10

Hit Song Science

François Pachet
Sony CSL

CONTENTS

Hit Song Science is an emerging field of investigation that aims at predicting the success of songs before they are released on the market. This chapter defines the context and goals of *Hit Song Science* (HSS) from the viewpoint of music information retrieval. In the first part, we stress the complexity of the mechanisms underlying individual and social music preference from an experimental psychology viewpoint. In the second part, we describe current attempts at modeling and predicting music hits in a feature oriented view of

popularity and, finally, draw conclusions on the current status of this emerging but fascinating field of research.

10.1 An Inextricable Maze?

Can someone predict whether your recently produced song will become a hit? Any pop song composer would probably laugh at this question and respond: How could someone predict the success of what took so much craft, pain, and immeasurable creativity to produce? I myself do not even have a clue!

This question raises a recurring fantasy in our culture: wouldn't it be thrilling to understand the "laws of attraction" that explain how this sort of preference system of music in human beings works, to the point of being able to predict the success of a song or any other cultural artifact before it is even released? This fantasy is elaborated in detail in Malcom Gladwell's story "The Formula" [14]. In this fiction, a—needless to say, fake—system is able to predict the success of movies by analyzing their script automatically. The system is even smart enough to propose modifications of the script to increase the success of the movie, with a quantified estimation of the impact in revenues. In the introduction, Gladwell begins by describing the reasons why we like a movie or not as resulting from a combination of small details. He writes:

> Each one of those ... narrative details has complicated emotional associations, and it is the subtle combination of all these associations that makes us laugh or choke up when we remember a certain movie... Of course, the optimal combination of all those elements is a mystery. [14]

This process is also true for music: what makes us like a song or not probably has to do with a complex combination of micro-emotions, themselves related to our personal history, to the specifics of the song and to many other elusive elements that escape our direct understanding. In spite of the many claims that writing hit songs is just a matter of technique (see, for example, Blume [5]), it is likely that, as the highly successful Hollywood screenwriter William Goldman said: "Nobody knows anything" [15].

Or is this the case? However daring, Hit Song Science attempts to challenge this assumption by precisely undertaking the task of making these kinds of predictions. Several companies now claim to be able to automatically analyze songs in order to predict their success (HSS, PlatiniumBlue) and to sell their results to record labels. Unfortunately, the exact mechanisms behind these predictions are not disclosed, and no reproducible data is provided to check the accuracy of these predictions. At the same time, the very existence of these services shows that hit prediction is taken seriously by the music industry.

Considering this hit song prediction fantasy from a scientific viewpoint raises issues in several disciplines, interrelated in complex ways that involve the following issues: (1) the psychology of music listening and the effects of repeated exposure, (2) the paradoxical nature of the Western media broadcasting system, radios in particular, and (3) the social influence human beings exert and receive from each other. Before describing the specific Music Information Retrieval (MIR) approach to Hit Song Science, each of these issues is first addressed.

10.1.1 Music Psychology and the Exposure Effect

Surprisingly, the question of "Why we like or not a particular song?" has received little attention from music psychology. Although music preference is recognized as a central aspect of modern identities, the field is "still in its infancy" [30]. The issue of *liking* per se is indeed difficult to study directly, and music psychologists have traditionally focused on less elusive, more directly measurable phenomena such as memorization, recognition or learning.

In our context, a central issue in trying to explain music hits is *exposure*, that is, the simple fact of listening to a musical piece. What is the effect of exposure on preference or liking? Studies on exposure show that there is indeed an impact of repeated exposure on liking, but also that this impact is far from simple. Parameters such as the context, type of music or listening conditions (focused or incidental), seem to influence the nature of this impact, and many contradictory results have been published.

The popular idea that repeated exposure tends to increase liking was put forward early [21] and was confirmed experimentally in a wide variety of contexts and musical genres [27]. The so-called *mere exposure* effect, akin to the *familiarity principle*, or *perceptual fluency*, is considered by many psychologists to be a robust principle, pervading many facets of music listening.

However, as noted by Schellenberg [29], this increase in liking may be restricted to musically impoverished or highly controlled stimuli. Indeed, other studies have shown a more subtle effect of repeated exposure. The study by Siu-Lan et al. [31] showed different effects of exposure on intact and patchwork compositions. An inverted U-curve phenomena was observed in particular by Szpunar et al. [33] and Schellenberg [30], itself explained in large part by the "two factor model" of Berlyne [3]. In this model, two forces compete to build up liking: (1) the *arousal* potential of the stimulus (the music), which decreases with repeated listening, thereby increasing liking (with the habituation to this arousal potential), and (2) *familiarity*, which tends to create boredom. These two forces combined produce typical inverted U-shapes that have been observed in many studies of preference. This model is itself related to the famous "Wundt curve" [36]. The Wundt curve describes the typical experience of arousal as being optimal when achieving a compromise between repetition/boredom and surprise (Figure 10.1). Interestingly, reaching such

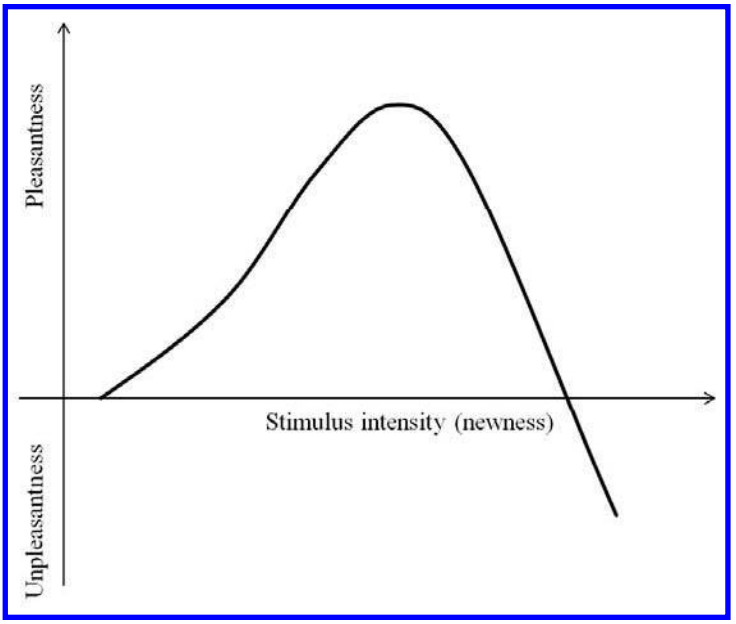

Figure 10.1
The Wundt curve describes the optimal "hedonic value" as the combination of two conflicting forces.

an optimal compromise in practice is at the root of the psychology of flow developed by Csíkszentmihályi [8].

Yet, other studies [35] show in contrast a *polarization* effect, whereby repeated exposure does not influence initial likings but makes them stronger, both positively or negatively. Finally, Loui et al. [19] studied exposure effects by considering exotic musical temperaments, to study the relation between learning and preference. They showed that passive exposure to melodies built in an entirely new musical system led to learning and generalization, as well as increased preference for repeated melodies. This work emphasizes the importance of learning in music preference.

These psychological experiments show that a relation between exposure and liking exists, but that this relation is complex and still not well understood, in particular for rich, emotionally meaningful pieces. It is therefore impossible to simply consider, from a psychological point of view, that repeated exposure necessarily increases liking: it all depends on a variety of factors.

10.1.2 The Broadcaster/Listener Entanglement

If the relation between exposition and preference is unclear in socially neutral contexts, the issue becomes even more confusing when considering the tangled interplay between the preference engine of individuals and the editorial choices of broadcasters, radio programmers, in particular.

Indeed, exposure is largely dependent upon the editorial strategies of programmers in the field of radio (in a broad definition of the term). Again, it is often said that it suffices to play a tune often enough to make it a hit, and that therefore hits are basically built by music marketing. However, the influence of radios on musical taste is paradoxical. On one hand, mass media (radio, television, etc.) want to broadcast songs that most people will like, in the hope of increasing their audience. Yet, what these media broadcasts actually influence, in turn, is the taste of audiences by means of repeated, forced exposition.

One process by which radios, for instance, maximize their audiences is so-called *radio testing*, which is performed regularly by various companies (e.g., musicresearch.com). Radio testing consists in playing songs to a selected panel that is representative of the radio audience, and then asking the listeners to rank songs. Eventually, only songs having received top ranks are kept and programmed. This radio testing phenomenon is more and more prevalent in Western society [17], but, strangely, has received so far little attention from researchers. Montgomery and Moe [22] exhibit a dynamic relationship between radio airplay and album sales, with vector autoregressive models, in an attempt to better inform marketing managers. Montgomery and Moe also stress the unpredictable evolution of this relationship as audiences may progressively evaluate these airplays critically by considering them as forms of advertisements.

This situation creates a paradox, also stressed by Montgomery and Moe [22]: "Not only is a radio station able to influence the public, but the public can also affect what is aired on the radio. Increased album sales may lead radio stations to play an album more." In turn, these songs are repeatedly exposed to a larger population with effects that are not completely clear, as seen above. As a result, if it is clear that radios do have an impact on musical taste, it is, again, difficult to assess exactly which one.

10.1.3 Social Influence

This is not the whole story. The situation is further complicated by the social influence that we all exert on one other. Knowing that a song is a hit, or at least preferred by others in our community, influences our liking. This phenomenon has been studied by Salganik et al. [28] in a remarkable experiment, which consisted in studying preferences in two groups of people: in the first group (*independent*), users had to rank individually, unknown songs. In the second group (*social influence*), users had the same task, with additional information

about what the other users of the group ranked. This information regarding the preferences of others had itself two strength levels.

The comparison between these two groups showed two interesting facts: (1) In the independent group, the distribution of preference was not uniform, showing that there are indeed songs that are statistically preferred to others, independently of social influence. This preference can only come from the songs themselves and can be considered as an indicator of their intrinsic quality. (2) The strength of the "social signal" increases the unpredictability of hits, that is, the more information about what others like, the less replicable are the experiments in terms of which songs become hits. This unpredictability, well-studied in network analysis, is referred to as the *cumulative advantage* effect. In the social influence group, hits are much more popular than in the independent group, but they are also different for each experiment, with the same initial conditions. One interesting argument put forward in this study is that the determination of which songs will become hits, in the social influence condition, eventually depend on "early arriving individuals" [34], in other words on initial conditions, which are themselves essentially random.

Under all of these conditions (unknown effects of repeated exposure, complex interplay between broadcasters and listeners, and the effects of social influence), is it reasonable even to attempt to program computers to predict hits in the first place?

10.1.4 Modeling the Life Span of Hits

Recognizing the importance of social pressure and the rich-get-richer effect, some works have attempted to predict hits using only social information, regardless of the intrinsic characteristics of songs.

For instance, Chon et al. [7] attempt to predict the popularity and life span of a jazz album given its entry position in the charts. This work used only charts information from Billboard, an American magazine maintaining music charts on a weekly basis. Analysis of the distribution of hits over time showed that the life span of a song tended to increase with its starting position in the charts. This result was interpreted as an encouragement to record labels to market albums before the sales, since the higher the starting position is, the longer it will stay in the charts. However, such a technique does not seem to be sufficient to yield more accurate predictions.

In the same vein, Bischoff et al. [4] attempted to identify critical early-stage effects of cumulative advantage. More precisely, this work posits that the success of a hit depends only on two factors: (1) its initial observed popularity after one week, as well as (2) contextual information such as the album, the general popularity of the artist, and the popularity of other tracks in the album. Similarly, this approach does not use any information concerning the actual content of the songs. Initial popularity and contextual information are converted into a 18 feature vector, and standard machine-learning techniques are then used to train and test a predictor (as described in detail in the next

section). Like in the previous work, ground-truth data is taken from Billboard. This experiment was conducted on a database of 210,350 tracks, performed by 37,585 unique artists. The results yield an improvement of 28% in AUC (area under ROC) compared to the work of Dhanaraj and Logan [9] described below.

These works show that there are patterns in the way social pressure generates hits. However, requiring initial popularity data, and being independent of both the characteristics of the songs and the listeners, they don't tell us much about why we like or not a given song. The following approaches take the opposite stance, trying explicitly to identify the features of songs that make them popular, regardless of social pressure effects.

10.2 In Search of the Features of Popularity

Several MIR researchers have recently attempted to consider hit prediction from a candid viewpoint. Like mathematicians trying to predict forecast or evolutions of financial markets, Hit Song Science has emerged as a field of predictive studies. Starting from the observation of the nonuniform distribution of popularity [12], the goal is to understand better the relation between intrinsic characteristics of songs (ignored in the preceding approaches) and their popularity, regardless of the complex and poorly understood mechanisms of human appreciation and social pressure at work.

In this context, popularity is considered as a *feature* of a song, and the problem, then, is to map this feature to other features that can be measured objectively. In other words, MIR sees Hit Song Science as yet another "feature problem," like genre or instrument classification.

It is important to stress the hypothesis at stake in this view, in light of the three difficulties described in the previous section. The attempt to directly model popularity with objective features ignores the difficulties that experimental psychology encounters in explaining the exposure effect. However, the yet unclear nature of human music habituation mechanisms does not imply that a predictor cannot be built. Of course, even if successful, such a predictor would probably not say much about the mysteries of the exposure effect.

The radio entanglement problem is related to the social influence issue: the final distributions of hits in a human community depend on random initial conditions which are not under control, from the choice of the members in the panel to the preferences of the "early arriving individuals." This intrinsic unpredictability in the hit distribution seems at first glance to threaten the whole Hit Song Science enterprise. An answer to this criticism consists in considering Hit Song Science as an idealistic attempt to determine the "objective causes" of individual music preference, independently of the effects of social influence.

Even if it is not realistic for individuals to listen and rate songs independently of each other, such an attempt is an important and sound approach for two reasons. First, if the works of Salganik et al. [28] aim at stressing the importance of social influence, they also show, in passing, that individual preferences do exist and are consistent and grounded. The difference between the nonuniform distribution of the independent group and a random one can precisely be seen as what *remains of individual preference*, once social influence is ignored. Because these remains of individuality are not random, it is worth trying to model them. Second, the search for the causes of our aesthetic experience, even partial ones, is a legitimate goal of cognitive science and should also be a goal of modern musicology. The remainder of this chapter focuses on this MIR view of hits, and more precisely on the following problem: under the absence of social pressure, which features of songs are able to explain their popularity?

10.2.1 Features: The Case of Birds

Before reviewing works that specifically address music features, we review here a fascinating and successful case of feature-based hit song prediction in a less complex area than human music: *bird songs*. Researchers in animal behavior have long been interested in the phenomenon of bird song production and its role in the mating process. In several bird species, male birds produce songs primarily to attract females. The issue of what makes a bird song more attractive than others has received particular attention in the recent years. This question echoes the Hit Song Science question (What are the features of popularity?), but in a simpler context, where social pressure is considered to be less significant.

Various results have indeed shown that specific features of songs can account for their popularity. For instance, great reed warbler females (*Acrocephalus arundinaceus*) were shown to prefer long over short songs in the wild [2].

More interestingly, the study by Draganoiu et al. [10] focused on the case of the domesticated canary (*Serinus canaria*). In this species, male bird songs have a specific phrase structure. Two features of these phrases were shown to significantly increase liking: (1) frequency bandwidth and (2) trill rate. However, it was also shown that these two features are somehow contradictory: a trade-off is observed in real phrases, due to the specific motor constraints of the bird vocal track.

The breakthrough experiment by Draganoiu et al. [10] consisted in synthesizing artificial phrases optimizing these two features in an unrealistic way, that is "beyond the limits of vocal production." The exposition of these artificial phrases to bird females showed unequivocally that females preferred these artificial phrases to the natural ones (see Figure 10.2). An interesting interpretation for this preference is that the production of "difficult" phrases

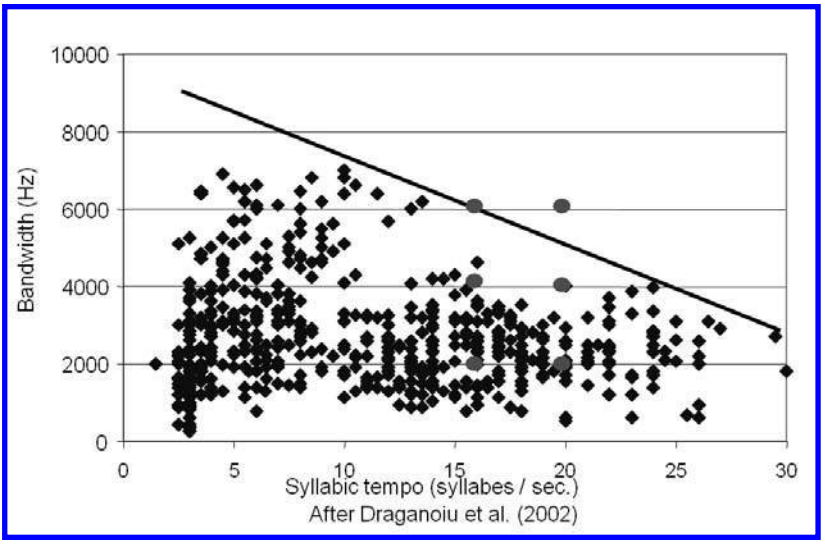

Figure 10.2
The distribution of canary phrases, in a bandwidth/tempo space, representing the natural trade-off between bandwidth and syllabic tempo. Circles represent the phrases used for the experiment. The artificial top right phrases optimizing the two features in unrealistic ways were the most successful [10].

maximizing both bandwidth and syllable rate may be a reliable indicator of male physical or behavioral qualities.

This evolutionary argument emphasizes the role of virtuosity in music appreciation. In popular music, virtuosity is explicitly present in specific genres (e.g., *shredding* in hard-rock, or melodic-harmonic virtuosity in bebop). However, it is probably a marginal ingredient of most popular styles (pop, rock), although virtuosity is still a largely understudied phenomenon. To which extent can these works be transposed to popular music?

10.2.2 The Ground-Truth Issue

In the MIR view of Hit Song Science, the nonuniform distribution of preferences is taken as *ground-truth* data. The problem is then to find a set of song features that can be mapped to song popularity. Once the mapping is discovered, the prediction process from a given, arbitrary new item (a song or a movie scenario) can be automated.

Considering the preceding arguments, a suitable database to conduct the experiment should ideally contain preference data which results from non-"socially contaminated" rankings. Such rankings can be obtained as in the experiment by Salganik et al. [28]. However, this process works only for a set of carefully chosen, unknown songs by unknown artists. In practice, there is

no database containing "normal songs" associated with such pure preference data. The experiments given in the following sections are based on databases of known music, with socially determined preference data, as described below. The impact of this approximation is not clear, and this approach *by default* could clearly be improved in future works.

10.2.3 Audio and Lyrics Features: The Initial Claim

The first attempt to model directly popularity in a feature-oriented view of music preference is probably the study by Dhanaraj and Logan [9]. This study consisted in applying the traditional machine-learning scheme, ubiquitous in MIR research, to the prediction of popularity. The features consisted both in traditional audio features Mel-frequency cepstral coefficients (MFCCs) extracted and aggregated in a traditional manner, as well as features extracted from the lyrics. The lyrics features were obtained by extracting an eight-dimensional vector representing the closeness of the lyrics to a set of eight semantic *clusters*, analyzed in a preliminary stage using a nonsupervised learning scheme.

The experiment was performed on a 1,700 song database, using Support Vector Machines (SVMs), and a boosting technique [13]. The conclusion of this study is that the resulting classifiers using audio or lyric information do perform better than random in a significant way, although the combination of audio and lyric features do not improve the accuracy of the prediction. However, a subsequent study described below showed contradictory results.

10.3 A Large-Scale Study

The studies by Pachet and Roy [23, 25] describe a larger-scale and more complete experiment designed initially to assess to which extent high-level music descriptors could be inferred automatically using audio features. A part of this study was devoted to the specific issue of popularity, seen as a particular high-level descriptor among many others. This experiment used a 32,000 song database of popular music titles, associated to fine-grained human metadata, in the spirit of the Pandora effort (http://www.pandora.com) as well as popularity data, obtained from published charts data like in the preceding approaches. To ensure that the experiment was not biased, three sets of different features were used: a generic acoustic set *à la* MPEG-7, a specific acoustic set using proprietary algorithms, and a set of high-level metadata produced by humans. These feature sets are described in the next sections.

10.3.1 Generic Audio Features

The first feature set was related to the so-called bag-of-frame (BOF) approach. The BOF approach owes its success to its simplicity and generality, as it can be, and has been, used for virtually all possible global descriptor problems. The BOF approach consists in modeling the audio signal as the statistical distribution of audio features computed on individual, short segments. Technically, the signal is segmented into successive, possibly overlapping frames, from which a feature vector is computed. The features are then aggregated together using various statistical methods, varying from computing the means/variance of the features across all frames to more complex modeling such as Gaussian Mixture Models (GMM). In a supervised classification context, these aggregated features are used to train a classifier. The BOF approach can be parameterized in many ways: frame length and overlap, choice of features and feature vector dimension, choice of statistical reduction methods (statistical moments or Gaussian Mixture Models), and choice of the classifier (decision trees, Support Vector Machines, GMM classifiers, etc.). Many articles in the Music Information Retrieval (MIR) literature report experiments with variations on BOF parameters on several audio classification problems [1, 11, 20, 26]. Although perfect results are rarely reported, these works demonstrate that the BOF approach is relevant for modeling a wide range of global music descriptors.

The generic feature set considered here consisted of 49 audio features taken from the MPEG-7 audio standard [18]. This set includes spectral characteristics (Spectral Centroid, Kurtosis and Skewness, High-Frequency Centroids, Mel-frequency cepstrum coefficients), temporal (Zero-Crossing Rate, Inter-Quartile Range), and harmonic (chroma). These features were intentionally chosen for their generality, that is they did not contain specific musical information nor used musically *ad hoc* algorithms. Various experiments (reported by Pachet and Roy [25]) were performed to yield the optimal BOF parameters for this feature set: localization and duration of the signal, statistical aggregation operators used to reduce dimensionality, frame size and overlap. The best trade-off between accuracy and computation time was achieved with the following parameters: 2,048 sample frames (at 44,100 Hz) with a 50% overlap computed on a two-minute signal extracted from the middle part of the title. The aggregated features were the two first statistical moments of this distribution (mean and variance) yielding eventually a feature vector of dimension 98 (49 means + 49 variances).

10.3.2 Specific Audio Features

The specific approach consisted in training the same (SVM) classifier with a set of "black-box" acoustic features developed especially for popular music analysis tasks by Sony Corporation [32]. These proprietary features have been used in commercial applications such as hard disk based Hi-Fi systems. Altogether, the specific feature set also yielded a feature vector of dimension

98, to guarantee a fair comparison with the generic feature set. As opposed to the generic set, the specific set did not use the BOF approach: each feature was computed on the whole signal, possibly integrating specific musical information. For instance, one feature described the proportion of perfect cadences (i.e., resolutions in the main tonality) in the whole title. Another one represented the proportion of percussive sounds to harmonic sounds.

10.3.3 Human Features

The last feature set considered was a set of human-generated features. We used the 632 Boolean labels provided by a manually annotated database (see the following section) to train the classifiers. This was not directly comparable to the 98 audio features as these labels were Boolean (and not floating point values). However, these features were good candidates for carrying high-level and precise musical information that are typically not well learned from features extracted from the acoustic signal.

10.3.4 The HiFind Database

10.3.4.1 A Controlled Categorization Process

Several databases of annotated music have been proposed in the MIR community, such as the RWC database [16], the various databases created for the MIREX tests [6]. However, none of them had the scale and number of labels needed to conduct this experiment. For this study the authors used a music and metadata database provided by the defunct HiFind Company. This database was a part of an effort to create and maintain a large repository of fine-grained musical metadata to be used in various music distribution systems, such as playlist generation, recommendation, or advanced music browsing. The HiFind labels were binary (0/1 valued) for each song. They were grouped in 16 categories, representing a specific dimension of music: Style, Genre, Musical setup, Main instruments, Variant, Dynamics, Tempo, Era/Epoch, Metric, Country, Situation, Mood, Character, Language, Rhythm, and Popularity. Labels described a large range of musical information: objective information such as the "presence of acoustic guitar," or the "tempo range" of the song, as well as more subjective characteristics such as "style," "character" or "mood" of the song. The Popularity category contained three (Boolean) labels: *low*, *medium*, and *high*, representing the popularity of the title, as observed from hit charts and records of music history. These three labels were mutually exclusive.

The HiFind categorization process was highly controlled. Each title was listened to entirely by one categorizer. Labels to be set to true were selected using an *ad hoc* categorization software. Label categories were considered in some specific order. Within a category, some rules could apply that prevented specific combinations of labels to be selected. The time taken, for a trained

categorizer, to categorize a single title was about six minutes. Categorized titles were then considered by a categorization supervisor, who checked consistency and coherence to ensure that the description ontologies were well understood and utilized consistently across the categorization team. Although errors and inconsistencies could be made during this process, the process nevertheless guaranteed a relative good "quality" and consistency of the metadata, as opposed for instance to collaborative tagging approaches with no supervision. As a consequence, the metadata produced was very precise (up to 948 labels per title), a precision difficult to achieve with collaborative tagging approaches.

The total number of titles considered in this study was 32,978, and the number of labels 632. Acoustic signals were given in the form of a wma file at 128 kbps. This database was used both for training and testing classifiers, as described in Section 10.3.5.3.

10.3.4.2 Assessing Classifiers

To avoid the problems inherent to the sole use of precision or recall, a traditional approach is to use F-measure to assess the performance of classifiers. For a given label, the recall R is the proportion of positive examples (i.e., the titles that are true for this label) that were correctly predicted. The precision P is the proportion of the predicted positive examples that were correct. When the proportion of positive examples is high compared to that of negative examples, the precision will usually be artificially very high and the recall very low, regardless of the actual quality of the classifier. The F-measure addresses this issue and is defined as:

$$F = 2 \times \frac{R \times P}{R + P}$$

However, in this specific case, the authors had to cope with a particularly unbalanced two class (True and False) database. Therefore, the mean value of the F-measure for each class (True and False) could be artificially good. To avoid this bias, the performances of classifiers were assessed with the more demanding *min F-measure*, defined as the minimum value of the F-measure for the positive and negative cases. A min-F-measure near 1 for a given label really means that the two classes (True and False) are well predicted.

10.3.5 Experiment

10.3.5.1 Design

The HiFind database was split in two "balanced" parts, Train and Test, so that Train contained approximately the same proportion of examples and counterexamples for each label as Test. This state was obtained by performing repeated random splits until a balanced partition was observed.

Three classifiers were then trained, one for each feature set (generic, specific, and human). These classifiers all used an SVM algorithm with a

radial-basis function (RBF) kernel. Each classifier, for a given label, was trained on a maximally "balanced" subset of Train, that is, the largest subset of Train with the same number of "True" and "False" titles for this label (popularity: Low, Medium, and High). In practice, the size of these individual train databases varied from 20 to 16,320. This train database size somehow represented the "grounding" of the corresponding label. The classifiers were then tested on the whole Test base. Note that the Test base was usually not balanced with regards to a particular label, which justified the use of the min-F-measure to assess the performance of each classifier.

10.3.5.2 Random Oracles

To assess the performance of classifiers, these were compared to that of random oracles defined as follows: given a label with p positive examples (and therefore $N - p$ negative ones, with N the size of the test set), this oracle returns true with a probability $\frac{p}{N}$. By definition, the min-F-measure of a random oracle only depends on the proportion of positive and negative examples in the test database.

For instance, for a label with balanced positive and negative examples, the random oracle defined as above has a min-F-measure of 50%. A label with 200 positive examples (and therefore around 16,000 negative examples) leads to a random oracle with a min-F-measure of 2.3%. So the performance of the random oracle was a good indicator of the size of the train set and could therefore be used for comparing classifiers as described below.

10.3.5.3 Evaluation of Acoustic Classifiers

The comparison of the performance of acoustic classifiers with random oracles showed that the classifiers did indeed learn something about many of the HiFind labels. More than 450, out of 632 labels, were better learned with the acoustic classifiers than with random oracle. Table 10.1 indicates, for each feature set, the distribution of the relative performance of acoustic classifiers with regards to random oracles.

Table 10.1 also shows that around 130 to 150 labels lead to low-performance classifiers, that is, acoustic classifiers that did not perform significantly better than a random oracle (the last row Table 10.1); approximately half of the labels led to classifiers that improve over the performance of a random classifier by less than 10; the rest (top rows) clearly outperformed a random oracle, that is, they were well modeled by acoustic classifiers.

It is interesting to see that the performance of these acoustic classifiers varied from 0% for both feature sets to 74% for the generic features and 76% for the specific ones. The statistical distribution of the performance was close to a power law distribution, as illustrated by the log-log graph of Figure 10.3.

These acoustic classifiers learned aspects of human musical categorization with a varying degree of success. The problem, as outlined below, is that popularity stands at the bottom line of this scale.

Improvement	Specific	Generic
50	8	0
40	12	15
30	43	20
20	111	79
10	330	360
0	128	158

Table 10.1
Number of Labels for Which an Acoustic Classifier Improves over a Random Classifier by a Certain Amount (Column "Improvement" reads as follows: there are 111 labels for which a specific acoustic classifier outperforms a random classifier by +20 [in min-F-measure].)

Not surprisingly, it could be observed that specific features performed always better than the generic ones (see Figure 10.4). Since the classifiers were both based on the same SVM/kernel, the difference in performance could only come from the actual features considered.

Last, the relationship between the performance and the size of the training set was studied. The trend lines in Figure 10.5 show that the performance of acoustic classifiers increase with the training data set size, regardless of the feature set. This was consistent with the acknowledged fact that machine-learning algorithms require large numbers of training samples, especially for high-dimensional feature sets.

These experiments showed that acoustic classifiers definitely learned musical information, with varying degrees of performance. It also showed that the subjective nature of the label did not seem to influence their capacity to be learned by audio features. For instance, the label "Mood nostalgic" was learned with 48% (specific features), and 43% (generic features), to be compared to the 6% of the random oracle. Similarly, label "Situation evening mood" was learned with 62% and 56% respectively, against 36% for random. Since *a priori* high-level features of songs could be learned with some success, why not popularity?

10.3.5.4 Inference from Human Data

This double feature experiment was complemented by another experiment with classifier trained using all the HiFind labels but the Popularity ones. Some pairs of HiFind labels were perfectly well correlated so this scheme worked obviously perfectly for those, but this result was not necessarily meaningful in general (e.g., to infer the country from the language). The same Train / Test procedure described above applied with the 629 nonpopularity labels as input yielded the following result (min-F-measure): 41% (Popularity-Low), 37% (Popularity-Medium), and 3% (Popularity-High).

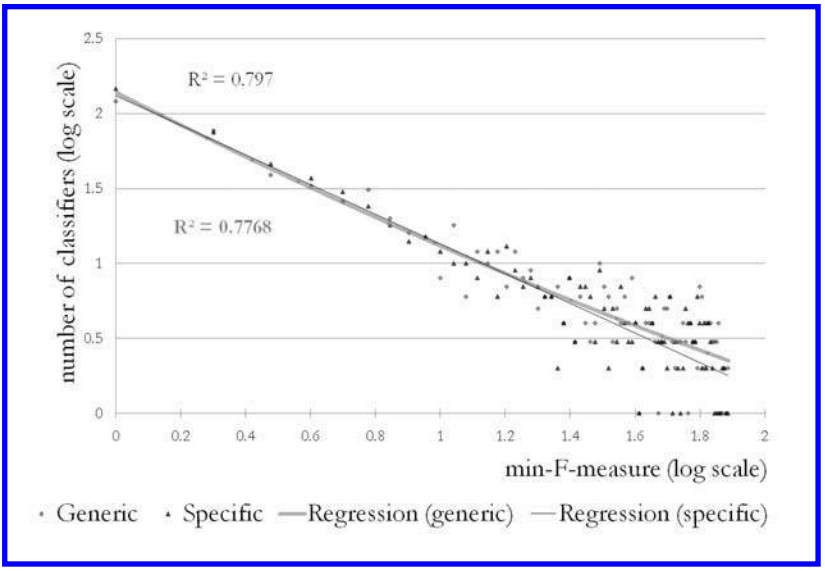

Figure 10.3
Log-log graph of the distribution of the performance of acoustic classifiers for both feature sets.

10.3.6 Summary

The results concerning the Popularity labels are summarized in Table 10.2. These results show clearly that the Popularity category was not well-modeled by acoustic classifiers: its mean performance was ranked fourth out of 16 categories considered, but with the second lowest maximum value among categories.

Although these results appear to be not so bad at least for the "Low" label, the comparison with the corresponding random classifiers shows that popularity is in fact not learned. Incidentally, the performance was not improved with the *correction scheme*, a method that exploits inter-relations between labels to correct the results [25]. Interestingly, human features (all HiFind labels) did not show either any significant improvement over random classifiers.

A last experiment was conducted with *a priori* irrelevant information: the letters of the song title, that is, a feature vector of size 26, containing the number of occurrences of each letter in the song title. The performances of the corresponding classifiers were respectively 32%, 28%, and 3%. (For the low-, medium-, and high-popularity labels, see Table 10.2.) This shows that even dumb classifiers can slightly improve the performance of random classifiers (by 5% in this case for the medium- and low-popularity labels). Obviously, this information does not teach us anything about the nature of hits and can be considered as some sort of noise.

These results suggest that there are no significant statistical patterns

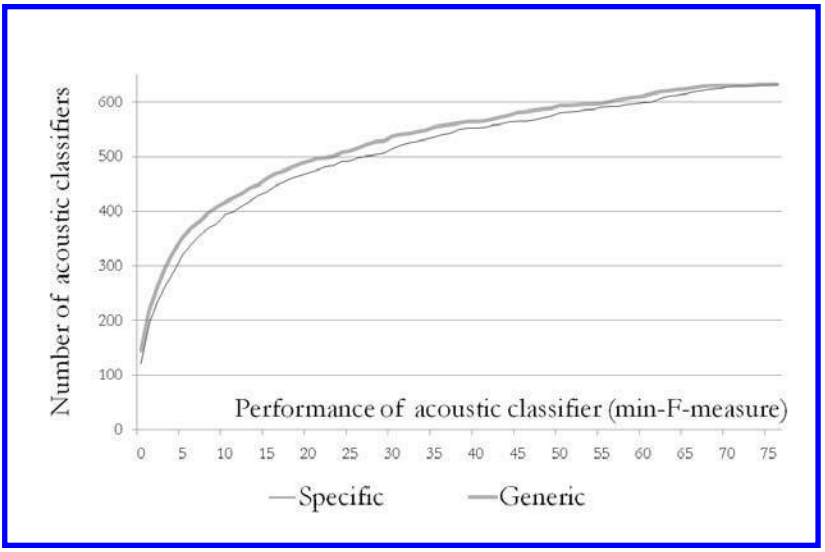

Figure 10.4
Cumulated distribution of the performance of acoustic classifiers for the generic and specific feature sets.

concerning popularity using any of the considered features sets (audio or humans). This large-scale evaluation, using the best machine-learning techniques available to date, contradicts the initial claims of Hit Song Science, that is that the popularity of a music title could be learned effectively from well-identified features of music titles. A possible explanation is that these early claims were likely based on spurious data or on biased experiments. This experiment was all the more convincing that other subjective labels could be learned reasonably well using the features sets described here (e.g., the "mood nostalgic" label).

The question remains: Do these experiments definitely dismiss the Hit Song Science project?

10.4 Discussion

The experiments described above show that current feature-oriented approaches to hit song prediction are essentially not working. This negative result does not mean, however, that popularity could not be learned from the analysis of a music signal or from other features. It rather suggests that the features used commonly for music analysis are not informative enough to grasp anything related to subjective aesthetic judgments.

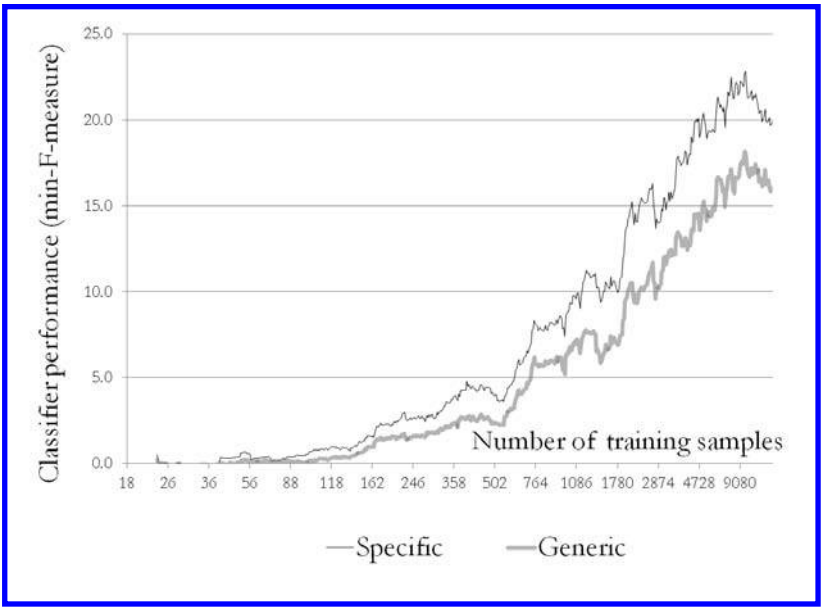

Figure 10.5
The relative performance of the 632 acoustic classifiers (i.e., the difference between the min-F-measures of the classifier and of the corresponding random oracle) for specific and generic features, as a function of the training database size. The performance of the acoustic classifiers increases with the size of the training database.

A natural way forward is to consider other feature sets. A promising approach is the use of *feature generation* techniques, which have been shown to outperform manually designed features for various audio classification tasks [24]. However, more work remains to be done to understand the features of subjectivity for even simpler musical objects such as sounds or monophonic melodies. Concerning the problem of social pressure, an interesting approach is to use music produced with exotic musical temperaments, an approach described by Loui et al. [19] to study the effects of exposure on musical learning and preference. This approach cannot be used on existing music, but has the great advantage of avoiding the biases of social pressure.

These negative results cast serious doubts on the predictive power of commercial Hit Song Science systems. Therefore, notwithstanding the limitations of current feature-based approaches, the arguments of social pressure effects are crippling: Hit Song Science cannot be considered, in its current state, as a reliable approach to the prediction of hits, because of the chaotic way individual preferences are mingled and propagated.

In spite of these negative results, we think that the main scientific interest of Hit Song Science, from a MIR viewpoint, lies precisely in the feature

Popularity	Generic Features	Specific Features	Corrected Specific	Human Features	Dummy Features	Random Oracle
Low	36	35	31	41	32	27
Medium	36	34	38	37	28	22
High	4	3	3	3	3	3

Table 10.2
The Performance (Min-F-Measures) of the Various Classifiers for the Three Popularity Labels (No significant improvement on the random oracle is observed.)

questions: Are there features of popularity, for an individual or for a community, and, if yes, what are they? From this perspective, Hit Song Science is a fascinating enterprise for understanding more what we like, and hence, what we are. The studies presented here have only started to scratch the surface of these questions: Hit Song Science is not yet a science but a wide open field.

Bibliography

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

[2] S. Bensch and D. Hasselquist. Evidence for active female choice in a polygynous warbler. *Animal Behavior*, 44:301–311, 1991.

[3] D.E. Berlyne. Novelty, complexity and hedonic value. *Perception and Psychophysics*, 8(5A):279–286, 1970.

[4] K. Bischoff, C. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Social knowledge-driven music hit prediction. In R. Huang, Q. Yang, J. Pei, J. Gama, X. Meng, and X. Li, editors, *Advanced Data Mining and Applications*, volume 5678 of *Lecture Notes in Computer Science*, pages 43–54. Springer, Berlin/Heidelberg, 2009.

[5] J. Blume. *6 Steps to Songwriting Success: The Comprehensive Guide to Writing and Marketing Hit Songs.* Billboard Books, New York, 2004.

[6] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack. ISMIR 2004 audio description contest. Technical Report MTG-TR-2006-02, University Pompeu Fabra, Italy, 2006.

[7] S.H. Chon, M. Slaney, and J. Berger. Predicting success from music sales data: A statistical and adaptive approach. In *AMCMM '06: Proceedings of the 1st ACM Workshop on Audio and Music Computing for Multimedia*, pages 83–88, ACM Press, New York, 2006.

[8] M. Csíkszentmihályi. *Beyond Boredom and Anxiety.* Jossey-Bass, San Francisco, 1975.

[9] R. Dhanaraj and B. Logan. Automatic prediction of hit songs. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 488–491, 2005.

[10] T.I. Draganoiu, L. Nagle, and M. Kreutzer. Directional female preference for an exaggerated male trait in canary (*Serinus canaria*) song. *Proceedings of the Royal Society of London B*, 269:2525–2531, 2002.

[11] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):68–80, 2006.

[12] R.H. Frank and P.J. Cook. *The Winner-Take-All Society.* Free Press, New York, 1995.

[13] Y. Freund and R. E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[14] M. Gladwell. The formula. *The New Yorker*, October 16, 2006.

[15] W. Goldman. *Adventures in the Screen Trade.* Warner Books, New York, 1983.

[16] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, Classical, and Jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, pages 287–288, 2002.

[17] M. Kelner. Heard the same song three times today? Blame the craze for testing tunes. *The Guardian*, May 19, 2008.

[18] H.G. Kim, N. Moreau, and T. Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval.* Wiley & Sons, New York, 2005.

[19] P. Loui, D.L. Wessel, and C.L. Hudson Kam. Humans rapidly learn grammatical structure in a new musical scale. *Music Perception*, 25(5):377–388, June 2010.

[20] D. Lu, L. Liu, and H. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–18, January 2006.

[21] M. Meyer. Experimental studies in the psychology of music. *American Journal of Psychology*, 14:456–478, 1903.

[22] A.L. Montgomery and W.W. Moe. Should record companies pay for radio airplay? Investigating the relationship between album sales and radio airplay. Technical Report, Marketing Dept., The Wharton School, University of Pennsylvania, June 2000.

[23] F. Pachet and P. Roy. Hit song science is not yet a science. In J. P. Bello, E. Chew, and D. Turnbull, editors, *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 355–360, Philadelphia, 2008.

[24] F. Pachet and P. Roy. Analytical features: A knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing*, 1:1–23, February 2009.

[25] F. Pachet and P. Roy. Improving multilabel analysis of music titles: A large-scale validation of the correction approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2):335–343, 2009.

[26] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In T. Crawford and M. Sandler, editors, *In Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 628–633, London, 2005.

[27] I. Peretz, D. Gaudreau, and Bonnel A.-M. Exposure effects on music preference and recognition. *Memory and Cognition*, 26(5):884–902, 1998.

[28] M.J. Salganik, P.S. Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

[29] E.G. Schellenberg. The role of exposure in emotional responses to music. *Behavioral and Brain Sciences*, 31:594–595, 2008.

[30] E.G. Schellenberg, I. Peretz, and S. Vieillard. Liking for happy- and sad-sounding music: Effects of exposure. *Cognition and Emotion*, 22(2):218–237, 2008.

[31] T. Siu-Lan, M.P. Spackman, and C.L. Peaslee. The effects of repeated exposure on liking and judgments of musical unity of intact and patchwork compositions. *Music Perception*, 23(5):407–421, 2006.

[32] Sony. 12-tone analysis technology, http://www.sony.net/SonyInfo/technology/technology/theme/12tonealalysis_01.html, 2010.

[33] K.K. Szpunar, E.G. Schellenberg, and P. Pliner. Liking and memory for musical stimuli as a function of exposure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):370–381, 2004.

[34] D.J. Watts. Is Justin Timberlake a product of cumulative advantage? *New York Times*, April 15, 2007.

[35] C.V.O. Witvliet and S.R. Vrana. Play it again Sam: Repeated exposure to emotionally evocative music polarizes liking and smiling responses, and influences other affective reports, facial EMG, and heart rate. *Cognition and Emotion*, 21(1):3–25, 2007.

[36] W. Wundt. *Outlines of Psychology.* Englemann, Leipzig, 1897.