AUDIO CHORD LABELING BY MUSIOLOGICAL MODELING AND BEAT-SYNCHRONIZATION

Björn Schuller, Benedikt Hörnler, Dejan Arsic, and Gerhard Rigoll

Institute for Human-Machine Communication,
Technische Universität München, Germany
schuller@tum.de

ABSTRACT

Automatic labeling of chords in original audio recordings is challenging due to heavy acoustic overlay by melody and percussion sections, detuning and arpeggios that demand for a measure-grid to assign notes to chords. Further chord labeling benefits from contextual information. In this respect we suggest applying an HMM framework incorporating a musiological model trained on 16k songs and synchronization with the measure grid by IIR comb-filter banks for tempo detection, meter recognition, and on-beat tracking. Features base on pitch-tuned chromatic information. Extensive evaluation on 11k chords of 7h of MP3 compressed popular music demonstrates effectiveness over traditional correlation analysis and single measure classification by Support Vector Machines.

Index Terms— Music, Hidden Markov models, Feature extraction

1. INTRODUCTION

The automatic recognition and transcription of musical chord progressions possesses a wide variety of applications: musicians can automatically transcribe their progression while jamming, or they can be offered a plug-in to media players to show them the current chord for play along. But chord knowledge can also be used as meta-information in many other Music Information Retrieval tasks, such as genre recognition (e.g. Jazz having many II-V-I successions, while e.g. Blues has many I-IV-V7s), musical mood recognition (e.g. ratio of major/minor or 7/maj7 chords), key recognition or structure analysis e.g. for chorus retrieval [1]. Also, DJs can be provided with automatic synthesis of additional fitting notes as sub-basses or arpeggios, or tools that blend music at matched key/chord. One final application is music similarity analysis or finding of plagiarism (e.g. chord progression of *Johann Pachelbel's "Canon in D"* (*"Canon per 3 Violini e Basso"*), which is found in multiple contemporary popular pieces, such as *"Go West"*, *"Streets of London"*, *"All Together Now"*, *"Basket Case"*, *"Big City Life"* or *"Volverte a Ver"*). To save cost-intensive and partly not feasible manual labeling, we introduce a beat-synchronous and data-driven approach in the ongoing. Already early works on chord recognition [2] use pitch class profiles. Using Hidden Markov Models (HMM) was proven beneficial e.g. in [3, 4]. It is also well known that context modeling improves recognition rates [5]. In this respect we show results on uniting these findings and add by highly reliable beat-tracking and a musiological model trained on a large corpus of 16k songs to show reachable results on a database of mixed original recordings with respect to interpret and style.

The paper is structured as follows: first we introduce our database or original audio recordings in sec. 2, then in sec. 3 we introduce our musiological model, in sec. 4 we shortly explain our tempo and down-beat detection. In sec. 5 we discuss the acoustic features before explaining the actual chord labeling process in sec. 6 and presentation of results and conclusion in sec. 7 and sec. 8.

2. CHORD DATABASE

In order to have sufficient data for machine learning and testing, we annotated a total of 100 musical pieces that cover a good selection of typically aired pop and rock music with the tempo in bpm, the key, and each chord. As ground truth reference original scores were used. The alignment was carried out by three experienced musicians. 64 different artists are comprised. On average, 1.6 pieces per artist are used, however, only 18 artists are found more than once in the set: the highest number of songs per artist resembles 5 for *Delta Goodrem*, *James Blunt*, *Robbie Williams*, followed by *Celine Dion*, *Coldplay* and *Enya* with 4 songs, each, *Bon Jovi*, *Bryan Adams*, *Cher* with 3, each, and *All Saints*, *Backstreet Boys*, *Britney Spears*, *Keane*, *Phil Collins*, *Roxette*, and *The Corrs* with 2, each. These pieces all have constant tempo. The list of songs can be found at [6]. The original recordings are compressed to 128 kbit/s MP3 for the oncoming tests. The total playtime resembles 6h 58min 12sec, and 10,702 bars are contained. This set is referred to as *Chord Recognition Database*, respectively *ChoRD*.

The chords have been annotated in the 7 main classes: major (Maj), minor (Min), Suspended (Sus2, Sus4), Aug-

mented (Aug), Diminished (Dim), and Power Chords (No3). Likewise we cover all typical triads consisting of root, second/third/fourth, and fifth. Note that not 7 x 12, but only 6 x 12 + 4 = 76 final chord classes are obtained, as only 4 different augmented chords exist. These classes have been clustered and re-mapped for testing due to partial sparse occurrence of rather unusual chords into the following two sets: *36 MajMinADPS* (Maj, Min, the other chords (augmented, diminished, power, and suspended mapped onto one group (ADPS)), and *24 MajMin* (Maj, Min). Note that the total of chords was kept constant by mapping chords that are not considered onto considered ones by their root and musical function (e.g. "C No3" is mapped onto "C Maj" if its function is accordingly).

In Table 1 the distribution of keys and chords within the ChoRD database is shown in detail for the classes major, minor, and others by root note.

Table 1. Distribution keys and chords in the ChoRD corpus.

Root	#Key	#Major	#Minor	#Other
A	7	511	459	57
A#	8	567	171	86
B	7	480	213	61
C	16	854	278	105
C#	5	312	315	61
D	3	557	349	94
D#	8	533	141	61
E	12	643	362	21
F	13	728	272	52
F#	4	407	209	44
G	12	719	287	103
G#	5	353	196	41
Sum	**100**	**6 664**	**3 252**	**786**

3. MUSIOLOGICAL MODEL

In order to model the context of a chord rather than recognize isolated chords, we employ a musiological model (MM), which resembles a typical language model (LM) as used in automatic speech recognizers. For training of the model we used the chord lead sheets of [7] after removal of doubles. These chord sheets are usually uploaded by users, which means that they are partly simplified, erroneous, or transposed into easily playable keys on guitar (e.g. G Major). However, for a statistical musiological model this is not too problematic, as we are only interested in typical chord successions. As the sheets often contain shortened progressions in a way that the chord succession is laid out only once, we use the following up-sampling rule: assuming 60 to 100 bars for a typical rock and pop piece, we strictly repeat whenever a song has below 30 bars until 60-100 bars are reached. Chords were translated into the used target set by rule-based parsing (e.g. elimination of bass-notes, clustering of different spelling variants). Overall, 19,025 songs, resulting in a total of 1,573,803 chords are used for the MM. Table 2 shows the top-ranked uni- and bi-grams by frequency.

Table 2. Top-ranked chord uni- and bigrams by frequency.

Rank	1-gram	#	2-gram	#
1	G	244 820	D-G	57 500
2	D	227 549	G-G	55 106
3	A	198 958	C-G	54 702
4	C	188 194	G-C	54 040
5	E	130 896	A-D	46 162
6	F	87 741	D-A	43 534
7	B	72 360	G-G	41 090
8	Am	58 929	A-A	40 161
9	Em	57 537	D-D	39 710
10	A#	32 583	E-A	36 659

4. RHYTHM INFORMATION

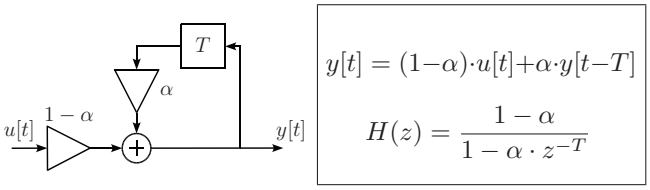

Fig. 1. Block diagram (left), difference equation (top) and transfer function (bottom) of an IIR comb filter

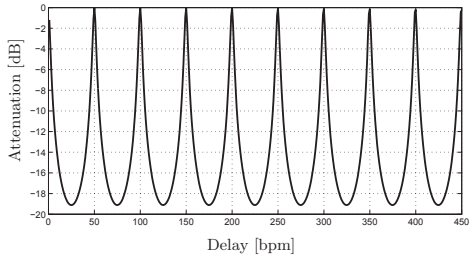

Fig. 2. Magnitude response for an IIR comb filter with gain $\alpha = 0.8$ and base tempo $50bpm$

We use our highly robust beat tracker introduced in [8, 9]to extract rhythmic structure. After a preprocessing step which involves down-sampling to 11,025 kHz and transforming into the frequency domain, the signal is filtered with the A-weighting function according to the human perception of sound. In order to reduce the number of bands without loosing rhythmic information the audio signal is split into frequency

bands using a bank of 24 overlapping triangular filters which are equidistant on the Mel-Frequency scale.

Next, the envelope of each band is extracted using a half wave raised cosine filter and processed by incorporating the moving average over the previous 10 and the following 20 samples due to the fact that humans perceive note onsets louder if they occur after a longer time of lower sound level. Hence, we determine the lowest metrical level referred to as tatum grid using a bank of 57 phase comb filters with gain $\alpha = 0.8$ and delays ranging from $\tau = 18$ to $\tau = 74$ envelope samples. A comb filter is able to extract a frequency and its multiples by adding to the signal a delayed version of itself specified by the gain α and the delay τ. An example for such a comb filter is depicted in figure 1, its magnitude response for $\alpha = 0.8$ and $\tau = 50$ bpm is illustrated in figure 2. Based on the tatum grid our beat tracker is able to determine meter and tempo features by setting up narrow comb filters centered on multiple tempos of the tatum grid.

5. HARMONIC INFORMATION

In order to incorporate the temporal harmonic structure of a song we use the chroma energy distribution normalized statistics introduced by Müller et al. [10]. These features are based on chroma features which are computed using a fast Fourier transform with a window length of 372 ms and an overlap of 0.5 by taking into account a psychoacoustic model using A-weighting filtering as within the beat tracking according to DIN EN 61672-1:2003-10 and by decomposing the audio signal into frequency bands representing the semitones which are defined for equal temperament as

$$f_i = f_0 \cdot 2^{i/12} \qquad f_0 = f(\mathrm{A0}) = 27.5\,\mathrm{Hz} \tag{1}$$

with $15 \leq i \leq 110$ (corrisponding to the notes C2–B9) and therefore covering 96 semitones (8 octaves). In order to overcome deficient recordings due to mis-arranged recording settings or intentional manipulations of the sound impression, pitch correction is applied. A long term frequency analysis computes the prominent frequency f_p and determines a factor c

$$c = \frac{f_p}{f_r} \tag{2}$$

with

$$f_r = \underset{f_i}{\operatorname{argmin}} \left\| \frac{f_p}{f_i} - 1 \right\| \tag{3}$$

Next, all semitones f_i are multiplied with the factor c to correct their pitch. In order to allocate the frequencies to the semitones a nearest neighbor approach is applied which implies the use of Gaussian bells $g_i(x)$ centered at f_i given by

$$g_i(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{\left(\frac{x-f_i}{f_i - f_{i-1}}\right)^2}{2\sigma^2}} \qquad \sigma = 0.125 \tag{4}$$

Now we normalize the resulting sub-bands s_i by dividing each one belonging to the same octave O by the sum of these sub-bands according to

$$\hat{s}_i = \frac{s_{i,O}}{\sum s_{i,O}} \qquad s_{i,O} = s_i \in O \tag{5}$$

In a final step we add up all sub-bands corresponding to the same relative pitch class, for example for the chroma C we compute $\overline{s}_1 = \hat{s}_{15} + \hat{s}_{27} + \ldots + \hat{s}_{99}$, and normalize the resulting values

$$v_i = \frac{\overline{s}_i}{\sum \overline{s}_i} \qquad 1 \leq i \leq 12 \tag{6}$$

Due to the fact that the local chroma features are too sensitive concerning articulation effects and local tempo deviations we extend the chroma features by applying to each component of $\mathbf{v} = (v_1, \ldots, v_{12})$ a quantization function Q as defined by Müller et al. [10]. In the next step, we convolve 11 consecutive quantized chroma vectors $Q(\mathbf{v}(i))$ component-wisely using a Hann window resulting in a weighted 12-dimensional features vector including temporal harmonic information. As the information changes due to the windowing being quite slow, down-sampling with a factor of 4 is applied. The resulting feature vectors are referred to as chroma energy distribution normalized statistics (CENS) which we will denote from now on as $\mathbf{v} = (v_1, \ldots, v_{12})$.

6. AUTOMATIC CHORD LABELING

First, a musical piece is converted from MP3 to a monophonic, 44.1 kHz, 16 Bit wave. Next, the tempo, meter, and down-beat position are determined by IIR-comb filtering as described in sec. 4. According to the tempo, the song is partitioned into consecutive bars. Per bar a 12-dimensional CHROMA-based C.E.N.S. vector is computed (cf. sec. 5). In this process audio data passes a spectral transformation, dB(A)-correction, compensation of detuning and mapping to pitch classes. The result of this cascade is a 12-dimensional vector containing the intensities for each semitone, taking temporal development into account. Note that dB(A)-correction for adaption to human perception according to norm IEC/DIN 651 and pitch tuning are not standard operations in C.E.N.S. feature computation. For pitch tuning, we acquire the prominent frequency during a long-term analysis of the piece in the range between 130 Hz and 1 kHz. Next, the nearest reference frequency to the measured prominent frequency is detected and the semi-tone filter-bank is shifted, accordingly.

For classification we consider a data-free cross-correlation (CC) with a hard template ("1" for each note that is contained in the chord, "0" for any other note in the scale) as reference. For the proposed data-driven processing we compare Support Vector Machines (SVM) with HMM with and without the language, respectively musiological model (MM). SVM

performed best with linear Kernel, pairwise multi-class discrimination and SMO learning. In the case of HMM one continuous model with one emitting state per beat is trained by 20 Baum-Welch iterations [11]. 1 mixture was found to be optimal. We use a context free grammar (word-loop) and Viterbi decoding to model the sequence of chords. If the MM is used (HMM+MM), Laplace smoothed class-based back-off-bigrams (Katz Back-Off, with cutoff 1) further proved optimal.

7. EXPERIMENTS

For evaluation we use song-independent cyclic "leave-one-song-out" (LOSO) training and testing. In Table 3 mean accuracies are summarized.

Table 3. Accuracies (and standard deviation) ChoRD corpus, LOSO evaluation. ADPS abbreviates the clustered group of augmented, diminished, power, and sustained chords.

Accuracy [%]	**CC**	**SVM**	**HMM**	**HMM +MM**
MajMinADPS	28.37 ±14.80	36.71 ±17.44	45.39 ±14.73	48.84 ±15.33
MajMin	39.41 ±16.99	40.24 ±17.52	58.57 ±19.54	**60.13** ±19.10

As can be seen, the data-driven approaches are superior, whereby HMM prevail. By language modeling a further gain is obtained, and the reduction to major and minor chords seems reasonable if appropriate.

8. CONCLUSION

A system was shown to fully automatically label chords by uniting the advantages from beat-synchronization, musiological modeling, de-tuning compensation, and data-driven processing. The combination of these was shown to be superior to merely knowledge-driven cross-correlation and single measure analysis. Moreover, impressive 60% accuracy could be reached on original MP3 compressed audio tracks that represent a broad mix of artists and styles rather than limitation e.g. to one artist. The integration of a musiological model trained on large data amounts was proven significantly beneficial. However, some variance was found throughout genres with respect to the difficulty of the task: some songs, as e.g. *"Enya - Silver Inches"* were recognized without mistake, while *"Prince - Purple Rain"* proved to be the toughest call: only every fourth chord was determined correclty. A particular advantage of the beat-synchrony is the ready-to-use lead-sheet character of the output. In future efforts we aim at investigation of benefits arising from chord enhancement by Non-Negative-Matrix-Factorization and use of stereophonic information. From an architectural point of view we will alternatively consider Bidirectional Long-Short-Term Recurrent Neural Networks that allow to model knowledge of the whole song for every chord decision.

9. REFERENCES

[1] B. Schuller, F. Dibiasi, F. Eyben, and G. Rigoll, "One day in half an hour: Music thumbnailing incorporating harmony- and rhythm structure," in *Proc. 6th Workshop on Adaptive Multimedia Retrieval, AMR 2008*, Berlin, Germany, 2008.

[2] T. Fujishima, "Real-time chord recognition of musical sound: A system using common lisp music," in *Proc. International Computer Music Conference*, 1999.

[3] K. Lee and M. Slaney, "A unified system for chord recognition and key extraction using hidden markov models," in *Proc. ISMIR 2007*, 2007.

[4] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using emtrained hidden markov models," in *Proc. ISMIR 2003*, Baltimore, Maryland, 2003, pp. 183–189.

[5] J. B. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proc. ISMIR 2005*, 2005, pp. 304–311.

[6] "Songlist chord data-set," in *http://www.mmk.ei.tum.de/sch/chord.txt*, 2006.

[7] "The on-line guitar archive," in *http://www.olga.net*, 2006.

[8] B. Schuller, F. Eyben, and G. Rigoll, "Tango or waltz?: Putting ballroom dance style into tempo detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, no. Article ID 846135, pp. 12 pages, 2008.

[9] B. Schuller, F. Eyben, and G. Rigoll, " Fast and Robust Meter and Tempo Recognition for the Automatic Discrimination of Ballroom Dance Styles," in *Proc. ICASSP*, April 2007, vol. 1, pp. 217–220.

[10] M. Müller, F. Kurth, and M. Clausen, "Chroma-Based Statistical Audio Features for Audio Matching," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005, pp. 275–278.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (v3.4)*, Cambridge University Press, Cambridge, UK, 2006.