AN ANALYSIS OF CHORUS FEATURES IN POPULAR SONG

Jan Van Balen[1] **John Ashley Burgoyne**[2] **Frans Wiering**[1] **Remco C. Veltkamp**[1]
[1]Utrecht University, Department of Information and Computing Sciences
[2]Universiteit van Amsterdam, Institute for Logic, Language and Computation
{j.m.h.vanbalen,f.wiering,r.c.veltkamp}@uu.nl, j.a.burgoyne@uva.nl

ABSTRACT

This paper presents a computational study of the perceptual and musicological audio features that correlate with the structural function of sections in pop songs, specifically the chorus. Choruses have been described as more prominent, more catchy and more memorable than other sections in a song, yet chorus detection applications have always been primarily based on identifying the most-repeated section in a song. Inspired by cognitive research rather than applied signal processing, this computational analysis compiles a list of robust and interpretable features and models their influence on the 'chorusness' of a collection of song sections from the Billboard dataset. This is done through the unsupervised learning of a probabilistic graphical model. We show that timbre and timbre variety are more strongly related to chorus qualities than harmony and absolute pitch height. A regression and a classification experiment are performed to quantify these relations.

1. INTRODUCTION

1.1 Chorus Analysis

The term *chorus* originates as a designation for those parts of a musical piece that feature a choir or other form of group performance. When solo performance became the norm in popular music, the term chorus was retained to indicate a repeated structural unit of musical form. In terms of musical content, the chorus has been referred to as the 'most prominent', 'most catchy' or 'most memorable' part of a song and 'the site of the more musically distinctive and emotionally affecting material' [4, 10].

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 [6]. On the other hand, as a frequent subject of study in the domain of Music Information Retrieval, the notion of chorus has been shown to correlate with a number of computable descriptors. Yet when studied more closely, 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 [4, 5]. The central research question of this paper is therefore: in which computable properties of popular music are choruses, when compared to other song sections, musically distinct?

The main motivations for a deeper study of the particularities of choruses are two-fold: first, the chorus being a central element of form in popular music, insight may be gained in the popular song as a medium, and conscious as well as unconscious choices in songwriting may be unveiled. Second, as choruses can be related to a catchy and/or memorable quality, to the notion of hooks, and perhaps to a general cognitive salience underlying these aspects, the nature of choruses may indicate some of the musical properties that constitute this salience, prominence or memorability.

This investigation relates to chorus detection as known in Music Information Retrieval, but it does not have the same goal. While chorus detection systems are built to locate the choruses given unsegmented raw audio for a song, this investigation aims to use similar and novel computational methods to improve our understanding of choruses.

1.2 Related Work

Existing work on chorus detection strongly relates 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. An overview of relevant techniques is given by Paulus et al. [13].

Definitions of the chorus in the MIR literature characterize the chorus 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 audio, including RefraiD by Goto and work by Eronen [4, 5]. RefraiD makes use of a scoring function that favors segments C occuring 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 $\frac{1}{4}$ of the way through the song and reoccur near $\frac{3}{4}$, 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.

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) [12, 14]. 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 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 [14]. 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 that this paper aims to address.

The contribution of this paper is to introduce the notion of chorusness, and a statistical model of this measure for MIR applications, for popular music understanding and for popular music perception and cognition, using a novel and rigorous take on corpus-scale audio music analysis.

2. METHODOLOGY

The research question formulated above is addressed by means of a statistical analysis of a selection of music descriptors, computed over a dataset of pop songs. The *Billboard* dataset, described in section 2.1, will be used as the ground truth. The expert structure annotations available for these data allow for parsing audio descriptors, detailed in part 2.2 of this paper, into per-section statistics. The analysis of the resulting variables will then be formalized by learning a probabilistic graphical model from the data, as explained in section 2.3.

2.1 The 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 [2]. The annotations include information about harmony, meter, phrase, and larger musical structure. The Billboard dataset is one of 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 both important commonalities and significant trends in popular music of the period of focus. It includes a wide variety of genres, and suits the goal of drawing broadly-applicable musicological conclusions, making it the best available dataset for analysis of popular music choruses. For the present study, the complete v1.2 release is used (649 songs), and 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 [18]. 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 time), chorus (24%), intro, solo, outro and bridge. The total number of sections, including the unlabeled ones, is 7762.

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

2.2 Perceptual Audio Features

The proposed corpus analysis–centered study requires somewhat different kinds of descriptors than traditionally used in machine-learning applications in Music Information Retrieval. Four important constraints are applied. First, all descriptors are demonstrated correlates of a relevant perceptual or cognitive attribute of music. Second, we favor musically interpretable descriptors. A classic example of an audio feature with demonstrated perceptual correlates but low interpretability are MFCC's [1]. Third, only transparent statistics over these features are considered. Higher order statistical moments and learned codewords can be very informative from an engineering perspective, but amount to highly uninterpretable descriptions. Finally, we also limit the set to a small number of handpicked descriptors since the amount of data required for the proposed analysis grows exponentially with the number of variables. All features are one-dimensional.

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 will make use of the implementation by Pampalk [11]. The model applies outer-ear filtering and a spreading function before computing specific loudness values (N_k in sones) per band k and summing these values over all bands to obtain the total loudness T:

$$S = \max_k(N_k) + 0.15 \cdot \sum_{k \neq \max} N_k \qquad (1)$$

where the factor 0.15 serves as a weighting that emphasizes the strongest band's contribution. 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 will make use of the specific loudnesses N_k as computed by Pampalk [11] and summing as formulated by Peeters [15]:

$$A = 0.11 \times \sum_k g(k) \cdot k \cdot N_k, \quad \text{where} \qquad (2)$$

$$g(k) = \begin{cases} 1 & k < 15 \\ 0.066 \times \exp(0.171 \cdot k) & k > 15 \end{cases} \qquad (3)$$

For every section, we use the *mean* sharpness.

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 consituent partials. We will make use of the MIRToolbox implementation by Lartillot et al. [9], which is based on a model by Plomp and Levelt [16]. Since the roughness feature is very nonlinearly distributed, the roughness feature is summarized for every section by taking its *median*.

MFCC MFCCs are established multidimensional correlates of several aspects of timbre, designed to be maximally independent. Typically around 13 MFCC coefficients are used. In this model, the descriptor of interest is the variety in timbre. This will be 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 [11], and the first component (directly proportional to energy) is discarded.

Chroma Variance Chroma features 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 *Chordino* chroma features distributed along with the Billboard dataset are used here. [2]

In this study, the variety in the section's harmony will be measured by modeling the normalized chroma features $\mathbf{p}$ as a 12-dimensional random variable and estimating a Dirichlet distribution to the total of all of the section's chroma observations. Note that estimating just the total variance, as done for MFCC, would neglect the normalisation constraint on chroma vectors and the dependencies it entails between pitch classes. The Dirichlet distribution $\mathcal{D}_{12}(\boldsymbol{\alpha})$, can be written:

$$f(\mathbf{p}) \sim \mathcal{D}_{12}(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{12} \alpha_k)}{\prod_{k=1}^{12} \Gamma(\alpha_k)} \prod_{k=1}^{12} p_k^{\alpha_k - 1}, \quad (4)$$

where Γ is the Gamma function, and can be seen as a distribution over distributions. We use the sum of the parameters $\boldsymbol{\alpha}$, commonly refered to as the Dirichlet precision s:

$$s = \sum_{k=1}^{12} \alpha_k \quad (5)$$

The Dirichlet precision quantifies the difference between observing the same combination of pitches thoughout the whole section (high precision) and observing many different distributions (low precision) [3]. There is no closed-form formula for s or $\boldsymbol{\alpha}$, but several iterative methods exist that can be applied to obtain a maximum-likelihood estimation (e.g. Newton iteration). Fast fitting can be done using the *fastfit* Matlab toolbox by Minka. [3]

Pitch Salience The notion of pitch salience exists in several contexts. Here, it will refer to the strength of (a discrete set of) pitches, i.e. a measure of the combined strength of a frequency and its harmonics, as in [17]. The *mean* of the strongest (per frame) pitch strength will be computed for every section.

Pitch Centroid The pitch height of polyphonic audio can be defined and computationally approximated in several ways. A predominant fundamental frequency in the classic sense is not always reliably found, especially for the case of polyphonic pop music. We therefore use the more robust *pitch centroid*. We define this as the average pitch height with all pitches weighted by their salience. Note that the pitch salience profile used here spans multiple octaves and exploits 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 inter-quartile range of this measure, represented by the **Pitch Centroid IQR**.

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 and 1.

The descriptors above have proven useful in a variety of other contexts and fall within the constraints listed above. The first three originate in psychoacoustics, the next three descriptors stem from MIR research. The pitch centroid is a novel descriptor designed with robustness in mind.

2.3 Analysis Method

2.3.1 PGM

The resulting descriptors 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. However, modeling all dependencies between a set of variables quickly leads to complex representations that are hard to manage.

Probabilistic graphical models (PGM) are graph-based representations of such complex higher-dimensional distributions that focus on modeling the direct probabilistic interactions between variables [8]. Unlike a correlation matrix, they encode which variables are conditionally dependent, i.e. correlate *given* the state of all other variables, regardless of any indirect effects from other influencing variables. Examples of the use of a PGM in music analysis can be found in [3].

Essential to a PGM is its graph structure. It is typically constructed using prior expert knowledge but, with enough data, can also be learned. 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 [7]. When not all directions are found, a partially directed graph is returned.

Given the limitations of currently available software packages, an important practical requirement for learning the

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

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

graph structure is that the variables follow similar distributions, e.g. all discrete, or all continuous Gaussian. In the analysis in the next section, all data are modeled as continuous. This means that also the **Section Type** variable will have to be modeled as continuous. We do this by introducing the notion of **Chorusness**.

2.3.2 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 with label T being labeled 'chorus' by an independent annotator. In terms of the annotations x_1 and x_2 of two independent listeners, $p_C(T)$ can be written:

$$p_C(T) = \frac{p(x_1 = C|x_2 = T) + p(x_2 = C|x_1 = T)}{2}, \tag{6}$$

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 [18]. The annotators' behaviour can therefore be modeled by means of a confusion matrix $M(T_1, T_2) \in [0, 1]^{22\times 22}$:

$$M(T_1, T_2) = f(x_1 = T_1 \cap x_2 = T_2) \tag{7}$$

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

$$M^\star = \frac{M + M^T}{2} \tag{8}$$

From here we can obtain the empirical Chorus Probability:

$$p_C(T) = \frac{M^\star(T, C)}{\sum_k M^\star(T, k)} \in [0, 1]. \tag{9}$$

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:

$$C(T) = \log\left(\frac{p_C(T)}{1 - p_C(T)}\right) \in (-\infty, \infty). \tag{10}$$

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

2.3.3 Implementation

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

$$S = \frac{s^\lambda - 1}{\lambda} = 1 - \frac{1}{s} \tag{11}$$

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 adjecent nodes and testing for conditional independence given all possible subsets of the remaining graph. [4] The procedure is applied to the 6462×12 dataset, with 'conservative' estimation of directionalities, i.e. no directionality is forced onto the edges where no V-structures were found indicating a specific direction.

3. ANALYSIS RESULTS

The resulting graphical model is shown in Figure 1. It is obtained with $p < 3.5 \times 10^{-5}$, the significance level required to bring the overall 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:

$$\alpha_{\text{CI}} = 1 - (1 - \alpha_{\text{PGM}})^{1/n} \approx \frac{\alpha_{\text{PGM}}}{n} \tag{12}$$

with $\alpha_{\text{PGM}} \ll 1$ (here 0.05) and n the number of tests performed (~ 1500). Note that $p \approx 10^{-5}$ is a conservative parameter setting for an individual test. As a result, we may choose to view the model as a depiction of dependencies rather than independencies, since the latter may always be present at a lower significance than required by the α.

3.1 Discussion

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. The importance of this distinction can be debated.

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).

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

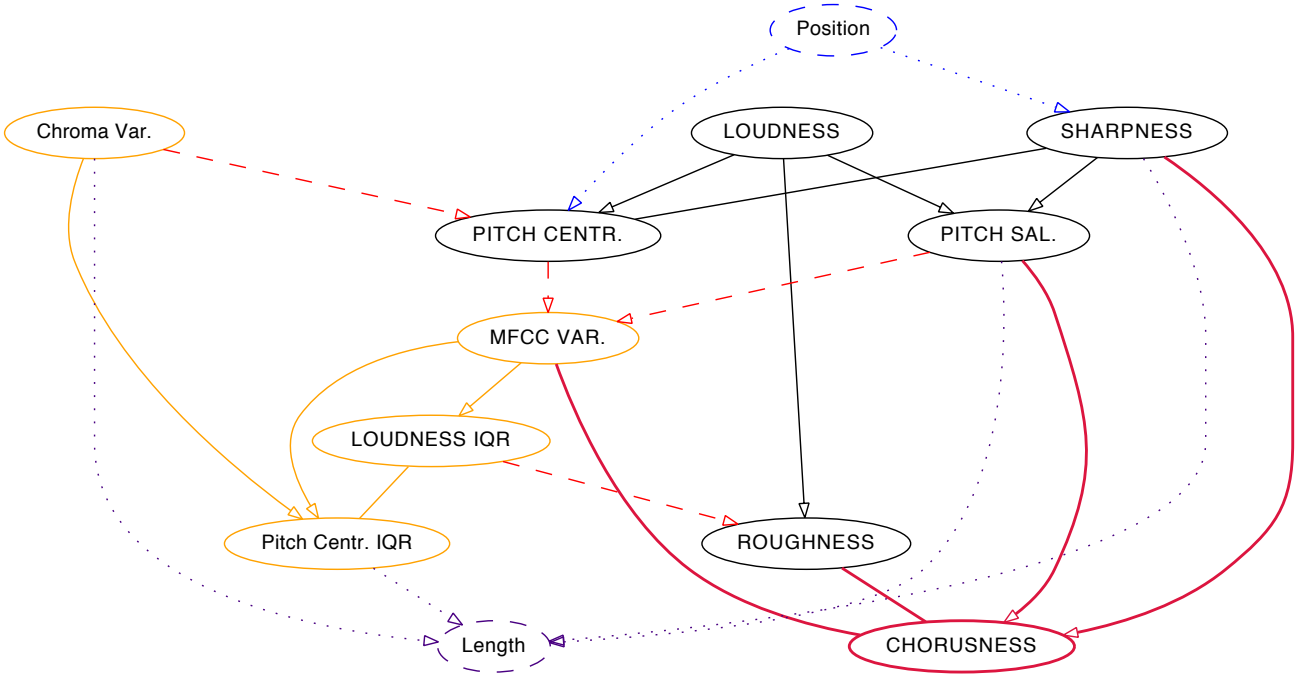

Figure 1. Graphical model of the 11 analyzed perceptual features and chorusness variable C. $\alpha_{\text{PGM}} = 0.05$.

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 directionality 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 [8].

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 is fair to say that it might not reflect the potentially complex relation of Chorusness variable and Section Position. However, the Position variable does predict Sharpness and Pitch Centroid to some extent (dotted edges). A simple regression also shows both variables correlate *positively*. This suggests some kind of overtime intensification along the frequency dimension exists in the songs of 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.

3.2 Regression

We are more interested to see in more detail how the set of Chorusness-related features 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. Note that, having defined Chorusness as a log odds ratio, this linear regression is *de facto* a common form of logistic regression on the section's original Chorus Probability $p_C \in [0, 1]$.

		95% CI	
	Coeff.	LL	UU
Sharpness	0.11	0.10	0.13
MFCC variance	0.12	0.09	0.15
Roughness	0.12	0.08	0.16
Pitch Salience ($\times 10$)	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 ($p < 10^{-15}$ for all coeff.). CI=confidence interval, LL=lower limit, UU=upper limit.

One can see that all features but the Loudness IQR have positive coefficients. We conclude that, in this model, sections with high Chorusness are louder, 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. VALIDATION

Finally, a validating 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 neighbour 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 are positive: 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 [14].

5. CONCLUSIONS

This paper presents a computational study of the musically interpretable and robust audio descriptors that correlate with the 'chorusness' of sections in pop songs. A selection of existing and novel perceptual and computational features is presented. The set has been analyzed using a probablistic 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 louder, sharper and more rough, and feature a higher and more salient pitch, a smaller dynamic range and greater variety of MFCC-measurable timbre than other sections.

The results obtained in a validating classification experiment show that our model does not reach the level of accuracy obtained by the state of the art techniques that incorporate repetition information. However, it demonstrates 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.

6. ACKNOWLEDGEMENTS

This research is supported by the NWO CATCH project COGITCH (640.005.004), and the FES project COMMIT/.

7. REFERENCES

[1] J. J. Aucoutourier and E. Bigand: "Mel Cepstrum & Ann Ova: The Difficult Dialog between MIR and Music Cognition," *Proc. of the Int. Society for Music Information Retrieval Conf.*, pp. 397–402, 2012.

[2] J. A. Burgoyne, J. Wild and I. Fujinaga: "An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis," *Proc. of the Int. Society for Music Information Retrieval Conf.*, pp. 633–638, 2011.

[3] J. A. Burgoyne: *Stochastic Processes and Database-driven Musicology*, PhD Thesis, McGill University, Montréal, Québec, 2011.

[4] A. Eronen and F. Tampere: "Chorus Detection with Combined Use of MFCC and Chroma Features and Image Processing Filters", *Proc. of the Int. Conf. Digital Audio Effects*, 2007.

[5] M. Goto: "A Chorus-section Detecting Method for Musical Audio Signals," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 437–440, 2003.

[6] H. Honing: "Lure(d) into listening: The Potential of Cognition-based Music Information Retrieval," *Empirical Musicology Review*, Vol. 5, No. 4, 2010.

[7] M. Kalisch et al.: "Causal Inference Using Graphical Models with the R Package pcalg," *Journal of Statistical Software*, Vol. 47, No. 11, pp. 1–26, 2012.

[8] D. Koller and N. Friedman: *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[9] O. Lartillot and P. Toiviainen: "MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio," *Proc. of the Int. Society of Music Information Retrieval Conf.*, pp. 127–130, 2007.

[10] R. Middleton: "Form," in: *Continuum Encyclopedia of Popular Music of the World*, eds. J. Shepherd, D. Horn, D. Laing, Continuum Int. Publishing Group, 2003.

[11] E. Pampalk: "A Matlab Toolbox to Compute Similarity from Audio," *Proc. of the Int. Society for Music Information Retrieval Conf.*, 2004.

[12] J. Paulus and A. Klapuri: "Labelling the Structural Parts of a Music Piece with Markov Models," in: *Computer Music Modeling and Retrieval: Genesis of Meaning in Sound and Music*, Berlin: Springer, pp. 166-176, 2009.

[13] J. Paulus et al.: "Audio-based Music Structure Analysis," *Proc. of the 11th Int. Society for Music Information Retrieval Conf.*, pp. 625–636, 2010.

[14] J. Paulus: "Improving Markov Model-Based Music Piece Structure Labelling with Acoustic Information," *Proc. of the 11th Int. Society for Music Information Retrieval Conf.*, pp. 303–308, 2010.

[15] G. Peeters: *A Large Set of Audio Features for Sound Description in the CUIDADO Project*, Tech. Rep., IRCAM, Paris, France, 2004.

[16] R. Plomp and W. J. M. Levelt: "Tonal Consonance and Critical Bandwidth," *Journal of the Acoustical Society of America*, Vol. 38, No. 4, pp. 548–560, 1965.

[17] J. Salamon and E. Gomez: "Melody Extraction from Polyphonic Music Signals Using Pitch Contour Characteristics," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 20, No. 6, pp. 1759–70, 2010.

[18] J. Smith et al.: "Design and Creation of a Large-Scale Database of Structural Annotations," *Proc. of the Int. Society for Music Information Retrieval Conf.*, pp. 555–560, 2011.