4

CHORUS ANALYSIS

This chapter presents a corpus analysis of the acoustic properties of the pop song chorus. We address the question: what makes a chorus distinct from other sections in a song?

Choruses have been described as more prominent, more catchy and more memorable than other sections in a song. Yet, in MIR, studies on chorus detection have always been primarily based on identifying the most-repeated section in a song.

Instead of approaching the problem through an application-centered lens, we present a first, rigorous, analysis-oriented approach.

4.1 INTRODUCTION

4.1.1 *Motivation*

The term *chorus* originates as a designation for the parts of a music piece that feature a choir or other form of group performance. In the popular music of the early twentieth century (e.g., Tin Pan Alley and Broadway in New York), solo performance became the norm and the term chorus remained in use to indicate a repeated structural unit of musical form. The same evolution was observed in European entertainment music [129].

In terms of musical content, the chorus has been referred to as the "most prominent", "most catchy" or "most memorable" part of a song [50] and "the site of the more musically distinctive and emotionally affecting material" [129]. It is also the site of the refrain, which features recurring lyrics, as opposed the more variable 'verse'. While agreement on which section in a song constitutes the chorus generally exists among listeners, attributes such as 'prominent' and 'catchy' are far from understood in music cognition and cognitive musicology [69].

This points to at least two motivations for a deeper study of the particularities of choruses. First, the chorus is a central element of form in popular music. In analyzing it we may gain insight into popular song as a medium, and conscious as well as unconscious choices in songwriting. The concept is also rather specific to popular music, so it may tell us something about where to look for the historical shifts and evolutions that have resulted in the emergence of a new musical style. Second, choruses may be related to a catchy or memorable quality, to the notion of hooks, and perhaps to a more general notion of cognitive salience underlying these aspects. The nature of choruses may indicate some of the musical properties that constitute this salience, prominence or memorability.

Recently, as a frequent subject of study in the domain of music information retrieval, systems have been proposed that identify the chorus in a recording; see also section 2.2.2. Yet, as we will show in the next section, the chorus detection systems that locate choruses most successfully turn out to rely on rather contextual cues such as the amount of repetition and relative energy of the signal, with more sophisticated systems also taking section length and position within the song into account [50, 57]. This suggests a third motivation for the proposed analysis: the potential to advance MIR chorus detection methods with a more informed approach.

The central research question of this analysis is therefore:

> In which measurable properties of popular music are choruses, when compared to other song sections, musically distinct?

4.1.2 *Chorus Detection*

Existing work on chorus detection is strongly tied to audio thumbnailing, music summarization and structural segmentation. Audio thumbnailing and music summarization refer to the unsupervised extraction of the most representative short excerpt from a piece of musical audio, and often rely on full structure analysis as a first step. The main ideas underlying the most important structure analysis methods are described in section 2.2.2. A more in-depth review of relevant techniques is given by Paulus et al. in [150].

Definitions of the chorus in the MIR literature characterize it as repeated, prominent and catchy. Since the last two notions are never formalized, thumbnailing and chorus detection are essentially reduced to finding the most often-repeated segment or section. A few chorus detection systems make use of additional cues from the song's audio, including RefraiD by Goto and work by Eronen [50, 57]. RefraiD makes use of a scoring function that favors segments C occurring at the end of a longer repeated chunk ABC and segments CC that consistently feature an internal repetition. Eronen's system favors segments that occur 1/4 of the way through the song and reoccur near 3/4, as well as segments with higher energy. In most other cases, heuristics are only used to limit the set of candidates from which the most frequent segment is picked, e.g., restricting to the first half of the song or discarding all segments shorter than 4 bars.

Some efforts have also been made in labeling structural sections automatically [148, 149, 205]. Xu and Maddage rely on a heuristic which 'agrees with most of the English songs', imposing the most likely of three typical song structures on the analyzed piece [205]. However, as Paulus and Klapuri show, the datasets that are typically used in structure analysis do not support the claim that a small number of structures recur very often [149]. In the *TUTstructure07* dataset for example, 524 out of 557 pop songs have a unique structure.

Paulus and Klapuri use a Markov model to label segments given a set of tags capturing which segments correspond to the same structural section (e.g., ABCBCD) [148, 149]. This approach performs well on *UPFBeatles*, a dataset of annotated Beatles recordings, and fairly well on a larger collection of songs (*TUTstructure07*).[1] An n-gram

[1] Dataset descriptions and links at `http://www.cs.tut.fi/sgn/arg/paulus/structure.html`

method with $n = 3$ and a variable-order Markov model come out as the best techniques. The same methods have also been enhanced by using limited acoustic information: section loudness and section loudness deviation [148]. This boosts the best performance (in terms of per-section accuracy) by up to 4 percent for *TUTstructure07*. Whether the model could be improved with more acoustic information remains an open question.

4.1.3 *Chorus Analysis*

The difference between the present investigation and the chorus detection methods above is both in the goals and in the execution. While chorus detection systems are built to locate the choruses given unsegmented raw audio for a song, this investigation aims to use computational methods to improve our understanding of choruses. And while the computational methods used in chorus analysis relate mostly to structure analysis techniques reviewed in section 2.2.2, we follow a corpus analysis approach—as described in chapter 3. Because structural boundary detection is not part of our goal, we can start from reliable manual annotations of the structural boundaries of a song.

We study the notion of chorus in two corpora: a newly created dataset of early Dutch popular music from the first half of the 20th century, and a large dataset of Western popular music from the second half of the 20th century. The focus on early Dutch choruses was included because it allows us to zoom in on a time period in which popular song developed as a style, and because of the interests of the Meertens Institute and the Institute of Sound and Vision (see section 1.1.2). The more recent dataset allows us to look at trends at a larger scale.

To find trends, we compile a list of appropriate features and model how they correlate with section labels in a collection of song sections. Expert structure annotations for the two datasets allow to parse audio descriptors (see section 4.2.2), into per-section statistics. The analysis of the resulting variables will be presented in sections 4.3 and 4.4.

The contributions of this chapter include the introduction of the 'chorusness' concept, a corpus analysis method to model it, and the resulting model, which we believe can serve MIR applications, popular music understanding and popular music perception and cognition.

4.2 METHODOLOGY

4.2.1 *Datasets*

The Dutch50 *Dataset*

The first dataset, the *Dutch50* dataset, was created especially for this study, and was conceived as a diverse and representative sample of the Netherlands' popular music as it sounded before the 1960s. The *Dutch50* dataset contains 50 songs by 50 different artists, all dated between 1905 and 1951. Figure 17 shows a histogram of the songs' year of release as provided by the publisher. The songs were obtained from compilation releases by the Dutch Theater Institute,[2] acquired by the Meertens Institute. Recurring styles include cabaret, colonial history-related songs, advertisement tunes released on record and early examples of the *levenslied* musical genre [92]. An expert on early Dutch popular music was consulted to validate the representativeness of the selected artists. Structural annotations were made by the author, indicating beginning and end of sections and labeling each with a section type chosen from a list of seven (intro, verse, chorus, bridge, outro, speech and applause).

The Billboard *Dataset*

The *Billboard* dataset is a collection of time-aligned transcriptions of the harmony and structure of over 1000 songs selected randomly from the *Billboard* 'Hot 100' chart in the United States between 1958 and 1991 [26]. The annotations include information about harmony, meter, phrase, and larger musical structure. The *Billboard* dataset is one of

[2] `http://www.tin.nl`

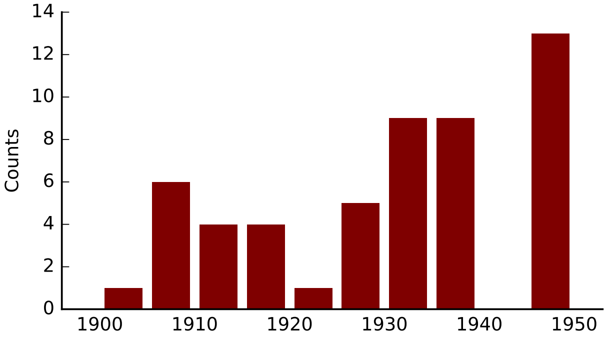


Figure 17.: The distribution of the years of release for the *Dutch50* dataset.

the largest and most diverse popular music datasets for which expert structure annotations exist and one of few to be consistently sampled from actual pop charts. It can be expected to reflect important commonalities and trends in the popular music of the period of focus. It includes a wide variety of popular music subgenres, and suits the goal of drawing musicological conclusions better than other datasets discussed so far, as it is representative of a relevant 'population' (the US charts), and carefully sampled from that population.[3]. This makes it the best available dataset for analysis of popular music choruses. For the present study, the complete v1.2 release is used (649 songs).[4]

Of the annotations, only the structural annotations are retained. The structural annotations in the dataset follow the format and instructions established in the *SALAMI* project [182]. The transcriptions contain start and end times for every section and section labels for almost all sections. The section labels the annotators were allowed to assign were restricted to a list of 22, some of which were not used. The most frequently recurring section labels are: verse (34% of total annotated

[3] E.g., by allowing for duplicates to give popular songs more weight, and by considering only chart notations up to 1991, to avoid some of the inconsistencies in how the Billboard charts themselves were compiled.

[4] `http://ddmal.music.mcgill.ca/billboard`

time), chorus (24%), intro, solo, outro and bridge. The total number of sections, including the unlabeled ones, is 7762.

Are the annotations as reliable as the sample? Here we should note that there could be a hint of bias. The annotators guide, the instructions the annotators received, defines: "chorus (aka refrain): in a song, a part which contrasts with the verse and which is repeated more strictly".[5] The emphasis on the more 'strict' repetition in a chorus may skew the set of cues used by the annotators to the perform the section labeling task, towards repetition–related information.

4.2.2 *Audio Features*

A corpus analysis–centered study requires different kinds of descriptors than traditionally used in machine-learning applications. The descriptors are therefore selected based on the constraints put forward in chapter 3: we would like a set of features that is robust, informative and limited in size. Limiting the set to a small number of hand-picked descriptors is especially important since, for a part of the analysis, the amount of data required grows exponentially with the number of variables. The following is a list of the features selected for this analysis, beginning with the features computed for the *Billboard* dataset. Many of the features appear in the feature overview in chapter 2, so we will focus the discussion on their implementation. All features are one-dimensional.

Psycho-acoustic Features & Timbre

Loudness
The loudness descriptor is the standard psychological analogy of energy. It is obtained through comparison of stimuli spectra and a standardized set of equal loudness curves. We use the implementation by Pampalk [142]. The model applies outer-ear filtering and a spreading function before computing specific loudness values (N_k in sones) per

[5] See http://www.music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf

Bark band k and summing these values over all bands to obtain the total loudness T:

$$S = \max_k(N_k) + 0.15 \cdot \sum_{k \neq \max} N_k \tag{22}$$

where the factor 0.15 serves as a weighting that emphasizes the contribution of the strongest band. For every section, the loudness *mean* is computed and stored, as well as the inter-quartile range (*Loudness IQR*), as a measure of the section dynamics.

Sharpness
The sharpness descriptor is the psychoacoustic analog of the spectral centroid. It characterizes the balance between higher- and lower-band loudness. We use the Bark-specific loudnesses N_k as computed by Pampalk [142] and summing as formulated by Peeters [154]:

$$A = 0.11 \times \sum_k g(k) \cdot k \cdot N_k, \quad \text{where} \tag{23}$$

$$g(k) = \begin{cases} 1 & k < 15 \\ 0.066 \times \exp(0.171 \cdot k) & k > 15 \end{cases} \tag{24}$$

For every section, we use the *mean* sharpness. Compared to loudness range, sharpness range has no direct informative psycho-acoustic interpretation, so it is not included.

Roughness
Like the loudness descriptor, roughness is a mathematically defined psychoacoustic measure. It characterizes a timbral property of complex tones, relating to the proximity of its constituent partials. We use the MIRToolbox implementation by Lartillot et al. [103], which is based on a model by Plomp and Levelt [159]. Since the roughness feature has a very skewed distribution, it is summarized for every section by taking its *median*.

MFCC
As discussed in section 2.1.2, MFCC's are established multidimensional spectral envelope descriptors, designed to be maximally inde-

pendent. Individual MFCC coefficients tend to have no particular interpretation. In this model, therefore, the descriptor of interest is the variety in timbre. This is modeled by computing the trace of the square root of the MFCC covariance matrix, a measure of the timbre *total variance*. The MFCCs are computed following [142], and the first component (directly proportional to energy) is discarded.

Pitch Features

Chroma variance

Chroma features, also discussed in chapter 2.1.3, are widely used to capture harmony and harmonic changes. In the most typical implementation, the chroma descriptor or pitch class profile consists of a 12-dimensional vector, each dimension quantifying the energy of one of the 12 equal-tempered pitch classes. These energies can be obtained in several ways. The *NNLS chroma* features distributed along with the *Billboard* dataset are used in this study.[6]

In this study, the variety in the section's harmony is measured. Chroma, unlike MFCC, isn't typically looked at as a vector in Euclidean space, but rather as a distribution (of energy over pitch classes). Estimating just the total variance, as done for MFCC, would neglect the normalization constraint on chroma vectors and the dependencies it entails between pitch classes. We therefore normalize the chroma features per frame and assume it is Dirichlet-distributed. With the normalized features p as a 12-dimensional random variable, we can estimate a Dirichlet distribution from all of the section's chroma observations.

The 12-dimensional Dirichlet distribution $\mathcal{D}_{12}(\alpha)$, can be written:

$$f(p) \sim \mathcal{D}_{12}(\alpha) = \frac{\Gamma(\sum_{k=1}^{12} \alpha(k))}{\prod_{k=1}^{12} \Gamma(\alpha(k))} \prod_{k=1}^{12} p(k)^{\alpha(k)-1}, \tag{25}$$

[6] http://www.isophonics.net/nnls-chroma

where Γ is the Gamma function. $\mathcal{D}_{12}(\alpha)$ can be seen as a distribution over distributions. We use the sum of the parameter vector $\alpha(k)$, commonly referred to as the Dirichlet precision s:

$$s = \sum_{k=1}^{12} \alpha(k) \tag{26}$$

It quantifies the difference between observing the same combination of pitches throughout the whole section (high precision) and observing many different distributions (low precision) [21]. There is no closed-form formula for s or α, but several iterative methods exist that can be applied to obtain a maximum-likelihood estimation (e.g., Newton iteration). Fast fitting was done using the *fastfit* Matlab toolbox by Minka.[7]

Pitch salience

The notion of pitch salience exists in several contexts. Here, it refers to the strength or energy of a pitch, specifically, the combined strength of a frequency and its harmonics, as in [168]. The *mean* of the strongest (per frame) pitch strength will be computed for every section.

Pitch centroid

As a last pitch-related feature, we include a notion of absolute pitch height, which is easy if the audio lends itself to reliable melodic pitch estimation. For the polyphonic pop music of the *Billboard* dataset, melody estimation is prone to octave errors. We therefore approximate pitch height in another way, using the more robust *Pitch centroid*. We define this as the average pitch height of all present pitches, weighted by their salience. Note that the pitch salience profile used here spans multiple octaves and involves spectral whitening, spectral peak detection and harmonic weighting in order to capture only tonal energy and emphasize the harmonic components. Our feature set includes the section *mean* of the pitch centroid as well as the interquartile range.

[7] `http://research.microsoft.com/en-us/um/people/minka/software/fastfit/`

Melody Features

Contrary to the *Billboard* dataset, the *Dutch50* dataset contains songs with a mostly prominent melody, and relatively little post-processing. Specifically, it is less affected by the kind of heavy dynamic range compression that is commonplace in more recent popular music. This allows for a reasonable reliable melody estimate to be extracted for those songs (see the melody extraction challenges listed in section 2.1.4). The following features will therefore only be used for the *Dutch50* corpus analysis.

For all songs in the *Dutch50* dataset, the melody is extracted using the *Melodia* Vamp plug-in (see section 2.1.4, [168]). The resulting pitch contours and pitch salience are segmented along the annotated boundaries. For each section, statistics on the contour are then computed and compared.

Pitch strength
The melodic pitch strength, also referred to as pitch salience or salience function, is a measure of the strength of the fundamental frequency of the melody and its harmonics. For each section, the *mean pitch strength* was computed and normalized by subtracting the average pitch strength for the complete song.

Pitch height
For each section, the *mean pitch height* is computed and again normalized. A measure of pitch range was computed as well, in this case, the *standard deviation* of the pitch height.

Pitch direction
Finally, the *pitch direction* is estimated. With this measure, we aim to capture whether the pitch contours in a section follow an up- or downward movement. It is computed simply as the difference between the pitch height of the section's last and first half.

Structure Descriptors

Section length
The length of the section in seconds.

Section position
The position of the section inside the song is included as a number between 0 (the beginning) and 1 (the end).

4.3 CHORUSES IN EARLY POPULAR MUSIC

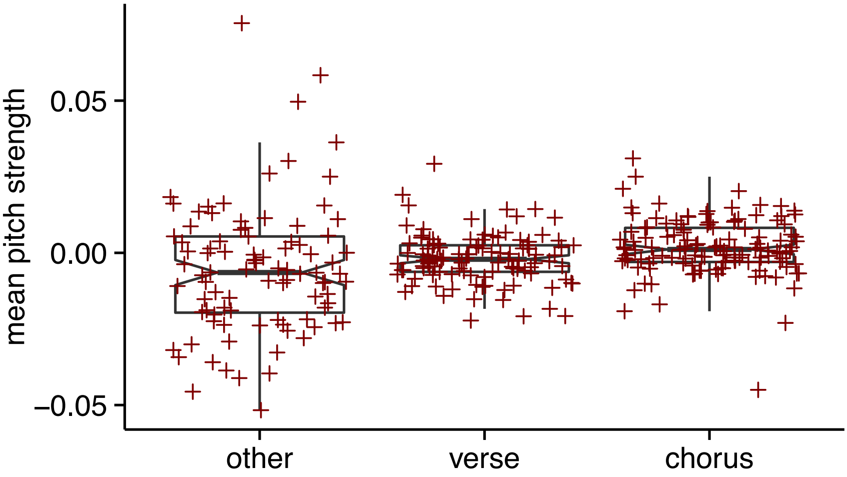
For the *Dutch50* dataset, we are more interested in melody than instrumentation and production, so in this first look at the features that make a chorus, the focus will be on melody.

The analysis will follow a simple approach, comparing raw feature differences between section types. This is easily manageable because of the constrained set of section labels. Nine songs were not considered as they contained only one type of section, in which case the labeling (verse or chorus or other) was found to be rather arbitrary. The remaining 41 songs contained a total of 330 sections, that were used to produce figures 18a – 19b.

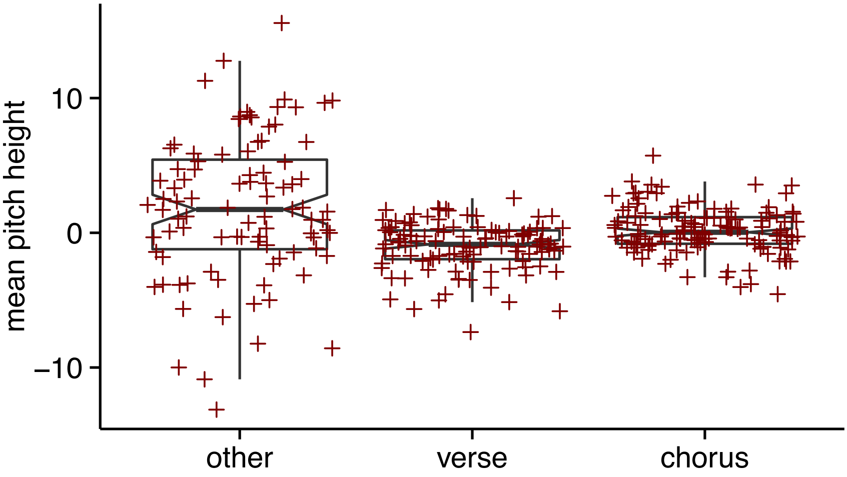
Figure 18a shows a scatter plot of the mean pitch strength values of all sections, over a box plot with estimates of the main quantiles (25%, 50% and 75%). A 95% confidence interval for the median is indicated by the notch in the box plot's sides. Remember that the mean pitch strength values for each section were normalized by subtracting the mean pitch strength over the complete song. The figure therefore illustrates how chorus pitch strengths do not significantly exceed the song average at 0, however, they are demonstrably higher than in verses and other sections. The former is confirmed in a t-test ($p < 0.001$) at a significance level, for this set of experiments, of 0.002.[8]

Figure 18b shows the *mean pitch height* for all sections, normalized by subtracting the overall song average. Correcting for multiple com-

[8] We correct for multiple comparisons based on a total of 23 comparisons: 12 box plots, and 11 tests. $\alpha_{23} = \frac{0.05}{23} = 0.002$.

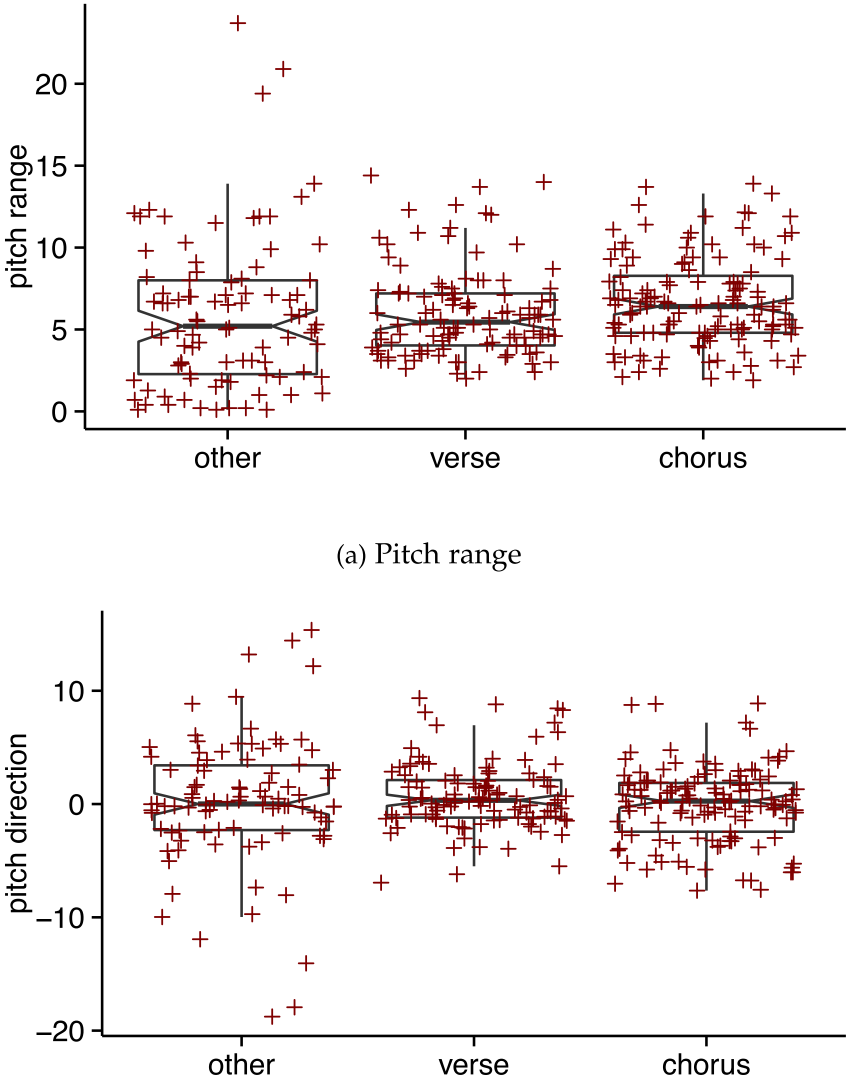


(a) Average pitch strength



(b) Average pitch height

Figure 18.: Pitch statistics per section type in the *Dutch50* dataset.



(a) Pitch range

(b) Pitch direction

Figure 19.: Pitch statistics per section type in the *Dutch50* dataset.

parisons, pitch in the chorus is not significantly different from the song average ($p = 0.030$), but it is higher then the pitch of the verse, with over a semitone difference in median and mean ($p < 10^{-6}$ in a two-sided t-test). It is not significantly different from the pitch of the bridge and other sections ($p = 0.004$, two-sided Welch's t-test).

Figures 19a and 19b show the pitch range and direction for all sections (not normalized). Pitch ranges are wider for choruses than for verses and other sections, but not significantly. Finally, average pitch direction shows no trend for choruses at all. The average direction for verses is greater than zero with $p = 0.003$, suggesting an upward tendency in pitch during the verse, but again, this is not significant when significance level is adjusted for multiple comparisons, so that no conclusions can be made from this observation.

Summing up the findings, the analysis shows how choruses in the *Dutch50* dataset have a stronger and higher pitch than verses. Note that several more trends would have emerged from these test statistics if the confidence level hadn't been corrected for. The correction is nonetheless crucial: the initial exploration using box plots, as well as the subsequent tests must be accounted for, as argued in Chapter 3.

4.4 CHORUSES IN THE BILLBOARD DATASET

For the *Billboard* analysis, all descriptors are used except for those pertaining to melody, since melody estimates are expected to the substantially less accurate then they were for the *Dutch50* data. The resulting features make up a dataset of 7762 observations (sections) and 12 variables (descriptors) for each observation: the above perceptual features and one section label. These data will be used to model what features correlate with a section being a chorus or not.

More specifically, they will be modeled using a PGM, or probabilistic graphical model. We now explain the concept of Probabilistic Graphical Models (PGM) by introducing three varieties of graphical models: correlation graphs, partial correlation graphs, and Bayesian networks.

4.4.1 *Graphical Models*

Graphs and networks can be very useful to conceptualize the relations between random variables. The easiest model to display relations between variables is the *correlation graph*. It's a graph in which all variables are nodes, and two variables are connected by an edge if and only if they are correlated:

$$E_c(i,j) = \begin{cases} 1 & \left|\rho(X_i, X_j)\right| > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

This is, the matrix E_c encoding the correlation graph's edges contains a 1 wherever the absolute correlation (e.g., Pearson correlation ρ) between the corresponding variables (e.g., X_i and X_j) is greater then some threshold ε.

Correlation graphs are useful as visualizations, but can be quite dense, with correlations between many of the variables. They also provide little information about the underlying multivariate distribution of $(X_1, \ldots, X_n)$.

A more widely used type of graphs are *partial correlation graphs*. The partial correlation between two variables X_i and X_j is the correlation between ϵ_i and ϵ_j: the errors for X_i and X_j after removing the effects of all other variables $\{X_k\}$. Specifically, the correlation between the errors of X_i and X_j after linear regression with $\{X_k\}$. This sheds some more light on each variables' contribution to the dependencies among the set of variables than a simple correlation graph.

$$E_{pc}(i,j) = \begin{cases} 1 & \left|\rho(\epsilon_i, \epsilon_j)\right| > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Finally, the graphs that are most often referred to when talking about Probabilistic Graphical Models are *Bayesian networks*. Bayesian networks encode *conditional independence*. When two variables in a PGM are *not* connected, they are conditionally independent given the other variables:

$$E_{PGM}(i,j) = \begin{cases} 0 & X_i \perp\!\!\!\perp X_j \text{ given } X_k \ \forall k \\ 1 & \text{otherwise} \end{cases} \tag{29}$$

If all variables are normally distributed (the distribution is multivariate Gaussian), and the graph is undirected, E_{PGM} will be the same as the partial correlation graph E_{pc} [200]. Generally, however, Bayesian networks contain directed edges (i.e., arrows), encoding a special kind of dependence in which some variables are *parents* of other variables. Bayesian networks are also acyclic, i.e., they contain no cycles. This allows us to identify for each variable, not just its parents and children, but also its *ancestors* and *descendants*. In a fully directed Bayesian network, a more specific independence property holds: any variable X_j is conditionally independent of its non-descendants, given its parents [21]. The latter implies a particular relationships between the conditional distributions of the variables and their joint distribution $p(X_1, \ldots, X_n)$:

$$p(X_1, \ldots, X_n) = \prod_{k=1}^{n} p(X_k | pa_k) \tag{30}$$

where pa_k denotes the set of parents of X_k. In other words, by defining pa_k, directed graphical models straightforwardly encode a factorization of the joint probability distribution p(X_k) underlying the graph.

Graph structures of Bayesian networks are typically constructed using prior expert knowledge, but they can also be learned from data. However, when considering a set of variables in the real world, it will not usually be possible to know the direction of every edge. When learning a PGM's graph structure from data, even if all conditional dependence relations are known, one set of conditional independences can often be represented by several Bayesian networks, with arrows pointing in different directions. For this and a number of other reasons, it is dangerous to interpret edge directions in Bayesian networks as *causal* relationships.

Yet, directed and partially directed graphical models can nonetheless be interesting: the absence of edges does encode independence with other variables controlled for, and sometimes, some edge directions may indeed be found, in which case an interpretation of the edge directions can help to assess whether the model makes sense. For more on the interpretation of Bayesian networks and examples for music analysis, see [21].

PGM Structure Learning

Learning the PGM structure generally requires a great amount of conditional independence tests. The *PC algorithm* optimizes this procedure and, in addition, provides information about the direction of the dependencies where they can be inferred [86]. When not all directions are found, a partially directed graph is returned.

One of the limitations of the PC algorithm, however, is that the variables must be either all discrete, or all continuous, following a normal distribution. In the analysis in the next section, all data are modeled as continuous. For most variables, this is straightforward, except for the *Section Type* variable, which will have to be remodeled as continuous. We do this by introducing the notion of *Chorusness*.

4.4.2 *Chorusness*

The Chorusness variable is derived from the *Chorus probability* p_C, a function over the domain of possible section labels. The chorus probability $p_C(T)$ of a section label T is defined as the probability of a section annotated with label T being labeled 'chorus' by a second annotator. In terms of the annotations T_1 and T_2 of two independent and unbiased, but 'noisy' annotators, $p_C(T)$ can be written:

$$p_C(T) = p(T_1 = C | T_2 = T) = p(T_2 = C | T_1 = T), \tag{31}$$

where C refers to the label 'chorus'.

The *Billboard* dataset has been annotated by only one expert per song, therefore it contains no information about any of the $p_C(T)$. However, in the *SALAMI* dataset, annotated under the same guidelines and conditions, two independent annotators were consulted per song [182]. The annotators' behaviour can therefore be modeled by means of a confusion matrix $M(T_1, T_2) \in [0,1]^{22 \times 22}$ between all 22 section types:

$$M(T_1, T_2) = f(x_1 = T_1 \cap x_2 = T_2) \tag{32}$$

with frequencies f in seconds (of observed overlapping labels T_1 and T_2). Since the identities of the two annotators have been randomized, M may be averaged out to obtain a symmetric confusion matrix $M^\star$:

$$M^\star = \frac{M + M^T}{2} \tag{33}$$

From here we can obtain the empirical Chorus probability:

$$p_C(T) = \frac{M^\star(T, C)}{\sum_k M^\star(T, k)} \in [0, 1]. \tag{34}$$

Chorus probability values for every section type were obtained from the *Codaich-Pop* subset of the *SALAMI* dataset (99 songs). Finally, the Chorus Probability is scaled monotonically to obtain the Chorusness measure $C(T)$, a standard *log odds ratio* of p_C:

$$C(T) = \log\left(\frac{p_C(T)}{1 - p_C(T)}\right) \in (-\infty, \infty). \tag{35}$$

It ranges from −8.41 (for the label 'spoken') to 0.83 (for the label 'chorus').

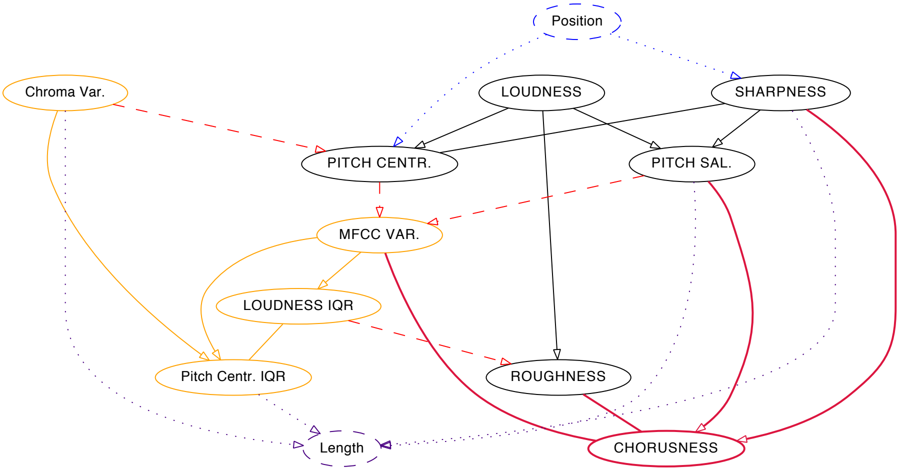
4.4.3 *Implementation*

Before the model learning, a set of Box-Cox tests is performed to check for rank-preserving transformations that would make any of the variables more normal. The Chroma variance s is found to improve with a power parameter $\lambda = -1$, and therefore scaled as:

$$S = \frac{s^\lambda - 1}{\lambda} = 1 - \frac{1}{s} \tag{36}$$

The Section length, Loudness IQR and Pitch centroid IQR are found to improve with a log transform. Weeding out divergent entries in the dataset leaves us with a subset of 6462 sections and 12 variables.

The R-package *pcalg* implements the PC-algorithm. Beginning with a fully connected graph, it estimates the graph skeleton by visiting all pairs of adjacent nodes and testing for conditional independence



Note. Bold edges highlight the features that correlate with Chorusness. Black edges are edges between features that are closely related on the signal level. The orange, lighter edges denote relations between features that represent some kind of variance. Dotted edges are used for the features that are not measured from the audio (Position and Length). The remaining edges are drawn as dashed, red lines.

Figure 20.: Graphical model of the 11 analyzed perceptual features and Chorusness variable C. $\alpha_{\mathrm{PGM}} = 0.05$. The edges are colored by the author to facilitate discussion

given all possible subsets of the remaining graph.[9] The procedure is applied to the 6462 × 12 dataset, with 'conservative' estimation of directionality, i.e. no direction is forced onto the edges when the algorithm cannot estimate them from the undirected graph structure.

4.4.4 *Analysis Results*

The resulting graphical model is shown in Figure 20. It is obtained with $p < 3.5 \times 10^{-5}$, the significance level required to bring the over-

[9] `http://cran.r-project.org/web/packages/pcalg/`

all probability of observing one or more edges due to chance, under 5 percent. In terms of the significance level α_{CI} of the conditional independence tests and α_{PGM} of the model:

$$\alpha_{\text{CI}} = 1 - (1 - \alpha_{\text{PGM}})^{1/n} \approx \frac{\alpha_{\text{PGM}}}{n} \tag{37}$$

with $\alpha_{\text{PGM}} \ll 1$ (here 0.05) and n the number of tests performed ($\sim$ 1500).

Note that $p \approx 10^{-5}$ is a conservative parameter setting for an individual test. As a result, it is best to view the model as a depiction of dependencies rather than independences, since the latter may always be present at a lower significance than required by the α.

The model is relatively stable with respect to α: it is unchanged when the learning procedure is repeated with more restrictive $\alpha_{\text{PGM}} =$ 0.01 and 0.005. Four additional edges appear with a more tolerant $\alpha = 0.10$.

4.4.5 *Discussion*

A number of observations can be made about the chorusness model in figure 20. First, the most expected dependencies in the model are highlighted.

At least three kinds of feature relations are expected. First, there are the correlations between features that are closely related on the signal level (black edges): Loudness and Pitch salience, for example, measure roughly the same aspects of a spectrum (and can be expected to be proportional to roughness), and so do Sharpness and Pitch centroid. Roughness is a highly non-linear feature that is known to be proportional to energy. The model reflects this.

The second kind of correlations are the relations between variance-based features and the Section length variable. Musically, it is expected that longer sections allow more room for an artist to explore a variety of timbres and pitches. This effect is observed for Chroma variance and Pitch centroid IQR, though not for MFCC variance and Loudness IQR. Interestingly, correlations with Section length point *towards* it rather than away (dotted edges): the length of a section length is a result of its variety in pitch and timbre content, rather than a cause. Note again, however, that directions of effects in a learned PGM are not always reliable enough to be taken at face value.

Third, some sections might just display more overall variety, regardless of the section length. This would cause different variances to relate, resulting in a set of arrows between the four variance features. Four such relations are observed (lighter, orange edges).

We now note that Sharpness, Pitch salience and Roughness predict Chorusness, as well as the MFCC variance (bold edges). All of these can be categorized as primarily timbre-related descriptors. Section length, Section position and Chroma variance are *d*-separated from Chorusness, i.e., no direct influence between them has been found. The status of Pitch centroid, Loudness, and Loudness IQR is uncertain. Depending on the true direction of the Chorusness, MFCC variance and Roughness relations, they may be part of the Chorusness *Markov blanket*, the set of Chorusness' parents, children, and parents of children, which *d*-separates Chorusness from all other variables, or they might be *d*-separated themselves (given the Markov blanket, no influence between these variables and Chorusness) [93].

Also interesting are the more unexpected dependencies. For example, two variables depend directly on the Section position, while Chorusness does not. This may be due to the limitations of the normal distribution by which all variables are modeled; it might not reflect the potentially complex relation of Chorusness variable and Section position. However, the Section position variable does predict Sharpness and Pitch centroid to some extent (dotted edges). A simple regression also shows both variables correlate *positively*, suggesting an increased presence of sharper and higher-pitched sections towards the end of the songs in the *Billboard* corpus.

Finally, the dashed red edges in the diagram indicate dependencies that are most unintuitive. Tentative explanations may be found, but since they have no effect on Chorusness, we will omit such speculations here.

		95% CI	
	β	LL	UL
Sharpness	0.11	0.10	0.13
MFCC variance	0.12	0.09	0.15
Roughness	0.12	0.08	0.16
Pitch salience (×10)	0.04	0.03	0.05
Loudness	0.03	-0.01	0.06
Loudness IQR	-0.33	-0.48	-0.18
Pitch centroid	0.10	0.07	0.12

Table 1.: Results of a multivariate linear regression on the Chorusness' Markov blanket.
CI=confidence interval, LL=lower limit, UL=upper limit.

4.4.6 *Regression*

We ran a regression model to see in more detail how the set of features related to Chorusness predict our variable of interest. Table 1 lists the coefficients of a linear regression on the Chorusness variable and its Markov blanket, i.e. those variables for which a direct dependency with Chorusness is apparent from the model. Since there is no certainty about the exact composition of the Markov blanket, all candidates are included, including Loudness, Loudness IQR and Pitch centroid. Note that, having defined Chorusness as a log odds ratio, this linear regression is in effect a logistic regression on the section's original Chorus probability $p_C \in [0, 1]$.

One can see that all features but the Loudness IQR have positive coefficients. Only loudness has no significant positive or negative correlation. We conclude that, in this model, sections with high Chorusness are sharper and rougher than other sections. Chorus-like sections also feature a slightly higher and more salient pitch, a smaller dynamic range and greater variety in MFCC timbre.

4.4.7 *Validation*

Finally, a classification experiment is performed. It consists of the evaluation of a 2-way classifier that aims to label sections as either 'chorus' or 'non-chorus'. A k-nearest neighbor classifier ($k = 1$) is trained on half of the available sections, and tested on the other half (randomly partitioned). This procedure is repeated 10 times to obtain an average precision, recall and F-measure.

The results confirm the trends found in the PGM: using just the Markov blanket features of table 1, the classifier performs better than random: $F = 0.52$, 95% CI $[0.51, 0.52]$ vs. a maximum random baseline of $F = 0.36$. The classifier also performs better than one that uses all features ($F = 0.48$), or only Loudness and Loudness IQR ($F = 0.48$), the features used in [148].

4.5 CONCLUSIONS

This chapter presents two computational studies into the robust and informative audio descriptors that correlate with the 'chorusness' of sections in pop songs. A selection of existing and novel perceptual and computational features is presented, and applied to two datasets. A small new dataset of early Dutch popular songs, *Dutch50*, is analyzed to reveal that choruses in the dataset have stronger and higher pitch then verses. A larger dataset, the *Billboard* dataset, has been analyzed using a probabilistic graphical model and a measure of Chorusness that is derived from annotations and an inter-annotator confusion matrix. The resulting model was complemented with a regression on the most important variables. The results show that choruses and chorus-like sections are sharper and rougher and, like the pre-1950 Dutch choruses, feature a higher and more salient pitch. They have a smaller dynamic range and greater variety of MFCC-measurable timbre than other sections. These conclusions demonstrate that, despite the challenges of audio corpus analysis presented in the previous chapter, musical insights can be gained from the analyses of readily available datasets. Moreover, having been guided by the desiderata in the previous section, we believe our insights are robust.

The results obtained in a classification experiment do not suggest that our model would read the level of accuracy obtained by the state-of-the-art techniques that incorporate repetition information. However, they demonstrate for the first time that there is a class of complementary musical information that, independently of repetition, can be used to locate choruses. This suggests that our model can be applied to complement existing structure analysis applications, while repetition information and section order can in turn enhance our model of Chorusness for further application in popular music cognition research and audio corpus analysis.