CORPUS ANALYSIS TOOLS FOR COMPUTATIONAL HOOK DISCOVERY

Jan Van Balen[1] **John Ashley Burgoyne**[2] **Dimitrios Bountouridis**[1]
Daniel Müllensiefen[3] **Remco Veltkamp**[1]
[1] Universiteit Utrecht [2] Universiteit van Amsterdam [3] Goldsmiths, University of London

ABSTRACT

Compared to studies with symbolic music data, advances in music description from audio have overwhelmingly focused on ground truth reconstruction and maximizing prediction accuracy, with only a small fraction of studies using audio description to gain insight into musical data. We present a strategy for the corpus analysis of audio data that is inspired by the FANTASTIC toolbox and optimized for interpretable results. The approach brings two previously unexplored concepts to the audio domain: audio bigram distributions to describe melodic and harmonic content, and the use of corpus-relative or "second-order" descriptors. To test the real-world applicability of our method, we present an experiment in which we model song recognition data collected in a widely-played music game. By using the proposed corpus analysis pipeline we are able to present a cognitively adequate analysis that allows a model interpretation in terms of the listening history and experience of our participants. We find that our corpus-based audio features are able to explain a comparable amount of variance to symbolic features for this task when used alone and that they can supplement symbolic features profitably when the two types of features are used in tandem. We discuss the further potential of audio features for corpus analysis, and highlight new insights into what makes music recognizable.

1. INTRODUCTION

This study addresses the scarcity of corpus analysis tools for audio data. Corpus analysis, by which we refer to any analysis of a collection of musical works in which the primary goal is to gain insight into the music itself, makes up only a small fraction of the music computing field, with much more research being done on classification, recommendation and retrieval [17], where the focus is often more on prediction accuracy than interpretability. Examples of corpus analysis studies include work on summarization and visualisation (e.g., [1]), specific hypothesis testing, (e.g., evidence for Western influence in the use of African tone scales in [11]), and discovery-based analysis (e.g., of the structural melodic features that predict performance in a music memory task [13]).

Strikingly, while audio data is by far the most widely researched form of information in the community [17], a brief review suggests that only a minority of corpus analysis studies used audio data. This includes the above work on visualisation [1], tone scales analysis [1, 11], and a number of recent studies on the structure and evolution of popular music [10, 16, 19]. Symbolic corpus analysis, in contrast, includes Huron's many studies [9], Conklin's work on multiple viewpoints and Pearce's extensions [6, 15], and studies of harmony by De Clercq & Temperley [7] and Burgoyne [5], to name a few, as well as toolkits such as Humdrum, [1] Idyom, [2] and FANTASTIC [12]. [3]

Although the music information retrieval community has made substantial progress in improving the transcription of audio to symbolic data, considerable hurdles remain [17]. We therefore aim to further the resources for audio analysis, by introducing a new set of tools. Specifically, we present a set of audio corpus description features that are founded on the use of three novel concepts. A new kind of melodic and harmonic interval profiles are used to describe melody and harmony, extending the notion of interval bigrams to the audio domain. We then propose three so-called "second-order" features, a concept that is proposed in the FANTASTIC toolbox but has yet to be applied to audio features. Finally, we define song-based and corpus-based second-order features.

We test our newly developed analysis pipeline in a case study on "hook discovery".

2. CORPUS-BASED AUDIO FEATURES

2.1 Harmony and Melody Description

We propose a novel set of harmony and melody descriptors. The purpose for these descriptors is to translate basic harmonic and melodic structures to a robust representation on which corpus statistics can be computed. They should be relatively invariant to other factors such as tempo and timbre, and have a fixed size.

In [18], the correlation matrix of the chroma features is used as a harmonic descriptor. The resulting 144-dimensional

JVB, JAB and DB are supported by COGITCH (NWO CATCH project 640.005.004) and FES project COMMIT/.

[1] http://www.musiccog.ohio-state.edu/Humdrum/
[2] https://code.soundsoftware.ac.uk/projects/idyom-project
[3] http://www.doc.gold.ac.uk/isms/m4s/

'chroma correlation features' measure co-occurrence of harmonic pitch. They capture more detail than a simple chroma pitch histogram, while preserving tempo and translation-invariance. The feature was shown to perform reasonably well in a small-scale cover song experiment. In this study we extend this and two related concepts to three new interval representations. Whereas pitch bigram profiles are expected to strongly correlate with the key of an audio fragment, interval bigrams are key-invariant, which allows them to be compared across songs.

The **Harmonic Interval Co-occurrence (HIC)** is based on the *triad profile*, which is defined as the three-dimensional co-occurrence matrix of three identical copies of the chroma time series $c_{t,i}$ (t is time, i is pitch class):

$$\text{triads}(c)_{i_1,i_2,i_3} = \sum_t c_{t,i_1}\ c_{t,i_2}\ c_{t,i_3}. \tag{1}$$

The pitch class triplets in this feature can be converted to interval pairs using the function:

$$\text{intervals}(X)_{j_1,j_2} = \sum_{i=0}^{12} X_{(i-j_1) \bmod 12,\ i,\ (i+j_2) \bmod 12}. \tag{2}$$

This essentially maps each triad (i_1, i_2, i_3) to a stack of intervals $(i_2 - i_1, i_3 - i_2)$. A major chord $(0, 4, 7)$ would be converted to $(4, 3)$, or a major third with a minor third on top. Applied to the triads matrix, the intervals function yields the *harmonic interval co-occurrence* matrix,

$$\text{HIC}(c)_{j_1,j_2} = \text{intervals}(\text{triads}(c_{t,i})) \tag{3}$$

It measures the distribution of triads in an audio segment, represented by their interval representation. For example, a piece of music with only minor chords will have a strong activation of $\text{HIC}_{3,4}$, while a piece with a lot of tritones will have activations in $\text{HIC}_{0,6}$ and $\text{HIC}_{6,0}$.

The same processing can be applied to the melodic pitch to obtain the **Melodic Interval Bigrams (MIB)**. We first define the three-dimensional *trigram profile* as an extension of the two-dimensional bihistogram in [18]:

$$\text{trigrams}(m)_{i_1,i_2,i_3} = \sum_t \max_\tau(m_{t-\tau,i_1})\ m_{t,i_2} \max_\tau(m_{t+\tau,i_3}), \tag{4}$$

with $\tau = 1 \ldots \Delta t$ and m the melody matrix, a binary chroma-like matrix containing the melodic pitch activations. The result is a three-dimensional matrix indicating how often triplets of melodic pitches (i_1, i_2, i_3) occur less than Δt seconds apart. The pitch trigram profile can be converted to an interval bigram profile by applying the intervals function (Eqn 2). This yields the *melodic interval bigrams* feature, a two-dimensional matrix that measures which pairs of pitch intervals follow each other in the melody:

$$\text{MIB}(X)_{j_1,j_2} = \text{intervals}(\text{trigrams}(m_{t,i})). \tag{5}$$

Finally, the *harmonisation feature* in [18] measures which harmonic pitches in the chroma c co-occurr with the melodic pitches in the melody m. We derive a **Harmonisation Interval (HI)** feature as:

$$\text{HI}(m, h)_j = \sum_t \sum_{i=0}^{12} m_{t,i}\ h_{t,i+j} \tag{6}$$

2.2 Second-order Features

One of the contributions of the FANTASTIC toolbox is to include *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. They help contextualize the values a feature can take. Is this 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? Examples of second-order features in the FANTASTIC toolbox include features based on document frequencies, i.e. how many songs (documents) in a large corpus contain an observed event or structure: *mtcf.mean.log.DF* computes the mean log document frequency over all melodic motives in a given melody.

2.2.1 Second-Order Audio Features in One Dimension

Like many audio features, most of the audio features discussed in this paper are based on frequency-domain computations, which are typically performed on short overlapping windows. As a result, the features discussed here represent continuous-valued, uncountable quantities. Symbolic features, on the other hand, operate on countable collections of events. This makes it impossible to apply the same operations directly to both, and alternatives must be found for the audio domain.

After comparison of several alternatives, we propose a non-parametric measure of typicality based on log odds. The second-order log odds 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 p-value, which measures the probability of observing a *more* extreme value, but we look at its complement, expressed as odds, and take the log.

We further propose a simple non-parametric approach to compute the above odds. By defining 'less extreme' as 'more probable', we can make use of density estimation (e.g., kernel density estimation) to obtain a probability density estimate $f(X)$ for the observed feature X, and look at the rank of each feature value's density in the reference corpus. Normalizing this rank by the number of observation gives us a pragmatic estimate of the probability we're looking for, and applying the logit function gives us the log odds:

$$Z(X) = \text{logit}\left[\frac{\text{rank}\,(f(X))}{N}\right] \tag{7}$$

where N is the size of the reference corpus. Since Z is non-parametric and based on ranks, the output always follows the same logistic distribution, which is bell-shaped, symmetric, and general very similar to a normal distribution. The feature can therefore be used out of the box for a variety of statistical applications.

Some caution is warranted when using Z where there are a limited number of observations. If the first order feature X is one-dimensional, some form of density estimation is typically possible even if few data are available. For multivariate features with independent dimensions (e.g.,

MFCC features), each dimension can be treated as a one-dimensional feature, and a meaningful density estimate can also be obtained. However, if the dimensions of a multidimensional feature are not de-correlated by design but highly interdependent (as is the case for chroma features), density estimates require more data. For such cases, a covariance matrix must typically be estimated, increasing the number of parameters to be estimated, and thereby the number of required data points for a fit.

2.2.2 Second-Order Audio Features in d Dimensions

For higher-dimensional features, such as the interval bigram and interval co-occurrence profiles *MIB* and *HIC*, we turn to other measures of typicalness. After comparison of several of the alternatives through inspection of the resulting distributions and correlations with other features, we adopt two approaches. The first measure, directly adopted from the FANTASTIC toolbox, is Kendall's rank-based correlation τ. The second measure is *information* (I), an information-theoretic measure of unexpectedness. This measures assumes that the multidimensional first order feature itself can be seen as a probability distribution F over possible observations in an audio excerpt (cf. term frequencies), and that a similar distribution F_c can be found for the whole of the reference corpus (cf. document frequencies). In this context, the total information in this feature is defined as the average of $-\log F_c$, weighted by F:

$$I(F) \;=\; -\sum_{i=1}^{d} F(i)\;\log F_c(i) \qquad (8)$$

The assumptions hold for HIC, BIM 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 FANTASTIC toolbox and highly correlated with *mtcf.mean.gl.weight*. Information is also used as a measure of surprise by Pearce [15].

2.3 Song- vs. Corpus-based Second-order Features

In a statistical learning perspective on expectation, 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 [9], Huron compares *veridical* and *schematic* expectations, as the expectation counterparts of episodic and semantic memory. The former describes expectations of a listener due to familiarity with a specific musical work. The latter refers to expectations that arise from the "auditory generalizations" that help us deal with novel, but broadly familiar situations.

For any corpus analysis in which the documents are song fragments rather than entire songs, second-order features create an opportunity to incorporate a crude model for both layers of expectation. By choosing the reference corpus to be a big collection of fragments spanning a large number of songs, the above typicality and surprise measures approximate schematic expectations: a value that is typical is representative of the reference corpus, and therefore expected. By choosing the reference corpus to be the set of all segments belonging to the same song, something closer to veridical expectations can be modeled.

In the experiment in the following section, we will describe the corpus-based second-order features as *conventionality*. The song-based second-order features indicate how representative a segment is for the work from which it is taken, and to a certain extent, how much a segment is repeated. We will therefore refer to this as *recurrence*.

3. HOOK DISCOVERY: A CASE STUDY

We tested the proposed approach to audio corpus analysis by examining data from the Hooked! experiment on long-term musical salience [3]. Using these data, we sought to address three questions: (i) how do the proposed audio features behave and what aspects of the music do they model, (ii) which attributes of the music, as measured by both an audio feature set and a selection of symbolic features, predict recognition rating differences within songs, and finally, (iii) how much insight do audio-based corpus analysis tools add when compared to the symbolic feature set?

3.1 Data

The Hooked! experiment used a broad selection of Western pop songs from the 1930s to the present. The experiment tested how quickly and accurately participants could recognise different segments from each song, based on the Echo Nest segmentation algorithm. [4] For each song segment, the data include an estimate of the *drift rate*, the reciprocal of the amount of time it would take a median participant to recognize the segment, based on linear ballistic accumulation, a cognitive model for timed recognition tasks [2,4]. To improve reliability, we excluded song segments that fewer than 15 serious participants had attempted to recognize (where a "serious" participant is defined to be a participant who attempted at least 15 segments). We further excluded all segments from songs from which fewer than 3 segments met the previous reliability criteria. After these exclusions, 1715 song segments remained, taken from 321 different songs, representing data from 973 participants. We were unable to obtain symbolic transcriptions of all songs, and so for comparing audio and symbolic features, we used a restricted set of 99 transcribed songs (536 segments).

3.2 Audio Features

For timbre description, we used a feature set that is largely the same as the one used in [19], where statistical analysis of an audio corpus is used to model pop songs choruses. 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. [5]

For each of these one-dimensional features, we then computed the corpus-based and song-based second-order

[4] http://www.echonest.com/

[5] http://mtg.upf.edu/technologies/melodia

features as described in Section 2.2.1 using Python. [6] 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. 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 melody and harmony description, we used the features described in Section 2.1, and compute the entropy H as a first-order measure of dispersion.The entropies were then normalized as follows:

$$H' = \log \frac{H_{\max} - H}{H_{\max}} \tag{9}$$

As second-order features, Kendall's τ and the information I were computed, as proposed in Section 2.2.2. Chroma features were based on HPCP. [7]

3.3 Symbolic features

The symbolic features used were 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.

3.4 Principal Component Analysis

Before going further with either the audio or the symbolic feature sets, we used principal component analysis (PCA) as a way to identify groups of features that may measure a single underlying source of variance and as a way to reduce the dimensionality of the feature spaces to a more manageable number of decorrelated variables. Features were centered and normalized before PCA, and the resulting components were transformed with a varimax rotation to improve interpretability. We selected the number of components to retain (12 in both cases) using parallel analysis [8].

3.5 Linear Mixed Effects Model

In order to fit the extracted components to the drift rates, we used a linear mixed-effects regression model. 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. Moreover, it is likely impossible to model between-song variation in recognisablity from content-based features alone: it may arise from differences in marketing, radio play, or social appeal.

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, one including audio components only, one including symbolic components only, and one including both feature types, and used a stepwise selection procedure at $\alpha = .005$ to identify the most significant predictors under each model. In all models, the dependent variable was the log drift rate of a song segment and the repeated measures (random effects) were handled as a random intercept, i.e., we added a per-song offset to a traditional linear regression (fixed effects) on song segments, with the assumption that these offsets be distributed normally:

$$\log y_{ij} = \boldsymbol{\beta}' \mathbf{x}_{ij} + u_i + \epsilon_{ij} \tag{10}$$

where i indexes songs, j indexes segments within songs, y_{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 $u_i \sim N(0, \sigma^2_{\text{song}})$, and the $\epsilon_{ij} \sim N(0, \sigma^2_{\text{residual}})$. To facilitate comparison, we fit the audio-only model twice: once using the full set of 321 songs and again using just the 99 songs with transcriptions.

4. RESULTS AND DISCUSSION

4.1 Audio Components

Table 1 displays the component loadings (correlation coefficients between the extracted components and the original features) for audio feature set. The loadings tell a consistent story. The 12 components we retain 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.

Component 9 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 not unreasonable 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.

Overall, the neatness of the above reduction attests to the advantage of using interpretable features, and to the potential of this particular feature set.

4.2 Recognizability Predictors

A look at the first column of results for the linear mixed effects model (Table 2) 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,

[6] code will be made available on `http://github.com/jvbalen`
[7] `http://mtg.upf.edu/technologies/hpcp`

	Component											
Feature	1	2	3	4	5	6	7	8	9	10	11	12
MIB | Song	.31	−.10	.12	.08	.05	**.66**	.05	.08	.23	.08	−.01	.14
HI | Song	−.25	−.08	.12	.06	.11	**.55**	.12	.35	−.06	.04	.01	−.02
MIB | Corpus	.15	−.03	−.02	.13	.00	**.77**	−.06	.00	.08	−.02	−.01	.05
HI | Corpus	−.28	−.09	−.05	−.01	.10	**.55**	.11	**.42**	−.15	−.02	.08	−.05
HIC | Song	.04	.13	.22	.04	.00	.13	−.04	**.58**	−.03	.06	−.02	−.03
HIC | Corpus	−.23	.11	.04	.32	.08	.15	−.07	**.66**	.03	−.06	.07	.00
HIC Entropy	**.88**	.06	.03	−.16	.02	.07	−.02	−.23	−.12	.02	−.00	−.10
MIB Entropy	**.83**	−.15	−.00	−.19	.04	.04	.08	.26	.26	.03	−.02	.20
HI Entropy	**.85**	−.06	.02	−.20	.01	−.01	.04	.15	.12	.02	−.02	.16
HIC Song Information	**.84**	.17	.06	.09	.11	.13	−.02	−.16	−.28	−.04	.10	−.13
MIB Song Information	**.79**	−.21	−.03	.01	.07	.05	.13	.25	.29	.07	−.02	.21
HI Song Information	**.90**	.18	.01	.11	.07	−.07	.00	−.17	−.03	−.02	.00	−.03
HIC Corpus Information	**.86**	.16	.06	.01	.10	.11	−.02	−.20	−.27	−.02	.09	−.13
MIB Corpus Information	**.79**	−.19	−.01	−.03	.07	.02	.14	.26	.31	.07	−.02	.21
HI Corpus Information	**.90**	.15	.02	−.01	.03	−.12	−.01	−.24	−.03	.00	−.02	−.03
HIB Entropy | Song	.03	.11	**.42**	.08	.03	.00	−.08	.15	.08	.19	.01	−.06
MIB Entropy | Song	.01	−.01	.07	.10	.03	−.01	.03	.02	−.01	**.82**	.00	.05
HI Entropy | Song	.03	.02	.11	.12	.06	.04	−.02	−.01	.02	**.81**	−.01	.02
HIB Entropy | Corpus	−.13	.08	.08	**.68**	.08	.15	−.06	.26	−.03	−.10	.07	−.02
MIB Entropy | Corpus	−.04	−.09	−.01	**.80**	.01	.06	.14	−.01	.05	.16	.00	.07
HI Entropy | Corpus	−.03	−.07	−.02	**.84**	.04	.04	.06	.04	.05	.19	−.02	.04
Loudness	−.04	**.92**	.07	−.06	−.05	−.05	−.07	.06	−.04	.02	−.07	.04
Roughness	.14	**.78**	.14	.01	.15	.09	.31	.06	−.08	.07	.06	.01
Melodic Pitch Height	.13	**.66**	−.05	−.03	.09	−.24	−.16	.09	.22	−.06	−.06	.00
MFCC Variance	.13	**−.51**	−.05	.08	−.26	.10	.05	−.02	**.48**	.02	−.22	−.10
Loudness | Song	−.03	−.05	**.67**	−.01	.06	.01	.07	−.04	.10	.03	.11	−.03
Roughness | Song	.04	.10	**.67**	−.03	−.01	.02	.11	.08	−.05	−.02	−.04	−.05
Mel. Pitch Height | Song	−.01	.02	**.46**	.03	.13	.14	−.12	−.15	.29	.07	.16	.03
MFCC Mean | Song	.07	.07	**.61**	−.04	.21	.12	.10	.10	−.07	.16	.11	.11
MFCC Variance | Song	.00	−.04	**.54**	.03	.01	−.06	.10	.08	−.10	−.09	−.06	.17
Loudness | Corpus	.04	−.23	.06	.07	.12	.08	**.76**	−.05	.22	.02	.10	−.05
Roughness | Corpus	.12	.34	.15	.03	.00	.01	**.71**	−.07	.05	.04	.03	−.07
Mel. Pitch Height | Corpus	.00	.04	.06	.06	.25	.06	.14	−.01	**.60**	.02	.14	−.09
MFCC Mean | Corpus	.21	.13	.12	.07	**.51**	.03	.31	.20	−.18	.05	.14	.08
MFCC Variance | Corpus	−.09	−.09	.08	.08	.25	−.02	**.40**	.05	−.13	−.13	−.12	.21
Sharpness	.23	.11	.03	.08	**.72**	.04	.29	.13	.08	−.01	.10	.05
Sharpness | Song	−.02	−.07	.24	−.04	**.50**	.06	−.14	−.07	.04	.15	−.08	−.04
Sharpness | Corpus	.08	.10	.03	.06	**.75**	.03	.03	−.02	.14	−.01	−.10	−.01
Loudness SD	.10	.38	.09	.06	−.06	.06	.22	.02	**.40**	.03	**−.61**	−.03
Loudness SD | Song	.04	.02	.22	.02	−.05	.00	−.05	.03	.14	.01	**.60**	.03
Loudness SD | Corpus	.03	.05	−.02	.05	−.03	.04	.19	.02	.04	.00	**.78**	.02
Mel. Pitch SD	.21	−.10	−.02	−.05	.04	−.19	.21	.18	−.27	.12	−.07	−.28
Mel. Pitch SD | Song	.01	.04	.11	.01	.04	.13	.00	−.15	.01	.14	.07	**.69**
Mel. Pitch SD | Corpus	.13	.03	−.02	.06	−.01	−.02	.01	.11	−.08	−.04	.00	**.74**
R^2	.16	.06	.05	.05	.05	.04	.04	.04	.04	.04	.04	.03

Note. MIB = Melodic Interval Bigram; HI = Harmonization Interval; HIC = Harmony Interval Co-occurrence. Loadings > .40 are in boldface. Collectively, these components explain 64 % of the variance in the underlying data. We interpret and name them as follows: (1) Melodic/Harmonic Entropy, (2) Timbral Intensity, (3) Timbral Recurrence, (4) Melodic/Harmonic Entropy Conventionality, (5) Sharpness Conventionality, (6) Melodic Conventionality, (7) Timbral Conventionality, (8) Harmonic Conventionality, (9) Vocal Prominence, (10) Melodic Entropy Recurrence, (11) Dynamic Range Conventionality, and (12) Melodic Range Conventionality.

Table 1. Loadings after varimax rotation for principal component analysis of corpus-based audio features.

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. Grouping by song, all models displayed are the optimal random-intercept models for the given feature types after step-wise selection using Satterthwaite-adjusted F-tests at $\alpha = .005$. Component scores – but not log drift rates – 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's technique for mixed-effects models [14]. The marginal coefficient reflects the proportion of variance in the data that is explained by the fixed effects alone and the conditional coefficient the proportion explained by the complete model (fixed and random effects together).

Table 2. Estimated prediction coefficients and variances for audio and symbolic components influencing the relative recognizability (log drift rate) of popular song segments.

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, as the audio-based model run on the reduced dataset also yields just two components. 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 DF, 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 relevant at $\alpha < .005$, the marginal R^2 now reaches .10, as opposed to .06 and .07 for the individual models. This answers the last of the questions stated in Section 3: 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.

5. CONCLUSIONS AND FUTURE WORK

We have presented a strategy for audio corpus description that combines a new kind of melodic and harmonic interval profiles, three general-purpose second-order features, and the newly introduced notion of song-based and corpus-based second-order features. Using these features to analyse the results of a hook discovery experiment, we show that all of the above contributions add new and relevant layers of information to the corpus description. We conclude that an audio corpus analysis as proposed in this paper can indeed complement symbolic corpus analysis, which opens a range of opportunities for future work. As possible future directions we would like to perform more experiments on the Hooked! data, exploring more first- and second-order descriptors and more powerful statistical or machine-learning models, to see if allowing for interactions and non-linearities helps to explain more of the variance in drift rates between sections. We also would like to extend the feature set to explore rhythm description and chord estimation, especially as more reliable transcription tools become available from the MIR community.

6. REFERENCES

[1] Mathieu Barthet, Mark Plumbley, Alexander Kachkaev, Jason Dykes, Daniel Wolff, and Tillman Weyde. Big chord data extraction and mining. In *Proceedings of the 9th Conference on Interdisciplinary Musicology*, Berlin, Germany, 2014.

[2] Scott Brown and Andrew Heathcote. The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3):153–78, 2008.

[3] John Ashley Burgoyne, Dimitrios Bountouridis, Jan Van Balen, and Henkjan J. Honing. Hooked: A game for discovering what makes music catchy. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, pages 245–50, Curitiba, Brazil, 2013.

[4] John Ashley Burgoyne, Jan Van Balen, Dimitrios Bountouridis, Themistoklis Karavellas, Frans Wiering, Remco C. Veltkamp, and Henkjan J. Honing. The contours of catchiness, or Where to look for a hook. Paper presented at the International Conference on Music Perception and Cognition, Seoul, South Korea, 2014.

[5] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. Compositional data analysis of harmonic structures in popular music. In Jonathan Wild, Jason Yust, and John Ashley Burgoyne, editors, *Mathematics and Computation in Music*, pages 52–63. Springer, Berlin, 2013.

[6] Darrell Conklin and Ian H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

[7] Trevor de Clercq and David Temperley. A corpus analysis of rock harmony. *Popular Music*, 30(1):47–70, 2011.

[8] John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–85, 1965.

[9] David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, MA, 2006.

[10] Matthias Mauch, Robert M MacCallum, Mark Levy, and Armand M Leroi. The evolution of popular music: USA 1960–2010. *Royal Society Open Science*, In press.

[11] Dirk Moelants, Olmo Cornelis, and Marc Leman. Exploring African tone scales. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 489–94, Kobe, Japan, 2009.

[12] Daniel Müllensiefen. FANTASTIC: Feature ANalysis Technology Accessing STatistics (in a corpus). Technical report, 2009.

[13] Daniel Müllensiefen and Andrea R Halpern. The role of features and context in recognition of novel melodies. *Music Perception: An Interdisciplinary Journal*, 31(5):418–435, 2014.

[14] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–42, 2013.

[15] Marcus Thomas Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, City University London, England, 2005.

[16] Joan Serrà, Alvaro Corral, Marián Boguñá, Martín Haro, and Josep Ll. Arcos. Measuring the evolution of contemporary Western popular music. *Scientific Reports*, 2(521), 2012.

[17] Xavier Serra, Michela Magas, Emmanouil Benetos, Magdalena Chudy, S. Dixon, Arthur Flexer, Emilia Gómez, F. Gouyon, P. Herrera, Sergi Jordà, Oscar Paytuvi, G. Peeters, Jan Schlüter, H. Vinet, and G. Widmer. *Roadmap for Music Information ReSearch*. MIReS Consortium, 2013.

[18] Jan Van Balen, Dimitrios Bountouridis, Frans Wiering, and Remco Veltkamp. Cognition-inspired descriptors for scalable cover song retrieval. In *Proceedings of the 15th Interational Society for Music Information Retrieval Conference*, pages 379–384, Taipei, Taiwan, 2014.

[19] Jan Van Balen, John Ashley Burgoyne, Frans Wiering, and Remco C. Veltkamp. An analysis of chorus features in popular song. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.