Musikpsychologie Bd. 25, 41–54, © Hogrefe Verlag, Göttingen 2015

Is it the Song and Not the Singer? Hit Song Prediction Using Structural Features of Melodies

Klaus Frieler, Kelly Jakubowski & Daniel Müllensiefen

Zusammenfassung

Diese Untersuchung versucht, kommerziell erfolgreiche Popsongs von kommerziell weniger erfolgreichen Popsongs mit Hilfe struktureller Features ihrer Hauptgesangsmelodien zu unterscheiden. Zu diesem Zweck wurden 266 Popsongs anhand ihres Erfolgs in der britischen Hitparade durch k-means Clustering als „Hits“ oder „Nicht-Hits“ eingestuft. Darüber hinaus wurde mit Hilfe der Software MeloSpySuite ein umfassender Satz von 152 intrinsischen Features für die Hauptmelodien berechnet, die einen weiten Bereich struktureller Dimensionen (Tonhöhe, Intervalle, Rhythmus, Metrum etc.) abdecken. Diese Features wurden als unabhängige Variable für eine Random-Forest-Klassifikation benutzt; zudem wurde für jede Variable ein Wilcoxon-Test berechnet, um die Klassifikationsergebnisse weiter zu stützen und zu beleuchten. Der Klassifikationserfolg war mit 52,6 % relativ gering und lag nur knapp über der Ratewahrscheinlichkeit. Die Ergebnisse der Wilcoxon-Tests entsprachen im Wesentlichen den Resultaten der Random-Forest-Prozedur. Interessanterweise beziehen sich die Variablen mit der höchsten Diskriminanzleistung alle auf den Intervallgehalt der Melodien. Ein zusätzlicher Klassifikationsbaum mit den wichtigsten Variablen der Random-Forest-Prozedur erreichte eine Klassifikationsgenauigkeit von 61 % mit einer einzelnen Variablen, die die Gleichverteiltheit von aufeinanderfolgen Paaren von Intervallrichtungen misst, und die für Hits größer war als für Nicht-Hits. Wir diskutieren mögliche Interpretationen unserer Ergebnisse und schlagen sich anschließende Forschungsvorhaben vor.

Abstract

This study aims at the classification of highly commercially successful versus less commercially successful pop songs using structural features of the song melodies. To this end, a set of 266 pop songs were classified into hits and nonhits according to success in the UK charts using k-means clustering. Subsequently, a comprehensive set of 152 intrinsic summary features spanning a wide range of structural dimensions (pitch, interval, rhythm, metre, etc.) were extracted using the software MeloSpySuite and subjected to a random forest classification procedure. Additionally, a battery of Wilcoxon tests was executed to supplement the findings from the classification procedure. Classification success was rather low; at 52.6 % this success rate just slightly exceeded chance level. Furthermore, the results from the Wilcoxon tests were in line with the results from the random forest classification. Interestingly, the most important variables in both analysis procedures were all related to interval content. An additional classification tree algorithm fed with the most important variables from the random forest analysis reached a classification accuracy of 61 % with only one decision variable–Parson's Code bigram entropy. This variable measures the uniformity of interval direction pairs and was higher for hits than for non-hits. A range of possible interpretations for these results are discussed, and further lines of research are proposed.

1 Introduction

Is there a proven formula for creating a No. 1 hit song? This is a question that musicians and music producers have been pondering for many years. There are a wide array of how-to books written on this subject for aspiring songwriters, mainly based on anecdotal evidence from other musicians' successes and failures (e.g., Blume, 2004; Leikin, 2008; Oliver, 2013). Despite the large number of popular publications in this area, it seems at the present time that the ultimate hit song formula still remains to be discovered. Scientific research has only recently begun to investigate this question in a more systematic way. This research is sometimes referred to as "Hit Song Science".

One common approach researchers have applied in their endeavors to analyze the commercial success of pop songs is to examine the acoustic features of previous hits versus non-hits and attempt to derive underlying commonalities that set hit songs apart. Dhanaraj and Logan (2005) employed support vector machines and boosting classifiers in an attempt to classify hit songs versus non-hit songs based on both acoustic and lyrical features of the music. The best classification rate achieved based on acoustic features of the songs was .66; this rate was slightly better for just lyrical features (.68). However, no improvements in classification rate were seen when the acoustic and lyrical features were combined. Ni et al. (2011) conducted a similar study that attempted to distinguish songs that reached the top 5 chart positions from songs in chart positions 30 to 4. The highest prediction accuracy achieved by the classifiers they employed was around .57. The authors also compared the acoustic features of music across a period of 50 years and concluded that current commercially successful music is louder (i.e., dynamically more compressed), harmonically simpler, and faster than in the past. Similarly, Serrà et al. (2012) analyzed the evolution of hit songs over a period of 55 years and concluded that music is getting louder, more timbrally homogeneous, and more restricted in terms of pitch patterns. Finally, Nunes and Ordanini (2014) explored the influence of song instrumentation on commercial success. The researchers hand-coded the instrumentation data for a set of 2,399 pop songs. They were able to deduce a number of patterns from this data, including the findings that more popular songs regularly included backing vocals and that songs containing an atypically low or high number of instruments tended to become hits. This study did not include any measure of hit prediction accuracy.

Despite some interesting initial findings, the approach of predicting songs based solely on acoustic features has been unable to achieve high rates of correct hit predictions or classifications. As such, this approach has also received some serious criticism. For instance, Pachet and Roy (2008) conducted a large-scale analysis of a database of 32,000 songs and reported that classifiers built using state-of-the-art machine learning techniques were not able to significantly predict the commercial success of pop songs based on the songs' acoustic features. The researchers thus concluded that "Hit Song Science" is not yet a science, but that there is scope to uncover new features that more accurately relate to human aesthetic judgments, which may be more useful for uncovering commonalities in commercially successful songs.

A few alternative approaches to the acoustic feature prediction approach have been investigated. For instance, some researchers have explored social-knowledge driven hit song predictions based on data obtained from music social networks (Bischoff et al., 2009) and conversations on Twitter (Kim, Suh, & Lee, 2014). Salganik, Dodds, and Watts (2006) investigated early market responses to new songs by creating an artificial "music market" in which participants were able to download previously unknown songs. The researchers found strong influences of social factors, such that in a condition where participants were provided information about how many other participants had downloaded particular songs they observed a "cumulative advantage" whereby early success of a song led to a significant overall success in the long term. Salganik and Watts (2008) conducted an online "music market" study along the same lines in order to investigate whether perceived success of a song could become a "self-fulfilling prophecy". They inverted the true early popularity of songs that were previously unknown to participants, and found that for several songs this perceived (but false) initial popularity strongly influenced how popular they became over time. However, for originally top-rated songs, these songs did tend to regain their popularity over time. This finding suggests that musical features of the songs themselves were able to combat the false social cues provided at the start of the study.

Another alternative approach is to investigate features of the compositional structure of hit songs, such as the use of harmonic progressions (Kramarz, 2006) or the sequential structure of song sections (Riedemann, 2012). Kramarz summarizes these approaches in a recent overview (Kramarz, 2014) and suggests an array of preferential formulae especially for harmonic but also for melodic structures, as well as for the overall construction of hit songs. In addition, melodic features of hit songs have been the focus of other studies investigating whether certain compositional aspects of a melody itself contribute to its relative success in the charts. Kopiez and Müllensiefen (2011) conducted a first exploration into this area by attempting to predict the commercial success of cover versions of songs from the Beatles' album *Revolver*. They were able to accurately predict 100 % of cases using a logistic regression model with just two melodic features–pitch range and pitch entropy. However, one should note the limitations of this project, as the sample of songs used (14 songs all composed by the same band) is very specific and such a simple classifier would unlikely be able to cope with the wide diversity of styles and artists represented in all of the pop music charts.

Following the methods employed by Kopiez and Müllensiefen (2011), the present study aimed to examine a larger selection of pop songs in terms of a wider variety of melodic summary features and to predict song popularity (measured by highest chart position attained and number of weeks in the charts) using state-of-the-art classification techniques. Thus, the aim of this study is to establish the existence of any links between simple mathematical and statistical features of melodic structure and the commercial success of pop tunes as a real-world outcome. A link between melodic features and commercial success would suggest a human preference toward certain melodic structures that can be captured by numeric features. Any features revealed as predictive of commercial success would then be prime candidates for follow-up studies under controlled lab conditions to investigate the effects of these features on related aspects of cognition, e.g., the memorability or pleasantness of tunes.

2 Method

The dataset comprised 266 pop songs taken from a related project on features of earworm tunes (Williamson & Müllensiefen, 2012; regarding the broader perspective on earworm research see Hemming, 2009). Half of these 266 songs were reported by participants as songs that they tended to have stuck in their heads as earworms on the "Earwormery" database (https://earwormery.wordpress.com), hosted by the Music, Mind, and Brain group at Goldsmiths University of London. All songs from this database used in the present project were reported as frequent earworms by at least three separate participants. The other half of the songs were songs selected specifically to be from similar artists and UK chart positions but that were never named as earworms on the "Earwormery" database. For each song, the song title, artist, UK chart data (weeks in the charts and highest entry in the charts), and genre was recorded. This information was obtained from the UK chart database at polyhex.com and the Geerdes midimusic database (http://www.geerdes.com). MIDI files for each of the 266 songs were also obtained from the Geerdes midimusic database. The melody line from the section of the song reported as most catchy by the Earwormery participants was manually extracted from the full MIDI file. In the case of the matched songs or for songs from the Earwormery for which no particular section was reported as being the catchiest part, the chorus of the song was extracted. The editing of all MIDI excerpts was done manually and all excerpts were checked aurally for correctness and integrity.

3 Analysis

The 266 songs were separated into a binary classification of "hits" and "non-hits" using k-means clustering, with the highest chart position and the number of weeks a song remained in the charts as clustering variables (see Table 1, for some examples).

The next step was to attempt to classify songs as hits or non-hits based on their melodic features. The predictor variables used in this step were 152 melodic summary features calculated using the software MeloSpySuite (Frieler, Abeßer, Zaddach & Pfleiderer, 2013). Examples and descriptions of these features are provided in section 3.1.

Due to the large number of predictor variables, the statistical procedure chosen to model the present data was a random forest classification (Breiman, 2001). This method has several advantages over traditional regression models: (1) it handles both categorical and numerical data, (2) it can cope with non-linear relationships, and (3) it can be used with a large number of predictor variables. Moreover, in a recent comparative study by Fernández-Delgado et al. (2014), random forests proved to be the most successful algorithm among an exhaustive set of classification methods when applied to a large and diverse collection of real-world datasets. For a brief summary of the random forest technique and tree-based classification and regression methods in general and their applications in music and psychology research see Pawley and Müllensiefen (2012) and Strobl, Mally, and Tutz (2009).

Tab. 1:
Examples of hits and non-hits from k-means clustering classification

Artist	Title	Genre	Highest Chart Entry	Weeks in Charts	Classification
Queen	I'm Going Slightly Mad	Rock	22	5	Non-hit
Queen	Bohemian Rhapsody	Rock	1	17	Hit
Britney Spears	Circus	Pop	13	18	Non-hit
Britney Spears	Toxic	Pop	1	14	Hit
Lionel Richie	Running with the Night	Pop	9	12	Non-hit
Lionel Richie	Hello	Pop Rock	1	15	Hit

3.1 Features

The features used in this study are calculated using simplified and abstract representations of melodies, which include onsets, durations, pitches (using MIDI numbers), as well as metrical annotations. All features used in this study can be classified as intrinsic or summary features that do not take any external information into account but are purely derived from the onset, duration, pitch and metrical information contained within the melodies. All intrinsic features can be calculated in a unique way directly from the abstract representations, e.g., as statistical descriptors of value distributions (see Müllensiefen & Halpern, 2014, for a general description of intrinsic features for melodic analysis and how they differ from extrinsic or corpus-based features that take cultural context into account). Most of the intrinsic features are not directly or only weakly based on genuine music-specific models or theories. Notable exceptions are metrical information, which is a specific musical property, and tonal pitch classes, which depend on the notion of tonality and the concept of octave equivalence.

3.2 Classification of Features

The 152 intrinsic features used in this study can be classified into 7 main groups, which will be explained briefly in the following section. All features used here are summary features in the sense that they describe a melody using a single number, e.g., the frequency of a certain pitch class or the entropy of its interval bigram distribution. Particularly of interest to the present study are entropy values and Zipf coefficients, as these features have been previously associated with melodic complexity (e.g., Eerola, Himberg, Toiviainen & Louhivuori, 2006). Information entropy (Shannon, 1948) is a well-known measure of information content in a probability distribution which estimates the amount of uniformity; the more uniform a distribution, the harder it will be to predict the next element in a sequence. A greater degree of uniformity in turn can be interpreted as being more complex to process but also as more "entertaining", since the melody is less predictable to the listener (Huron, 2006). The Zipf coefficient (Zanette, 2006; Zipf, 1949) is a rather similar measure that is often correlated with information entropy. It measures the dominance of certain few elements in a probability distribution. For example, a pitch sequence has a high Zipf coefficient if a subset of certain pitches is used much more often than the rest of the pitches. Besides these and other properties of distributions of single elements, short subsequences (N-Grams) are of special interest, since melodies are essentially successions of elements unfolding in time. For example, bigrams are subsequences of two elements, which capture more detailed sequential aspects of a melody ("which follows what"). The bigrams of a sequence form a distribution of nominal values, which enable the calculation of entropy and Zipf coefficients, but not metrical statistical descriptors. A sequence with a uniform distribution of single elements ("unigrams") does not necessarily also have a uniform bigram distribution and vice versa, e.g., consider the sequence "abcabcabc", where the bigrams “ac”, “ba”, and “cb” are not found, but each single element occurs identically often.

Contour features. These features aim to capture the up and down motion of a melody in pitch space. There are many different ways to do so, e.g., by counting the pitch maxima and minima, or using a compact code such as the Huron Contour (Huron, 1996; for an overview of contourization options see Müllensiefen & Frieler, 2004). In the present study, only the ratio of the number of pitch turning points to the total number of notes was used as a feature.

Interval features. Semitone intervals are a common and important aspect of pitch movement. In this study, distributions of semitone intervals and certain reductions of them were used. The two reductions employed are Parson’s code (aka contour) and fuzzy interval (aka refined contour). In the latter, each interval is classified into one of five distinct classes (unisons, steps, leaps, jumps, and big jumps) along with its direction (up or down), giving eleven fuzzy interval classes. Parson’s code classifies intervals with respect to their basic direction: up, repetition or down. From the interval distributions, statistical descriptors such as the range, mean, median, standard deviation, entropy, Zipf coefficient, etc., as well as single densities, were calculated and used as features.

Pitch features. For raw pitch distributions, a similar set of statistical descriptors was used (pitch range, mean, median, standard deviations, entropy, Zipf coefficients, etc.). Additionally, absolute pitch classes were calculated by reducing a pitch to one of the twelve (enharmonic) pitch classes C, C#, … , B, while disregarding octave position. Since this calculation yields circular distributions, several circular statistical descriptors (Mardia & Jupp, 2000) were computed along with relative frequencies of single pitch classes.

Rhythm features. Rhythm was operationalized using note durations and inter-onset intervals (IOI) that were both subjected to the same classification process. This process maps durations/IOIs onto one of five classes (“very short”, “short”, “medium”, “long”, “very long”) using graded intervals of duration with respect either to an absolute reference time of .5 s or the beat duration of the melody. This results in four different classifications in total, for which statistical descriptors and class densities of the according distributions were used as features. Finally, the pairwise variability index (Patel & Daniele, 2003) and the coefficient of variation (the ratio of the standard deviation of tone duration to the mean tone duration) were added to the set of rhythm features.

Metrical features. Meter is a very specific and important dimension of music. The metrical features used here are circular statistics and single relative frequencies of 48 classes derived from a Metrical Circle Map (Frieler, 2007) applied to the melodies. The Metrical Circle Map divides each measure into N (here: N=48, abbreviated MCM48) identical segments, and each melodic event is mapped to its corresponding segment.

Sequence features. Since melodies are time series, it is straightforward and meaningful to design features that capture aspects of their sequential nature. We used two different types: Mean run-lengths (for five IOI and three Parson’s Code classes) and entropies of bigram distributions. The latter were calculated based on several underlying basic representations: semitone intervals, fuzzy interval, Parson's code, IOI classes, raw pitch and pitch class. Run-lengths are the lengths of sub-sequences with identical items, i.e., "runs".

Finally, basic global features such as the number of notes were also included in the feature set.

4 Results

The feature values for all melodies were subjected to a random forest procedure (from the "party" package in R; Hothorn, Hornik, & Zeileis, 2006), using an unbiased version to avoid over-fitting, which could easily occur with data where the number of cases is similar to the number of predictor variables. We conducted an analysis using the binary hit index ("hit" or "non-hit", based on the k-means clustering solution). From the random forest procedure we calculated confusion matrices (across five different runs due to the random nature of the algorithm) and used the normalized sum of diagonals as a measure of mean classification accuracy (cf. Table 2). Furthermore, we computed mean variable importances across five random forest calculations of the classification task, and used the four most important variables (cf. Table 3) as input for a binary classification tree (cf. Figure 1; see Strobl et al., 2009).

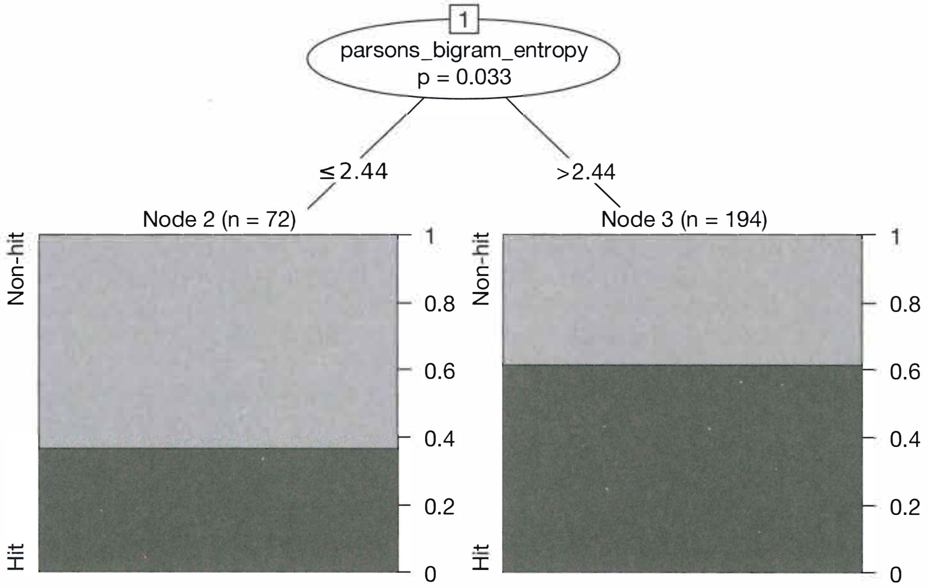

Fig. 1:

Classification tree for hit songs using the 3 most important variable.

Comments: Only one variable survived (*parsons_bigram_entropy*, which measures the uniformity of interval direction pairs). Classification accuracy is 61.7 %.

Tab. 2:
Confusion matrix of predicted versus real hit songs over five Random Forest runs

Confusion matrix	Non-hit	Hit
Non-hit (predicted)	513 (38.6%)	424 (31.9%)
Hit (predicted)	207 (15.6%)	186 (14.0%)

Comment: Mean classification accuracy is 52.6 %.

Tab. 3:
Mean variable importance and standard error for hit song classification averaged over five Random Forest runs with 1 000 trees each

Variable	Mean Importance (x1 000)	Standard Error (x1 000)
parsons_bigram_entropy	1.87	.017
parsons_bigram_entropy_norm	1.71	.027
int_zipf	1.40	.019
parsons_entropy	1.21	.015

Comment: Only importances with absolute values greater than .001 are shown. For an explanation of the variables please refer to the main text.

Tab. 4:
Significant differences between hits and non–hits from Wilcoxon rank tests.

Variable	p	Cohen’s d
number_notes	.008	–.22
total_duration	.017	–.21
parsons_bigram_entropy	.022	–.31
parsons_bigram_entropy_norm	.025	–.31
fuzzyint_hist_step_down	.032	–.26
mcm_zipf	.033	–.26
abs_int_zipf	.037	–.26
int_zipf	.049	–.22

Comment: Only variables with p < .05 are shown. For an explanation of the variables please refer to the main text.

4.1 Predicting Hit songs

As can be seen in Table 2, the mean classification accuracy of 52.6 % is only slightly above chance level, i.e., intrinsic melodic features do not seem well suited to predicting hit songs. Using the four most important variables as input to a classification tree, however, gives a slightly better classification accuracy of 61 %, but it is not expected that this result will generalize well to other song samples. The single surviving variable for the hit song classification tree is the entropy of the Parson's Code bigram distribution *(parsons_bigram_entropy)*. This can be interpreted to mean that hit songs tend to use all nine possible combinations of the three contour values (down, repetition, up) of two subsequent pitch intervals in equal proportion. The only other variables that attain a mean importance value of greater than .001 are *int_zipf, parsons_entropy* and *parsons_bigram_entropy_norm* (a normalized version of *parsons_bigram_entropy*, providing no additional information). The variable *int_zipf* measures the slope of the log-log distribution of rank-ordered semitone intervals and is an indicator as to whether a distribution is dominated by a few values. A higher Zipf coefficient indicates greater predominance of a smaller number of intervals, as is the case for the hit songs in our sample. Parson's Code entropy *(parsons_entropy)* is a measure of the uniformity of the interval directions. Though the overall classification accuracy is very low, it is still noteworthy that all variables with high importance values measure the complexity of interval content.

Additionally, we conducted a set of Wilcoxon rank tests using each feature as a dependent variable and the hit vs. non-hit classification as a binary predictor (cf. Table 4). Expectedly, the largest differences were found for those features that came out as most important in the random forest model. Additionally, *number_notes* (note count) and *total_duration_bar* (length of melody measured in bars) showed large differences with considerable effect sizes, despite the fact that these features did not play an important role in the random forest classification model. Interestingly, the relevant parts of hit songs (mainly the chorus, as implemented in this study) are significantly shorter than the relevant parts of non-hit songs. Three more variables are present on the list of important variables as indicated by the results of the Wilcoxon tests: *fuzzyint_hist_step_down, mcm_zipf*, and *abs_int_zipf*. The latter is closely related to *int_zipf* (Spearman's rho = .76, $p < .000$). The second *(mcm_zipf)* is the Zipf coefficient of the MCM48 distribution, which is lower for hit songs, indicating higher metrical variability. The first *(fuzzyint_hist_step_down)* is the density of downward steps (semi- and whole tones), which occur less often in hits. This finding is interesting since generally downward steps are the most frequent interval in many relevant corpora (pop songs, e.g., Müllensiefen, Wiggins, & Lewis, 2008; European folk songs, e.g., Huron, 2006; and jazz solos, Frieler et al., 2013).

The *p*-values in the Wilcoxon rank tests were not corrected for multiple testing. At a significance level of .05 and 152 variables, one would expect to see 7.6 tests coming up significant by chance alone, which is closely fulfilled here. Hence, no general conclusion should be drawn from the results of the Wilcoxon tests alone, but it is reassuring that these results align at least partially with the variable importance scores from the random model, thus strengthening those results.

5 Discussion and Outlook

The classification accuracy of the random forest using all 152 melodic features is surprisingly low and only marginally exceeds the chance level of 50 %. Random forests are generally considered to be a very powerful statistical classification and prediction method (cf. Fernández-Delgado et al., 2014), thus the weak results obtained here are surprising. Additionally, the battery of Wilcoxon tests revealed only a rather small set of significant differences, which could have occurred by chance alone and supports the weak classification results from the random forest model.

For hit song classification, one could argue that there are very many factors involved in producing a hit, most of them probably extra-musical. If we trust the statistical and classification methods, the most obvious explanation would be that the types of intrinsic features we employed are not well suited to capturing psychologically important characteristics of melodies (e.g., memorizability, simplicity/complexity, etc.), at least not for those songs involved in this study. In line with Pachet and Roy's (2008) findings, we would not expect that an even larger set of features, constructed in similar ways as done here, would raise the classification success rate significantly. Hence, two possible explanations come to mind. First, the basic and very abstract representation of melodies as sequences of note onsets, durations and pitches is too sparse, since all moments of expressivity and timbre are excluded. Maybe, it is really the singer and the performance of a song that drives its popularity and commercial success, rather than the structural features of the song's melody. This would be in line with the findings by Pawley and Müllensiefen (2012), who reported that the features of the musical performance (and not the features of musical structure) proved to be the most important ones for inciting people to sing along to pop music in leisure contexts. A counter-argument to this explanation would be that very successful songs normally occur in a large variety of expressive renditions (cover songs, ordinary people singing the song on the street, different instrumentations), making them partially independent of their concrete realization, which is corroborated by the fact that they are easily recognizable even from dead-pan MIDI versions. Thus following this line of argument, expressivity and timbre might not be entirely crucial. The second potential explanation concerns the fact that all features used here are intrinsic features. Hence, they do not make use of information of the cultural context in which the melodies occur, which is assumed to form the basis for the processing of melodic information in listeners. We conjecture, thus, that extrinsic corpus-based features, such as the ones described in Müllensiefen & Halpern (2014), might significantly boost classification performance. This is a readily testable hypothesis that we plan to address in the near future.

References

Bischoff, K., Firan, C. S., Georgescu, M., Nejdl, W. & Paiu, R. (2009). Social knowledge-driven music hit prediction. In R. Huang, Q. Yang, J. Pei, J. Gama, X. Meng, & X. Li (Eds.), *Advanced data mining and applications* (pp. 43–54). Berlin: Springer.

Blume, J. (2004). *Six steps to songwriting success: The comprehensive guide to writing and marketing hit songs*. New York: Billboard Books.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32. http://doi.org/1.1023/A:1017934522171

Dhanaraj, R. & Logan, B. (2005, August). Automatic prediction of hit songs. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR), Barcelona 2004* (pp. 488–491).

Eerola, T., Himberg, T., Toiviainen, P. & Louhivuori, J. (2006). Perceived complexity of western and African folk melodies by western and African listeners. *Psychology of Music, 34*(3), 337–371. http://doi.org/1.1177/0305735606064842

Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181.

Frieler, K. (2007). Visualizing music on the metrical circle. In S. Dixon, D. Bainbridge & R. Typke (Eds.) *Proceedings of the 8th International Symposium on Music Information Retrieval, ISMIR2007* (pp. 291–292). Wien: OCG.

Frieler, K., Abeßer, J., Zaddach, W.-G. & Pfleiderer, M. (2013). Introducing the Jazzomat Project and the Melo(S)py Library. In P. van Kranenburg, C. Anagnostopoulou & A. Volk (Eds.), *Proceedings of the Third International Workshop on Folk Music Analysis* (pp. 76–78). Utrecht, NL: Meertens Institute and Utrecht University Department of Information and Computing Sciences.

Hemming, J. (2009). Zur Phänomenologie des „Ohrwurms". In W. Auhagen, C. Bullerjahn & H. Höge (Hrsg.), *Musikpsychologie – Musikalisches Gedächtnis und musikalisches Lernen* (Jahrbuch der Deutschen Gesellschaft für Musikpsychologie, Bd. 20, S. 184–207). Göttingen: Hogrefe.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674. http://doi.org/1.1198/106186006X133933

Huron, D. (1996). The melodic arch in Western folksongs. *Computing in Musicology, 10*, 3–23.

Huron, D. (2006). *Sweet anticipation*. Cambridge, MA: MIT Press.

Kim, Y., Suh, B. & Lee, K. (2014). #nowplaying the Future Billboard: Mining Music Listening Behaviors of Twitter Users for Hit Song Prediction. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis* (pp. 51–56). New York: ACM.

Kopiez, R. & Müllensiefen, D. (2011). Auf der Suche nach den „Popularitätsfaktoren" in den Song-Melodien des Beatles-Albums Revolver: eine computergestützte Feature-Analyse. In S. Meine & N. Noeske (Hrsg.), *Musik und Popularität. Beiträge zu einer Kulturgeschichte zwischen 1500 und heute* (S. 207–225). Münster: Waxmann Verlag.

Kramarz, V. (2006). *Die PopFormeln: [die Harmoniemodelle der Hitproduzenten]*. Bonn: Voggenreiter.

Kramarz, V. (2014). *Warum Hits Hits werden. Erfolgsfaktoren in der Popmusik*. Bielefeld: transcript. http://doi.org/1.14361/transcript.9783839427231

Leikin, M. A. (2008). *How to write a hit song* (5th ed.). Milwaukee, WI: Hal Leonard Books.

Mardia, K. V. & Jupp, P. (2000). *Directional Statistics* (2nd ed.). Hoboken, NJ: John Wiley and Sons Ltd.

Müllensiefen, D. & Frieler, K. (2004). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology, 13*, 147–176.

Müllensiefen, D. & Halpern, A. (2014). The role of features and context in recognition of novel melodies. *Music Perception, 31*(5), 418–435. http://doi.org/1.1525/mp.2014.31.5.418

Müllensiefen, D., Wiggins, G. & Lewis, D. (2008). High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. In A. Schneider (Hrsg.), *Hamburger Jahrbuch für Musikwissenschaft* (Bd. 24, S. 133–155). Frankfurt/Main: Peter Lang.

Ni, Y., Santos-Rodríguez, R., Mcvicar, M. & De Bie, T. (2011). Hit song science once again a science? In *4th International Workshop on Machine Learning and Music*.

Nunes, J. C. & Ordanini, A. (2014). I like the way it sounds: The influence of instrumentation on a pop song's place in the charts. *Musicae Scientiae, 18*(4), 392–409. http://doi.org/1.1177/1029864914548528

Oliver, B. (2013). *How [not] to write a hit song! 101 common mistakes to avoid if you want songwriting success.* Montreal, CA: Rapido Books.

Pachet, F. & Roy, P. (2008). Hit song science is not yet a science. In J. P. Bello, E. Chew & D. Turnbull (Eds.), *ISMIR 2008. Proceedings of the 9th International Conference on Music Information Retrieval* (pp. 355–360). Raleigh, NC: Lulu.com.

Patel, A. D. & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*, B35–B45. http://doi.org/1.1016/S0010-0277(02)00187-7

Pawley, A. & Müllensiefen, D. (2012). The science of singing along: A quantitative field study on sing-along behavior in the North of England. *Music Perception, 30*(2), 129–146. http://doi.org/1.1525/mp.2012.3.2.129

Riedemann, F. (2012). Computergestützte Analyse und Hit-Songwriting. In D. Helms & T. Phleps (Hrsg.), *Black Box Pop* (S. 43–56). Bielefeld: transcript.

Salganik, M. J., Dodds, P. S. & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*(5762), 854–856. http://doi.org/1.1126/science.1121066

Salganik, M. J. & Watts, D. J. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly, 71*(4), 338–355. http://doi.org/1.1177/019027250807100404

Serrà, J., Corral, A., Boguná, M., Haro, M. & Arcos, J.L. (2012). Measuring the evolution of contemporary Western popular music. *Scientific Reports, 2*, 521. http://doi.org/1.1038/srep00521

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656. http://doi.org/1.1002/j.1538-7305.1948.tb00917.x

Strobl, C., Malley, J. & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348. http://doi.org/1.1037/a0016973

Williamson, V. J. & Müllensiefen, D. (2012, July). *Earworms from three angles: Situational antecedents, personality predisposition and the quest for a musical formula.* Paper presented at the 12th International Conference on Music Perception and Cognition, Thessaloniki, Greece.

Zanette, D. H. (2006). Zipf's law and the creation of musical context. *Musicae Scientiae, 10*, 3–18. http://doi.org/1.1177/102986490601000101

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, MA: Addison-Wesley.