3

AUDIO CORPUS ANALYSIS

This chapter will provide more context for the research in this thesis, though not in terms of its technological context, like the previous chapter, but in terms of its methodology: *audio corpus analysis*. A large majority of the studies in computational music analysis center around transcription, classification, recommendation and retrieval, effectively limiting themselves to the reconstruction of a ground truth and rarely leveraging the power of computation to mine music collections for novel musical insights. We discuss the challenges and pitfalls in applying MIR's technology in the pursuit of a better understanding of music. At the end of this section, a selection of desiderata for dedicated audio corpus analysis technology is given.

3.1 AUDIO CORPUS ANALYSIS

3.1.1 *Corpus Analysis*

In this thesis, we define corpus analysis as: any analysis of a collection of musical works in which the primary goal is to gain insight into the music itself. Consider, as an example, Huron's study of melodic arcs in Western folk song [77]. In this study, Huron used the *Humdrum* toolkit and a corpus of 6251 folk songs from the Essen Folksong Collection to show a tendency towards arch-shaped melodic contours. Particularly, he demonstrated that, of nine simple contour-types, a convex shape was most common, and that there is a significant tendency for ascending and descending phrases to be linked together in pairs. As we argued in chapter 1, corpus studies like this form part of 'empirical musicology', distinguish themselves from a large body of other computational music research by answering a musicological question and aiming at new musical insights.

3.1.2 *Audio Corpus Analysis*

Audio corpus analysis can now be defined simply as corpus analysis on audio data, as opposed to symbolic data (such as scores) or manual annotations. In practice, music data may not always come in an unambiguously unimodal form, but it is safe to say that there is a striking prevalence of symbolic datasets in corpus analysis. A review in the next section will show that audio is used only in a minority of the studies, despite its potential for corpus analysis as argued in section 1.2.4: despite its availability and despite being, by far, the most widely used and researched form of information in the music computing community.

Additionally to the arguments presented in chapter 1, recent audio corpus analysis results have gathered a wide interest in the press, notably, since work on this thesis began, Serrà and Mauch's studies of the evolution of popular music [122, 175] (see section 3.4) and the first results from the *Hooked!* game (see Chapter 7). Meanwhile at a more general level, too, the promise of leveraging bigger datasets in science and the humanities, has drummed up popular interest in data-rich research across the sciences and on the intersection of disciplines, see e.g., Leroi in the New York Times [106]. There is a vast, unexplored potential in using MIR technologies to answer questions about the increasingly abundant resource that are audio collections.

Now that the notion of corpus analysis and audio corpus analysis have been outlined, a selection of prior research will be critically reviewed in the next sections, focused on linking the first, the most influential, and most recent contributions.

3.2 Review: Corpus Analysis in Music Research

We structure this review by distinguishing between three data formats: manual annotations, symbolic data, and audio data, where the category 'audio data' includes any recording of a piece of music, and 'symbolic data' roughly corresponds to machine-readable music notation (including scores, chord labels, digitized tablature and MIDI). Manual annotations refer to any set of manually assigned labels. The distinction between audio and symbolic, symbolic and encoding, may be somewhat artificial at times, but it is useful enough to guide us through the history of music corpus research. Finally, any overview like this is necessarily incomplete. The work that is included is selected to represent a variety of subdomains of music research. We explicitly exclude work with an important retrieval (e.g., search or classification) component, even if much of it has been very important to corpus analysis, e.g., classification of folk songs into tune families. We also haven't included much work from performance studies, a discipline that frequently employs music corpora as well.

3.2.1 *Corpus Analysis Based on Manual Annotations*

One pioneer who envisioned what can be considered the first 'data-driven' approach to the study of music culture, was Alan Lomax. Working as as ethnographer in the USA during the 1930's and 40's, and in England and Europe during the 1950's, Lomax made field-recordings of folk singers and musicians. Later, in the 1960's, he contributed to the foundations of ethnomusicology and performance studies with the *Cantometrics* methodology, constructed by Lomax and Victor Grauer in 1963. Cantometrics is a *coding* system in which recordings of sung performances from around the world are assigned scores on the basis of their stylistic properties and the social context in which the music is performed [112]. These annotations include how many singers participate in a performance, the melodic complexity, and how much vocal embellishments were used, among many other things. Lomax' intention was to correlate these ratings to other aspects of culture.

In one study, for instance, data from over 4000 songs out of the Cantometrics program were used to show that the prevalent performance style of a culture reflects the 'degree and kind of group integration that is appropriate and necessary to the culture's adaptive structure' [113]. Since the study of musical cultures in the world evolved, however, Lomax' perspectives on the evolutionary hierarchy of human cultures have been criticized, as well as his choices of cultural-area units and his recurring assumption that each of these cultural areas repertoire can be represented by a single song or style [170].

In a much more recent study, Savage et al. present the results of an analysis in which the aspirations of Lomax are strongly echoed. A carefully curated sample of music recordings is examined for musical universals, properties of music that can be found in each of the recordings [171]. Using comparative methods from evolutionary biology, historical relationships between related cultures are controlled for. Though no absolute universals were found among the 32 features that were tested, many *statistical universals* were found, indicating that there are indeed properties of music that apply to 'almost' the entire sample of vocal and non-vocal music. These include Lomax' and Grauer's definition of a song as a 'vocalization using discrete pitches or regular rhythmic patterns or both'. In addition to the statistical universals, eight 'universal relationships' between musical features are identified, i.e., pairs of features that consistently occur together. All these pairs are connected in a network that centers on *synchronized group performance* and *dancing*. The network also contains features related to *drumming* but, somewhat surprisingly, excludes pitch-related features. Altogether, the results are read as the first confirmation of a recent hypothesis by Fitch [52], proposing song, drumming, dance and social synchronization as the 'four core components of human musicality', and a starting point for future cross-cultural comparisons of musical features.

In popular music, Schellenberg's study on emotional cues in a sample of Top 40 records uses manual annotations of two well-established cues of emotion in music, the mode (major vs. minor) and the tempo (fast vs. slow) [173]. These annotations are used to test the hypothesis that popular music has become more sad-sounding and emotionally ambiguous over time. The dataset of 1010 songs was sampled from the Billboard Hot 100 list, taking the top 40 for each year of the second half of each decade between 1960 and 2010. The tempo was measured by expert assistants who were asked to tap along, from which the tempo was calculated using a software tool. The mode was also determined by experts and defined as the mode of the tonic triad. (Some songs, mostly of the hip hop genre, were considered to have an 'indeterminate mode'.) In the subsequent analysis, it was found that songs have evolved to make use of the minor mode more often over time, with the discrete variable mode accounting for 7.0% of the variance in recording year, and the proportion of minor songs doubling over five decades of data. With regard to tempo, it was found that both major and minor songs have decreased in tempo significantly: major-mode songs by 6.3 beats per minute (BPM) per decade, on average, and minor-mode songs by 3.7 BPM, per decade. The conclusion frames this as an increase in sad-sounding and emotionally ambiguous songs –where emotional ambiguity of a song is equated to minor-mode songs being fast and major-mode songs being slow– because of the stronger change in tempo on major-mode songs. Though the authors cite other research in support of this finding, including a study on the negativity of lyrics, the conclusions in [173] generally point to interpretations that depend very much on one's reading of the prior literature on mode and tempo as emotional cues (most dealing with instrumental music), and how well the conclusions therein carry over to music with lyrics.

3.2.2 *Corpus Analysis Based on Symbolic Data*

Alan Lomax in the 1970's was an early adopter of computers for the statistical analysis of his data. However, with the widespread availability of personal computing at the end of the twentieth century, increasingly complex statistics could be computed. The hand-coding of audio features was no longer necessary, and more advanced computational approaches could now be pursued. This was first and foremost an opportunity for those working with the first digitized scores. Hence, symbolic corpus analysis goes back much longer than its audio counterpart, which was made possibly only after audio researchers in music information retrieval disseminated developments first made in speech processing, to the music domain.

In the field of music cognition, like in Lomax's field, several authors have analyzed collections of Western and non-Western music, in search of pervasive, cross-cultural trends. Huron reviews much of his research on this topic in his book *Sweet Anticipation: Music and the Psychology of Expectation*, and in an earlier lecture series on the same topic [79, 80]. Huron reviews theories of expectation by Leonard Meyer, Eugene Narmour (Implication-Realization, I-R), and Lerhdahl and Jackendoff (the Generative Theory of Tonal Harmony, or GTTM) and contrasts them with empirical findings by Henry Watt in the 1920's, Vos and Troost in the 1980's, his student Paul von Hippel and himself, many of them based on corpus studies. Synthesizing these results, Huron proposes a set of five 'robust melodic tendencies', statistical properties of melodies that are shown to hold in various musical cultures. The five patterns include step declination (the tendency of large intervals to go up and small intervals to go down), melodic regression (the tendency of melodies to return to the median pitch) and the melodic arch.

Conklin and Witten, in the 1990's, devised a model of music expectation that is entirely based on statistical learning [204]. This *multiple viewpoints* model proposes an account of how multiple representations of a stimulus—series of note lengths, series of pitch classes, series of melodic intervals...—maintained in parallel, each contribute to an estimate of the most probable next event. Ten years later, the multiple viewpoints model was further formalized in terms of information theory by Pearce and Wiggins as the *IDyOM* model of musical expectation. The model is validated in several corpus studies, examining how well specific trends in the corpus can be explained [151].

A frequent collaborator of Pearce and Wiggins, Müllensiefen used a set of symbolic music features to analyze several interesting symbolic corpora, including melodies off the Beatles album *Revolver* [95], and a set of stimuli used in a music memory experiment [134]. The feature set draws on inspiration from natural language processing (N-gram models and latent semantic analysis in particular), and descriptors first proposed by Huron, among others.

In musicology, harmony has been a popular subject of corpus studies. Two notable analyses were done De Clercq & Temperley [41], and Burgoyne et al. [27]. In [41], De Clercq & Temperley transcribe the chords for a corpus of 99 rock songs, about 20 for every decade between 1950 and 2000, and analyze the transcriptions in terms of chord root transitions and co-occurrence as they evolved over time. In their findings, they highlight the strong (but decreasing) prominence of the IV chord and the IV-I progression. In [27], Burgoyne presents an analysis of 1379 songs out of the *Billboard* dataset of popular songs (the complete set), in terms of chord composition. The compositional analysis centers around the representation of the dataset as a hierarchical clustering of the 12 possible roots, a *balance tree*, derived from each chord roots' occurrence in each of the 1379 songs. The resulting structure is reportedly consistent with De Clerq and Temperley's analysis. The balances, i.e., the log odds ratios of the branches at each node of the tree, and their inter-correlations, are compared to decade of release and popularity. The findings include a trend towards minor tonalities, a decrease in the use of dominant chords, and a positive effect of 'non-core' roots (roots other than I, V, and IV) on popularity.

Other examples of analyses of distributions of symbolic data include the studies of pitch class and scale degree usage by Krumhansl [100]. Examples of corpus-based analyses of rhythmic patterns include Mauch's analysis of 4.8M individual bars of drum patterns sampled from around 48,000 songs, and Volk and Koops' analyses of syncopation patterns in a corpus of around 11,000 ragtime MIDI files [94, 121, 195].

Seeking to connect a musicological interest in Western classical style to perceptual theory, Rodriguez-Zivic et al. performed a statistical analysis of melodic pitch in the *Peachnote* corpus[1] in [166]. The dataset contains music from 'over 65,000 scores', automatically digitized us-

[1] http://www.peachnote.com/info.html

ing OMR. A dictionary based on pairs of melodic intervals is used to represent each 5-year period between 1730 and 1930 as a single, compact distribution. $k = 5$ factors are then identified using k-means clustering, four of which are observed to coincide with the historic periods of baroque, classical, romantic and post-romantic music, and can be read as a description of their stylistic properties. The four periods are roughly characterized by, respectively, use of the diatonic scale, repeated notes, wide harmonic intervals, and chromatic tonality.

A number of the above contributions have resulted in toolboxes dedicated to the analysis of symbolic music corpora. Huron worked with the *Humdrum* toolkit and its associated *Kern* representation of scores, both of which are still used and supported.[2] Pearce made a Lisp implementation of the Idyom model available on the Soundsoftware repository[3] and Müllensiefen's FANTASTIC toolbox, written in R, is also available on line.[4] The *Peachnote* corpus can be accessed through an API at `www.peachnote.com`.

3.2.3 *Corpus Analysis Based on Audio Data*

Much of the existing work involving audio corpus analysis has focused on popular music and non-Western music—two big clusters of music for which scores or other symbolic representations are not often a musical work's most authoritative form. (In both of these groups of styles, music notation is typically either unavailable, or only available because a recording has been transcribed.)

In an example of non-Western music analysis, Moelants et al. describe in [130] a procedure of the automatic analysis of automatically extracted pitch histograms. The procedure is applied to a collection of historic African music recordings, and show evidence for Western influence in the use of African tone scales. Also using tone scale analysis, Panteli and Purwins compare theory and practice of scale intonation

[2] `http://www.musiccog.ohio-state.edu/Humdrum/`,
`http://github.com/humdrum-tools/humdrum-tools`

[3] `http://code.soundsoftware.ac.uk/projects/idyom-project`

[4] `http://www.doc.gold.ac.uk/isms/m4s/`

in contemporary (liturgical) Byzantine chant. Analyzing 94 recordings of performances by 4 singers in terms of the tuning and prominence of scale degrees in 8 different modes, they find that smaller scale degree steps tend to be increased, while large gaps are diminished [146].

In an example of popular music analysis, Deruty and Tardieu test a number of hypotheses about the evolution of dynamics in popular music [42]. The hypotheses are formulations of a recurring intuition among producers and consumers, hypothesizing a 'loudness war', a speculative trend in which the loudness of pop songs has gradually increased, in a race between producers of new releases to stand out on the radio. In the study, 2400 recordings released between 1967 and 2011, sampled from a list of critically-acclaimed popular music albums, are analyzed in terms of their energy (root mean square energy or RMS), loudness, loudness range (measuring macro-dynamics) and peak-to-RMS factor (measuring micro-dynamics). They conclude that the energy and loudness have indeed increased, and that micro-dynamics have indeed decreased. Macro-dynamics, however, were not found to evolve significantly.

In the domain of music cognition (specifically, embodied music cognition) one recent study uses corpus analysis to identify the acoustic properties of music that affect walking speed (in m/s) [104]. Leman et al. had 18 participants walk freely to a precompiled playlist of 52 songs, all with a fixed tempo of 130 BPM, and measured their walking speed using wireless accelerometers. The acoustic correlates of walking speed were assessed in a two-stage statistical analysis involving feature selection from a candidate set of 190 audio descriptors, and a model selection stage, using the 10 best features, in which the best fitting linear model was found via (nested) cross-validation. The best performing model involves 4 features and is found to explain 60% of the variance after a second ('outer') cross validation. The features are said to capture variations in pitch and loudness patterns at periods of three, four and six beats.

A few contributions in the domain of performance studies have also involved the analysis of a dedicated audio corpus. In [97], for example, Kosta et al. compare loudness dynamics across 239 piano performances of a selection of 5 Chopin mazurkas. They find that pairs of dynamic markings in the score don't always correspond to an expected change in decibel levels, and expose further non-trivial dependencies between loudness, note density and dynamics.

Finally, two relatively recent studies have focused on the topic of popular music evolution. In [175], pitch, timbre and loudness features are analyzed for a sample of songs, with dates, from the Million Song Dataset (MSD). In [122], songs sampled from the Billboard charts of US popular music are analyzed using techniques from text mining and bio-informatics. Given their topic—popular music—and their relevance to other results presented as part of this thesis, we will review these two studies as a case study in section 3.4.

A note on Automatic Transcriptions

While the music information retrieval community has made substantial progress in its efforts to improve the transcription of audio to symbolic data, considerable hurdles remain [179]. To our knowledge, no corpus analysis studies have yet been proposed that rely on the complete polyphonic transcription of an audio corpus. And understandably so, since the assumptions of the transcription model would have a considerable impact on the quality of the data, and, worse, most certainly introduce biases in the data itself.

One approach that illustrates the inherent risks in the analysis of transcribed corpora, is Barthet et al.'s study on chord data mining [8]. In an analysis of one million automatically transcribed chord sequences, Barthet et al. acknowledge the drawback that is the chord recognition system's error rate. However, they 'assume that the most frequent patterns emerging from the analysis should be robust to noise'. Even if this is the case, no mention of potential structural biases is made. Many of the state-of-the-art chord transcription systems rely on a form of priors that govern in which order the system expects to see chords. For such systems, a simple count of root transitions would already return biased results.

Barthet's chord extraction was performed using Chordino, which only involves frame-by-frame matching of chroma features to a dictionary of chord profiles, followed by 'heuristic chord change smoothing' [120].[5] The system lacks a language model, so it puts fewer restrictions on its output. Nevertheless, it has been trained on or optimized for a particular collection of music, and the patterns present and not present in that particular dataset (e.g., the very popular Beatles dataset) will be reflected in its output. A similar argument applies to other kinds of transcription (e.g. melodic transcription), as well as for Rodriguez-Zivic's study described above, as it relies on OMR for the transcription from images of scanned scores.

Most existing work on audio corpus analysis has therefore focused on the analysis of audio features rather than automatically transcribed melodies or chords.

3.3 methodological reflections

Studying music through the analysis of a collection comes with its own particular set of challenges. What are important issues in audio corpus analysis that are not typical issues in MIR, and how can they be addressed? This section seeks to answer that question by looking at the methods reviewed above, and by reviewing existing commentaries on the use of music information retrieval technologies in interdisciplinary scientific research. Note that several of the points below apply to corpus analysis in general as much as they apply to audio.

To structure the discussion, we start from the observation that a majority of studies follow variations of the same procedure, involving the choice of a research question or hypothesis, a dataset, a feature set and an analysis method. Each of these steps will now be discussed.

[5] http://isophonics.net/nnls-chroma

3.3.1 *Research Questions and Hypotheses*

Outside of MIR, the prevailing scientific practice of addressing a research question using data involves hypothesis testing. Generally, a prior intuition or theory is followed so as to arrive at a prediction or *hypothesis*, describing a certain trend. A good theory leads to a hypothesis that can be falsified or *rejected* in a *statistical test*. A statistical test looks at data and decides whether the data contradicts the hypothesis, and the hypothesis must be rejected, or not. If the hypothesis is not rejected, it is not considered proven, rather, there is no evidence that it is wrong. Tests are performed at a certain significance level α, specifying the probability one allows for a trend being found due to chance, i.e., due to a coincidence in the sample. Lowering α decreases the chance of false discoveries (type I error), but increases the chance of rejecting an existing effect (type II error).

Much of this carries over to music research, and has been applied in countless studies, but it bears repeating how vastly different the procedure is from prediction-evaluation paradigm seen in classic MIR, where hypotheses are rarely stated explicitly. It is also important to look at some challenges that come with hypothesis-based research that are not often acknowledged.

In [81], Huron points to the importance of taking care when choosing hypotheses in music research. He stresses that, historically, hypotheses were typically formed before any data could be acquired. With the arrival of large datasets, however, it is tempting to formulate hypotheses based on an initial exploration of ones dataset, causing the data to be used twice. This increases the chance of confirming, in subsequent tests, a trend that was spurious to begin with, an artifact due to sampling that wouldn't be present if new data were collected. Huron therefore advices against such exploratory activities, calling for the treatment of datasets as finite resources that lose value every time a correlation is computed. If needed, exploratory studies should be done with idiosyncratic rather than representative data.

Several of the above corpus studies do not follow a hypothesis-driven approach. They try to answer questions like: 'What are the salient patterns in this particular genre?' (Mauch, Koops) and 'How does the use of patterns evolve over time?' (Burgoyne, Volk). These are common musicological questions that cannot be formulated in terms of a single hypothesis. Some studies therefore explicitly center on what could be considered exploratory analysis.

Burgoyne, in [21], presents the results of an experiment in which a Bayesian network or probabilistic graphical model (PGM, see also section 4.4) is learned from a set of variables relating to the harmony and chart position of the songs in the Billboard dataset. This approach aims to expose pairwise correlations between variables that are significant after the effect of other variables is removed, without a prior hypothesis as to which of them are expected or why. Leman et al., in [104], use cross-validation to make the most of the data they have collected, as a time consuming experiment like theirs cannot easily be repeated to collect more data.

Because these studies are not strictly followed by a confirmation on new, independently collected data, their approach is at odds with Huron's advice. Are they therefore invalid? Not necessarily, many of these are respected, peer-reviewed results. Exploratory analysis, with proper use of statistics, can be useful and valid if it is accounted for using appropriate significance levels.

This suggests a spectrum of methods practiced, on which Huron's position represents a rather conservative perspective, which allows, when taken to its extreme, only for yes-or-no research questions, and not for questions of the what/when/where kind (e.g., what makes a song popular, or, when did ragtime syncopation patterns change most), or at least not with a single dataset. Analysis methods will be discussed later in this section, but we can conclude for now that prior hypotheses are a valuable, but not the only option, and that several alternative analysis methods have been developed enough for more open-ended questions to be asked and answered, with appropriate precaution, in a statistically rigorous way.

3.3.2 *Choice of Data in Corpus Analysis*

The intricacies of proper dataset curation make for a PhD topic of their own, see e.g., Burgoyne's account in [26] and Smith in [182]. Though referring to a study as 'corpus analysis' may make it seem as if the corpus is a given, that should be analyzed to answer a particular question. Ideally, of course, the choice of dataset will *follow* the research question: the set of musical works is chosen that allows the question to be addressed most reliably. Compiling a dataset generally requires careful demarcation of the kind of music the research question pertains to, and careful sampling to adequately represent this population.

In the context of corpus analysis, it should be stressed that many of the datasets that are used in music computing have *not* been compiled to be representative of a particular music, but to serve as a test bed for various MIR technologies. The content varies accordingly. The Beatles dataset, often used for chord extraction evaluation, contains a wealth of rare and challenging harmonies, but draws on the work of just a single group of artists. The later Billboard dataset is a much more representative sample of popular music, as it is sampled from the Billboard Hot 100 chart. For example, it includes duplicates to reflect the varying number of weeks songs stayed in the charts [26]. But it was also constructed with large-scale harmonic analysis in mind. Furthermore, it contains only music released up to 1991, when Billboard's own measurement strategies changed. As a result, new genres such as Hip-Hop, that cannot be characterized in terms of chords and modes as easily as earlier genres, are missing from the corpus. The result is a potential bias towards songs with harmonies that can be parsed in terms of traditional music theory. Finally, the Million Song Dataset (MSD) was compiled using a variety of criteria: by downloading the music of the 100 artists for each of The Echo Nest's 200 most-used tags, plus any artists reached by a random walk starting from the most familiar ones, according to The Echo Nest.[6] It is explicitly biased towards challenging rather than representative musical material (e.g., by including music relating to an intentionally broad range of

[6] see `http://labrosa.ee.columbia.edu/millionsong/faq`

tags) [14]. It follows that current MIR datasets aren't necessarily suitable for corpus analysis as defined here.

In [81], Huron also notes how large datasets, theoretically, allow for trends to be found with low error rates, both of type I and type II. However, this also makes statistical significance of spurious trends due to a biased sample more likely. Therefore, it is always good practice to validate findings derived from a corpus with new, independently gathered data. However, of all suggested practices, this is one of the most difficult and potentially expensive ones. And, as almost all music data are per definition historic, there is often a fundamental upper limit to how much new data can be acquired.

3.3.3 *Reflections on Audio Features*

Just like the datasets used in music information retrieval aren't necessarily appropriate for corpus analysis, audio features can be inappropriate too. One prominent voice of reflection in MIR, Sturm has argued that a lot of studies in the audio-based music information retrieval field have focused excessively on flawed evaluation metrics, resulting in vast over-estimations of the modeling power of many widely used technologies, including audio features.

In [184], Sturm inspects a large number of studies that have all used the GTZAN genre classification dataset, a dataset for genre recognition training and evaluation compiled by Tzanetakis in 2002 [188], as well as the dataset itself, and observed that there is a hard ceiling to the performance numbers that can be realistically obtained. The ceiling is due to mistaken tags, repeated entries, and other issues. Despite this ceiling, several systems report near-100% accuracies. The study then shows how some of these impossible performance numbers can be attributed to errors in the evaluation, while others cannot be replicated at all.

However, rather than putting the blame with the authors for the quality of their contributions, Sturm examines the evaluation pipeline itself, to conclude that the evaluation of classification systems just based on their accuracies, is flawed. In essence, much of the apparent progress as reported using the above dataset, should be seen as fitting systems to the dataset rather than the task, even when cross-validation is used to avoid overfitting.

What does this imply regarding the use of audio features? As a result of the above practices, Sturm suggests, the current state-of-the-art systems in genre recognition do not listen to the music as much as they listen to a set of largely irrelevant factors that turn out to be proxies for the genre labels as they have been assigned in the GTZAN dataset. In short, features that have been shown to do well in predictive, classification-based MIR, aren't necessarily meaningful descriptions of the music. Or, again in other words, it is not because a feature works in a MIR system, that it is meaningful.

Sturm's concern is echoed in some of the arguments made by Aucouturier and Bigand in [6]. Aucouturier and Bigand, an MIR researcher and a cognitive scientist, examined some of the possible reasons for the MIR communities' limited success in gaining interest from music cognition and neuroscience. As one of seven problems they have identified, the authors stress that many of the audio features used by MIR may seem, at first, to have some cognitive of perceptual basis. Yet often enough, they do not. Whereas the use of the spectral centroid as a timbre descriptor can be justified using evidence from psycho-acoustics, spectral skewness, for example, is mostly a convenient extension of the spectral centroid (see Section 2.1.2), rather than a realistic perceptual attribute of timbre. Likewise, MFCC features may build on some perception-inspired manipulations of an acoustic signal, like the use of the Mel scale and the use of a logarithm for compression. But the discrete cosine transform, that is used next in the computational pipeline, is simply unlikely to have a neural analog.

Similarly, Haas and others have noted that a worrisome number of data-oriented MIR systems completely neglect time [29, 61]. They show that the so-called bag-of-frames approach to music description (audio, especially) is very widespread. In this approach, features are computed over short frames, and frames are pooled by taking the mean or variance, or some other summary statistic that is invariant to order. The efforts that have been made to re-introduce time in music description have largely been focused on symbolic data (see examples in [61]), leaving audio features behind. In general, Haas points to a variety of opportunities in incorporating more musical knowledge and music cognition in music description.

Naturally, this is not to say that everyone has been doing everything wrong. It is rarely the intention of MIR researchers to develop realistic models of neural mechanisms. In a typical goal-oriented setting, features will be used if they help improve the precision of an algorithm, regardless of whether they have a verified psychological or neural underpinning. Aucouturier and Bigand are right to point out that this procedure is fundamentally at odds with scientific practice in the natural sciences, where variables aren't added to a model because they improve its performance, but because they correspond to a theory or hypothesis that is being tested. But, as most MIR researcher would attest, 'science' may just often not be the goal [6].

And then there is another appropriate nuance that isn't often discussed: music cognition and neuroscience are themselves at times divided, on topics such the learned and culturally mediated nature of mental representations [71], and the neural basis of apparent cognitive 'modularities' [158]. Yet, the above illustrates why it is important to exert caution whenever a feature that was originally developed for some MIR application is used in the context of scientific music research, even if it is widely used.

To conclude, existing commentators point to a tendency among researchers to choose convenience and prevalence over relevancy and cognitive or perceptual validity of features. While efforts in feature design have resulted in an impressive canon of powerful audio features, most are *a priori* uninformative, and therefore of little use in interdisciplinary research. There is a lot of room for the perceptual validation of existing features and the design of novel cognition-inspired features that better align with cues that are known to be important in human music perception and cognition.

3.3.4 *Reflections on Analysis Methods*

Similar arguments can be heard in discussions on the analysis and learning algorithms that integrate features to make predictions. Aucouturier and Bigand, in the second out of their seven problems, criticize the algorithms used in MIR with a very similar argument as they first made about features: algorithms are presented as if they reflect cognitive mechanisms, but they do not. Even if all the features in a model were accurate measurements of plausible perceptual or cognitive correlates of a musical stimulus, most of the commonly used statistical models wouldn't reveal much about how these attributes are combined into more complex judgements on e.g., the perceived valence of the emotions conveyed by the music (as in the example given in the article). Even if a feature is only incorporated into an algorithm if its individual predicting power is tested and validated, it may be unclear "what sub-part of a problem that feature is really addressing", especially when modeling a highly cognitive construct like genre or emotion [6].

To Huron, statistical practices are a recurring concern. Along with his recommendations on choice of hypotheses and data (reviewed above), he reviews the statistical caveats that come with the use of big datasets in musicology [81]. As it has become easier than ever before to undertake a large number of experiments, thresholds of significance should adapt: if, in a typical study, well over 20 relationships have been tested for significance, some will end up being spurious, and a significance level well below the traditional 0.05 should be considered. The Bonferroni correction for multiple tests, an adjustment of the significance level α based on the number of tests, is traditionally used to address this issue:

$$\alpha_B = 1 - (1 - \alpha)^{1/n} \approx \alpha / n \tag{21}$$

with n the number of tests and α the overall significance level of the study (e.g., 0.05). It is an effective measure against overfitting to a sample. Usually, however, n only counts formal tests. When exploratory processing of a dataset is involved, α_B should reflect the substantial amount of visual exploration and eyeballing of potentially interesting relationships that is often done prior to any formal testing. And as others have argued in the debate on the use and misuse of p-values in science, reporting effect sizes also helps to communicate results convincingly: as datasets get bigger, it is increasingly easy to find significant, but small effects [81].

In standard machine learning-style analysis, significance is even more difficult to asses. An SVM classifier may tell you if a feature is helpful or not, but it doesn't reliably quantify the significance of that feature's contribution, let alone its effect size. It is often even unclear in what 'direction' a feature contributes—suppose a classifier predicts a strong influence of tempo on whether or not a song is perceived as happy, it often won't tell you what range of tempos make a song more happy, especially if the classifier is a 'distributed' model like those based on boosting, forests or neural networks.

Another statistical modeling issue that is not typically deemed relevant in machine learning is the role of variable intercorrelations and confounding effects. A confounding effect occurs when some trend is attributed to one variable while it is in reality due to another (observed or unobserved) variable. When dealing with a set of correlated features, it may turn out that some of the features that correlate with a dependent variable of interest, contribute little explanation in the presence of other features; they are, as is said, 'explained away'.

An ideal statistical analysis that is focused on not just correlations, but on 'effects', allows to control for obvious confounding correlations, and acknowledges the possible effect of correlations that couldn't be controlled for. This kind of 'causal modeling' is far from trivial. The common perspective is that it requires *interventions* that allow some variable to be willfully adjusted, as in a randomized controlled trial. When dealing with historic data, such as any music collection, intervention is not typically an option, and the feasibility of causal modeling can be disputed.[7] Others have argued that, under certain restrictions, causal relationships may still be obtained. Probabilistic graph-

[7] See, e.g., Sturm's thoughts on the subject `http://highnoongmt.wordpress.com/2015/07/30/home-location-and-causal-modeling/`

ical models, for example, model conditional relationships between variables, allowing for a certain amount of insight into the confounding effects between the observed variables [93]. Most other statistical methods, however, don't model the effects of individual variables, and therefore don't account for interactions in the causal sense.

Studies with many variables, hypothesis- or discovery driven strategies alike, face an additional statistical challenge: the amount of data required to fit a reliable model may scale unfavorably with the number of variables. This problem is sometimes referred to as the 'curse of dimensionality'. It denotes a wide range of issues that arise because data get sparse quickly as the dimensionality of the feature space goes up. Mathematically: distances between uniformly distributed data points in a high number of independent dimensions tend to lie mostly on a relatively thin shell around any given reference data point, and data points tend to have a very similar degree of dissimilarity to each other [66].[8]

This forms a recurring challenge in statistics and statistical learning. Many statistical models involve $O(m^2)$ parameters. An arbitrarily structured multivariate normal distribution in m dimensions, for example, requires $m^2 + m$ parameters to be fit. Any fit with less than m m-dimensional data points will therefore fail, as there are more degrees of freedom in the parameters than in the data points. A good fit will take several times that number. This dependency will show up in any model that acknowledges correlations, i.e., any model that doesn't treat its dimensions as completely independent—a very strong assumption, most of the time.

The problem gets worse for models that account not just for binary interactions between features, but between any combination of variables, e.g. learned graphical models. Because the number of dif-

[8] The technical argument is given by Hastie as follows: for inputs uniformly distributed in a m-dimensional unit hypercube, "suppose we send out a hyper-cubical neighborhood about a target point to capture a fraction r of the other observations. Since this corresponds to a fraction r of the unit volume, the expected edge length will be $e(r) = r^{\frac{1}{m}}$. In ten dimensions: $e(0.01) = 0.63$ and $e(0.10) = 0.80$, when the entire range for each input is only 1. So, to capture 1% or 10% of the data to form a local average, we must cover 63% or 80% of the range of each input variable."

ferent graph structures grows super-exponentially with the number of nodes, many tests are typically required to find the best candidate, even if heuristics are used [16]. All of these issues, while less cumbersome in machine learning and prediction, put unfortunate limitations on the complexity of statistical analyses of high-dimensional data.

While this last point reads like an argument against large-feature-set analyses at large, there are exceptions and alternatives. For example, if the number of data points is low or the number of dimensions very high, heuristics and regularization may be used. In the specialized literature, techniques for the estimation of structures in sparse, correlated datasets are emerging, e.g., regularized covariance matrix estimation [15]. If the number of data points is sufficient for a reliable estimate of the covariance matrix, dimensionality reduction (e.g., PCA) may be applied to facilitate further model selection, ideally accompanied by steps taken to avoid compromising interpretability in the process.

Overall, choosing a good modeling approach seems to elicit the same problems as choosing the right audio features: models that have been shown to work fall short on requirements that are inessential for information retrieval but crucial in scientific modeling, including simplicity (in terms of the number of parameters), accounting for correlating variables, and accounting for multiple tests. Furthermore, the analogy of analysis methods with perceptual and cognitive modularities is often flawed. When, finally, a statistical model seems appropriate, there may not be enough data for the model to be fit, or for any true effects to surface after significance levels are adjusted to reflect all testing involved in fitting the model.

3.4 case study: the evolution of popular music

Having reviewed a range of issues that come up in the corpus analysis of audio data, we present a case study that illustrates how some of these issues have been addressed in practice, and others have not.

The case study focuses on two publications on the topic of popular music evolution, by Serrà et al. in 2012, and Mauch et al. in

2015 [122, 175]. The common question they address is, roughly, 'has Western popular music, over the past decades, become more or less diverse'? The studies are of particular interest because of their focus on popular music, but also because they arrive at partially contradicting conclusions on the supposed decline of diversity in popular music. Both studies are also the work of researchers with a considerable authority on the subject of audio analysis, and have been given wide attention in the popular press.[9]

First, each study's methods and results will be summarized and compared. A discussion section will then discuss the differences between the studies in terms of data, features and models. Along the way, some additional pitfalls that haven't been brought up in the methodology literature will be identified. Finally, the most important differences and pitfalls will be summarized at the end. A side-by-side comparison of the studies' analysis pipelines is given by the diagram in figure 16.

3.4.1 *Serrà, 2012*

In the first study, by Serrá et al., pitch, timbre and loudness features are analyzed, to answer a number of questions that includes the one above [175]. The dataset is a sample of 464,411 songs from the MSD, all released between 1955 and 2010. The features correspond to precomputed pitch, timbre and loudness features as provided by The Echo Nest[10], computed over 10 million consecutive frames for every year of data, sampling from a five-year window. For each feature, a codeword dictionary is then extracted, yielding a vocabulary of pitch, timbre and loudness codewords for each year.[11] The studies hypothe-

[9] E.g., http://graphics.latimes.com/music-evolution-hip-hop-rap/, http://www.theguardian.com/music/2012/jul/27/pop-music-sounds-same-survey-reveals

[10] the.echonest.com

[11] Codewords are simplified, discrete representations of multidimensional feature vectors. The mapping of feature vectors to codewords is often found by applying clustering to a dataset, after which each data point is mapped to the closest cluster center. Here, a simpler heuristic was used to discretize the features [175].

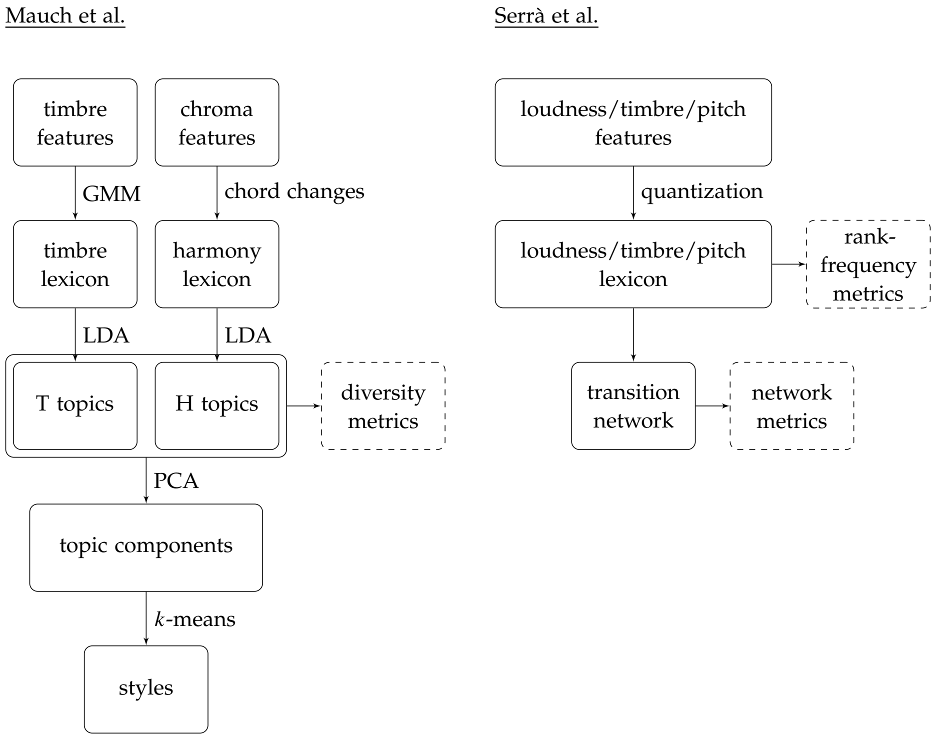

Figure 16.: Diagram comparing Mauch's and Serrà's analysis approaches.

sis questions are addressed through a statistical analysis of the distribution and transition network of these codewords.

In a short analysis prior to computing loudness codewords, the empirical mean of loudness values is found to have increased, from -22dB_{FS} to -13dB_{FS}, or about 0.13dB_{FS} per year, while dynamic range hasn't evolved significantly—findings that are loosely consistent with Deruty et al.'s later results described in section 3.2. In terms of loudness codewords and transitions, the network's topology is maintained throughout the decades.

When timbre is analyzed, the codewords are modeled using a rank-frequency distribution based on Zipf's law. Zipf's law states that an event's frequency of occurrence can be modeled as a logarithmic function of its rank when all events are sorted by frequency. The rank-frequency distribution that is used is parametrized by the exponent parameter β. For timbre codewords, it is found that β decreases over time since 1965, indicating a homogenization of the timbre palette. Like loudness codewords, no changes in the topology of timbre codeword transitions could be found.

In the pitch domain, no change in β is found. However, some trends have been discerned when looking at the pitch transition network. The most obvious indicator of diversity of transitions, the median degree of the network k, is unchanged. However, the clustering factor C, assortativity Γ and the network's average shortest path l are found to change significantly, in a way that, together, constitutes a decrease in 'small-worldness' of the network, showing a restriction of the possible transitions, and thus a decrease in the variety of observed transitions [175].

To sum up, Serrà et al. find a progressive increase in the predictability of pitch use, a tendency towards mainstream sonorities in the timbre domain, and an increase in loudness of productions.

3.4.2 *Mauch, 2015*

Mauch et al. follow a very different approach, leading to a rather different conclusion, in another quantative analysis of popular music evolution [122].

Like Serrà, Mauch et al. analyze audio data for a large corpus of popular music sampled from the last 50 years. Specifically, their corpus extends from 1960 to 2010 in 17,094 30-*s* segments from the same number of songs. Instead of Echo Nest features, freely available tools were used to compute chroma and timbre descriptors from these segments. The study also employs a quantisation step to convert each of these segments to a sequence of *words* out of a newly constructed lexicon, or in this case, two: one for timbre and one for harmony. The timbre lexicon is obtained using unsupervised clustering based on Gaussian mixture models (GMM), of the 14-dimensional audio feature space (12 MFCC coefficients, one delta-MFCC coefficient, and the zero-crossing-rate). In GMM, data points (here: frames) are assigned to an optimized number of Gaussian-shaped clusters. The optimal number number of clusters is found at 35. The harmonic lexicon consists of 192 possible intervals between the most common chord types.[12].

However, the text data approach is taken further than it is in [175]. With each 30-s frame of audio converted to a single word, *topics* are extracted from the data, using a topic modeling technique from text mining called latent Dirichlet allocation (LDA). This hierarchical model regards documents as distributions over a set of topics, which are themselves distributions over the lexicon. 16 topics are found, 8 for timbre and 8 for harmony. Following an analogy with evolutionary biology, these topics can be regarded as *traits* or *character expressions* associated with the 'genome' that is, in this study, the string of words of which a song is composed.

The documents' distributions over these topics are used to compute four measures of genetic diversity, borrowed from bio-informatics. The diversity measures show substantial fluctuations over time, most

[12] Four modes (major, minor, major 7, minor 7) × four modes (for the second chord) × twelve root intervals = 192, plus one label for ambiguous harmonies [122]

notably a drop in the early 1980's, followed by an increase to a maximum in the early 2000's. Interestingly, however, no evidence is found for the progressive homogenization of the music in the charts as posed by [175], neither in the timbre domain, nor in the harmony domain.

In a second and third part of the study, one more layer of abstraction is added when the topics distributions are grouped into *styles*, similar to populations in genetics. First, the topic space is further reduced to 14 dimensions, using PCA with standardization of the components. The styles are then found using k-means clustering on the timbre and harmony topics. The best fitting number of styles is found to be 13. At each stage, the feature spaces employed in this paper, be it the lexicon, the topic space, or the musical styles, are checked for interpretability. This is achieved using a combination of expert annotations on a representative mixture of short listening examples, interpretation of the chord labels by the authors, and enrichment of the styles with tags from *Last.fm*.[13]

The second question to be addressed is: when did popular music change the most? To address this, the 14-dimensional topic space is first used to compute a distance matrix over all analyzed years, and a novelty curve can be computed. A novelty function, as introduced in Section 2.2.2, tracks discontinuities in the time series. It is found that three years brought significant change to the topic structure of popular music: 1991, 1964, and 1983, of which the one in 1991 is the biggest. Using a similar analysis of the 13 extracted 'styles', it is found that these years coincide with the moments that soul and rock took over from doo-wap (in 1964), the year that soft-rock, country, soul and R&B made place for new wave, disco and hard rock (in 1983), and the rise of hip hop-related genres to the mainstream in 1991.

3.4.3 *Discussion*

The two studies above have both received wide exposure in the specialized and popular press, despite conflicting conclusions on the supposed decline of diversity in popular music. Where does this contra-

[13] `www.last.fm`

diction arise? We discuss the most important differences between the studies in light of the concerns raised in the previous section (section 3.3). We distinguish between choices of research questions, data, audio features and analysis methods, and focus on the common part of each study's research question: is there evidence that popular music has become more homogeneous, more predictable or less varied over time?

Data

The first critical difference is the music sample of choice. Mauch et al. aim to use the complete Billboard Hot 100 as their sample and manage to include about 86% of the complete list. Serrà et al. choose to construct a sample that includes a large portion of the Million Song Dataset.

Both datasets have their advantages and drawbacks. The MSD is much larger, but sampled rather arbitrarily (see earlier). Therefore it is neither controlled for popularity, nor a complete picture. Neither Serrà or Mauch discuss the option of controlling for popularity. A popular music corpus, ideally, gives more weight to songs that were listened to more often. Mauch do this to a primitive extent, by sampling only songs from the Billboard Hot 100, but much more could be done: the sampling procedure used to compile the Billboard dataset by Burgoyne, for example, allowed for songs to be included several times if they stayed in the charts longer [26].

But the Billboard Hot 100 also has other flaws. For instance, it is known that in 1991, the method of measuring popularity as a function of radio play and sales, was automated, and as a result, drastically changed [26]. This calls for some caution when interpreting the claim that popular music's biggest moment of change came in 1991—one should at least consider the possibility that this effect is in fact, sample noise, an effect due to the way the Billboard Hot 100 list was compiled. The paper, making no mention of the measurement procedures of the Billboard organisation, does not address that possibility. It goes to show that, as discussed above, a consistent sampling strategy is crucial in corpus-based studies.

Thus, neither of the studies seem to have properly considered the issues we formulated regarding datasets for corpus analysis, though Mauch et al. make a somewhat stronger case by not choosing the charts over the Million Song Data. The different approaches and outcomes are reminiscent of several anecdotes that are used to illustrate a common 'big data' fallacy, in which a sample is deemed reliable because it comprises almost all of the population, and its biases are dismissed as if the dataset's size could somehow make up for it.[14]

Features

The studies also differ in their choice of features and statistical measurements on descriptors. While Serrà et al. focus on networks of transitions between code words, Mauch et al. group them into topics and look at the evolution of those topics. While it may seem that Mauch et al. disregard the time component that is very often overlooked, but somewhat included in Serrà's analysis, it must be noted that, in the latter, changes in the transition network's topology only drive the homogenization effect in the pitch domain, not in timbre or loudness. Furthermore, Mauch et al. effectively do include some time information in their representation of pitch, as the harmony features used to extract the harmonic topics are based on chord transitions rather than just the chords themselves.

Other description-related differences remain. The descriptors used by Mauch do not include melody, whereas the features used by Serrà arguably could, and Mauch narrows harmony down to a space defined in terms of chords (and specifically: triads), which, as other have noted, are perhaps not appropriate for the description of recent 'urban' music (hip hop and related genres) [26, 173]. On the other hand, Serrà et al.'s network representation only considers binary counts: whether

[14] see, e.g., the often recounted case of Literary Digest's predictions for the 1936 US presidential elections. Their poll, one of the largest in history at the time, failed in the end as the result of a bias in both sample and response [183].

or not a code word or transition appears, regardless of its proportion in the sample.

Analysis Methods

A third set of differences and potential issues, ultimately, appear in the analysis methods. The studies use a different set of diversity measures: network statistics (Serrà) vs. bio-informatics measures (Mauch), of which only Mauch's have been validated in other studies with similar research questions.

Serrà's method also runs into an issue related to confounding variables. In his study, both loudness and timbre are reported to homogenize over time. One obvious question that is not addressed is: does increased loudness not affect the range of possibilities left in the timbre palette? If a substantial amount of the timbre-related trend can be explained by an increase in loudness, the results of the paper would look quite different. The conclusions do not acknowledge this. Mauch et al. don't run into this issue, because no trends are found in either the timbre-related set of topics, or the harmony-related set.

Finally, Mauch et al. look for trends only after transforming the code word representation of their data set to the strongly dimensionally-reduced abstraction that are the topics. The study could have included some measurements of diversity on the codeword representation itself, to see if a homogenization can be observed earlier in the analysis.

3.4.4 *Conclusion*

The above analysis brings up a range of substantial differences between the two studies, of which choices in data and analysis methods seem the most salient. Serrà et al. primarily expose their approach to criticism by working with an uncontrolled sample, and by not controlling for loudness in their analysis of the evolution of timbre (and vice versa). Mauch et al. work with a sample that is more convincing, but similarly lacking some control over what makes exactly makes it representative, due to the procedures by which the Billboard charts are compiled. It is impossible to know which of these differences contribute more to the discrepancy in conclusions without running experiments to test particular variations of their methods. But together, such differences could explain some or all of the disagreement in the results. Conclusions on which approaches should have been followed instead, if any, won't be made here. Section 3.5, however, will list some general recommendations distilled from the observations made above.

On a positive note, both studies deal very thoroughly with a host of other issues: they start from a clear hypothesis, thoroughly motivate their analyses, and refrain from making claims on the cause of the observed effects. In addition, both studies acknowledge potential limitations of their conclusions, e.g. as exemplified by Mauch et al.'s comment that their conclusion is limited to the features they have studied, and that their measures only capture a fraction of the actual complexity of the music in their dataset.

A Note on Interpretability

We close the case study with an observation about the reception of each studies' results, in the general press and among researchers in related fields. Between the two studies, there is large gap in the amount of effort spent on interpretation of the audio features and their abstractions used in the models. Serrà et al. use Echo Nest features, which is a proprietary technology for which the mathematical specifications haven't been published. The abstractions used in the model aren't qualified in terms of musical domain language, but in terms of network statistics. Meanwhile, Mauch et al. use openly available features and collect human annotations for each of the topics in their model, and social tags for each of the styles.

In following the broader reception of both articles and in discussions among colleagues, this discrepancy became especially apparent. Results are seen as uninformative if they are the result of a method that is convoluted or opaque. In contrast, the importance of interpretation of audio features is not widely discussed in the methodology literature reviewed above. This suggests that interpretable audio features and analysis methods are perhaps more important than authors in the field acknowledge.

Problematically, however, *interpretability* is not easily defined. What constitutes the interpretability of, for example, an audio feature? If we were to define it, we could say it is a feature's property of having an agreed-upon interpretation, where an *interpretation* is an unbiased and sufficiently detailed mapping from the signal or computational domain to natural language or domain language (perceptual, cognitive or music-theoretic). In other words, features that can only be interpreted in terms of computations on the audio signal, carry no information outside of the computational domain, whereas a properly informative feature allows to translate a mathematical trend or pattern into natural language or domain language information.

As a definition, this is rather subtle. MFCC features, as a whole, have an agreed-upon, empirically validated correlation with some subspace of Western musical timbres [186]. Yet, individual MFCC coefficients have no particular interpretation: it doesn't mean much for one or more coefficients to be high or low to most people (except for, perhaps, the first one, a correlate of energy). Moreover, whether or not a feature or analysis method is interpretable is inherently subjective, depending very much on the background knowledge of the audience the results of a study are reported to.

In short, feature interpretability is a quality that is generally considered helpful and important. Yet it cannot be prescribed in exact, universal terms, in part because it is very difficult to define, in part because it is highly dependent on context and audience. It is mostly useful as a predictor for the degree to which a research result may convince scholars outside its immediate domain. We adopt a very pragmatic stance: researchers should adopt the methods that best allow them to communicate with the audience they wish to persuade. Mauch's study, therefore, does have the added benefit, over Serrà's, of offering feature interpretations at every step of the analysis, thus making a stronger case for its results despite the high degree of abstraction of its representation.

3.5 SUMMARY AND DESIDERATA

From the review of corpus analysis research in section 3.2, the methodological reflections outlined in section 3.3, and from the above case study, we distill a number of desiderata. These are desired properties of audio features and analysis techniques, for the context of audio corpus analysis.

3.5.1 *Research Questions and Hypotheses*

Preceding any analysis is the choice of research question. The literature review above roughly leaves room for three options:

1. - an external hypothesis, established before any data are collected, or
   - a hypothesis based on an idiosyncratic sub-sample of the data, or
   - no hypothesis, but an explicit strategy for exploratory or discovery-based analysis

with significance levels set accordingly.

3.5.2 *Data*

When choosing a corpus, it is crucial to aim for

2. a representative dataset, carefully sampled from a clearly defined population.

Many existing MIR datasets haven't been compiled to represent anything in particular, and should be used with care, or not at all. When compiling one's own dataset, representative sampling should prioritized over dataset size.

3.5.3 *Audio Features*

To guide the choice of audio features, we put three criteria forward, taken from several of the sections in this chapter.

3. robust features: features can be reliably computed for the entire corpus.

Features that are still very difficult to accurately compute from audio include the precise onsets of a vocal melody or any other source in a complex polyphonic mix, as well as any features based on this information (e.g. inter-onset intervals), and any features that rely on complete transcription of the piece.

When features cannot be treated as independent, it is wise to work with

4. a total number of feature dimensions that stays well below the size of the dataset.

Having a feature set that is easy to oversee aids the transparency of the analysis. However, a modest feature set also helps to avoid the 'curse of dimensionality' as explained in section 3.3.4. Larger feature sets may still be useful for audio corpus analysis, e.g., in conjunction with a dimensionality reduction that allows a meaningful interpretation, and is robust and stable.

Finally, the kind of insight that can be obtained through corpus analysis depends not only on the quality, but also on the perceptual validity or interpretability of features. We should therefore favor:

5. informative features: features have an agreed-upon and validated natural language or domain language interpretation that is accessible to the intended audience of the study.

Ideally, only features are used that are empirically demonstrated correlates of some one-dimensional perceptual, cognitive or musicological quantity. Any summary statistics should be robust and interpretable as well.

3.5.4 *Analysis methods*

Regarding analysis methods, we specify four main desiderata, as guiding principles in selecting an adequate statistical analysis method. First of all, a good model that generalizes to unseen data requires

6. a strategy to avoid overfitting to the sample

by ensuring that no data gets re-used.

Analyses with an interest in estimating causal relationships, require

7. a model that accounts for correlations between measured variables.

See also our earlier comment on Serrà's treatment of timbre and loudness as independent in section 3.4. Consequently, a good analysis also includes as many of the potentially confounding variables as possible, without compromising on its ability to test all resulting interactions.

Analyses with an interest in quantitative relationships, require

8. a model that explains how much each feature contributes.

An ideal quantitative model explains, for each feature, if it contributes positively or negatively, which of the features contribute more, and in the most perfect circumstances: each feature's absolute effect size.

It should be clear by now that such information is the most difficult to obtain. Effect sizes can only be trusted if the underlying model is reliable, which in turn requires all features to be reliable and all potentially confounding interactions to be accounted for.

Finally, when discussing results, it is essential to

9. acknowledge potential issues with all of the above constraints.

e.g., shortcomings of the features, the possibility of not having observed important factors, the possibility of having too many variables or not having seen enough data, and assumptions on the distributions of variables. Any results must be read with care, and effect sizes more so than anything else.

3.6 TO CONCLUDE

Compared to studies with symbolic music data, advances in music description from audio have overwhelmingly focused on ground truth reconstruction and maximizing prediction accuracy, and only a handful of studies have used audio description to learn something about the music.

In this chapter we defined corpus analysis as the analysis of a music collection with the aim of gaining insights into the music itself. We reviewed the most import work in corpus analysis, and the most relevant literature on the subject of modeling music with audio data. Based on this review and a case study of two analyses of popular music evolution, we proposed several guidelines for the corpus analysis of audio data. In short, every step in the choice of hypothesis and dataset and the construction of the feature set and analysis pipeline should be considered carefully. To do this well, a good understanding of the perspective of cognitive science and statistics is desirable.

The above recommendations should be a first step towards this goal. These are not definitive guidelines, but suggestions based on the most relevant literature and an in-depth analysis of two example studies. In the following chapters of this thesis, we will aim to extend the set of available tools that satisfy these criteria.