two-dimensional representation that has the time dimension on the abscissa and the lag time (repetition rate or tempo) on the ordinate. The beat spectrogram visualizes how the tempo changes over time and allows for a detailed analysis of the rhythmic structures and variations.

Note that the beat spectrogram represents similar information as the joint acoustic and modulation frequency feature (see above). Both representations capture rhythmic content of a signal. However, the beat spectrogram represents the variation of tempo over time while the joint acoustic and modulation representation reveals rhythmic patterns independently of time. The difference between both representations is that the beat spectrogram provides temporal information while it neglects the distribution of acoustic frequencies and the modulation spectrogram preserves acoustic frequencies and neglects time. Both complement each other.

The beat spectrum serves as a basis for onset detection and the determination of rhythmically similar music. It may be used for the segmentation of pieces of music into rhythmically different parts, such as chorus and verse.

Cyclic beat spectrum. A related representation to the beat spectrum is the cyclic beat spectrum (CBS) [89]. The CBS is a compact and robust representation of the *fundamental* tempo of a piece of music. Tempo analysis with the beat spectrum reveals not only the fundamental tempo but also corresponding tempos with a harmonic and subharmonic relationship to the fundamental tempo (e.g. 1/2-, 1/3-, 2-, 3-,... fold tempo). The cyclic beat spectrum groups tempos belonging to the same fundamental tempo into one tempo class. This grouping is similar to the grouping of frequencies into chroma classes (see Section 5.4.5).

The CBS is derived from a beat spectrum. Kurth et al. first low-pass filter the signal (to remove timbre information that may be neglected for tempo analysis) and compute a spectrogram by short-time Fourier transform. They derive a novelty curve by summing the differences between adjacent spectral vectors. The novelty curve is then analyzed by a bank of comb filters where each comb filter corresponds to a particular tempo. This analysis results in a beat spectrogram where peaks correspond to dominant tempos. The beat spectrum is divided into logarithmically scaled tempo octaves (tempo classes) similarly to pitch classes in the context of chroma. The CBS is obtained by aggregating the beat spectrum over all tempo classes.

The CBS robustly estimates one or more significant and independent tempos of a signal and serves as a basis for the analysis of rhythmic structures. Kurth et al. employ the beat period (derived from the CBS) together with more complex rhythm and meter features for time-scale invariant audio retrieval [89].

Beat tracker. An important rhythm feature is Scheirer's beat tracking algorithm which enables the determination of tempo and beat positions in a piece of music [160, 161]. The algorithm starts with a decomposition of the input signal into subbands. Each subband envelope is analyzed by a bank of comb filters (resonators). The resonators extract periodic modulations from the subband envelopes and are related to particular tempos. The resonator's outputs are summed over all subbands in order to obtain an estimate for each tempo under consideration. The frequency of the comb filter with the maximum energy output represents the tempo of the signal.

An advantage of using comb filters instead of autocorrelation methods for finding periodic modulations is that they allow for the detection of the beat positions and thus enable beat-tracking. Scheirer tracks beat positions by analyzing the phase information preserved by the comb filters. The author empirically shows that the proposed technique approximates the beat-tracking abilities of human listeners. See [160] for a comparison of comb filters with autocorrelation methods and more details on the beat-tracking algorithm.

Beat histogram. The beat histogram is a compact global representation of the rhythm content of a piece of music [180, 183]. It describes the repetition rates of main beat and sub beats together with their strength. Similarly to other rhythm features, the computation is based on periodicity analysis in multiple frequency bands. The authors employ Wavelet transform in order to obtain an octave-frequency decomposition. They detect the most salient periodicities in each subband and accumulate them into a histogram. This process is similar to that of pitch histograms in Section 5.4.4.

Each bin of the histogram corresponds to a beat period in beats-per-minute where peaks indicate the main- and sub beats. The beat histogram compactly summarizes all occurring beat periods (tempos) in a piece of music. The beat histogram is designed for music information retrieval, especially genre classification. A measure for the beat strength may be easily derived from the beat histogram as in [184]. Grimaldi et al. introduce a derivation of the beat histogram in [56] that builds upon the discrete Wavelet packet transform (DWPT) [112].

Rhythm patterns. Rhythm patterns are proposed for music similarity retrieval by Pampalk et al. in [139]. They build upon the specific loudness sensation in sone (see Section 5.4.3). Given the spectrogram (in specific loudness) the amplitude modulations are extracted by a Fourier analysis of the critical bands over time. The extracted modulation frequencies are weighted according to the fluctuation strength to approximate the human perception [204]. This results in a two-dimensional representation of acoustic versus modulation frequency. A detailed description of the computation is given in [155]. Note that rhythm patterns are similar to the joint acoustic and modulation frequency features mentioned above.

5.7 Eigendomain Features

Features in this group represent long-term information contained in sound segments that have a duration of several seconds. This leads to large amounts of (redundant) feature data with low expressiveness that may not be suitable for further processing (e.g. classification).

Statistical methods may be applied in order to reduce the amount of data in a way that preserves the most important information. The employed statistical methods usually decorrelate the feature data by factorization. The resulting representation allows for dimensionality reduction by removing factors with low influence. Methods such as Principal Components Analysis (PCA) and Singular Value Decomposition (SVD) are standard techniques for this purpose.

Rate-scale-frequency (RSF) features. Ravindran et al. introduce RSF features for general purpose sound recognition in [156]. The computation of the features relies on a model of the auditory cortex and the early auditory model, used for noise-robust audio features (see NRAF in Section 5.5.2). RSF features describe modulation information for selected frequency bands of the auditory spectrum. Ravindran et al. apply a two-dimensional Wavelet transform to the auditory spectrum in order to extract temporal and spatial modulation information resulting in a three-dimensional representation. They perform PCA for compression and decorrelation of the data in order to obtain an easily processable fingerprint.

MPEG-7 audio spectrum basis/projection. The MPEG-7 standard defines the combination of audio spectrum basis (ASB) and audio spectrum projection (ASP) descriptors for general purpose sound recognition [73, 82]. ASB is a compact representation of the short-time spectrogram of a signal. The compression of the spectrogram is performed by Singular Value Decomposition. ASB contains the coarse frequency distribution of the *entire* spectrogram. This makes it suitable for general purpose sound recognition. The ASP descriptor is a projection of a spectrogram against a given audio spectrum basis. ASP and ASB are usually combined in a retrieval task as described in [83].

Distortion discriminant analysis (DDA). DDA features are used for noise-robust fingerprinting [19]. Initially, the signal is transformed using a modulated complex lapped transform (MCLT) which yields a time-frequency representation [113]. The resulting spectrogram is passed to a hierarchy of oriented Principal Component Analyses to subsequently reduce the dimensionality of the spectral vectors and to remove distortions. This hierarchical application of the oriented Principal Component Analysis yields a compact time-invariant and noise-robust representation of the entire sound.

DDA generates features that are robust to several types of noise and distortions, such as time-shifts, frequency distortions, and compression artifacts. Burges et al. point out that DDA is even robust against types of noise that are not present in the training set [20].

5.8 Phase Space Features

In speech production non-linear phenomena, such as turbulence have been observed in the vocal tract [87]. Features in the domains mentioned so far (temporal, frequency, cepstral, etc.) are not able to capture non-linear phenomena. The *state space* represents a domain that reveals the non-linear behavior of a system. However, in general it is not possible to extract the state space for an audio signal, since not all necessary variables may be derived from the audio signal. Alternatively, the *reconstructed phase space*, an approximation that shares important properties with the state space, may be computed. For phase space reconstruction the original audio signal is considered to be a one-dimensional projection of the dynamic system. The reconstructed phase space is built by creating time-lagged versions of the original signal. The original signal is shifted by multiples of a constant time lag. Each dimension of the reconstructed phase space relates to a delayed version of the original signal. The dimension of the reconstructed phase space corresponds to the number of time-lagged versions of the original signal. The critical steps in phase space reconstruction are the determination of embedding dimension and time lag. An extensive description of phase space reconstruction is given in [1]. The possibly high-dimensional attractor of the system unfolds in the phase space if time-lag and embedding dimension are properly selected. Several parameters of the attractor may serve as audio features.

The *Lyapunov exponents* of the attractor measure the "degree of chaos" of a dynamic system. Kokkinos and Maragos employ Lyapunov exponents for the distinction of different phonemes in speech [87]. They observe that phonemes, such as voiced and unvoiced fricatives, (semi)vowels, and stop sounds may be characterized by their Lyapunov exponents due to the different degree of chaos in these phonemes.

Lindgren et al. employ the *natural distribution* of the attractor together with its first derivative as features for phoneme recognition [100]. The natural distribution describes the spatial arrangement of the points of the attractor, i.e. the coarse shape of the attractor. The first derivative characterizes the flow or trajectory of the attractor over time.

Further features derived from reconstructed phase space are dimension measures of the attractor, such as *fractal dimension* [87] and *correlation dimension* [149].

Bai et al. show that phase space features are well-suited for musical genre classification [10]. They compute the angles between vectors in phase space and employ the variance of these angles as features.

Phase space features capture information that is orthogonal to features that originate from linear models. Experiments show that recognition solely based on phase space features is poor compared to results of standard features, such as MFCCs [100]. Consequently, phase space features are usually combined with traditional features in order to improve accuracy of recognition.

6 Related Literature

6.1 Application Domains

In the following we briefly present the application domains that we cover in this article together with selected references to relevant publications. The major research areas in audio processing and retrieval are automatic speech recognition (ASR), music information retrieval, environmental sound recognition (ESR), and audio segmentation. Audio segmentation (often called audio classification) is a preprocessing step in audio analysis that separates different types of sound e.g. speech, music, environmental sounds, silence, and combinations of these sounds [79, 156]. Subdomains of audio segmentation address silence detection [14, 145], the segmentation of speech and non-speech [68], and the segmentation of speech and music [140].

The segmented audio stream may be further analyzed by more specific analysis methods. ASR is probably the best investigated problem of audio retrieval [151]. However, there is still active research on audio features for ASR [6, 30, 87]. Related fields of research are speaker recognition and speaker segmentation [91, 199]. Speaker recognition deals with the identification of the speaker in an audio stream. Applications of speaker identification are authentication in safety systems and user recognition in dialog systems. Speaker segmentation determines the beginning and end of a speech segment of a particular speaker [108]. Another discipline dealing with speech is language identification where systems automatically predict the language of a speaker [45, 58, 135, 176].

Recent approaches aim at the recognition and assessment of stress and other emotions in spoken language which may help to design mood driven human computer interfaces [33, 70, 137, 157]. Further domains of speech processing are gender detection and age detection from speech [119, 181]. A novel approach is speech analysis in medical applications for the detection of illnesses that affect human speech [15].

This article further focuses on ESR-related techniques. A typical application is the classification of general-purpose sounds, such as dog barks, flute sounds or applause, which require specialized audio features [28, 102, 138]. Typical ESR tasks are surveillance applications where the environment is scanned for unusual sounds [153]. Furthermore, video analysis and annotation is a popular domain that deals with environmental sounds. Important tasks are violence detection in feature films [146] and highlight detection in video. Highlight detection addresses identification of key scenes in videos, for example in sports videos [22, 189]. Multimodal approaches improve the detection rate by combining auditory and visual information [27]. Another application is the analysis of affective dimensions in the sound track of feature films (e.g. arousal, valence) [25].

Additionally, ESR covers pattern recognition in bioacoustics. Bioacoustic pattern recognition deals among others with acoustic monitoring of animals in the wild and the discrimination and retrieval of animal sounds, such as bird song and whale sounds [32, 123].

This article further addresses features related to music information retrieval (MIR). MIR is a rapidly growing field of scientific interest due to the growing number of publicly available music databases. The main research areas of music analysis are recognition of instruments, genres, artists, and singers [42, 44, 57, 95, 117, 127, 167, 180, 200]. Music similarity retrieval addresses the identification of pieces of music that sound similar [9, 64, 75, 94]. A related task is music identification (or music recognition) where different interpretations or versions of a single piece of music are matched [34, 130]. Furthermore, research focuses on emotion detection in music. The goal of emotion detection is to classify music into categories, such as *cheerful* and *depressive* [94].

A related field is structural music analysis which addresses the extraction of repeated patters, such as chorus and verse of a piece of music [54, 107]. Additionally, the analysis of structures such as rhythm and tempo is a popular task [59, 163]. A related topic is music transcription that deals with the extraction of notes and key(s) from a piece of music [48, 202]. Music summarization and thumbnailing address the extraction of the most significant part(s) in a piece of music [13, 35, 56].

Query-by-humming (QBH) is a very popular MIR application. In a QBH application a user can search for music in a database by humming the melody of the piece of music. The matching between the hummed query and the music database usually employs content-based audio features [141, 156]. Additionally, content-based music visualization, organization and browsing techniques employ audio features for the representation of audio signals [17, 139].

We review a variety of audio features that originate from audio fingerprinting. Audio fingerprinting addresses matching of audio signals based on fingerprints [154, 173]. A fingerprint is a compact numeric representation that captures the most significant information of a signal. A popular application are information systems that retrieve the artist and title of a particular piece of music given only a short clip recorded with a mobile phone.

This article covers the most active domains of audio processing and retrieval. We have systematically reviewed the most important conference proceedings and journals that are related to audio retrieval and signal processing. The result of the literature survey is a collection of more than 200 relevant papers that address audio feature extraction.

6.2 Literature on Audio Features

The literature survey yields a large number of publications that deal with feature extraction and audio features. We organize the publications according to the addressed audio features in order to make them manageable for the reader. Tables 8 and 9 list relevant publications for each audio feature in alphabetical order and help the reader to get an overview of the literature in the context of an audio feature.

We have tried to identify the base paper for each feature. This is not always possible, since some features do not seem to have a distinct base paper, as in the case of zero crossing rate and short-time energy. In cases where no base paper exists, we have tried to identify an early paper, where the feature is mentioned. Base papers and early papers are printed in boldface.

6.3 Relevant Published Surveys

Audio feature extraction and audio retrieval both have a long tradition. Consequently several surveys have been published that cover these topics. Most related surveys focus on a single application domain, such as MIR or fingerprinting and cover a relatively small number of features. In the following, we briefly present important surveys in the field of audio feature extraction.

Audio Feature	**Selected References**
4 Hz Modulation Energy	[**162**], [31, 104]
4 Hz Modulation Harmonic Coef.	[**31**]
Adapt. Time-Frequency Transform	[**185**]
Amplitude Descriptor	[**123**]
Auditory filterbank temp. envelopes	[114]
Autocorrel. MFCCs	[**164**]
Band Periodicity	[**106**], [109]
Bandwidth	[4, 22, 32, 109, 114, 127, 154, 167, 194]
Bark-scale Frequency Cepstral Coef.	[51, 127]
Beat Histogram	[**183**], [56, 57, 98, 99, 116]
Beat Spectrum (Beat Spectrogram)	[**49, 50**], [26, 116]
Beat Tracker	[**160**],[161]
Chroma CENS Features	[**130**]
Chromagram	[**12**], [13, 54, 75, 128]
Cyclic Beat Spectrum	[**89**]
Daubechies Wavelet Coef. Histogr.	[**98**], [94, 95, 97]
Distortion Discriminant Analysis	[**19**], [20]
DWPT-based rhythm feature	[**56**], [57]
(Multi-resolution) Entropy	[15, 122, 154, 167]
(Modified) Group Delay	[**198**], [6, 61, 132, 134, 163]
Harmonic Coefficient	[**31**], [200]
Harm. Concentration	[**167**]
Harmonic Derivate	[**167**]
Harm. Energy Entropy	[**167**]
Harmonic Prominence	[**22**]
Inharmonicity	[**3**], [4, 142]
Integral Loudness	[**144**], [147]
Line Spectral Frequencies	[41, 76, 88, 106, 108]
Linear Prediction Cepstral Coef.	[**8**], [76, 81, 88, 93, 110, 131, 195]
Linear Prediction ZCR	[**41**]
Linear Predictive Coding	[78, 79, 102, 123, 152]
Mel-scale Frequency Cepstral Coef.	[**18**], [11, 27, 30, 37, 97, 127, 153, 191]
Modulation Frequency Features	[**172**], [39, 85, 138, 173]
MPEG-7 Audio Fundamental Freq.	[**73**], [83]
MPEG-7 Audio Harmonicity	[**73**], [83]
MPEG-7 Audio Power	[**73**], [83]
MPEG-7 Audio Spectrum Basis	[**73**], [82, 83]
MPEG-7 Audio Spectrum Centroid	[**73**], [83]
MPEG-7 Audio Spectrum Spread	[**73**], [83, 142, 150]

Table 8: This table contains selected references for each audio feature. Base papers and early papers are typeset in bold font.

Audio Feature	**Selected References**
MPEG-7 Audio Waveform	[**73**], [83]
MPEG-7 Harmonic Spec. Centroid/Deviation/Spread/Variation	[**73**], [83, 143]
MPEG-7 Log Attack Time	[**73**], [83, 142]
MPEG-7 Spectral Centroid	[**73**], [83]
MPEG-7 Temporal Centroid	[**73**], [83, 142]
Noise-Robust Auditory Feature	[**156**]
Perceptual Linear Prediction (PLP)	[**62**], [55, 81, 93, 110, 122]
Phase Space Features	[87, 100, 148, 149]
Pitch	[4, 25, 27, 33, 104, 180, 194]
Pitch Histogram	[**179**], [98, 99, 180, 182]
Pitch Profile	[**202**]
Pitch Synchronous ZCPA	[**52**], [53]
Psychoacoustic Pitch	[**115**]
Pulse Metric	[**162**]
Rate-scale-frequency Features	[**156**]
Relative Spectral PLP	[**63**]
Rhythm Patterns	[**139**], [155]
Sharpness	[**204**], [64, 142]
Short-Time Energy (STE)	[22, 25, 76, 32, 90, 109, 163, 168, 195]
Sone	[**139**], [127, 155]
Spectral Center	[64, 114, 163]
Spectral Centroid	[4, 22, 109, 114, 127, 154, 162, 180, 194]
Spectral Crest	[5, 10, 64, 90, 127, 142, 154, 185]
Spectral Dispersion	[163]
Spectral Flatness	[**74**], [5, 64, 69, 90, 142, 154]
Spectral Flux	[68, 76, 78, 79, 95, 162, 181, 200]
Spectral Peaks	[**186**], [187]
Spectral Peak Struct.	[**167**]
Spectral Rolloff	[76, 97, 114, 142, 150, 162, 167, 181]
Spectral Slope	[127, 142]
Subband Energy Ratio	[22, 32, 76, 102, 109, 114, 127, 154]
Subband Spectral Flux	[**22**]
Volume	[10], [76, 102, 104, 114, 127, 140]
Zero Crossing Peak Amplitudes (ZCPA)	[**80**], [51, 81]
Zero Crossing Rate (ZCR)	[4, 22, 27, 41, 76, 79, 127, 140, 162, 168]

Table 9: This table contains selected references for each audio feature. Base papers and early papers are typeset in bold font.

Lu et al. provide a survey on audio indexing and retrieval techniques in [105]. The survey describes a set of traditional time- and frequency domain features, such as harmonicity and pitch. The authors focus on feature extraction and classification techniques in the domains of speech and music. Furthermore, the survey discusses concepts of speech and music retrieval systems.

In [192] the authors present a comprehensive survey of features for multimedia retrieval. The survey covers basic short-time audio features, such as volume, bandwidth, and pitch together with aggregations of short-time features. The authors extract audio features together with video features from a set of TV programs and compute the correlation between the features in order to show redundancies.

A bibliographical study of content-based audio retrieval is presented in [38]. The survey covers a set of seven frequently used audio features in detail. The authors perform retrieval experiments in order to prove the discriminant power of the features.

Tzanetakis surveys a large set of music-related features in [179]. The author describes techniques for music analysis and retrieval, such as features for beat tracking, rhythm analysis, and pitch content description. Additionally, the author surveys traditional features that mainly originate from ASR. Finally, the survey presents a set of features that are directly computed from compressed MPEG signals.

Compressed-domain features are also presented in [188]. The authors discuss features for audio-visual indexing and analysis. The survey analyzes the applicability of traditional audio features and MPEG-7 descriptors in the compressed domain. However, the major part of the paper addresses content-based video features.

A survey of audio fingerprinting techniques is presented in [23]. Fingerprints are compact signatures of audio content. The authors review the most important recent feature extraction techniques for fingerprinting.

Peeters summarizes a large set of audio features in [142]. The author organizes the features among others in global and frame-based descriptions, spectral features, energy features, harmonic features, and perceptual features. The feature groups in [142] are similar to the groups of the taxonomy we present in Section 4.

There has been extensive research done in the field of audio feature extraction in recent years. However, we observe that most surveys focus on a small set of widely used traditional features while recent audio features are rarely addressed. In contrast to existing surveys we solely focus on feature extraction which allows us to cover a richer set of features and to introduce some structure in the field. Additionally, the survey presented in this paper covers a wide range of application domains. The advantage of this approach is that it brings features from different domains together, which facilitates the comparison of techniques with different origins.

7 Summary and Conclusions

This paper presents a survey on state-of-the-art and traditional content-based audio features originating from numerous application domains. We select a set of 77 features and systematically analyze their formal and structural properties in order to identify organizing principles that enable a categorization into meaningful groups. This leads to a novel taxonomy for audio features that assists the user in selecting adequate ones for a particular task. The taxonomy represents a novel perspective on audio features that associates techniques from different domains into one single structure.

The collection of features in this paper gives an overview of existing techniques and may serve as reference for the reader to identify adequate features for her task. Furthermore, it may be the basis for the development of novel features and the improvement of existing techniques.

Additionally, we conclude that most of the surveyed publications perform retrieval tasks on their own audio databases and ground truths. Hence, the results are not comparable. We stress that the entire field of audio retrieval needs standardized benchmarking databases and ground truths specified by domain experts who have an unbiased view on the field. Although attempts of standardized benchmarking databases in the domains of speech and music retrieval have been made more work has to be directed towards this task.

8 Acknowledgements

We want to express our gratitude to Werner A. Deutsch (Austrian Academy of Sciences, Vienna) for his suggestions that lead to considerable improvements of this work. This work has received financial support from the Vienna Science and Technology Fund (WWTF) under grant no. CI06 024.

References

[1] H. Abarbanel. *Analysis of Observed Chaotic Data*. Springer, New York, New York, 1996.

[2] A. Adami and D. Barone. A speaker identification system using a model of artificial neural networks for an elevator application. *Information Sciences*, 138(1-4):1–5, Oct. 2001.

[3] G. Agostini, M. Longari, and E. Pollastri. Musical instrument timbres classification with spectral features. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pages 97–102, Cannes, France, Oct. 2001. IEEE, IEEE.

[4] G. Agostini, M. Longari, and E. Pollastri. Musical instrument timbres classification with spectral features. *EURASIP Journal on Applied Signal Processing*, 2003(1):5–14, 2003.

[5] E. Allamanche, J. Herre, O. Helmuth, B. Frba, T. Kasten, and M. Cremer. Content-based identification of audio material using mpeg-7 low level description. In *Proceedings of the International Symposium of Music Information Retrieval*, 2001.

[6] L.D. Alsteris and K.K. Paliwal. Evaluation of the modified group delay feature for isolated word recognition. In *Proceedings of the International Symposium on Signal Processing and Its Applications*, volume 2, pages 715–718, Sydney, Australia, Aug. 2005. IEEE, IEEE.

[7] ANSI. *Bioacoustical Terminology, ANSI S3.20-1995 (R2003).* American National Standards Institute, New York, 1995.

[8] B.S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312, Jun. 1974.

[9] J.-J. Aucouturier, F. Pachet, and M Sandler. The way it sounds: timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, Dec. 2005.

[10] L. Bai, Y. Hu, S. Lao, J. Chen, and L. Wu. Feature analysis and extraction for audio automatic classification. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 767 – 772, Big Island, Hawaii, Oct. 2005. IEEE, IEEE.

[11] L. Baojie and K. Hirose. Speaker adaptive speech recognition using phone pair model. In *Proceedings of the 5th International Conference on Signal Processing*, volume 2, pages 714–717, Beijing, China, Aug. 2000.

[12] M.A. Bartsch and G.H. Wakefield. To catch a chorus: using chroma-based representations for audio thumbnailing. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 15–18, New Platz, New York, Oct. 2001. IEEE, IEEE.

[13] M.A. Bartsch and G.H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, Feb. 2005.

[14] R. Becker, G. Corsetti, J. Guedes Silveira, R. Balbinot, and F. Castello. A silence detection and suppression technique design for voice over ip systems. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, pages 173–176, Victoria, Canada, Aug. 2005. IEEE, IEEE.

[15] R. Behroozmand and F. Almasganj. Comparison of neural networks and support vector machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation. In *Proceedings of the IEEE International Symposium on Signal Processing and*

Information Technology, pages 844–849, Athens, Greece, Dec. 2005. IEEE, IEEE.

[16] B. Bogert, M. Healy, and J. Tukey. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe-cracking. In *Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed.)*, pages 209–243. New York: Wiley, 1963.

[17] E. Brazil, M. Fernström, G. Tzanetakis, and P. Cook. Enhancing sonic browsing using audio information retrieval. In *Proceedings of the International Conference on Auditory Display*, Kyoto, Japan, Jul. 2002.

[18] J.S. Bridle and M.D. Brown. An experimental automatic word recognition system. JSRU Report No. 1003, Ruislip, England: Joint Speech Research Unit, 1974.

[19] C.J.C. Burges, J.C. Platt, and S. Jana. Extracting noise-robust features from audio data. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 1021–1024, Orlando, FL, May 2002. IEEE, IEEE.

[20] C.J.C. Burges, J.C. Platt, and S. Jana. Distortion discriminant analysis for audio fingerprinting. *IEEE Transactions on Speech and Audio Processing*, 11(3):165–174, May 2003.

[21] D. Byrd and T. Crawford. Problems of music information retrieval in the real world. *Information Processing & Management*, 38(2):249–272, Mar. 2002.

[22] R. Cai, L. Lu, A. Hanjalic, H.J. Zhang, and L.H. Cai. A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on Speech and Audio Processing*, 14:1026–1039, May 2006.

[23] P. Cano, E. Batle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pages 169–173, St. Thomas, Virgin Islands, Dec. 2002. IEEE, IEEE.

[24] J.P. Champbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, Sep. 1997.

[25] C.G. Chan and G.J.F. Jones. Affect-based indexing and retrieval of films. In *Proceedings of the annual ACM International Conference on Multimedia*, pages 427–430, Singapore, Singapore, 2005. ACM Press.

[26] X. Changsheng, N.C. Maddage, S. Xi, C. Fang, and T. Qi. Musical genre classification using support vector machines. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 429–432, Hong Kong, China, Apr. 2003. IEEE, IEEE.

[27] C.C. Cheng and C.T. Hsu. Fusion of audio and motion information on hmm-based highlight extraction for baseball games. *IEEE Transactions on Multimedia*, 8(3):585–599, Jun. 2006.

[28] Y.C. Cho and S.Y. Choi, S.and Bang. Non-negative component parts of sound for classification. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, pages 633–636, Darmstadt, Germany, Dec. 2003. IEEE, IEEE.

[29] Y.D. Cho, M.Y. Kim, and S.R. Kim. A spectrally mixed excitation (smx) vocoder with robust parameter determination. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 601–604, May 1998.

[30] E.H.C. Choi. On compensating the mel-frequency cepstral coefficients for noisy speech recognition. In *Proceedings of the Australasian Computer Science Conference*, pages 49–54, Hobart, Australia, 2006. Australian Computer Society, Inc.

[31] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 865–868, Salt Lake City, Utah, May. 2001. IEEE, IEEE.

[32] W.T. Chu, W.H. Cheng, J.Y.J Hsu, and J.L. Wu. Toward semantic indexing and retrieval using hierarchical audio models. *Multimedia Systems*, 10(6):570–583, May 2005.

[33] Z.J. Chuang and C.H. Wu. Emotion recognition using acoustic features and textual content. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 53–56, Taipei, Taiwan, Jun. 2004. IEEE, IEEE.

[34] M. Clausen and F. Kurth. A unified approach to content-based and fault-tolerant music recognition. *IEEE Transactions on Multimedia*, 6(5):717–731, Oct. 2004.

[35] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130, New Paltz, New York, Oct. 2003. IEEE, IEEE.

[36] M. Cowling and R. Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907, Nov. 2003.

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

[38] M. Davy and S.J. Godsill. Audio information retrieval a bibliographical study. Technical Report, Feb. 2002.

[39] D. Dimitriadis, P. Maragos, and A. Potamianos. Modulation features for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 377–380, Orlando, FL, May 2002. IEEE, IEEE.

[40] J. Stephen Downie. Music information retrieval (chapter 7). *Annual Review of Information Science and Technology*, 37:295–340, 2003.

[41] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 2445–2448, Istanbul, Turkey, Jun. 2000. IEEE, IEEE.

[42] S. Esmaili, S. Krishnan, and K. Raahemifar. Content based audio classification and retrieval using joint time-frequency analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 665–668, Montreal, Canada, May 2004. IEEE, IEEE.

[43] S. Essid, G. Richard, and B. David. Inferring efficient hierarchical taxonomies for mir tasks, application to musical instruments. In *Proceedings of the International Conference on Music Information Retrieval*, Sept. 2005.

[44] A.M. Fanelli, L. Caponetti, G. Castellano, and C.A. Buscicchio. Content-based recognition of musical instruments. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, pages 361–364, Rome, Italy, Dec. 2004. IEEE, IEEE.

[45] J. Farinas, F.C. Pellegrino, J.-L. Rouas, and F. Andre-Obrech. Merging segmental and rhythmic features for automatic language identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 753–756, Orlando, FL, May 2002. IEEE, IEEE.

[46] H. Fastl. Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. *Hearing Research*, 8(1):59–69, Sep. 1982.

[47] H. Fletcher and W.A. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5(2):82–108, Oct. 1933.

[48] S.W. Foo and W.T. Leem. Recognition of piano notes with the aid of frm filters. In *Proceedings of the International Symposium on Control, Communications and Signal Processing*, pages 409–413, Hammamet, Tunisia, Mar. 2004. IEEE, IEEE.

[49] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 452–455, New York, NY, Aug. 2000. IEEE, IEEE.

[50] J. Foote and S. Uchihashi. The beat spectrum: a new approach to rhythm analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 881–884. IEEE, IEEE, 2001.

[51] B. Gajic and K.K. Paliwal. Robust speech recognition using features based on zero crossings with peak amplitudes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 64–67, Hong Kong, China, Apr. 2003. IEEE, IEEE.

[52] M. Ghulam, T. Fukuda, J. Horikawa, and T. Nitta. A noise-robust feature extraction method based on pitch-synchronous zcpa for asr. In *Proceedings of the International Conference on Spoken Language Processing*, pages 133–136, Jeju Island, Korea, Oct. 2004.

[53] M. Ghulam, T. Fukuda, J. Horikawa, and T. Nitta. Pitch-synchronous zcpa (ps-zcpa)-based feature extraction with auditory masking. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517– 520, Philadelphia, Pennsylvania, Mar. 2005. IEEE, IEEE.

[54] M. Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 437–440, Hong Kong, China, Apr. 2003. IEEE, IEEE.

[55] S. Greenberg and B.E.D. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1647–1650. IEEE, IEEE, Apr. 1997.

[56] M. Grimaldi, P. Cunningham, and A. Kokaram. A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. In *Proceedings of the ACM SIGMM international workshop on Multimedia information retrieval*, pages 102–108, Berkeley, California, 2003. ACM Press.

[57] M. Grimaldi, P. Cunningham, and A. Kokaram. Discrete wavelet packet transform and ensembles of lazy and eager learners for music genre classification. *Multimedia Systems*, 11(5):422–437, Apr. 2006.

[58] Q.R. Gu and T. Shibata. Speaker and text independent language identification using predictive error histogram vectors. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 36–39, Hong Kong, China, Apr. 2003. IEEE, IEEE.

[59] E. Guaus and E. Batlle. Visualization of metre and other rhythm features. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, pages 282–285, Darmstadt, Germany, Dec. 2003. IEEE, IEEE.

[60] M. Hegde, H.A. Murthy, and V.R. Gadde. Significance of joint features derived from the modified group delay function in speech processing. *EURASIP Journal on Applied Signal Processing*, 15(1):190–202, Jan. 2007. doi:10.1155/2007/79032.

[61] R.M. Hegde, H.A. Murthy, and G.V.R. Rao. Application of the modified group delay function to speaker identification and discrimination. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, Montreal, Quebec, Canada, May 2004. IEEE, IEEE.

[62] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, Apr. 1990.

[63] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2:578–589, 1994.

[64] J. Herre, E. Allamanche, and C. Ertel. How similar do songs sound? towards modeling human perception of musical similarity. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 83–86, New Paltz, NY, Oct. 2003. IEEE, IEEE.

[65] J. Herre, E. Allamanche, and O. Hellmuth. Robust matching of audio signals using spectral flatness features. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130, New Paltz, NY, Oct. 2001. IEEE, IEEE.

[66] W. Hess. *Pitch determination of speech signals : algorithms and devices*. Springer, Berlin, Germany, 1983.

[67] T. Houtgast and H.J. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77(3):1069–1077, Mar. 1985.

[68] R. Huang and J.H.L. Hansen. High-level feature weighted gmm network for audio stream classification. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1061–1064, Jeju Island, Korea, Oct. 2004.

[69] Y.C. Huang and S.K. Jenor. An audio recommendation system based on audio signature description scheme in mpeg-7 audio. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 639– 642, Taipei, Taiwan, Jun. 2004. IEEE, IEEE.

[70] Z. Inanoglu and R. Caneel. Emotive alert: Hmm-based emotion detection in voicemail messages. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 251–253, San Diego, California, USA, 2005. ACM Press.

[71] ISMIR. International conference on music information retrieval. http://ismir2004.ismir.net, 2004. last visited: September, 2009.

[72] International Organization for Standardization ISO. International standard 226, acoustics - normal equal-loudness level contours, 1987.

[73] ISO-IEC. *Information Technology - Multimedia Content Description Interface - part 4: Audio.* Number 15938. ISO/IEC, Moving Pictures Expert Group, 1st edition, 2002.

[74] N.S. Jayant and P. Noll. *Digtial Coding of Waveforms Principles and Applications to Speech and Video.* Prentice-Hall signal processing series. Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

[75] T. Jehan. Hierarchical multi-class self similarities. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 311–314, New Paltz, New York, Oct. 2005. IEEE, IEEE.

[76] H. Jiang, J. Bai, S. Zhang, and B. Xu. Svm-based audio scene classification. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 131–136, Wuhan, China, Oct. 2005. IEEE, IEEE.

[77] B. Kedem. Spectral analysis and discrimination by zero-crossings. *IEEE Proceedings*, 74:1477–1493, 1986.

[78] M. Kashif Saeed Khan, Wasfi G. Al-Khatib, and Muhammad Moinuddin. Automatic classification of speech and music using neural networks. In *MMDB '04: Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 94–99. ACM Press, 2004.

[79] M.K.S. Khan and W.G. Al-Khatib. Machine-learning based classification of speech and music. *Multimedia Systems*, 12(1):55–67, Aug. 2006.

[80] D-S. Kim, J-H. Jeong, J-W. Kim, and S-Y. Lee. Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 61–64. IEEE, IEEE, October 1996.

[81] D.S. Kim, S.Y. Lee, and R.M. Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on Speech and Audio Processing*, 7(1):55–69, Jan. 1999.

[82] H. Kim, N. Moreau, and T. Sikora. Audio classification based on MPEG-7 spectral basis representations. *In IEEE Trans. on Circuits and Systems for Video Technology*, 14:716–725, 2004.

[83] H. Kim, N. Moreau, and T. Sikora. *MPEG-7 audio and beyond.* Wiley, West Sussex, England, 2005.

[84] B. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25:117–132, 1998.

[85] T. Kinnunen. Joint acoustic-modulation frequency for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 665–668, Toulouse, France, May 2006. IEEE, IEEE.

[86] A. Klapuri and M. Davy. *Signal processing methods for music transcription.* Springer, New York, NY, 2006.

[87] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6):1098–1109, Nov. 2005.

[88] A.G. Krishna and T.V. Sreenivas. Music instrument recognition: from isolated notes to solo phrases. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 265–268, Montreal, Canada, May 2004. IEEE, IEEE.

[89] F. Kurth, T. Gehrmann, and M. Müller. The cyclic beat spectrum: Tempo-related audio features for time-scale invariant audio identification. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 35–40, Victoria, Canada, Oct. 2006.

[90] R. Lancini, F. Mapelli, and R. Pezzano. Audio content identification by using perceptual hashing. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 739–742, Taipei, Taiwan, Jun. 2004. IEEE, IEEE.

[91] K.Y. Lee. Local fuzzy pca based gmm with dimension reduction on speaker identification. *Pattern Recogn. Lett.*, 25(16):1811–1817, 2004.

[92] M.S. Lew. *Principles of visual information retrieval.* Springer, London, Great Britain, Jan. 2001.

[93] Q. Li, F.K. Soong, and O. Siohan. An auditory system-based feature for robust speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 619–622, Aalborg, Denmark, Sep. 2001. International Speech Communication Association.

[94] T. Li and M. Ogihara. Content-based music similarity search and emotion detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 705–708, Montreal, Quebec, Canada, May 2004. IEEE, IEEE.

[95] T. Li and M. Ogihara. Music artist style identification by semi-supervised learning from both lyrics and content. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 364–367, New York, NY, 2004. ACM Press.

[96] T. Li and M. Ogihara. Music genre classification with taxonomy. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 197–200. IEEE, IEEE, Mar. 2005.

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

[98] T. Li, M. Ogihara, and Li Q. A comparative study on content-based music genre classification. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 282–289, Toronto, Canada, 2003. ACM Press.

[99] T. Li and G. Tzanetakis. Factors in automatic musical genre classification of audio signals. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 143–146, New Paltz, New York, Oct. 2003. IEEE, IEEE.

[100] A.C. Lindgren, M.T. Johnson, and R.J. Povinelli. Speech recognition using reconstructed phase space features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 60–63, Hong Kong, China, Apr. 2003. IEEE, IEEE.

[101] M. Liu and C Wan. Feature selection for automatic classification of musical instrument sounds. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 247–248. ACM Press, 2001.

[102] M. Liu and C. Wan. A study on content-based classification and retrieval of audio database. In *Proceedings of the International Symposium on Database Engineering and Applications*, pages 339–345, Grenoble, France, Jul. 2001. IEEE Computer Society.

[103] Z. Liu, J. Huang, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene classification. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pages 343–348, Princeton, NJ, Jun. 1997. IEEE, IEEE.

[104] Z. Liu, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. *The Journal of VLSI Signal Processing*, 20(1-2):61–79, Oct. 1998.

[105] G. Lu. Indexing and retrieval of audio: A survey. *Multimedia Tools and Applications*, 15(3):269–290, Dec. 2001.

[106] L. Lu, H. Jiang, and H.J. Zhang. A robust audio classification and segmentation method. In *Proceedings of the 9th ACM international conference on Multimedia*, pages 203–211, Ottawa, Canada, 2001. ACM Press.

[107] L. Lu, M. Wang, and H.J. Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 275–282, New York, NY, 2004. ACM Press.

[108] L. Lu and H.J. Zhang. Unsupervised speaker segmentation and tracking in real-time audio content analysis. *Multimedia Systems*, 10(4):332–343, Apr. 2005.

[109] L. Lu, H.J. Zhang, and S.Z. Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems*, 8(6):482–492, Apr. 2003.

[110] C. Lvy, G. Linars, and P. Nocera. Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems. In *Proceedings of the Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan, Apr. 2003.

[111] N. Maddage, C. Xu, M. Kankanhalli, and X. Shao. Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the ACM International Conference on Multimedia*, pages 112–119. ACM, 2004.

[112] S. Mallat. *A wavelet tour of signal processing*. Academic Press, San Diego, California, 1999.

[113] H. Malvar. A modulated complex lapped transform and its applications to audio processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1421–1424, Phoenix, AZ, Mar. 1999. IEEE, IEEE.

[114] M.F. McKinney and J. Breebaart. Features for audio and music classification. In *Proceedings of the International Conference on Music Information Retrieval*, Oct. 2003.

[115] R. Meddis and L. O'Mard. A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, 102(3):1811–1820, Sep. 1997.

[116] A. Meng, P. Ahrendt, and J. Larsen. Improving music genre classification by short time feature integration. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 497–500, Philadelphia, Pennsylvania, Mar. 2005. IEEE, IEEE.

[117] A. Mesaros, E. Lupu, and C. Rusu. Singing voice features by time-frequency representations. In *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, volume 1, pages 471–475, Rome, Italy, Sep. 2003. IEEE, IEEE.

[118] I. Mierswa and K. Morik. Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58(2-3):127–149, Feb. 2005.

[119] N. Minematsu, M. Sekiguchi, and K. Hirose. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 137–140, Orlando, FL, May 2002. IEEE, IEEE.

[120] MIREX. Music information retrieval evaluation exchange. http://www.music-ir.org/ mirexwiki, 2007. last visited: September, 2009.

[121] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky. Spectral entropy based feature for robust asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 193–196, Montreal, Canada, May 2004. IEEE, IEEE.

[122] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard. Multi-resolution spectral entropy feature for robust asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 253–256, Philadelphia, Pennsylvania, Mar. 2005. IEEE, IEEE.

[123] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder. Discrimination and retrieval of animal sounds. In *Proceedings of IEEE Multimedia Modelling Conference*, pages 339–343, Beijing, China, Jan. 2006. IEEE, IEEE.

[124] B.C.J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, Amsterdam, The Netherlands, 5th edition, 2004.

[125] B.C.J. Moore and B.R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753, Sep. 1983.

[126] C.J. Moore, R.W. Peters, and B.R. Glasberg. Auditory filter shapes at low center frequencies. *Journal of the Acoustical Society of America*, 88(1):132–140, 1990.

[127] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken. Modeling timbre distance with temporal statistics from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):81–90, Jan. 2006.

[128] F. Mörchen, A. Ultsch, M. Thies, I. Löhken, M. Nöcker, C. Stamm, N. Efthymiou, and M. Kümmerer. Musicminer: Visualizing timbre distances of music as topographical maps. Technical Report, 2005.

[129] M. Müller. *Information retrieval for music and motion*. Springer, Berlin, Germany, 2007.

[130] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 288–295, London, Great Britain, Sep. 2005.

[131] R. Muralishankar and A.G. Ramakrishnan. Pseudo complex cepstrum using discrete cosine transform. *International Journal of Speech Technology*, 8(2):181–191, Jun. 2005.

[132] H.A. Murthy and V. Gadde. The modified group delay function and its application to phoneme recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 68–71, Hong Kong, China, April 2003. IEEE, IEEE.

[133] H.A. Murthy, K.V.M. Murthy, and B. Yegnarayana. Formant extraction from fourier transform phase. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 484–487, May 1989.

[134] T. Nagarajan and H.A. Murthy. Subband-based group delay segmentation of spontaneous speech into syllable-like units. *EURASIP Journal on Applied Signal Processing*, 2004(17):2614–2625, 2004.

[135] J. Navratil. Spoken language recognition-a step toward multilinguality in speechprocessing. *IEEE Transactions on Speech and Audio Processing*, 9(6):678–685, Sep. 2001.

[136] A.M. Noll. Short-time spectrum and ”cepstrum” techniques for vocal-pitch detection. *The Journal of the Acoustical Society of America*, 36(2), 1964.

[137] T.L. Nwe, S.W. Foo, and L.C. De Silva. Classification of stress in speech using linear and nonlinear features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 9–12, Hong Kong, China, Apr. 2003. IEEE, IEEE.

[138] L. Owsley, L. Atlas, and C Heinemann. Use of modulation spectra for representation and classification of acoustic transients from sniper fire. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 1129–1132, Philadelphia, Pennsylvania, Mar. 2005. IEEE, IEEE.

[139] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 570–579. ACM Press, 2002.

[140] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on Multimedia*, 7(1):155–166, Feb. 2005.

[141] S. Pauws. Cubyhum: A fully operational query by humming system. In *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, Oct. 2002. IRCAM - Centre Pompidou, IRCAM - Centre Pompidou.

[142] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical Report, 2004.

[143] G. Peeters, S. McAdams, and P. Herrera. Instrument description in the context of mpeg-7. In *Proceedings of International Computer Music Conference*, Berlin, Germany, Aug. 2000.

[144] S. Pfeiffer. The importance of perceptive adaptation of sound features for audio content processing. In *Proceedings SPIE Conferences, Electronic Imaging 1999, Storage and Retrieval for Image and Video Databases VII*, pages 328–337, San Jose, California, Jan. 1999.

[145] S. Pfeiffer. Pause concepts for audio segmentation at different semantic levels. In *Proceedings of the ACM International Conference on Multimedia*, pages 187–193, Ottawa, Canada, 2001. ACM Press.

[146] S. Pfeiffer, S. Fischer, and E. Effelsberg. Automatic audio content analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 21–30, Boston, Massachusetts, 1996. ACM Press.

[147] S. Pfeiffer, R. Lienhart, and W. Effelsberg. Scene determination based on video and audio features. *Multimedia Tools and Applications*, 15(1):59–81, Sep. 2001.

[148] V. Pitsikalis, I. Kokkinos, and P. Maragos. Nonlinear analysis of speech signals: Generalized dimensions and lyapunov exponents. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 817–820, Geneva, Switzerland, Sep. 2003.

[149] V. Pitsikalis and P. Maragos. Speech analysis and feature extraction using chaotic models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 533–536, Orlando, FL, May 2002. IEEE, IEEE.

[150] T. Pohle, E. Pampalk, and G. Widmer. Evaluation of frequently used audio features for classification of music into perceptual categories. In *Proceedings of the 4th International Workshop Content-Based Multimedia Indexing*, Riga, Latvia, 2005.

[151] L. Rabiner and B. Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.

[152] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1978.

[153] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 158–161, New Paltz, New York, Oct. 2003. IEEE, IEEE.

[154] A. Ramalingam and S. Krishnan. Gaussian mixture modeling using short time fourier transform features for audio fingerprinting. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1146–1149, Amsterdam, The Netherlands, Jul. 2005. IEEE, IEEE.

[155] A. Rauber, E. Pampalk, and D. Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity. In *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, Oct. 2002. IRCAM - Centre Pompidou, IRCAM - Centre Pompidou.

[156] S. Ravindran, K. Schlemmer, and D. Anderson. A physiologically inspired method for audio classification. *EURASIP Journal on Applied Signal Processing*, 2005(9):1374–1381, 2005.

[157] A.A. Razak, M.H.M. Yusof, and R. Komiya. Towards automatic recognition of emotion in speech. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, pages 548–551, Darmstadt, Germany, Dec. 2003. IEEE, IEEE.

[158] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613620, 1975.

[159] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 993–996, Atlanta, GA, May 1996. IEEE, IEEE.

[160] E. Scheirer. Tempo and beat analysis of acoustic musical signals. *Joint Acoustic Society of America*, 103(1):588–601, Jan. 1998.

[161] E. Scheirer. *Music-Listening Systems*. PhD. Thesis, Program in Media Arts and Sciences. MIT, Cambridge, MA, 2000.

[162] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1331–1334, Munich, Germany, Apr. 1997.

[163] W.A. Sethares, R.D. Morris, and J.C. Sethares. Beat tracking of musical performances using low-level audio features. *IEEE Transactions on Speech and Audio Processing*, 13(2):275–285, Mar. 2005.

[164] B.J. Shannon and K.K. Paliwal. Mfcc computation from magnitude spectrum of higher lag autocorrelation coefficients for robust speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 129–132, Oct. 2004.

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

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

[167] H. Srinivasan and M. Kankanhalli. Harmonicity and dynamics-based features for audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 321–324, Montreal, Canada, May 2004. IEEE, IEEE.

[168] S. Srinivasan, D. Petkovic, and D. Ponceleon. Towards robust features for classifying audio in the cuevideo system. In *Proceedings of the 7th ACM international conference on Multimedia (Part 1)*, pages 393–400. ACM Press, 1999.

[169] S.S. Stevens. The relation of pitch to intensity. *The Journal of the Acoustical Society of America*, 6(3):150–154, 1935.

[170] S.S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153–181, May 1957.

[171] S.S. Stevens, J. Volkmann, and Newman. E.B. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, Jan. 1937.

[172] S. Sukittanon and L.E. Atlas. Modulation frequency features for audio fingerprinting. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1773–1776, Orlando, FL, May 2002. IEEE, IEEE.

[173] S. Sukittanon, L.E. Atlas, and W.J. Pitton. Modulation-scale analysis for content identification. *IEEE Transactions on Signal Processing*, 52(10):3023 – 3035, 2004.

[174] H. Terasawa, M. Slaney, and J. Berger. Perceptual distance in timbre space. In *Proceedings of Eleventh Meeting of the International Conference on Auditory Display*, pages 61–68, Limerick, Ireland, Jul. 2005.

[175] E. Terhardt. Zur Tonhöhenwahrnehmung von Klängen. I. Psychoakustische Grundlagen. *Acoustica*, 26:173–186, 1972.

[176] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *Proceedings of the International Conference on Spoken Language Processing*, pages 89–92, Denver, CO, Sep. 2002.

[177] J.Y. Tourneret. Statistical properties of line spectrum pairs. *Signal Processing*, 65(2):239–255, Mar. 1998.

[178] T. Tremain. The government standard linear predictive coding algorithm: Lpc-10. *Speech Technology Magazine*, 1:40–49, Apr. 1982.

[179] G. Tzanetakis. *Manipulation, analysis and retrieval systems for audio signals.* PhD. Thesis. Computer Science Department, Princeton University, 2002.

[180] G. Tzanetakis. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, Jul. 2002. 2002.

[181] G. Tzanetakis. Audio-based gender identification using bootstrapping. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 432– 433, Victoria, Canada, Aug. 2005. IEEE, IEEE.

[182] G. Tzanetakis, A. Ermolinskyi, and P. Cook. Pitch histograms in audio and symbolic music information retrieval. *Journal of New Music Research*, 32(2):143 – 152, Jun. 2003.

[183] G. Tzanetakis, G. Essl, and P. Cook. Audio analysis using the discrete wavelet transform. In *Proceedings of the International Conference on Acoustics and Music: Theory and Applications*, Malta, Sep. 2001.

[184] G. Tzanetakis, G. Essl, and P. Cook. Human perception and computer extraction of musical beat strength. In *Proceedings of the International Conference on Digital Audio Effects*, pages 257–261, Hamburg, Germany, Sep. 2002.

[185] K. Umapathy, S. Krishnan, and S. Jimaa. Multigroup classification of audio signals using time–frequency parameters. *IEEE Transactions on Multimedia*, 7(2):308–315, Apr. 2005.

[186] A. Wang. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval*, pages 7–13, Baltimore, Maryland, Oct. 2003.

[187] A. Wang. The shazam music recognition service. *Communications of the ACM*, 49(8):44–48, Aug. 2006.

[188] A. Wang, A. Divakaran, A. Vetro, S.F. Chang, and H. Sun. Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, 14(2):150–183, Jun. 2003.

[189] K. Wang and C. Xu. Robust soccer highlight generation with a novel dominant-speech feature extractor. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 591–594, Taipei, Taiwan, Jun. 2004. IEEE, IEEE.

[190] X. Wang, Y. Dong, J. Hakkinen, and O. Viikki. Noise robust chinese speech recognition using feature vector normalization and higher-order cepstral coefficients. In *Proceedings of the 5th International Conference on Signal Processing*, volume 2, pages 738–741, Aug. 2000.

[191] X. Wang, Y Dong, J. Häkkinen, and O. Viikki. Noise robust chinese speech recognition using feature vector normalization and highter-order cepstral coefficients. In *Proceedings of the 5th International Conference on Signal Processing*, volume 2, pages 738–741, Beijing, China, Aug. 2000.

[192] Y. Wang, Z. Liu, and J.C. Huang. Multimedia content analysis-using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, Nov. 2000.

[193] R.L. Wegel and C.E. Lane. The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physical Review*, 23:266 – 285, Feb. 1924.

[194] T. Wold, D. Blum, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):2736, 1996.

[195] C. Xu, N.C. Maddage, and X. Shao. Automatic music classification and summarization. *IEEE Transactions on Speech and Audio Processing*, 13(3):441–450, May 2005.

[196] M. Xu, L. Duan, L. Chia, and C. Xu. Audio keyword generation for sports video analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 758–759, 2004.

[197] X. Yang, K. Wang, and S. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38(2):824–839, Mar. 1992.

[198] B. Yegnanarayan and H.A. Murthy. Significance of group delay functions in spectrum estimation. *IEEE Transactions on Signal Processing*, 40(9):2281–2289, Sep. 1992.

[199] K.H. Yuo, T.H. Hwang, and H.C. Wang. Combination of autocorrelation-based features and projection measure technique for speaker identification. *IEEE Transactions on Speech and Audio Processing*, 13(4):565–574, Jul. 2005.

[200] T. Zhang. Automatic singer identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 33–36. IEEE, IEEE, Jul 2003.

[201] T. Zhang and C. C. J. Kuo. *Content-Based Audio Classifcation and Retrieval for Audiovisual Data Parsing.* Kluwer Academic Publishers, Boston, Massachusetts, 2001.

[202] Y. Zhu and M.S. Kankanhalli. Precise pitch profile feature extraction from musical audio for key detection. *IEEE Transactions on Multimedia*, 8(3):575–584, Jun. 2006.

[203] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33:248, 1961.

[204] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models.* Springer, Berlin, Heidelberg, Germany, 2nd edition, 1999.