CHAPTER 8

MUSICAL GENRE, SIMILARITY, AND MOOD

The automatic recognition of the musical genre or the *musical style* from an audio signal is one of the oldest subjects of ACA and can be regarded as one of the key areas leading to today's research field MIR.

The classification into musical genre can in some ways be seen as a special case of a more generalized music similarity measure in the sense that the similarity measure in musical genre classification is restricted to dimensions which are meaningful for genre. The signals are sorted into pre-defined similarity clusters, the musical genres.

Applications for these technologies can be found mainly in the annotation, sorting, and retrieval of (related) audio files from large databases or the Internet. For music similarity there are also somewhat more creative use cases such as the automatic generation of playlists and the generation of so-called mash-ups, mixes of two or more "compatible" songs.

This chapter will cover musical genre classification, music similarity measures, mood classification as well as instrument recognition. All these systems have in common that they try to represent a property (such as the genre) of the audio signal as either a feature vector or a (time-) series of feature vectors; these feature vectors are then used either for classification or for a distance measure to retrieve the required result.

8.1 Musical Genre Classification

The first publications on the automatic recognition of musical style appeared in the 1990s; at that time, the focus was mainly on the discrimination of speech and music signals [275–

279], although algorithms aiming at classifying a broader range of signals can be found during that decade as well [280, 281]. In a way, the task of automatic musical genre classification is a classic machine learning task — suitable features are extracted from the audio signal and with these features a classification system is trained and used for the classification task.

As pointed out above, the classification into musical genre can be interpreted as a measure of music similarity restricted to certain genre-defining dimensions and categorized to pre-defined classes, the genres. Similarity can in principle have other dimensions as well which are unimportant for the genre definition (see Sect. 8.2.1.1).

General surveys of approaches to musical genre classification have been published by Scaringella et al. and Fu et al. [282, 283].

8.1.1 Musical Genre

At first glance the meaning of the term *musical genre* appears to be self-explanatory and its intuitive definition obvious. On a more thorough investigation, however, it has to be concluded that an objective definition of the term is hardly possible. Some of the reasons for this have been summarized by Pachet and Cazaly [284] and later by Scaringella et al. [282]:

- *Scope of the genre label*: Can an individual song be classified into a genre or does the context of album and performing artist influence or overrule the classification decision? Or may even different parts of a piece of music have different genres?

- *Non-agreement of taxonomies*: The number and definition of genre labels can strongly vary; in the year 2000 Pachet and Cazaly compared genre taxonomies from the three web sites Allmusic, Amazon, and MP3.com. The overall number of genres varied from 430 (MP3.com) to 531 (Allmusic) to 719 (Amazon), and these labels had only 70 terms in common. Furthermore, the hierarchy of the taxonomies differed.

  In practical applications of ACA the number of genres has to be more restricted due to technical limitations of the classification systems used, but the underlying problems of defining a taxonomy remain the same. Figure 8.1 shows two taxonomies defined in the context of early musical genre classification systems.

- *Ill-defined genre labels*: There is a semantic confusion between genre labels. They can be geographically defined (*Indian music*), related to an era in music history (*baroque*), refer to technical requirements (*barbershop*), the instrumentation (*symphonic music*), or usages (*Christmas songs*).

- *Scalability of genre taxonomies*: A specific genre may be split into a variety of subgenres (e.g., *Hip Hop* into *Gangsta Rap*, etc.). The number of subgenres might evolve over the years.

- *Non-orthogonality of genre categories*: One piece of music can possibly be sorted into multiple genre categories at the same time.

Given these observations it becomes clear that deriving a complete and musicologically consistent taxonomy is practically impossible [218]. Without a clear definition of the term *musical genre* it also comes as no surprise that the performance of humans annotating pieces with genre labels is far from perfect [285, 286]. Automatic systems can be hardly expected to outperform humans at this task.

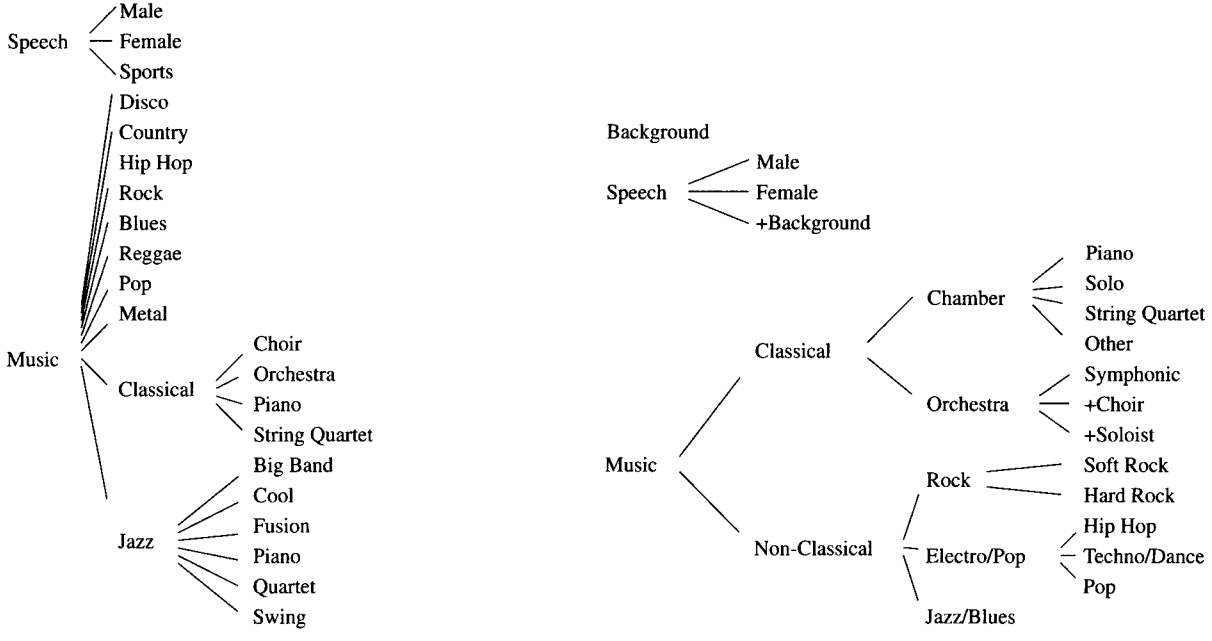

Figure 8.1 Two examples for author-defined genre taxonomies. Left: Tzanetakis and Cook [60], Right: Burred and Lerch [218]

8.1.2 Feature Extraction

From a perceptual point of view, musical genre categorization may be influenced by

- *temporal characteristics* such as the tempo, the time signature, and rhythmic patterns,
- *dynamic characteristics* such as the loudness range, the change of loudness over time, accents,
- *tonal characteristics* such as melodic properties, the complexity of harmony (progression), and prominent pitch classes,
- *production-related characteristics* such as a specific sound quality and volume relations between instruments, and stereo panning properties, and
- *instrumentational characteristics* such as the number and type of instruments used.

If for simplicity's sake the latter two are grouped into a timbre category, the resulting categories are basically tonal, temporal, intensity, and timbre (plus technical) signal properties which were introduced in Sect. 1.1.

The early audio classification systems utilized only a restricted set of features; they mainly used the zero crossing rate (see Sect. 3.4.3) and intensity-related features (see Chap. 4) [275, 276, 287, 288], although timbre-related spectral features quickly got added to the set of features [280, 281, 289]. Pitch-related features which included properties of the fundamental frequency variation could also be found in the early speech/music classification systems [278, 290, 291] but are usually not used for systems targeting polyphonic audio input.

Over the years the number of features grew more and more, eventually covering more or less all features presented in Chap. 3 and adding even more instantaneous features [61, 292].

The most common features remained related to intensity and timbre but, although the classification results with these features seem to be surprisingly good, other feature dimensions got added to the set of features. These additional features include temporal and rhythmic features derived from a beat histogram (see Sect. 6.4) [60, 218, 293], simple tonal features such as pitch histogram features [157], and stereo panning features [294].

8.1.2.1 Texture Window

The instantaneous features, i.e., the features which do not require a long time window such as histogram-based features, are usually processed per *texture window* as described in Sect. 3.5 (low-pass, derivative, subfeatures, etc.). A classification decision can be made for each texture window.

The impact of the texture window length on the classification accuracy has been studied by Tzanetakis and Cook [60], Burred and Lerch [218], and Ahrendt et al. [295] with different results: Tzanetakis and Cook found that the classification accuracy does not improve after a texture window length of 1 s, Burred and Lerch identified a texture window of approximately 15 s to yield satisfactory results, and Ahrendt et al. found a window length of 5 s to be sufficient in most cases. No conclusions can be drawn from this result except that the optimal texture window size depends strongly either on the features and subfeatures used or on the test data set and its categories.

It is worth noting that humans seem to excel at quickly identifying the musical genre; in a data set with 10 genres test subjects were able to identify the genre reliably after listening to audio snippets significantly shorter than 1 s [296].

A hierarchical form of the texture window approach is the default feature extraction mode in the software environment *Marsyas*[1] [297]: statistical subfeatures are computed for the texture window but are not directly used as classifier input but are in turn subjected to the computation of statistical subfeatures (or "*sub*subfeatures") in a longer "super"-texture window.

8.1.3 Classification

From a machine learning point of view it is obvious that any arbitrary classifier can be trained on and applied to the previously extracted features. The differences between the different approaches can be mainly found in the choice of the classifier type and in the genre categorization, mainly through the number of classes but also through the definition of a flat or a hierarchical genre tree. The latter has the advantage of branch-specific feature weighting at the cost of multiple classification steps [218].

Over the years basically all known classification approaches have been evaluated for musical genre classification. The most common are

- the *K-Nearest Neighbor (KNN)* classifier which evaluates the number of the closest training examples in the feature space,
- the *Gaussian Mixture Model (GMM)* which models the classes' feature distribution with multiple Gaussians,
- *Artificial Neural Networks (ANNs)* which are computational models inspired by the structure of biological neural networks [127], and finally
- *Support Vector Machines (SVMs)*, state-of-the-art classifiers transforming the features into a high dimensional space and finding the optimal separating hyperplane between the closest data points [298]; SVMs can nowadays be considered a standard tool in musical genre classification.

A typical way of assessing classification accuracy is *N-fold cross validation*. It is a method to ensure that training set and test set are not identical in order to avoid unrealistically good evaluation results. A standard value for N would be 10, meaning that the data is partitioned into 10 subsets. Of these subsets, 9 are used to train the classification system and the remaining subset is used for testing. This process is repeated until all 10 subsets have been used once for testing, and the overall performance can be estimated by averaging the 10 individual results. Leave-one-out cross validation is an extreme case where N equals the number of data observations minus 1.

Evaluation results of the classification performance strongly depend on the number of classes and the diversity, noisiness, and other general properties of the test set. The MIREX results indicate a classification accuracy of 50–80% for state-of-the-art systems and 10 genre categories. It is self-evident that a 2-class system will yield better results than, for instance, a system with 20 classes. During the first few years of evaluating musical genre classification systems, the aspect of having several songs by the same artist in the same class was neglected, leading to an artist-related training and comparably high classification performance [299].

[1] Marsyas. http://http://marsyas.info. Last retrieved on Dec. 1, 2011.

8.2 Related Research Fields

There are fields of MIR that share so many similarities with musical genre classification that from a engineer's point of view they can be summarized here rather then being individual chapters. Besides a general music similarity detection these technologies include mood classification, instrument recognition, and artist identification. While there exist perceptual and musicological differences between such systems, the technical approaches to solving those problems are closely related as they use similar feature sets and similar classification systems. In the following, we will focus on instructions to music similarity detection, mood classification, and instrument recognition.

8.2.1 Music Similarity Detection

As pointed out above, *music similarity detection* is similar to musical genre classification since the genre itself is a grouping of songs with similar acoustic or musical properties. From this point of view, music similarity detection differs from musical genre classification in the replacement of the classification itself by a distance or similarity measure or by a grouping rule.

Reducing music similarity to genre similarity, however, would be too simplifying since genre similarity is a subset of music similarity. This complicates the definition and evaluation of the latter even more than the definition and evaluation of musical genre. Musical genre classification can at least have a somewhat verified ground truth generated, for example, by manual annotation and categorization of music databases; there can be no such database for similarity measures as long as the term *music similarity* should mean more than just *genre similarity*. This is due to the multi-dimensional and probably associative character of music similarity; since the meaning of music similarity largely depends on both the individual user and the task at hand, this problem probably cannot be solved systematically in general. There remains a gap between what music psychologists and empirical musicologists know to be the music similarity and the simplified similarity definitions that signal processing and machine learning experts attempt to train.

8.2.1.1 Music Similarity

The similarity of two pieces of music can have many facets, and there is research on the number and characteristics of individual perceptual dimensions of music similarity. In the following, only a few publications will be named to exemplify the different approaches and the number of dimensions. Two pieces of music may, for example, be similar with respect to rhythm [198, 300], structure [301], surface and texture [302], melody and motives [303, 304], harmony [305], as well as with respect to performance attributes such as articulation, tempo variation, and dynamics [305, 306]. MacAdams et al. categorized these types of similarity into three clusters, surface and texture, figural, and structural [307]. There are indications that subjective music similarity ratings depend amongst other things on the familiarity of the test subject with the music [308]. They might also depend on the listener's expert level as well.

The associative nature of memory cannot be neglected in real-world applications as editorial data will also be influencing a human similarity decision. Editorial data refers to data that cannot be extracted from the audio signal such as recording date and studio, artists and producers who participated in other albums, the label, etc.

From an application developer point of view it is also possible to define music similarity on a technical level; one may, for instance, simply look for songs with the same tempo or related musical key.

8.2.1.2 Features

The type and number of features used in music similarity detection is closely related to those for musical genre classification. The early systems utilize very small (timbre-related) feature sets such as MFCCs [281, 309–311], then other features such as loudness-related and rhythm-related features are added [58, 312], and nowadays basically all low-level features and mid-level features introduced in the preceding chapters are investigated for music similarity detection [313, 314].

The representation of similarity features per texture window or recording differs in some cases from the statistical subfeatures known from musical genre classification. Logan and Salomon use a K-means clustering algorithm to find a good average description of the spectral envelope with MFCCs [309], Aucouturier and Pachet model this spectral envelope with GMMs [311], and Pampalk et al. use a histogram containing level classes per frequency band [58].

8.2.1.3 Similarity Measure

The pieces of music are represented as (normalized) feature vectors. The simplest approach to a similarity measure is to compute a simple vector distance between those feature vectors. Typical pairwise distances would be the Euclidean distance, the Manhattan distance, and the cosine distance.

Instead of computing the pairwise distance it is common to either automatically group the vectors or to map them to a lower-dimensional space by means of unsupervised machine learning algorithms. Two examples of such methods are

- *K-means clustering*: K-means clustering aims at clustering the vectors into K groups by minimizing the intra-cluster variance. The standard approach to K-means clustering has the following algorithmic steps:
  1. *Initialization*: randomly select K vectors from the data set as initialization.
  2. *Update*: compute the mean for each cluster.
  3. *Assignment*: assign each observation to the cluster with the mean of the closest cluster.
  4. *Iteration*: go to step 2 until the clusters converge.
- *Self-Organizing Map (SOM)*: A SOM is a form of an ANN which produces a two-dimensional map as a representation of the (higher dimensional) training samples. In its simplest form, it features the following algorithmic steps:
  1. *Initialization*: specify (randomly or deterministically) a set of nodes spanning a net. Each node is represented by a weight vector with the same dimension as feature vectors.
  2. *Update*: pick a training sample (feature vector) and compute the distance to all nodes. Update the weight vectors of nodes at a close distance to the training sample to move them toward the training vector. The amount of increment depends on the proximity of the node and the training sample and possibly decreases with an increasing number of iterations.
  3. *Iteration*: go to step 2 until the maximum number of iterations is reached.

8.2.2 Mood Classification

The mood of a piece of music is one of the cues a typical user finds helpful in finding and browsing music [315]. This has lead to an increasing amount of research targeted at automatically recognizing the emotional characteristics of recordings of music. Terms for this field are, for instance, (audio) *mood classification* and *music emotion recognition*.

8.2.2.1 Emotion and Mood in Music

Understanding the meaning of the terms *emotion* and *mood* seems to be essential for the successful design of mood classification systems. Unfortunately, there is no established understanding of what emotion and mood actually are. This is true not only in the context of music but also in general. Kleinginna and Kleinginna, for example, reviewed more than a hundred scientific definitions of emotion [316] without being able to identify a consensus. There might be better or worse definitions but in the end it is just not possible to prove the correctness of an individual definition.

Weld described the difference between emotion and mood by characterizing emotion as temporary and evanescent in contrast to mood which is more permanent and stable [317]; mood can also be seen as a diffuse affect state which can emerge without apparent cause [318]. However, whether the term *emotion* or the term *mood* is more fitting in a musical context is unclear.

Researching the relation of emotion and music exposes some of the typical problems in psychological research; one problem is, for example, the process of verbalization which confines the description of subtle and varied emotional states to the standardized words used to denote them [319]. Meyer also points out that descriptions of emotions are usually apocryphal and misleading since emotions are named and distinguished largely in terms of the external circumstances in which the response takes place; music itself, however, presents no external circumstances [319]. Scherer criticizes the tendency to assume that music evokes basic emotions such as anger, fear, etc. [320]. This led him to propose the differentiation between *utilitarian* emotions, which are the emotions usually studied in emotion research (anger, fear, joy, disgust, sadness, shame, guilt, etc.), and *aesthetic* emotions which are not driven by external influences or personal goals but rather by the appreciation of the intrinsic qualities of a work of art [321]. Zentner et al. found that negative emotions such as guilt, shame, disgust, anger, fear, etc. are practically never aroused by music [322].

Recent research indicates that the "description of musical emotions requires a more nuanced affect vocabulary and taxonomy than is provided by current scales and models of emotion" [322].

It is of importance to distinguish between the emotion aroused in the listener and the conveyed emotion perceived by a listener without particularly feeling it [319]. Ratings of *perceived* emotion differ significantly from ratings of *felt* emotion [322].

To complicate matters further, it might be of interest to differentiate between score-inherent emotions and performance-inherent emotions. To study this, however, seems to be only possible in very controlled environments [323–325]; a real-world scenario does not allow for such a differentiation as the performance is an integral part of the "music" (see also Chap. 10). Also, the differentiation between score and music performance is mainly made for non-popular or classical music as opposed to popular music.

When assessing the mood of a musical piece, the usual approach is to rely either on models or on label categories. Russel's two-dimensional emotion space is one of the frequently

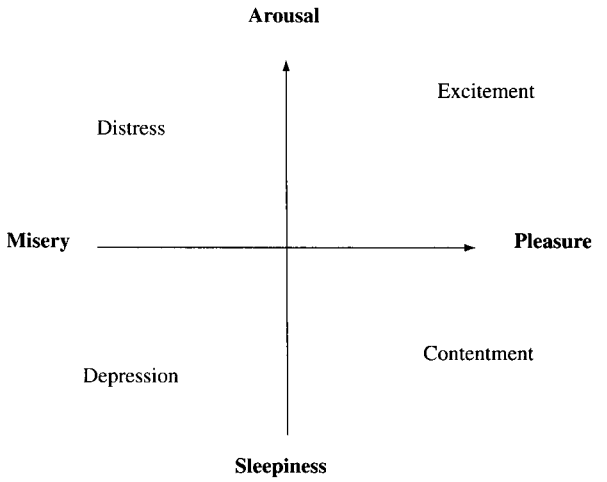

Figure 8.2 Russel's two-dimensional model of mood

used models [326]. It describes mood by the two dimensions *pleasure-misery* (horizontal) and *arousal-sleepiness* (vertical) and is displayed in Fig. 8.2.

A third dimension (for example, *dominance* or *interest*) is sometimes added to this model, but the usefulness of this dimension is less obvious than with the first two dimensions [327, 328].

Defining a set of mood categories or clusters is a way to reduce the complexity of the model significantly. Based on preceding reasearch on measuring emotional response to music, Schubert grouped mood labels into nine clusters as shown in Table 8.1 [329]. Deriving mood clusters by statistical analysis from large publicly available meta data collections has been done by Hu and Downie [330]. The resulting five clusters as shown in Table 8.2 are also used in the MIREX automatic mood classification task.

In research on music performance, there is strong evidence of a relation between moods and both the tempo and the loudness of the performance, as reported by Juslin [323], Kantor [331], Sloboda and Lehmann [332], Schubert [333], and Timmers et al. [334]. The mode (major vs. minor) has also been reported to have influence on the mood [335].

8.2.2.2 Recognition

The features and classification algorithms used for mood classification are very similar to the ones used in musical genre classification. Some mood-specific features which are not common in musical genre classification are articulation-based features estimating the smoothness of "note transitions" [336]. Mood classification systems also use tempo and rhythm-related features more frequently than systems for musical genre classification [337–339].

While the features for detecting the mood of music are relatively similar among researchers, the models and mood categories vary. Feng classifies music into the four classes *happiness*, *anger*, *sadness*, and *fear* [337], and Li uses the three dimensions *cheerful-depressing*, *relaxing-exciting* and *comforting-disturbing* [313]. Frequently used is Russell's two-dimensional arousal/valence-model as shown in Fig. 8.2 [326]; sometimes this model is simplified to define every quadrant as a category or cluster (*contentment*, *depression*,

Table 8.1 Mood Clusters as presented by Schubert

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*	*Cluster 6*	*Cluster 7*	*Cluster 8*	*Cluster 9*
Bright	Humorous	Calm	Dreamy	Dark	Heavy	Tragic	Agitated	Dramatic
Cheerful	Light	Delicate	Sentiment	Depressing	Majestic	Yearning	Angry	Exciting
Happy	Lyrical	Graceful		Gloomy	Sacred		Restless	Exhilarated
Joyous	Merry	Quiet		Melancholy	Serious		Tense	Passionate
	Playful	Relaxed		Mournful	Spiritual			Sensational
		Serene		Sad	Vigorous			Soaring
		Soothing		Solemn				Triumphant
		Tender						
		Tranquil						

Table 8.2 Mood clusters derived from meta data and used in MIREX

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
Rowdy	Amiable/Good Natured	Literate	Witty	Volatile
Rousing	Sweet	Wistful	Humorous	Fiery
Confident	Fun	Bittersweet	Whimsical	Visceral
Boisterous	Rollicking	Autumnal	Wry	Aggressive
Passionate	Cheerful	Brooding	Campy	Tense/Anxious
		Poignant	Quirky	Intense
			Silly	

exuberance, *anxious/frantic*) [338, 340]. By using continuous values for arousal and valence a categorical taxonomy can be avoided and each piece of music can be represented by a point in the two-dimensional space [341, 342]. Other options are to use a fuzzy classification that assigns probabilities to each class [340] and a multi-class or multi-label classification system which assigns a group of labels to a test sample [313, 338, 343].

Two additional aspects have also been the subject of research for mood classification; first, the possible mood change with the temporal evolution of music leading to the requirement of a non-stationary approach to mood classification [338, 344] and, second, the personalization of the recognition systems allowing to model the emotional concepts or responses of individuals or groups [339].

The evaluation of mood classification systems is — due to the fuzzy and subjective nature of mood — even more complicated than it was in the case of musical genre classification. An overview of different approaches to generating ground truth data is given by Kim et al. [345].

For the five mood clusters defined for the MIREX evaluation, the mood classification accuracy ranges between 40 and 60%. A recent study by Huq et al. presents evidence that with the currently used features and classification approaches further improvement of classification accuracy is questionable [342].

8.2.3 Instrument Recognition

An algorithm for *instrument recognition* attempts to identify the musical instruments which compose a sound or are present in a musical recording. In contrast to the classification into genres or moods it is straightforward to find ground truth data for training and evaluation even if the problem of defining instrument taxonomies cannot be considered to be ultimately solved either [346].

The two basic forms of instrument recognition can be distinguished by their type of input signal; some systems require a single note with no other pitches or instruments present; the signal has to be properly edited to eliminate preceding or succeeding notes or noises. Other systems work on a complex mixture of different instruments and estimate the (number and) type of the instruments present in the mixture.

Since the latter case is obviously algorithmically harder to handle, it comes as no surprise that most of the early publications on instrument recognition work on monophonic snippets of sound containing only one note. Herrera et al. give a good survey on the literature on monophonic instrument recognition [347].

The restriction to short monophonic input signal snippets allows the definition and usage of a new specialized feature set that extends the set of features introduced in Chap. 3. More specifically, it allows the system to use features that cannot be used in the context of polyphonic music until it will be possible to separate the sources into individual monophonic subsignals. The additional features can be structured in two categories:

- *Temporal envelope features* are features describing the temporal evaluation of the sound. Simple examples are the sound's duration, its temporal centroid, and the (logarithmic) attack time.
- *Pitch-based features* utilize the fundamental frequency of the sound. Examples include the energy ratio of even and odd harmonics, the inharmonicity of the harmonics, and the onset asynchrony between harmonics.

Kaminskyj et al. presented one of the first systems for automatic instrument recognition [348, 349]. It used a short time RMS and some harmonicity and spectral onset asynchrony features; the classification is done with either an ANN or a nearest-neighbor classifier. Similar to the systems for music similarity and musical genre classification, cepstral coefficients and MFCCs are frequently used for automatic instrument recognition, for instance, by the systems presented by Brown [350] and Marques and Moreno [351]. With time, the number of features and the diversity of instrument classes increased, while as classification approaches basically the same systems KNN, ANN, GMM, and SVM are used [352–355].

The next level in the history of instrument recognition was reached when the input signals did not have to be individual notes anymore; while the input still needed to be monophonic, it could now contain phrases and whole melodies [356–362].

There are only a few systems estimating the instrumentation of polyphonic signals. Eggink and Brown presented a system based purely on spectral features with a GMM-based classification that automatically masks out temporary "unreliable" features [363–365]. Eisenberg proposed a system for detecting instruments using a so-called *harmonic peak spectrum* by modeling instrument sounds with harmonic sinusoidal peaks [35]. It extracts the most salient component in the input signal so that it usually detects only the most prominent instrument; according to Eisenberg the system is able to detect accompanying instruments during pauses of the solo instrument. Heittola et al. attempt to decompose the signal into a sum of spectral bases and detect the individual sound sources [366]. The classification is done with GMMs on MFCCs.