9 CONCLUSIONS

9.1 CONTRIBUTIONS

The goal of this thesis, as stated at the beginning of chapter 1, is to make a number of contributions to the scientific study of music based on audio corpus analysis. We set out to address three sets of research problems in particular. The first problem is that there is little information on what makes an audio descriptor a good descriptor for corpus analysis research, and that more such adequate descriptors may have to be developed. The second problem is that audio corpus analysis methods themselves, too, haven't been charted and may need to be improved. Third, we addressed two goals that are central to the COGITCH project. The first is a lack of audio description techniques and similarity models for retrieval and research on musical heritage collections. The second is the ambition to gain insight on the notion of 'hooks'. We now review the contributions made to address each of these problems.

9.1.1 *Audio Description*

The first set of contributions has been to map and extend the pool of adequate audio description techniques that are available for audio corpus analysis research. To this end, we began chapter 2 with a review of existing audio features for the description of melody, harmony and musical timbre. In chapter 3, based on a review of the corpus analysis literature, a list of guidelines was deduced to guide the choice of audio features for corpus analysis research, related to robustness, dimensionality and interpretability. In chapter 4, we then presented a set of features that can be used to describe popular song sections, including psychoacoustic features, and simple harmony and melody descriptors.

In chapters 5 and 6, we introduced 'audio bigrams', a new family of multidimensional harmony and melody descriptors. They were defined in chapter 6 as measuring the co-occurrence of salient pitch events. Six examples were introduced in chapter 5. They are inspired by the notion of bigrams and trigrams in text and symbolic music analysis. Mathematically, all six relate to distributions over pairs of melodic and harmonic pitches and pitch intervals that occur close together in an audio excerpt.

9.1.2 *Audio Corpus Analysis*

The second set of contributions has been to review and extend the available methods for audio corpus analysis. First of all, this called for a review of the disciplines and interdisciplinary research contexts in which empirical music research takes place, allowing us to position audio corpus analysis research among them (chapter 1). In chapter 3, we then reviewed the most important work done in corpus analysis research. This resulted in a set of guidelines for future choices of research questions, data, audio descriptors and analysis methods in section 3.5.

In chapter 4, we presented the first use of a probabilistic graphical model in the analysis of audio features. We showed that it is possible to study the relation between a variable of interest—chorusness—and a selection of reliable audio features while controlling for confounding correlations between the audio features.

Expanding on the features evaluated in chapters 4 and 5, and inspired by methods from latent semantic analysis and symbolic corpus analysis, we also proposed the first 'second order' or corpus-relative audio features (chapter 8), quantifying the distinctiveness and recurrence of audio feature values in a corpus. In a corpus analysis of song sections and hook annotations, a selection of first and second-order audio features were shown to be both competitive and complementary to symbolic first- and second-order features.

Finally, in the same corpus study of hooks, we also introduced two methods for the statistical modeling of within-song variation. The notion of song-based second-order features can quantify distinctiveness and recurrence within a song. And a statistical model of song sections can quantify differences between song sections within songs, while controlling for differences between the songs themselves.

9.1.3 *Music Similarity and Hooks*

The third set of contributions pertains to the goals of the COGITCH project: improving audio similarity models for music heritage collections, and uncovering the properties of hooks.

Music Similarity

The challenges of modeling music similarity at scale were explained in chapter 1, contrasting alignment and non-alignment-based solutions to cover song detection. In chapters 5 and 6, we have presented several new pitch description methods. Throughout both chapters, the proposed description techniques were evaluated by applying them to the song similarity problem of cover detection.

The six novel pitch descriptors proposed in chapter 5 are all fixed-size representations of pitch use in a song or song segment. This makes them a useful asset in the design of efficient music similarity models. Three features were evaluated in a series of cover song experiments on two datasets: a dataset of early to mid-20th century translated songs from the *S&V* record collection, and a dataset of more recent cover songs used for benchmarking performance. Results showed a reasonable performance on either task when compared to other scalable systems.

Chapter 6 shows how audio bigrams relate to a range of existing 'soft audio fingerprinting' algorithms, algorithms for content-based identification of music recordings. We also showed that in its most general formulation, the computation of audio bigram features can be fully vectorized, and even formulated entirely in terms of neural network components (convolutions, matrix multiplications and non-linearities, such as rectified linear units) making highly efficient implementations possible. Finally, we have presented PYTCH, an implementation of the audio bigram features paradigm in Python, for use in audio description, song similarity models and retrieval.

Analysis of Hooks

The second project goal was to use new audio description and corpus analysis methods to gain insight into the phenomenon of catchiness and hooks.

The analysis of choruses in chapter 4 served as an experiment to prepare for this eventual goal. Choruses are a recurring object of study in music information retrieval, and are often said to have a catchy and memorable quality. By studying choruses, we could use the readily available datasets used for structure analysis and segmentation, to get a first insight into what makes a piece of music catchy. The analysis showed that choruses in the Billboard charts are perceptually sharper and rougher than other sections. They also have a smaller dynamic range and greater variety of timbre. Finally, choruses feature a higher and more salient pitch, a trend that is already present in choruses of songs from the first half of the twentieth century.

For a deeper understanding of the properties of hooks, a definition of hooks was first required. In chapter 7, we introduced the notion of catchiness, and defined hooks as the part of a song that is most recognizable. After reviewing the prevailing hypotheses on what makes music catchy, we then described Hooked, a game we made to collect a dataset of hooks. The analysis of the music and recognizability estimates gathered using Hooked followed in chapter 8. It involved several of the audio descriptors introduced in chapters 4–6, and the novel concept of corpus-based and song-based second order features discussed above. A principal component analysis of the selected features revealed twelve interpretable dimensions that could be used to match the audio features to recognizability. The results, controlling for differences between songs, show how sections with vocals and sections that were most representative of the song in terms of timbre, are generally more recognizable. Recognizable song sections also have a more typical, expected sound, as measured by several corpus-based second order features. In other words, hooks are characterized by the presence of vocals, and components that suggest repetition and conventionality. This is confirmed in an analysis of melodic transcriptions using similar, symbolic features: recognizable melodies are more repetitive and contain less atypical motives.

9.2 LOOKING BACK

Having reviewed contributions, we can now look back and critically assess them in light of the goals set in chapter 1 (section 9.2.1) and the methodological guidelines formulated in chapter 3 (section 9.2.2).

9.2.1 *Research Goals*

Have the research goals set at the beginning of this thesis been reached? For the most part, we believe they have. The above discussion lists several of the new approaches to audio description that have been have been proposed and tested as part of this thesis. However, not all opportunities to leverage the full potential of MIR for corpus analysis, were seized. We give a brief overview.

First, *rhythm description* has largely been absent from this thesis. This is unfortunate—strategies for rhythm description would be a valuable extension of the corpus analysis tool set. As chapter 5.1.1 explains: at the level of distributions of basic patterns, rhythm description proves to be significantly more challenging than harmony and pitch description. In the context of polyphonic audio, rhythm perception relies on two inference processes that are notably hard to model: note onset detection (especially of non-percussive onsets) and streaming. A lot of music perception and signal processing work is still to be done when it comes to modeling these two perceptual skills, more than we could have done as part of this thesis.

Second: melody, harmony and timbre, too, have more facets than can be measured by the audio bigram descriptors introduced in this thesis—e.g., we talked about melodies but not about melodic contours, we talked about harmony but not about different voicings. Many other description methods could have been implemented, tested or developed to further characterize melody and harmony for corpus analysis research but we decided to leave these for future work.

Similarly, some of the statistical analysis methods encountered in the literature review of corpus analysis studies have not been applied or evaluated. In the two sets of original corpus studies we presented (chorus analysis and hook analysis) we used hypothesis testing, graphical models and classification (chapter 4), and a linear mixed effects model with principal components analysis (chapter 8). Methods we have not yet explored include large feature set analyses with feature selection (e.g., using cross validation, as in Leman's study on walking speed [104]), or Bayesian models (as we are currently using in an ongoing analysis of the *Hooked on Music* data).

Finally, we acknowledge that our efforts to improve content-based similarity, which we re-framed as 'soft audio fingerprinting' in chapter 6, have not yet yielded the powerful solutions we aimed for in the COGITCH project. However, we have provided a new theoretical perspective on an array of soft audio fingerprinting approaches. The resulting audio bigram paradigm was formally defined and implemented, and is ready to be tested in applications like large-scale content-based matching.

In short, many of the goals were achieved. The work that we havent been able to address centers on four issues: (i) the continued lack of validated rhythm description methods, (ii) the many possibilities of extending our approaches to melody, harmony and timbre description, (iii) opportunities to apply several more statistical methods to corpus analysis, and (iv) the application of audio bigram-based features to large scale document matching in musical heritage collections.

9.2.2 *Methodology*

Have the contributions listed above given sufficient consideration to the methodological guidelines in chapter 3? The guidelines or 'desiderata' for a good audio corpus analysis strategy pertain to choices in research questions, data, descriptors, analysis methods.

Have we used only robust, low-dimensional and informative features, as prescribed? Our pitch description features (chapter 5) mostly comply. They require a strategy for melody extraction, but not for note segmentation or other music transcription hurdles. They are rather high-dimensional, but they were complemented in chapter 8 with first and second-order aggregation functions that provide a useful one-dimensional summary of its dimensions. And they are informative: each of the dimensions can be interpreted as a probability of observing some combination of pitch classes or pitch class intervals.

In our choice of statistical analysis methods we have been cautious about false positives and overfitting. Significance levels were always set according to the number of tests in an experiment, and the number of parameters to each model was consistently kept in check with the amount of data, partially due to the aforementioned summarization of feature dimensions.

We have also devised, on two occasions, explicit strategies to control for confounding variables. In the analysis of choruses, we used a probabilistic graphical model—a statistical model of conditional independences rather than just correlations. This allowed us to identify different kinds of relationships between chorusness and audio features, even if those features were correlated. In the analysis of hooks, we presented a statistical analysis of song sections in which we used a mixed effects model to find trends in a corpus of song sections while controlling for differences between songs.

Finally, in each set of 'experiments' in part ii (chorus analysis and cover detection), two different datasets were used, which allowed us to corroborate the most important conclusions—not just different samples drawn from a larger dataset, but new, 'idiosyncratic' data.

A similar approach will be followed in the further analysis of hooks: chapter 8 presented the analysis of the *Hooked!* data, an analysis of the *Hooked on Music* data is still underway.

An example of a case in which, arguably, better choices could have been made, is in the choice of data. The Billboard dataset, used in chapter 4, has been very carefully sampled from the clearly defined population that is the Billboard charts. The Billboard charts themselves, however, were shown in chapter 3 to suffer from more biases and discontinuities than desirable for a supposedly authoritative metric of song popularity, mostly because of the way sales were measured by Billboard over the decades since the beginning of the charts. In general, popularity is an elusive concept which, many musicologists would argue, cannot be captured in a single number. The same music can be wildly popular in one place and unknown in another. And some songs are popular at the time of their release but hardly known today—many chart-topping songs from the past decades haven't made it into the collective memory.

Biases and inconsistencies therefore also exist in the Top 2000 list from which the Hooked data were sampled—possibly problematic, even if we control for some of the effects of inter-song differences: not all songs can be expected to have the same kind of hooks. Making an unbiased selection of well-known songs is a nearly impossible task, but it is worth reflecting on how the song collections could nonetheless have been more carefully sampled, as it is important for transparency of research results and, therefore, the integration of findings in the musicological discourse.

Finally, one could argue that we could have made a larger contribution by assessing the output of not just one, but several analysis methods per dataset, and comparing the findings. Each time a music corpus was analyzed, we have focused on showing that there exists a viable corpus analysis strategy that can be used to gain insight into the data.

However, in choosing a good audio description or a corpus analysis method, we are skeptical that simple, widely applicable answers exist. Each corpus analysis problem may call for new audio description approaches or a different corpus analysis strategy. Conclusions about evaluation of descriptors and analysis methods might therefore not generalize to other contexts if they are based on a limited number of cases, e.g., the ones presented in this thesis. What makes our investigation valuable, then, are not the just the corpus studies and their results, but the literature review and methodological guidelines in chapter 3 and, above all, the description and analysis methods themselves. New methods were tested, but in the spirit of the new empirical method, not to benchmark them with respect to some measure of performance, but to show how they can be used and demonstrate their potential in one or two real-world music analysis problems.

9.3 LOOKING AHEAD

We now look ahead at the planned, upcoming work (section 9.3.1) and the most promising future work after that (section 9.3.2).

9.3.1 *Ongoing Work*

The CATCHY *Toolbox*

In the near future, we plan to release the code we used to compute second-order audio features as a small toolbox that can be used on top of PYTCH. It will be released on Github under the same name as the paper in which second-order audio features and the *Hooked!* analysis were introduced: CATCHY, or 'corpus analysis tools for computational hook discovery'.[1] The CATCHY toolbox will be tested together with colleagues at Goldsmiths University of London, as part of their research on the properties of earworms.

Hooked on Music *Data Analysis*

An important unfinished element of the COGITCH project is the analysis of the response times and accuracies gathered in the *Hook on Music*

[1] http://www.github.com/jvbalen/catchy

game, the UK version of Hooked, as the data collection stage of this experiment has only recently closed. In the coming months, we will finalize the analysis of the participant data and the music.

Rather than using the exact same analysis strategy, we now aim to integrate the LBA model of memory retrieval into the mixed effects model. The result we aim for is a hierarchical Bayesian model in which all parameters can be estimated at once. Having access to the responses of up to 100 times more participants than were available in the *Hooked!* dataset will, hopefully, yield precise estimates of each of these parameters.

The results of the analysis will then be compared to the results obtained for the *Hooked!* data to see if our earlier findings are confirmed. Trends that are very significant but not shared may shed some light on the difference between the music in both datasets, and between the Dutch and UK music listeners.

9.3.2 *Future work*

Section 9.2 reviewed the contributions of this thesis in terms of the goals and the methodological perimeter set in part i. Here, we distill from this discussion three important and promising avenues for future work.

Rhythm Description

As said in above in section, the analysis of rhythm has been given very little attention in this thesis, even though we believe it should be part of a the corpus analysis toolbox. In future work, we believe rhythm description should be a priority.

Corpus studies such as those by Serrà et al., Mauch et al. (see section 3.4), and ourselves (chapter 8) paint a skewed picture of popular musing by measuring diversity, change, distinctiveness and repetition only in terms of melody, harmony and timbre. For example, claims that popular music has become increasingly homogeneous solely on the basis of those three musical facets ignore the possibility that harmonic and melodic complexity might have been replaced with rhythmic complexity, e.g., due to the rise of hip hop and electronic music in the last decades. Even though hypothetical—we don't know how rhythmic complexity evolved—this example illustrates the potential impact of methodological decisions. First, the overall conclusion might have looked very different if rhythm was taken into account. Second, it illustrates how a mildly Euro-centric methodology (giving priority to melody, harmony and timbre over rhythm) can lead to conclusions with decidedly Euro-centric connotations (music based on loops, rap and sampling is somehow less complex). It reminds us to consider the concerns of new musicology at the end of last century: researchers bring their own cultural biases into the lab.

Audio Bigrams and Learned Fingerprints

In chapter 6 we proposed the umbrella task of soft audio fingerprinting. Essentially, soft audio fingerprinting is any kind of content identification in which a fixed-size representation is used to efficiently compare documents. At the end of the chapter, it was suggested that, given a specific soft audio fingerprinting problem (e.g., sample detection, remix detection or efficient cover song identification) and a ground truth of related documents, it may be possible to learn an optimal audio bigram representation of the music that is to be analyzed.

This approach has not been tested. However, we consider this a very promising path to efficient and versatile fingerprinting. The potential of feature learning techniques to outperform traditional information retrieval methods has already been shown in several other MIR tasks, so it can be expected that strategies exist in which feature learning works for fingerprinting applications as well. What makes a task like fingerprinting or cover detection difficult is that, unlike chord detection or tag prediction (see section 2.2), it cannot be trivially reduced to a typical classification problem. First, any given pair of remixes or cover songs tends to differ in length. Current feature learning systems tend to crop or scale their input data rather arbitrarily, in a way that would be problematic for a remix or cover song recognition system (often, selecting a 30 second fragment at random). Second, the order in which the musical material appears is very important in cover detection, as the success of alignment-based systems suggests. And finally, training a classifier typically requires a rather large number of examples per class. In remix detection and cover detection, there are often only one or two cover versions of a song.

An audio bigram-based fingerprinting system, if implemented using the convolutional neural network components as described in section 6.3.3, gives us a way to deal with the first two problems: the matrix product reduces each song to a set of n $k \times k$ matrices (where n and k are predefined constants) that represent ordered occurrences of events. A possible strategy to overcome the third problem is to approach fingerprinting first as a binary classification problem, classifying pairs of documents as either the same or not the same (e.g., in a way similar to the recently proposed 'siamese' network architecture used in [163]). Then, having trained such a model, one of the learned, intermediate representations can be used as the basis of a fingerprint for matching at scale.

The current implementation of the audio bigram feature in the PYTCH toolbox (also presented in chapter 6) makes use of Numpy and Scipy, two Python toolboxes for numerical modeling that are well suited for vectorized computations, but not ideal for learning representations. Tools that are naturally suited for this problem are hybrid (symbolic and numerical algebra) toolboxes like Theano[2] and Tensorflow[3], which use techniques from symbolic computing to optimize the numerical manipulation of vectors, matrices and tensors. An adaptation of the audio bigram toolbox to work with one of these toolboxes could make it possible to test the potential of learning fingerprints from a dataset of examples.

In the longer term, we hope to be able to use this approach to improve efficient document matching, so that it can be used in musical heritage-related soft audio fingerprinting problems such as cover song

[2] htp://www.deeplearning.net/software/theano

[3] htp://www.tensorflow.org

detection in the S&V collection, and matching the MI's recordings of monophonic folk songs with S&V's digitized 78 RPM records.

The Future of Hooked

The data obtained with the Hooked games has already been put to good use. However, we believe much more research can be done, even after the first analysis of the *Hooked on Music* data is completed. In the analysis of each dataset, we have incorporated a notion of 'listening history' of the participants, which was used to compute distinctiveness of individual song fragments, and estimated using the whole of the music corpus used in each of the games.

A powerful extension of this approach could be pursued by modeling a more detailed listening history for each of the participants. Since we have data on how well they were able to recognize each of the stimuli, this can even be done fairly straightforwardly. The most straightforward approach would be to perform clustering of the users based on their response times per song. Advanced variants of this approach exist in music recommendation, specifically designed to deal with the sparsity of information that occurs when not all users listen to all music. The underlying idea is to factorize the matrix of users' response times into a matrix of taste profiles (each represented by a weighed subset of all the songs) and a matrix of participants listening preferences (representing each user's preferences as a weighted combination of taste profiles); see e.g, [73]. Integrating this kind of listener profiles may help in exposing interactions between distinctiveness and recognizability that our models can currently not observe.

One similar line of research has already been initiated by Burgoyne et al. [24]. In a 2015-2016 update to the *Hooked on Music* game, players are presented, as they progress, with songs they are increasingly likely to know. The updates are based on songs they have already recognized. The aim of this variant of Hooked is to test whether an adaptive version of the game can be used as a tool to guide players towards the music they are most familiar with. This would make it a valuable instrument in helping music listeners who suffer from memory loss reconnect with their preferred music even if they have forgotten titles and names of artists.

9.4 THE FUTURE OF AUDIO CORPUS ANALYSIS

In this thesis, we have tried to make a useful contribution to the available methods and technologies for audio description and audio corpus analysis. We have strived to make these methods and technologies transparent and flexible. Along the way, we have gained new insights into choruses, catchiness and hooks.

By presenting concrete applications of the proposed technologies, following the proposed methods, we believe we have shown that rigorous audio corpus analysis is possible and that, even though there is more work left to do, the technologies to engage with it are available.

In pursuing a methodology that is transparent about the origin and limitations of data and algorithms, and technologies that leverage the context and cognition of listening, we make it easier to research music from a critical angle: as a product of the mind, with a strong cultural dimension. We believe this makes our contributions important steps towards the integration of audio analysis into the new empirical method, and, ultimately, musicology.

Hopefully, our efforts can be a stepping stone and an inspiration for future empirical, audio-based research. We encourage researchers to take up this approach, learn from our experiences, and use the wealth of audio available today to discover more about music and our intriguing relation to it as listeners.