8

HOOK ANALYSIS

In this chapter, we build on insights on corpus analysis from chapters 3 and 4, and on the audio bigram features proposed in section 5, to present a corpus analysis of the song recognition data described in chapter 7.

As part of the analysis, we propose a new set of *second-order* features (section 8.1. The notion of second-order music descriptors is inspired by latent semantic analysis methods from text retrieval. They encode typicality and distinctiveness of feature values. By adapting the concept to audio features, we are able to present a cognitively adequate analysis of the music and participant data of the *Hooked!* game that allows for findings to be interpreted in terms of listening expectations of the participants.

Section 8.2 presents the analysis itself, and compares the new features to a set of symbolic features. We find that our corpus-based audio features are able to explain a comparable amount of variance to symbolic features. When the two types of features are used together, they supplement each other profitably. Along the way, we discuss the newly gained insights into what makes music recognizable, as revealed by the *Hooked!* data.

8.1 SECOND-ORDER AUDIO FEATURES

We begin by introducing the notion of second-order features, and proposing a fully specified adaptation of this idea for audio descriptors.

8.1.1 *Second-Order Features*

Second-order features are derivative descriptors that reflect, for a particular feature, how an observed feature value relates to a reference corpus [132]. There are several motivations for the use of second-order features.

First, they help in quantifying similarity and relevance of documents in the context of information retrieval. In information retrieval from text data, a common document description paradigm based on word counts involves *weights* that depend on the frequency of each word in a large corpus. Uncommon words are typically given more weight, as for example in 'TF×IDF' weighting, where TF (for term frequency) is the frequency of each term in the document, and IDF (for inverse document frequency) relates to the number of documents in the corpus that contain the term. Feature weighting helps estimating the relevance of a document in a retrieval context, and has been used as such for a very long time. The text analysis field of latent semantic analysis (LSA) is concerned with this kinds of text description [162].

Another motivation for the use of corpus-relative features could be to make the resulting feature more interpretable. They help contextualize the values a feature can take. Is 18.25 a high number? Is it a common result? Or if the feature is multivariate: is this combination of values typical or atypical, or perhaps representative of a particular style?

Finally, we shall show in section 8.1.4 that second-order features can be more cognitively plausible descriptions than the features we have been using so far. By giving us a means to approximately quantify expectations, distinctiveness and recurrence (the importance of which has been discussed in chapter 7), second-order features can be particularly useful in the analysis of recognizability and hooks.

8.1.2 *Second-Order Symbolic Features*

Second-order features have been used in symbolic music analysis, both of the retrieval and the corpus analysis kind. Like in text mining,

many features use the notion of document frequency, e.g., the number of songs in a large corpus that contain a given pitch interval.

The FANTASTIC toolbox by Müllensiefen implements many of these features. For example, the mtcf.mean.log.DF feature represents the average document frequency of all melodic motives or 'm-types' in a given melody, given a melody corpus.

M-types, inspired by the concept of *types* (entries in a dictionary) used in computational linguistics, are short sequences of symbols encoding pitch intervals and duration ratios of neighboring notes. M-types relate closely to the musical concept of melodic motives.

There is a strong parallel between the M-type counts used in the FANTASTIC toolbox, and the audio bigram features proposed in chapter 5 of this thesis. Both encode the relative occurrence of pitches in a specific order. We now discuss how the notion of document frequencies, and second-order features in general, can be adapted and used with audio bigrams and other audio descriptors.

8.1.3 *Second-Order Audio Features*

A fundamental difference between symbolic and audio representations of music, is that symbolic representations represent music as a streams of discrete event (e.g., notes, chords), while digital audio represents continuous, uninterrupted signals. This also applies to features: symbolic features operate on countable collections of events. Audio representations, even if they are discrete time series, based on frequency-domain computations or otherwise measured over short windows, represent continuous, uncountable quantities. This makes it impossible to apply the same operations directly to both, and alternatives must be found for the audio domain.

We define three types of second-order features. All represent how typical an observation is in a certain reference corpus. In statistical terms, the typicality of an observation in some feature space—how often does this value occur?—corresponds to the frequency density of this feature at the location of the observation. We distinguish between

one-dimensional descriptors such as loudness, and multivariate features such as audio bigrams.

Second-Order Audio Features in One Dimension

In one dimension, the most straightforward measure of typicality uses a density estimation method to estimate, for an observed feature value, the frequency density of the feature in the corpus. This approach of 'replacing feature values with densities' is also followed in the FANTASTIC toolbox.

Figure 31 shows, top left, a scatter plot of 100 simulated feature value observations (x-axis) and their associated frequency density (y-axis). The original 'first order' feature values are drawn from a standard -normal distribution ($\mu = 0$, $\sigma = 1$, $N = 100$). This has a particular downside, however: since by definition, more observations in the corpus will be associated with a higher feature density, the resulting distribution of the second-order feature will be heavily skewed towards higher values of typicality. The histogram on the top right of figure 31 shows the distribution of this second-order features based on density, for the same 100 simulated observations. This skew is a potential obstacle when the feature is to be used in statistical models, many of which assume normally distributed features, or at least minimally skewed distributions.

A typical method of dealing with this is to find a monotonous transformation that removes the skew from the distribution. Here, we propose a transformation based on *log odds*. Log odds are an alternative measure of probability. The log odds associated with the probability $p \in [0, 1]$ is given by the *logit* function:

$$\log \text{odds} = \text{logit}(p) \tag{72}$$

$$= \log(\frac{p}{1-p}) \tag{73}$$

Our log odds-based second-order feature, the 'logit ranked density' of a feature value x can formally be defined as the *log odds of observing a less extreme value in the reference corpus*. It is conceptually similar to a

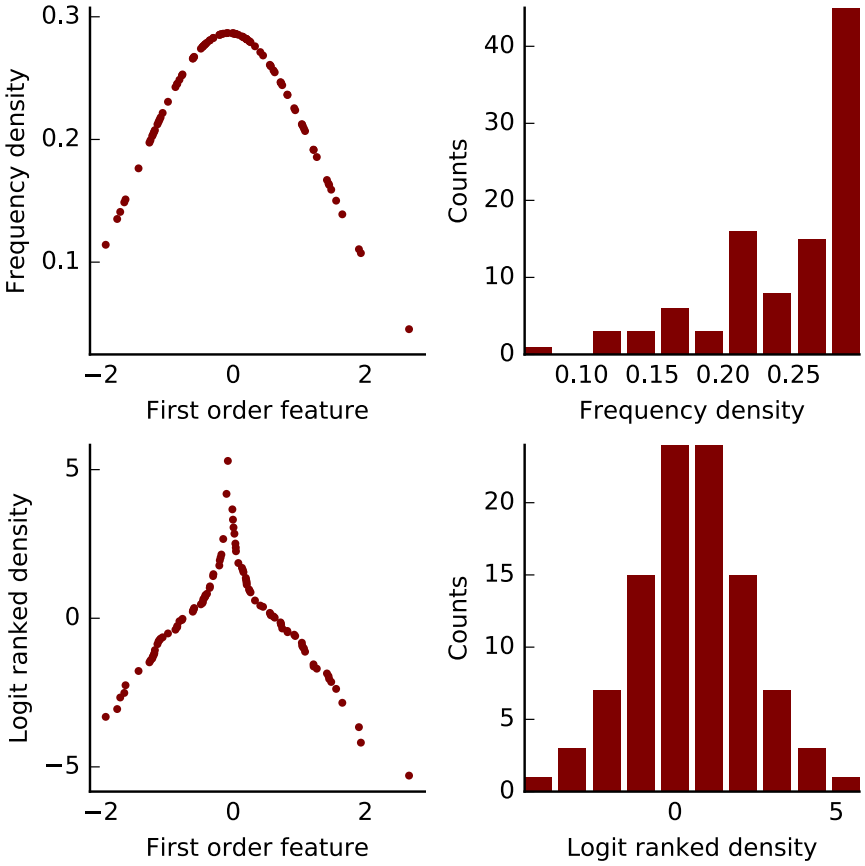

Figure 31.: Left: scatter plot of first order vs. second-order feature values for a sample of 100 simulated observations, and two second-order feature types: density (top) and logit ranked density (bottom). Right: histogram of second-order feature values for the same two feature types.

p-value, which measures the probability of observing a *more* extreme value, but we look at its complement, expressed as log odds.

We further propose a simple non-parametric approach to compute the above odds, based on ranking. By defining 'less extreme' as 'more probable', we can follow the density estimation approach described above to obtain probability density estimates $f(X)$ for all corpus values X and the observed feature value x. We then sort both $f(X)$ and $f(x)$ to find the rank of the feature value's density $f(x)$, and normalize it by the number of items in the corpus. Applying the logit function gives us the *logit ranked density*, hereafter, Z:

$$Z(X) = \text{logit}\left[\frac{\text{rank}\left(f(X)\right) - 0.5}{N}\right] \tag{74}$$

where N is the size of the reference corpus.

The lower half of figure 31 shows, on the left, a scatter plot for 100 simulated feature values (x-axis) and their second-order logit ranked density Z (y-axis). The distribution is somewhat unstable around the $x = 0$, but it is not divergent: as it is based on ranking, the distribution of $Z(X)$ is always bound to $\max(Z) = \text{logit}((N - 0.5)/N) = \log(2N - 1)$.

When we look at the distribution of $Z(X)$, shown on the right, we see that it is perfectly symmetrical. Indeed, again because of its definition based on ranks, Z always follows the same logistic distribution, which is bell-shaped, and generally very similar to a normal distribution. The feature can therefore be used out of the box for a variety of statistical applications.

If the first order feature X is one-dimensional, some form of density estimation is typically possible even if few data are available. Some caution is warranted when using Z where there are a limited number of observations, compared to the number of dimensions. The difficulty of multidimensional density estimation is widely acknowledged in statistics. In short, when the dimensions of a multidimensional feature are expected to be correlated (as is the case for chroma features), a covariance matrix must typically be estimated as part of any kind of density estimation. This increases the number of parameters to be estimated, and therefore the number of required data points—which may not always be available. In the FANTASTIC toolbox, too, "densities are

only computed for one-dimensional features because of the additional conceptual complexity and the high computational resources needed to estimate densities" [132].

Second-Order Audio Features in d Dimensions

For features with more than two or three dimensions, we now propose three methods of computing an alternative second-order feature. The first alternative is very simple: when a multivariate features has relatively independent dimensions by design (e.g., MFCC features), each dimension may be treated as a one-dimensional feature, and a meaningful Z based on density estimation can still be obtained. In practice this amounts to following equation 74 for Z above, but using a diagonal covariance matrix in the density estimation step.

The audio bigram features *MIB* and *HIC* (chapter 5) have 144 dimensions. We may be able to treat these as independent dimensions and get a useful estimate of typicality, however, since the audio bigram features can generally be understood as a probability distribution themselves, other measures of typicality may be more relevant. As a balanced compromise between a range of different options, we adopt two complementary measures of which the distributions are well-behaved.

The first approach is to compute *information* (I), an information-theoretic measure of *unexpectedness*. This measure assumes that the multidimensional first order feature itself can be seen as a frequency distribution F over possible observations in an audio excerpt (cf. term frequencies), and that a similar distribution F_{corpus} can be found for the full reference corpus. We define the $I(F)$ as the average $-\log F_{\text{corpus}}$ weighted by F:

$$I(F) \;=\; -\sum_{i=1}^{d} F(i)\ \log F_{\text{corpus}}(i) \tag{75}$$

The assumptions hold for *MIB*, *HIC* and *HI*, and produce well-behaved second-order feature values. The result is similar to *mean.log.TFDF*, *mtcf.mean.log.DF* and *mtcf.mean.entropy* in the FANTAS-

TIC toolbox. Information is also used as a measure of surprise, or prediction error, by Pearce and others in computational models of (music) cognition [53, 153].

The second measure, a measure of *expectedness* also used in the FANTASTIC toolbox, is a pragmatic, non-parametric measure of similarity between two vectors: Kendall's rank-based correlation τ, computed for the 'term frequencies' F and 'document frequencies' F_{corpus}. Kendall's τ counts the difference in the number of concordant and discordant pairs when the two vectors are sorted and the ranks are compared for each dimension of F. Both $I(F)$ and τ can be computed even for a small reference corpus.

8.1.4 *Song- vs. Corpus-based Second-order Features*

We can expand the notion of second-order features once more when we have access to a corpus of song *sections* rather than songs, as is the case for the Hooked data. Specifically, when a first-order description is available for several sections per song, we can define two reference corpora for every section: the large reference corpus, containing many sections from many songs, and a small, local reference corpus consisting only of sections from the same song. This in turn allows for two types of second-order features: corpus-based and song-based second-order features. In this section, we discuss the advantages of both types of features.

In a statistical learning perspective, expectations arise from statistical inference by the listener, who draws on a lifetime of listening experiences to assess whether a particular stimulus is to be expected or not. In chapter 7, we introduced Huron's three types of musical expectation, including *schematic* expectations, analogous to episodic and semantic memory, and veridical expectations. In short, schematic expectations arise from the 'auditory generalizations' that help us deal with novel, but broadly familiar situations. Veridical expectations are due to familiarity with a specific musical work. Finally, Huron also describes 'adaptive' expectations, which arise dynamically, upon lis-

tening. A repeated motive in a song you never heard before would generate this kind of adaptive expectations.

As the statistical learning paradigm goes, patterns that are more representative of the listener's listening history are more expected. The corpus-based second order features defined above measure typicality and surprise using a large corpus to approximate listening history. Therefore, we can use them to incorporate a crude approximation of schematic expectation in our analysis of hooks, or any other corpus analysis in which expectation plays a role. In the following section, we will refer to corpus-based second-order features as *conventionality*.

The song-based second-order features, by choosing as the reference corpus the set of all segments belonging to the same song, can be said to approximate 'local' expectations. Whether these are more adaptive or more veridical in nature is not entirely obvious—perhaps song-based second-order features cannot distinguish between the expectedness or surprise in some part of an unknown song as it comes along, and the expectedness of a song fragment to a listener who is familiar with the song (veridical expectations). In either interpretation, the features indicate how representative a segment is for the song, and to some extent, how much a segment is repeated. We will therefore also refer to song-based second-order features as *recurrence*.

8.2 DISCOVERY-DRIVEN HOOK ANALYSIS

Now that we have introduced second-order features, we can analyze our corpus of hooks. Using the data from the *Hooked!* experiment, we address the questions:

1. which attributes of the music, as measured by first and second-order features, predict the recognizability of sections of popular music?

This question will be approached by modeling differences between song sections of the same song, as will be argued later in this section.

Additionally, we consider this experiment a good test case to evaluate the newly proposed second-order descriptors described above. Therefore, we would like to know:

2. how do the proposed audio features behave and what aspects of the music do they model?

3. how much insight do audio-based corpus analysis tools add when compared to a symbolic feature set?

We first discuss data and features in sections 8.2.1—8.2.3, before introducing the statistical modeling approach in section 8.2.4.

8.2.1 *Data*

The Hooked experiment and its first implementation, *Hooked!* were described in chapter 7. The experiment tested how quickly and accurately participants could recognize different segments from each song in a collection.

For each song segment and each participant, the *Hooked!* data include a *recognition time* r. The recognition times of all trials in which a user knew the song fragment were combined into a *drift rate*, a single estimate of its recognizability roughly equal to the the reciprocal of the amount of time it would take a median participant to recognize the segment. Stimulus drift rates are commonly used as a measure of recognizability in timed recognition tasks.

The *drift-diffusion model* of memory retrieval, by Ratcliff, was the first cognitive model to propose such a measure [164]. The model assumes that, in a memory retrieval task with two possible answers, responses are driven by evidence accumulating over time, in a way that can be modeled by a continuous random walk process. Figure 32 shows a representation of this process. Here, time is shown on the x-axis, and the (non-monotonic) accumulation of evidence is shown on the y-axis. The random walk begins at a bias level z at a drift rate with mean u and variance s^2. A positive response ('I know this song!') is reached when the evidence hits the top match boundary a. A negative

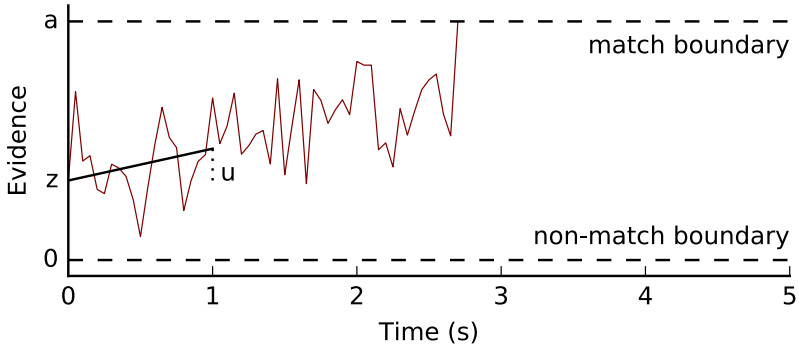

Figure 32.: Diagram of the Ratcliff drift diffusion model of memory retrieval [164]. The decision process is modeled as a random walk, beginning at a bias level z and ending when one of the the match boundaries 0 and a is reached. The mean rate of the random walk is the drift rate u.

response is reached when the accumulated evidence dives below the non-match boundary (x-axis).

The estimation of drift rates in the *Hooked!* dataset is based on a simplified, linear version of Ratcliff's model: the linear ballistic accumulator (LBA) [20]. Linear ballistic accumulator models are easier to fit to data than Ratcliff's original stochastic model. The LBA model, shown in figure 33, associates with each possible response a different 'accumulator', each with their own drift rate distribution (normal with mean v_i and variance s) and bias distribution (uniform between 0 and A). A response is reached when one the accumulators reaches the common match boundary b. LBA allows for more than two accumulators, so it can be applied to tasks with more than two possible responses.

This allowed us to adapt the LBA model and include three accumulators: one for trials in which a participant didn't know the song, one for trials in which the verification question was answered correctly,

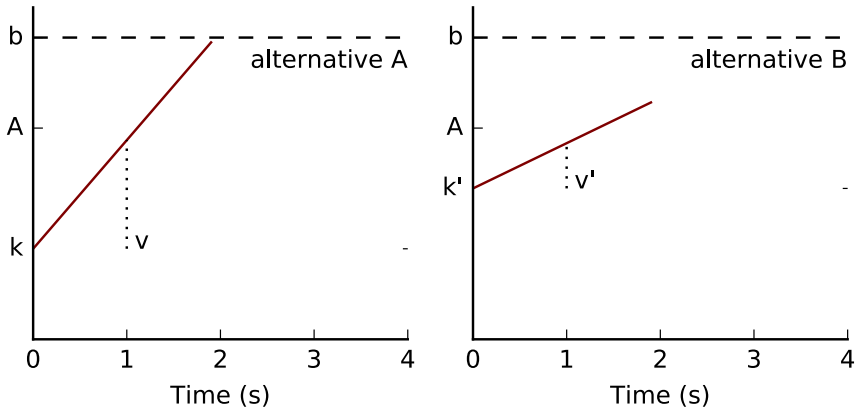

Figure 33.: Diagram of the linear ballistic accumulator model of memory retrieval [20]. The decision process is modeled by one 'accumulator' for each alternative response, each with their own drift rate distribution (normal with mean v_i and variance s) and bias distribution (uniform between 0 and A). A response is reached when one the accumulators reaches the common match boundary b.

				Dispersion		
		2nd-order			2nd-order	
	1st-order	Corpus	Song	1st-order	Corpus	Song
loudness	mean	Z(mean)	Z(mean)	std. dev.	Z(std.dev.)	Z(std.dev.)
sharpness	mean	Z(mean)	Z(mean)			
roughness	mean	Z(mean)	Z(mean)			
MFCC		Z(mean)	Z(mean)	total var.	Z(tot.var.)	Z(tot.var.)
pitch	mean	Z(mean)	Z(mean)	std. dev.	Z(std.dev.)	Z(std.dev.)
MIB		I, τ	I, τ	entropy	Z(entropy)	Z(entropy)
HIC		I, τ	I, τ	entropy	Z(entropy)	Z(entropy)
HI		I, τ	I, τ	entropy	Z(entropy)	Z(entropy)

Table 5.: Overview of the audio feature set used in the *Hooked!* data analysis.

and one for trials in which the participant failed the verification question [25]. In other words, we fit three drift rates per stimulus. All other parameters were set to depend on the participant (e.g., bias), or fixed.

To ensure a reliable fit, we iteratively excluded all song segments with fewer than 15 responses, and participants with fewer than 15 trials. We further excluded all segments from songs with fewer than 3 segments left. After these exclusions, 1715 song segments remained, taken from 321 different songs, representing data from 973 participants. An additional subset was created from 99 songs (536 segments) for which we were able to obtain symbolic transcriptions of the melody and bass line. This subset was used for the symbolic feature model, and to compare audio and symbolic features.

8.2.2 *Audio Features*

Two sets of audio descriptors were combined: first- and second order timbre descriptors, and first- and second-order pitch (melody and

harmony) descriptors. The total number of features is 44. All features were computed over 15-s segments starting from the beginning of each segment, as participants in the experiment were given a maximum of 15 s for recognition.

For timbre description, we used a feature set that is largely the same as the one used in chapter 4. Specifically, we computed the loudness (mean and standard deviations) for each segment, mean sharpness and roughness, and the total variance of the MFCC features. Instead of the pitch centroid feature, we obtained an estimate of pitch height using the *Melodia* melody extraction algorithm and computed the mean and standard deviation.[1]

For each of these one-dimensional features, we then computed the corpus-based and song-based second-order features Z as described in section 8.1.3 using a Python implementation.[2] Finally, we added song and corpus-based $Z(X)$ features based on the mean of the first 13 MFCC components. First-order features based on the MFCC means were not included because of their limited interpretability. An overview of the audio feature set is given in table 5.

For melody and harmony description, we used three of the features described in chapter 5: Melodic Interval Bigrams (*MIB*), Harmonic Interval Co-occurrence (*HIC*) and Harmonization Intervals (*HI*). HPCP were used as chroma features.[3] From these descriptors, we compute the entropy H as a first-order measure of dispersion.[4]

$$H = \sum_{i_1} \sum_{i_2} F(i_1, i_2) \log F(i_1, i_2) \tag{76}$$

The entropies were normalized as follows:

$$H' = \log \frac{H_{\max} - H}{H_{\max}} \tag{77}$$

As second-order features, the information I, and Kendall's τ were computed, as proposed in section 8.1.3.

1 `http://mtg.upf.edu/technologies/melodia`

2 code will be made available at `http://github.com/jvbalen`

3 `http://mtg.upf.edu/technologies/hpcp`

4 To capture as much variance as possible, entropy computation was performed on triads and trigrams before converting them to interval profiles.

8.2.3 *Symbolic Features*

For the symbolic reference feature set, we used a subset of 19 first-order and 5 second-order features from the FANTASTIC toolbox , computed for both melodies and bass lines. Second-order features were computed with both the song and the full dataset as a reference, yielding a total of 58 symbolic descriptors. Table 6 lists all features, with a short description. For exact definitions, see [132].

8.2.4 *Statistical Analysis*

There are two main particularities about the statistical analysis method that will be used in the analysis of the *Hooked!* data: first, it is a discovery-driven analysis, and second, it will be restricted to the analysis of within-song differences. We will now explain what both of these things mean.

Principal Component Analysis

Which attributes of music predict recognizability? Answering the question raised at the beginning of this section calls for a discovery-driven analysis method. This approach to corpus analysis is one of three types of research questions identified in section 3.5.1. It is an exploratory approach in which no particular hypothesis is tested. Typically, we are interested to know which of a candidate set of features correlates with a particular variable of interest. Examples are the approach followed in the analysis of choruses in chapter 4, Leman's analysis of audio features that predict walking speed, and Müllensiefen's analysis of oldness ratings in a melodic memory task [104, 134].

A challenge that arises with this approach, one of several reviewed in chapter 3, is that it typically requires many tests to assess the correlation of several feature with the variable of interest. As a result, a sound strategy is needed to minimize *false positives*, discoveries due to chance. In the three examples above, three strategies are followed: in chapter 4, a probabilistic graphical model is learned, with significance

first-order feature	description
d.median	median note duration
d.range	note duration range (maximum − minimum)
d.entropy	entropy of the note durations distribution
p.std	standard deviation of the pith distribution
p.range	pitch range
p.entropy	entropy of the pitch distribution
i.abs.mean	mean absolute pitch interval
i.abs.std	standard deviation of absolute pitch interval
i.abs.range	absolute pitch interval range
i.entropy	entropy of the pitch interval distribution
len	length of the melody in notes
glob.duration	global duration of the melody
note.dens	number of notes per second
int.cont.grad.mean	mean gradient of the pitch contour
int.cont.grad.std	standard deviation of the gradient of the contour
tonalness	highest of 24 correlations with Krumhansl's key profiles
tonal.clarity	ratio of highest and second-highest key correlation
mean.entropy	mean entropy of the distributions of length-n m-types
mean.productivity	mean of the fraction of length-n m-types appearing only once in the melody
second-order feature	**description**
mtcf.mean.log.DF	mean document frequency of the melody's m-types
mtcf.mean.log.TFDF	TF-weighted mean document frequency of the melody's m-types
mtcf.mean.productivity	mean of the fraction of length-n m-types appearing only once in the corpus
mtcf.TFIDF.m.entropy	entropy of TF-IDF weights of the melody's m-types
mtcf.TFDF.kendall	Kendall τ for TF and DF of the melody's m-types

Table 6.: List of symbolic features used in the *Hooked!* data analysis. All features from the FANTASTIC toolbox [132].

levels for each of the tests adjusted based on the total number of tests involved. In [104], a set of linear models is fit, and cross validation is used to perform model selection on the results. In [134], partial least squares regression (PLSR) is used to combine features into components before fitting a linear model.

In a simpler variation on the PLSR approach by Müllensiefen, we will use principal component analysis (PCA) before fitting the features to the drift rates, as a way of identifying groups of features that may measure a single underlying source of variance. PCA reduces the feature space to a more manageable number of decorrelated variables. This reduces the number of tests required in the next stap of the analysis, a linear model, and thereby the risk of false positive discoveries.

PCA was applied to both the audio and symbolic features, separately. Features were centered and normalized before PCA, and the resulting components were transformed with a varimax rotation to improve interpretability. This orthogonal transformation of the principal components finds rotations in which components have just a few highly-loading parameters, and variables load onto just a few of the components. We selected the number of components to retain (12 in both cases) using parallel analysis, a heuristic method that identifies the number of components needed to model most of the information in the data, by comparing the ranked principal components to those of a randomly sampled dataset with the same number of variables and observations [72]. Not all 12 components representing the symbolic features will be discussed here, but they were considered coherent and interpretable enough to proceed with the analysis. The audio feature components will be discussed as part of the results.

Linear Mixed Effects Model

The second main idea behind our approach to statistical analysis, is that we want to exploit the structure of the *Hooked!* dataset: as we have drift rate estimates for each of the songs sections, we can perform an analysis that looks only at differences between sections of the same song. This allows us to ignore between-song variation, a component

of recognizability that may be dominated by the effects of a variety of extramusical factors, e.g., difference in age of the song, marketing, radio play or social appeal. Instead, we focus on within-song variation, which is much more related to our definition of hooks as the most recognizable part of the song, regardless of its absolute 'catchiness'.

We use a linear mixed-effects regression (LMER) model to fit the feature principal components to the drift rates. Mixed-effects models can handle 'repeated-measures' data where several data points are linked to the same song and therefore have a correlated error structure. The *Hooked!* data provide drift rates for individual sections within songs, and one would indeed expect considerably less variation in drift rates within songs than between them: some pop songs are thought to be much catchier than others overall. Linear mixed-effects models have the further advantage that they are easy to interpret due to the linearity and additivity of the effects of the predictor variables. More complex machine-learning schemes might be able to explain more variance and make more precise predictions for the dependent variable, but this usually comes at the cost of the interpretability of the model.

We fit three models: two including audio components only, one including symbolic components only, and one including both feature types, and used a stepwise selection procedure at $\alpha = 0.005$ to identify the most significant predictors in each model. The audio-only model is fit twice to facilitate comparison between audio and symbolic features: once using the full set of 321 songs and again using just the 99 songs with transcriptions.

In all models, the dependent variable was the *log* drift rate of a song segment and the repeated measures (random effects) are handled as a random intercept, i.e., we add a per-song offset to a traditional linear regression (fixed effects) on song segments, with the assumption that these offsets be distributed normally:

$$\log v_{ij} = \boldsymbol{\beta}\mathbf{x}_{ij} + u_{i0} + \epsilon_{ij} \tag{78}$$

where i indexes songs, j indexes segments within songs, v_{ij} is the drift rate for song segment ij, $\mathbf{x}_{ij}$ is the vector of standardized feature

component scores for song segment ij plus an intercept term, the $y_i \sim N(0, \sigma^2_{\text{song}})$, and the $\epsilon_{ij} \sim N(0, \sigma^2_{\text{residual}})$.

8.3 RESULTS AND DISCUSSION

8.3.1 *Audio Components*

The results of the principal components analysis of the audio features set, the component loadings, are shown in a table in appendix A. The component loadings (correlation coefficients between the extracted components and the original features) tell a consistent story. The first 11 components break the audio feature set down into three timbre components (first order, conventionality, and recurrence) and three entropy components (idem), two features grouping conventionality and recurrence for melody and harmony, respectively, and three more detailed timbre components correlating with sharpness, pitch range and dynamic range.

The last component (component 9 in the table in appendix A) is the most difficult to interpret. It is characterized by an increased dynamic range and MFCC variance, and a typical pitch height. We hypothesize that this component correlates with the presence and prominence of vocals. It is reasonable to assume that the most typical registers for the melodies in a pop corpus would be the registers of the singing voice, and vocal entries could also be expected to modulate a section's timbre and loudness. This hypothesis is also consistent with our own observations while listening to a selection of fragments at various points along the component 9 scale. The high end includes a number of *a capella* or minimally accompanied vocal segments along with a few prominent guitar solos in the vocal register. The verse section from Alicia Keys's No One (2007) is a representative example, with very prominent vocals and only sparse accompaniment. The low end consists primarily of instrumental breaks with relatively undefined melodic content or segments with notably faded vocals, as in the instrumental break of Foo Fighters' Everlong (1997).

Overall, the neatness of the above reduction attests to the advantage of using interpretable features, and to the potential of this particular feature set. Specifically, the tendency of the components to distinguish between conventionality and recurrence suggest that the distinction between song-based and corpus-based second-order features is indeed informative.

8.3.2 *Recognisability Predictors*

Results

Table 7 contains the results of all four linear mixed effects models, showing the fixed effect coefficients, the random intercepts and R^2 values for each model. As expected, the random intercepts per song explain a large amount of variance in the drift rates: between 37 and 40%. However, we are mostly interested in the within-song differences. The coefficients for these fixed effects can roughly be interpreted as percent increase in drift rate per unit of standard deviation in the component (because the dependent variable is logarithmically scaled and the correlation coefficients are relatively low), which makes interpretation easier.

A look at the first column of results for the linear mixed effects model confirms that the audio features are indeed meaningful descriptors for this corpus. Eight components correlate significantly, most of them relating to conventionality of features. This suggests a general pattern in which more recognizable sections have a more typical, expected sound. Another component, timbral recurrence, points to the role of repetition: sections that are more representative of a song are more recognizable. Finally, the component with the strongest effect is Vocal Prominence.

The model based on symbolic data only, in the third column, has just two components. This is possibly due to the reduced number of sections available for fitting. The results, in the second column, for an audio-based model fit on the reduced dataset of 99 songs supports this explanation, as it also yields just two components. In the presence

Parameter	Audio[a] $\hat{\beta}$	Audio[a] 99.5 % CI	Audio[b] $\hat{\beta}$	Audio[b] 99.5 % CI	Symbolic[b] $\hat{\beta}$	Symbolic[b] 99.5 % CI	Combined[b] $\hat{\beta}$	Combined[b] 99.5 % CI
Fixed effects								
Intercept	−0.84	[−0.91, −0.77]	−0.67	[−0.78, −0.56]	−0.62	[−0.73, −0.51]	−0.63	[−0.74, −0.53]
Audio								
Vocal Prominence	0.14	[0.10, 0.18]	0.11	[0.04, 0.17]			0.08	[0.01, 0.15]
Timbral Conventionality	0.09	[0.05, 0.13]						
Melodic Conventionality	0.06	[0.02, 0.11]						
M/H Entropy Conventionality	0.06	[0.02, 0.10]						
Sharpness Conventionality	0.05	[0.02, 0.09]						
Harmonic Conventionality	0.05	[0.01, 0.10]						
Timbral Recurrence	0.05	[0.02, 0.08]						
Mel. Range Conventionality	0.05	[0.01, 0.08]	0.07	[0.02, 0.13]			0.07	[0.01, 0.12]
Symbolic								
Melodic Repetitivity					0.12	[0.06, 0.19]	0.11	[0.05, 0.17]
Mel./Bass Conventionality					0.07	[0.01, 0.13]	0.08	[0.01, 0.14]
Random effects								
$\hat{\sigma}_{\text{song}}$	0.39	[0.34, 0.45]	0.35	[0.26, 0.45]	0.34	[0.25, 0.44]	0.32	[0.24, 0.42]
$\hat{\sigma}_{\text{residual}}$	0.48	[0.45, 0.50]	0.40	[0.37, 0.44]	0.39	[0.35, 0.43]	0.38	[0.34, 0.42]
R^2_{marginal}[c]	.10		.06		.07		.10	
$R^2_{\text{conditional}}$[c]	.47		.46		.47		.47	
$-2 \times$ log likelihood	2765.61		699.81		576.74		558.11	

Note. Random-intercept models, grouping by song, for the given feature types after step-wise selection using Satterthwaite-adjusted F-tests at $\alpha = .005$. Component scores were standardized prior to regression.

[a] Complete set of 321 songs ($N = 1715$ segments).

[b] Reduced set of 99 songs with symbolic transcriptions ($N = 536$ segments).

[c] Coefficients of determination following Nakagawa and Schielzeth [138].The marginal and conditional coefficients reflect, respectively, the proportion of variance in the data that is explained by the fixed effects alone and the proportion explained by the complete model (fixed and random effects together).

Table 7.: Estimated coefficients and variances for audio and symbolic components predicting the relative recognizability of popular song segments.

of less data, only the most important components stand out. The symbolic and reduced audio model seem to be comparable in power with a marginal R^2 of 0.06 and 0.07, respectively (see figure caption for definitions).

The top symbolic features that make up the first of the significant components are melodic entropy and productivity, both negatively correlated, suggesting that recognizable melodies are more repetitive. The top features that make up the second components are *mtcf.mean.log.DF*, for the melody (song-based and corpus-based), and negative *mtcf.mean.productivity* (song-based and corpus-based for both bass and melody). This suggests that recognizable melodies contain more typical motives (higher codument frequencies, lower second-order productivity).

The last column shows how the combined model, in which both audio and symbolic components were used, retains the same audio and symbolic components that make up the previous two models. The feature sets are, in other words, complementary: not only are all four components still predictive at $\alpha < 0.005$, the marginal R^2 now reaches 0.10, as opposed to 0.06 and 0.07 for the individual models. This answers the last of the questions stated in section 8.2: for the data in this study, the audio-based corpus analysis tools contribute substantial insight, and make an excellent addition to the symbolic feature set.

Discussion

We briefly discuss our findings on the properties of hooks. First, the presence of vocals appears to be the strongest predictor of recognizability. We see this as an unsurprising but important result: it suggests that vocal melodies are very important in the recognition of popular music.

Second, sections that are more conventional in terms of melody, harmony, bass and timbre are more recognizable: 7 conventionality components in total, across both the audio and the symbolic model, show a consistent positive correlation between conventionality and recog-

nizability. In other words, if there were to exist a positive effect of *distinctiveness* on recognizability, as suggested in some of the hook hypotheses in chapter 7, no evidence is found for it in this analysis. The analysis suggests rather the opposite: recognizable sections are more typical than they stand out.

Finally, the data suggest that recognizable song sections are more repetitive (as measured by symbolic melody repetitiveness), and more repeated (as measured by timbre recurrence). This does align with some of the related findings reviewed in chapter 7: repeated exposure and recognizability go hand in hand—at least, as measured using these two sets of features.

8.4 CONCLUSIONS AND FUTURE WORK

In this chapter, we have presented a new approach to corpus-level audio description, and a new discovery-driven analysis of popular music hooks. We introduced three general-purpose second-order audio descriptors: the 'logit ranked density' Z, information I and Kendall's τ, and the notion of song-based and corpus-based second-order features. In the hook discovery experiment, two features sets were compiled: an audio feature set based on the new the audio description methods and a symbolic reference feature set. We then used PCA and LMER to predict, from these features, recognizability of song fragments in the *Hooked!* dataset.

From the results and discussion of the statistical analysis we conclude that the harmony and melody descriptors, the corpus-based second-order features and the song-based second-order features contribute new and relevant layers of information to the corpus description. From the results of the audio analysis, we conclude that sections with vocals and sections that are most representative of the song in terms of timbre, are better recognized. Recognizable song sections also have a more typical, expected sound. From the symbolic results, we conclude that recognizable melodies are more repetitive, and contain less atypical motives. In short: vocals, conventionality and repetition best predict recognizability.

Finally, we conclude that an audio corpus analysis as proposed in this paper can indeed complement symbolic corpus analysis, as the experiment sees both kinds of features explaining an important share of the variance in the data. This opens up a range of opportunities for future work. These will described in the last chapter of this thesis.