A DISCRETE MIXTURE MODEL FOR CHORD LABELLING

Matthias Mauch and Simon Dixon
Queen Mary, University of London,
Centre for Digital Music.
matthias.mauch@elec.qmul.ac.uk

ABSTRACT

Chord labels for recorded audio are in high demand both as an end product used by musicologists and hobby musicians and as an input feature for music similarity applications. Many past algorithms for chord labelling are based on chromagrams, but distribution of energy in chroma frames is not well understood. Furthermore, non-chord notes complicate chord estimation. We present a new approach which uses as a basis a relatively simple chroma model to represent short-time sonorities derived from melody range and bass range chromagrams. A chord is then modelled as a mixture of these sonorities, or subchords. We prove the practicability of the model by implementing a hidden Markov model (HMM) for chord labelling, in which we use the discrete subchord features as observations. We model gamma-distributed chord durations by duplicate states in the HMM, a technique that had not been applied to chord labelling. We test the algorithm by five-fold cross-validation on a set of 175 hand-labelled songs performed by the Beatles. Accuracy figures compare very well with other state of the art approaches. We include accuracy specified by chord type as well as a measure of temporal coherence.

1 INTRODUCTION

While many of the musics of the world have developed complex melodic and rhythmic structures, Western music is the one that is most strongly based on harmony [3]. A large part of harmony can be expressed as chords. Chords can be theoretically defined as sets of simultaneously sounding notes, but in practice, including all sounded pitch classes would lead to inappropriate chord labelling, so non-chord notes are largely excluded from chord analysis. However, the question which of the notes are non-chord notes and which actually constitute a new harmony is a perceptual one, and answers can vary considerably between listeners. This has also been an issue for automatic chord analysers from symbolic data [16]. Flourishing chord exchange websites [1] prove the sustained interest in chord labels of existing music. However, good labels are very hard to find, arguably due to the tediousness of the hand-labelling process as well as the lack of expertise of many enthusiastic authors of transcriptions. While classical performances are generally based on a score or tight harmonic instructions which result in perceived chords, in Jazz and popular music chords are often used as a kind of recipe, which is then realised by musicians as actually played notes, sometimes rather freely and including a lot of non-chord notes. Our aim is to translate performed pop music audio back to the chord recipe it supposedly has been generated from (*lead sheet*), thereby imitating human perception of chords. A rich and reliable automatic extraction could serve as a basis for accurate human transcriptions from audio. It could further inform other music information retrieval applications, e.g. music similarity. The most successful past efforts at chord labelling have been based on an audio feature called the *chromagram*. A chroma frame, also called pitch class profile (PCP), is a 12-dimensional real vector in which each element represents the energy of one pitch class present in a short segment (frame) of an audio recording. The matrix of the chroma frame columns is hence called chromagram. In 1999, Fujishima [5] introduced the chroma feature to music computing. While being a relatively good representation of some of the harmonic content, it tends to be rather prone to noise inflicted by transients as well as passing/changing notes. Different models have been proposed to improve estimation, e.g. by tuning [6] and smoothing using hidden Markov models [2, 11]. All the algorithms mentioned use only a very limited chord vocabulary, consisting of no more than four chord types, in particular excluding silence (*no chord*) and dominant 7th chords. Also, we are not aware of any attemps to address chord fragmentation issues.

We present a novel approach to chord modelling that addresses some of the weaknesses of previous chord recognition algorithms. Inspired by word models in speech processing we present a chord mixture model that allows a chord to be composed of many different sonorities over time. We also take account of the particular importance of the bass note by calculating a separate bass chromagram and integrating it into the model. Chord fragmentation is reduced using a duration distribution model that better fits the actual chord duration distribution. These characteristics approximate theoretic descriptions of chord progressions better than

[1] e.g. http://www.chordie.com/

previous approaches have.

The rest of this paper is organised as follows. Section 2 explains the acoustical model we are using. Section 3 describes the chord and chord transition models that constitute the hierarchical hidden Markov model. Section 4 describes how training and testing procedures are implemented. The result section 5 reports accuracy figures. Additionally, we introduce a new scoring method. In section 6 we discuss problems and possible future developments.

2 ACOUSTIC MODEL

2.1 Melody and Bass Range Chromagrams

We use mono audio tracks at a sample rate of 44.1 kHz and downsample them to 11025 kHz after low-pass filtering. We calculate the short-time discrete Fourier transform for windows of 8192 samples ($\approx 0.74s$) multiplied by a Hamming window. The hop-size is 1024 samples ($\approx 0.09s$), which corresponds to an overlap of $7/8$ of a frame window. In order to map the Fourier transform at frame t to the log-frequency (pitch) domain magnitudes $Q_k(t)$ we use the constant Q transform code written by Harte and Sandler [6]. Constant Q bin frequencies are spaced $33\frac{1}{3}$ cents (a third of a semitone) apart, ranging from 110 Hz to 1760 Hz (four octaves), i.e. the k^{th} element of the constant Q transform Q_k corresponds to the frequency

$$2^{\frac{k-1}{36}} \cdot 110\,\text{Hz}, \tag{1}$$

where $k \in 1, \ldots, (4 \cdot 36)$. In much the same way as Peeters [15], we smooth the constant Q transform by a median filter in the time direction (5 frames, $\approx 0.5s$), which has the effect of attenuating transients and drum noise.

For every frame t we wrap the constant Q magnitudes $Q(t)$ to a chroma vector $\mathbf{y}^*(t)$ of 36 bins by simply summing over bins that are an octave apart,

$$y^*_j(t) = \sum_{i=1}^{4} |Q_{36\cdot(i-1)+j}(t)|, \quad j = 1, \ldots, 36. \tag{2}$$

Similar to Peeters [15], we use only the strongest of the three possible sets of 12 semitone bins, e.g. $(1, 4, 7, \ldots, 34)$, thus “tuning” the chromagram and normalise the chroma vector to sum to 1,

$$y_k(t) = y^*_{3k+\nu}(t) \bigg/ \sum_{i=1}^{12} y^*_{3i+\nu}(t)\ , \tag{3}$$

where $\nu \in \{0, 1, 2\}$ indicates the subset chosen to maximise $\sum_t \sum_k y^*_{3k+\nu}(t)$. A similar procedure leads to the calculation of the bass range chromagrams. The frequency range is 55 Hz to 220 Hz. The number of constant Q bins per semitone is 1, not 3. We linearly attenuate the bins at the frequency range borders, mainly to prevent a note just above the bass frequency range from leaking into the bass range.

2.2 Data

Harte has provided chord transcriptions for 180 Beatles recordings [7], the entirety of the group’s 12 studio albums. Some of the songs have ambiguous tuning and/or do not pertain to Western harmonic rules. We omit 5 of these songs [2]. In a classification step similar to the one described by Mauch et al. [13] we map all chords to the classes *major*, *minor*, *dominant*, *diminished*, *suspended*, and *no chord* (which account for more than 94% of the frames) as well as *other* for transcriptions that do not match any of the classes. We classify as *dominant* the so-called dominant seventh chords and others that feature a minor seventh. We exclude the chords in the *other* class from all further calculations. Hence, the set of chords has $n = 12 \times 6$ elements.

2.3 Subchord Model

We want to model the sonorities a chord is made up of mentioned in Section 1 and call them *subchords*. Given the data we have, it is convenient to take as set of subchords just the set of chords introduced in the previous paragraph, denoting them S_i, $i = 1, ..., n$. In this way, we have a heuristic that allows us to estimate chroma profiles for every subchord [3]; in fact, for every such subchord S_i we use the ground truth labels G_t to obtain all positive examples

$$\mathbf{Y}_i = \{\mathbf{y}_t : G_t = S_i\}$$

and calculate the maximum likelihood parameter estimates $\hat{\theta}_i$ of a Gaussian mixture with three mixture components by maximising the likelihood

$$\prod_{\mathbf{y} \in \mathbf{Y_i}} L(\theta_i|\mathbf{y}).$$

Parameters are estimated using a MATLAB implementation of the EM algorithm by Wong and Bouman [4] with the default initialisation method. From the estimates $\hat{\theta}_i$, we obtain a simple subchord score function

$$p(S_i|\mathbf{y}) = \frac{L(\hat{\theta}_i, \mathbf{y})}{\sum_j L(\hat{\theta}_j, \mathbf{y})} \tag{4}$$

and hence a subchord classification function

$$s(\mathbf{y}) := \underset{S_i}{\operatorname{argmax}}\, p(S_i|\mathbf{y}) \in \{S_1, ..., S_n\}. \tag{5}$$

These will be used in the model with no bass information.

[2] *Revolution 9* (collage), *Love You Too* (Sitar-based), *Wild Honey Pie* (tuning issues), *Lovely Rita* (tuning issues), *Within You Without You* (Sitar-based)

[3] We only fit one Gaussian mixture for each chord type (i.e. *major*, *minor*, *diminished*, *dominant*, *suspended*, and *no chord*) by rotating all the relevant chromagrams, see [15]).

[4] http://web.ics.purdue.edu/˜wong17/gaussmix/gaussmix.html

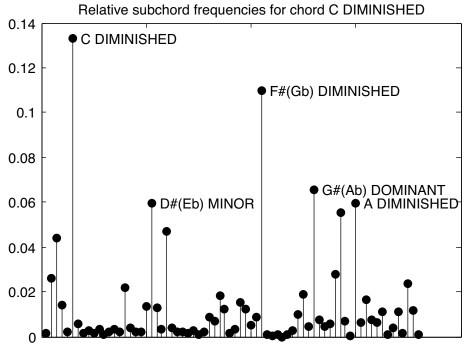

Figure 1. Example of subchord feature relative frequencies $b_{S|C}$ for the *diminished* chord. The five most frequent features are labelled. The subchord corresponding to C *diminished* most likely to be the best-fitting feature is indeed C *diminished*.

2.4 Subchord Model including Bass

In order to model the bass from the bass range chromagrams, every subchord S_i has a set $B_i \subset \{1, \ldots, 12\}$ of valid pitch classes coinciding with chord notes. The score for the bass range chromagram of subchord S_i at a bass chroma frame $\mathbf{y}^{\text{b}}$ is the maximum value

$$p^{\text{b}}(S_i|\mathbf{y}^{\text{b}}) = \frac{\max_{j \in B_i} \{y_j^{\text{b}}\}}{\sum_k \max_{j \in B_k} \{y_j^{\text{b}}\}} \in [0, 1], \tag{6}$$

the bass chromagram assumes in any of the pitch classes in B_i, b stands for *bass range*.

In order to obtain a model using both melody range and bass range information the two scores are combined to a single score

$$p(S_i|\mathbf{y}, \mathbf{y}^{\text{b}}) = p^{\text{b}}(S_i|\mathbf{y}^{\text{b}}) \cdot p(S_i|\mathbf{y}). \tag{7}$$

Analogous to Equation 5 we obtain a second subchord classification function

$$s(\mathbf{y}, \mathbf{y}^{\text{b}}) := \operatorname*{argmax}_{S_i} p(S_i|\mathbf{y}, \mathbf{y}^{\text{b}}) \in \{S_1, ..., S_n\}. \tag{8}$$

2.5 Discrete HMM Observations

We discretise the chroma data $\mathbf{y}$ (and $\mathbf{y}^{\text{b}}$) by assigning to each frame with chroma $\mathbf{y}$ the relative subchord, i.e. $s(\mathbf{y}, \mathbf{y}^{\text{b}})$, or $s(\mathbf{y})$ depending on whether we want to consider the bass chromagrams or not. That means that in the HMM, the only information about a frame $\mathbf{y}$ we keep is which subchord fits best.

3 LANGUAGE MODEL

In analogy to speech processing the high-level processing in our model is called language modelling, although the language model we are employing is a hidden Markov model (HMM, see, e.g. [9]). Its structure can be described in terms of a chord model and a chord transition model.

3.1 Chord Model

The chord model represents one single chord over time. As we have argued above, a chord can generate a wealth of very different subchords. The HMM takes the categorical data $s(\mathbf{y}) \in \{S_1, \ldots, S_n\}$ as observations, which are estimations of the subchords. From these, we estimate the *chords*. The chords $C_1, \ldots, C_n$ take the same category names (C *major*, C# *major*,...) as subchords, but describe the perceptual concept rather than the sonority [5] . Given a chord C_i the off-line estimation of its emission probabilities consists of estimating the conditional probabilities

$$P(C_i|S_j),\ i, j \in 1, \ldots, n \tag{9}$$

of the chord C_i conditional on the subchord being S_j. The maximum likelihood estimator is simply the relative conditional frequency

$$b_{i|k} = \frac{|\{t : s(\mathbf{y}_t) = S_i \wedge C_k = G_t\}|}{|\{t : C_k = G_t\}|}, \tag{10}$$

where G_t is the ground truth label at t. These estimates are the (discrete) emission distribution in the hidden Markov model. A typical distribution can be seen in Figure 1, where C_k is a C *diminished* chord.

In hidden Markov models, state durations follow an exponential distribution, which has the undesirable property of assigning the majority of probability mass to short durations as is shown in Figure 2. The true distribution of chord durations is very different (solid steps), with no probability assigned to very short durations, and a lot between one and three seconds. To circumvent that problem we apply a variant of the technique used by Abdallah et al. [1] and model one chord by a left-to-right model of three hidden states with identical emission probabilities $b_{i|k}$. The chord duration distribution is thus a sum of three exponential random variables with parameter λ, i.e. it is gamma-distributed with shape parameter $k = 3$ and scale parameter λ. Hence, we can use the maximum likelihood estimator of the scale parameter λ of the gamma distribution with fixed k:

$$\hat{\lambda} = \frac{1}{k}\bar{d}_N, \tag{11}$$

where $\bar{d}_N$ is the sample mean duration of chords. The obvious differences in fit between exponential and gamma modelling are shown in Figure 2. Self-transitions of the states in the left-to-right model within one chord will be assigned probabilities $1 - 1/\lambda$ (see also Figure 3).

[5] In fact, the subchords could well be other features, which arguably would have made the explanation a little less confusing.

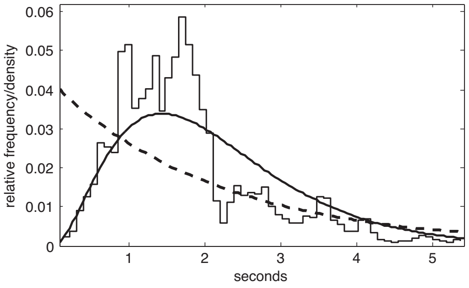

Figure 2. Chord duration histogram (solid steps) and fitted gamma density (solid curve) with parameters $\hat{\gamma}$ and $k = 3$ used in our model. Exponential density is dashed.

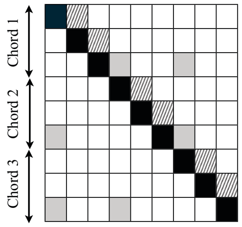

Figure 3. Non-ergodic transition matrix of a hypothetical model with only three chords. White areas correspond to zero probability. Self-transitions have probability $1 - 1/\hat{\lambda}$ (black), inner transitions in the chord model have probability $1/\hat{\lambda}$ (hatched), and chord transitions (grey) have probabilities estimated from symbolic data.

3.2 Chord Transition Model

We use a model that in linguistics is often referred to as a bigram model [9]. For our case we consider transition probabilities

$$P(C_{k_2}|C_{k_1}) \tag{12}$$

employing the estimates $\{a'_{k_1 k_2}\}$ derived from symbolic data smoothed by

$$a_{k_1 k_2} = a'_{k_1 k_2} + \max_{k_1,k_2}\{a'_{k_1 k_2}\}. \tag{13}$$

increasing probability mass for rarely seen chord progressions. The chord transition probabilities are symbolised by the grey fields in Figure 3. Similar smoothing techniques are often used in speech recognition in order not to under-represent word bigrams that appear very rarely (or not at all) in the training data [12].

The initial state distribution of the hidden Markov model is set to uniform on the starting states of the chord, whereas we assign zero to the rest of the states.

4 IMPLEMENTATION

4.1 Model Training

We extract melody range and bass range chromagrams for all the songs in the Beatles collection as described in Section2.1. The four models that we test are as follows:

no bass, no duplicate states	no bass, duplicate states
bass, no duplicate states	bass, duplicate states

We divide the 175 hand-annotated songs into five sets, each spanning the whole 12 albums. For each of the four models we performe a five-fold cross-validation procedure by using one set in turn as a test set while the remaining four are used to train subchord, chord and chord transition models as described in sections 2.3 and 3.1.

4.2 Inference

For a given song from the respective test set, subchord features for all frames are calculated, thus obtaining a feature sequence $s(\mathbf{y}_t)$, $t \in T_{\text{song}}$, and the resulting emission probability matrix is

$$B_k(y_t) = b_{s(y_t)|k}, \tag{14}$$

where $b_{s(y_t)|k} = b_{i|k}$ with $i : S_i = s(y_t)$. In order to reduce the chord vocabulary for this particular song we perform a simple local chord search: B is convolved with a 30 frame long Gaussian window, and only those chords that assume the maximum in the convolution at least once are used. This procedure reduces the number of chords dramatically, from 72 to usually around 20, resulting in a significant performance increase. We use Kevin Murphy's implementation [6] of the Viterbi algorithm to decode the HMM by finding the most likely complete chord sequence for the whole song.

5 RESULTS

We calculate the accuracy for the set of chord classes. As we have six chord classes (or types), rather than two [11] or three [10] we decided to additionally provide results in which *major*, *dominant*, and *suspended* chords are merged. The calculation of accuracy is done by dividing summed duration of correctly annotated frames by the overall duration of the song collection. Similarly, in the case of one particular chord type (or song), this has been done by dividing the summed duration of correctly annotated frames of that chord type (or song) by the duration of all frames pertaining to that chord type (or song).

[6] http://www.cs.ubc.ca/˜murphyk/Software/HMM/hmm.html

5.1 Song-Specific Accuracy

It is obvious that any chord extraction algorithm will not work equally well on all kinds of songs. Table 1 shows overall accuracy figures in both the merged and full evaluation mode for all four models. The models including bass information perform slightly better, though not significantly, with a mean chord recognition rate (averaged over songs) of 66.84% / 51.6% in the case of merged / full evaluation modes. The use of duplicate states has very little effect on the accuracy performance.

		without bass		with bass	
		std.	dupl.	std.	dupl.
merged	mean	64.74	64.96	66.46	**66.84**
	std. deviation	11.76	13.21	11.59	13.00
	max	86.35	89.15	86.99	88.81
full	mean	49.87	49.37	**51.60**	51.17
	std. deviation	13.70	14.85	13.93	15.65
	max	79.55	82.19	78.82	81.93

Table 1. Accuracy with respect to songs. Full and merged refer to the evaluation procedures explained in Section 5. The labels "without bass" and "with bass" denote if information from the bass chromagrams has been used or not, whereas "dupl." denotes the model in which the duplicated states have been used (see Section 3).

5.2 Total and Chord-specific Accuracy

Our top performance results (50.9 % for full evaluation mode, 65.9 % for merged evaluation mode) lie between the top scoring results of Lee and Slaney [11] (74 %) and Burgoyne et al. [4] (49 %). This is encouraging as we model more chord classes than Lee and Slaney [11], which decreases accuracy for either of the classes, and their figures refer to only the first two Beatles albums, which feature mainly *major* chords. Unfortunately, we cannot compare results on individual chords. We believe that such an overview is essential because some of the chord types appear so rarely that disregarding them will increase total accuracy, but delivers a less satisfying model from a human user perspective.

5.3 Fragmentation

For a human user of an automatic transcription not only the frame-wise overall correctness of the chord labels will be of importance, but also—among others properties—the level of fragmentation, which would ideally be similar to the one in the ground truth. As a measure for fragmentation we used the relative number of chord labels in the full evaluation mode. One can see in Table 3, the gamma duration modelling has been very successful in drastically reducing the fragmentation of the automatic chord transcription. This sheds a new light on the results as presented in Tables 1 and 2: the new duration modelling retains the level of accuracy but reduces fragmentation.

		without bass		with bass	
		std.	dupl.	std.	dupl.
merged	**total**	63.85	64.04	65.59	**65.91**
	major (merged)	70.31	72.04	72.58	**74.43**
	minor	48.57	43.93	**50.27**	45.63
	diminished	**14.63**	13.22	11.51	10.35
	no chord	**34.58**	27.42	25.59	19.48
full	**total**	49.17	48.64	**50.90**	50.37
	major	52.16	52.92	54.56	**55.45**
	minor	48.57	43.93	**50.27**	45.63
	dominant	44.88	46.42	**46.51**	46.42
	diminished	**14.63**	13.22	11.51	10.35
	suspended	**16.61**	11.04	13.22	9.04
	no chord	**34.58**	27.42	25.59	19.48

Table 2. Accuracy: Overall relative duration of correctly recognised chords, see also Table 1.

	without bass		with bass	
	std.	dupl.	std.	dupl.
fragmentation ratio	1.72	**1.12**	1.68	1.13

Table 3. Fragmentation.

6 DISCUSSION

6.1 Different Subchord Features

In the model presented in this paper, the subchord features coincide with the chords and emission distributions are discrete. This is not generally necessary, and one could well imagine trying out different sets of features, be they based on chromagrams or not. Advances in multi-pitch estima-

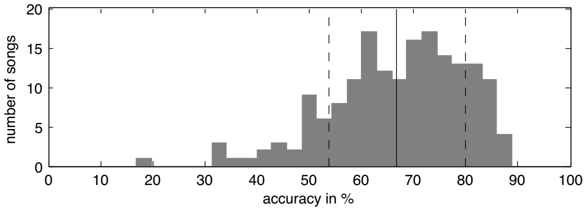

Figure 4. Histogram of recognition accuracy by song in the model using both gamma duration modelling and bass information, merged *major*, *minor*, and *suspended* chords, with mean and standard deviation markers.

tion[7] may make it feasible to use features more closely related to the notes played.

6.2 Hierarchical Levels and Training

While our duration modelling is a very simple form of hierarchical modelling, additional approaches are conceivable. Modelling song sections is promising because they could capture repetition, which is arguably the most characteristic parameter in music [8, p. 229]. Another option is key models, and a combination of the algorithms proposed by Noland and Sandler [14] and Lee and Slaney [11] is likely to improve recognition and enable key changes as part of the model. Such higher level models are needed to make on-line training of transition probabilities sensible as otherwise frequent transitions will be over-emphasised.

7 CONCLUSIONS

We have devised a new way of modelling chords, based on the frequency of *subchords*, chord-like sonorities that characterise a chord by their frequency of occurrence. A hidden Markov model based on this chord model has been implemented to label chords from audio with 6 chord classes (resulting in an overall vocabulary of 6 × 12 chords), while previous approaches never used more than four. The algorithm has shown competitive performance in five-fold cross-validation on 175 Beatles songs, the largest labelled data set available. In addition to the chord model we used a bass model, and more sophisticated state duration modelling. The use of the latter results in a reduction of the fragmentation in the automatic trancription while maintaining the level of accuracy. We believe that the novelties presented in this paper will be of use for future chord labelling algorithms, yet improvement in feature and model design provide plenty of room for improvement.

References

[1] Samer Abdallah, Mark Sandler, Christophe Rhodes, and Michael Casey. Using duration models to reduce fragmentation in audio segmentation. *Machine Learning*, 65:485–515, 2006.

[2] Juan P. Bello and Jeremy Pickens. A Robust Mid-level Representation for Harmonic Content in Music Signals. In *Proc. ISMIR 2005, London, UK*, 2005.

[3] Herbert Bruhn. *Allgemeine Musikpsychologie*, volume 1 of *VII Musikpsychologie*, chapter 12. Mehrstimmigkeit und Harmonie, pages 403–449. Hogrefe, Göttingen, Enzyklopädie der Psychologie edition, 2005.

[4] John Ashley Burgoyne, Laurent Pugin, Corey Kereliuk, and Ichiro Fujinaga. A Cross-Validated Study of Modelling Strategies for Automatic Chord Recognition in Audio. In *Proceedings of the 2007 ISMIR Conference, Vienna, Austria*, 2007.

[5] Takuya Fujishima. Real Time Chord Recognition of Musical Sound: a System using Common Lisp Music. In *Proceedings of ICMC 1999*, 1999.

[6] Christopher Harte and Mark Sandler. Automatic Chord Identifcation using a Quantised Chromagram. In *Proceedings of 118th Convention*. Audio Engineering Society, 2005.

[7] Christopher Harte, Mark Sandler, Samer A. Abdallah, and Emilia Gomez. Symbolic representation of musical chords: A proposed syntax for text annotations . In *Proc. ISMIR 2005, London, UK*, 2005.

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

[9] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts, 1997.

[10] Kyogu Lee and Malcolm Slaney. Acoustic Chord Transcription and Key Extraction From Audio Using Key-Dependent HMMs Trained on Synthesized Audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2), February 2008.

[11] Kyogu Lee and Malcolm Slaney. A Unified System for Chord Transcription and Key Extraction Using Hidden Markov Models. In *Proceedings of the 2007 ISMIR Conference, Vienna, Austria*, 2007.

[12] Christopher D. Manning and Hinrich Schütze. *Foundations of Natural Language Processing*. MIT Press, 1999.

[13] Matthias Mauch, Simon Dixon, Christopher Harte, Michael Casey, and Benjamin Fields. Discovering Chord Idioms through Beatles and Real Book Songs. In *ISMIR 2007 Conference Proceedings, Vienna, Austria*, 2007.

[14] Katy Noland and Mark Sandler. Key Estimation Using a Hidden Markov Model. In *Proceedings of the 2006 ISMIR Conference, Victoria, Canada*, 2006.

[15] Geoffroy Peeters. Chroma-based estimation of musical key from audio-signal analysis. In *ISMIR 2006 Conference Proceedings, Victoria, Canada*, 2006.

[16] David Temperley and Daniel Sleator. Modeling Meter and Harmony: A Preference-Rule Approach. *Computer Music Journal*, 25(1):10–27, 1999.

[7] e.g. http://www.celemony.com/cms/index.php?id=dna