Technical Report

EHU-KZAA-TR-2012-02

UNIVERSITY OF THE BASQUE COUNTRY
Department of Computer Science and Artificial Intelligence

Oracles for Audio Chord Estimation

Thomas Rocher
Darrell Conklin

June 2012

San Sebastian, Spain
www.ccia-kzaa.ehu.es
https://addi.ehu.es/handle/10810/4562

ORACLES FOR AUDIO CHORD ESTIMATION

Thomas Rocher[1] **Darrell Conklin**[1,2]
[1]Department of Computer Science and Artificial Intelligence
University of the Basque Country UPV/EHU, San Sebastián, Spain
[2]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
rocher@labri.fr conklin@ikerbasque.org

ABSTRACT

This paper explores how audio chord estimation could improve if information about chord boundaries or beat onsets is revealed by an oracle. Chord estimation at the frame level is compared with three simulations, each using an oracle of increasing powers. The beat and chord segments revealed by an oracle are used to compute a chord ranking at the segment level, and to compute the cumulative probability of finding the correct chord among the top ranked chords. Oracle results on two different audio datasets demonstrate the substantial potential of segment versus frame approaches for chord audio estimation. This paper also provides a comparison of the oracle results on the Beatles dataset, the standard dataset in this area, with the new Billboard Hot 100 chord dataset.

Keywords: audio, music information retrieval, harmony, chroma, chord

1. INTRODUCTION

Audio chord estimation has been a very active field for the Music Information Retrieval community for several years. Existing methods in this area differ in a variety of ways. While some rely on music theory and pattern recognition [3, 10–12] others use data-driven approaches [7, 8, 13]. Most existing methods can be described as *frame-based*, attempting to estimate a chord label for each frame of signal.

Methods for audio chord estimation have the opportunity for comparative evaluation in the MIREX [5], an annual community-based framework for the evaluation of MIR systems and algorithms. Since 2009, the most accurate methods for the audio chord estimation achieved similar scores, compete at roughly the same level of frame labelling accuracy. Moreover, it seems that in recent years, the accuracy of chord estimation methods has reached a plateau of around 80% on triadic chords.

A major problem in all audio chord estimation methods is the problem of *fragmentation*: since labelling takes place at the frame level, there is little to prevent chords being split apart by short segments which do not correspond to a real chord. This problem can be addressed by subsequent smoothing of the estimated frame labels, and some recent methods have considered using time segmentation and chord duration explicitly in the audio chord estimation process [7, 12].

This paper investigates the impact of time segmentation on audio chord estimation in more detail, exploring the central question: what could be achieved in terms of chord labelling accuracy if the time segmentation of the audio stream into beats or chords was provided? This question is explored through a detailed oracle simulation. The results suggest a new direction for audio chord estimation, redefining the problem as primarily one of segmentation rather than one of frame labelling.

2. METHODS

This section presents the theoretical foundations of the oracle simulations, reviewing the problem of audio chord estimation, presenting the oracles and the different segmentation informations that they can reveal, and showing how chord labels are propagated from time segments down to the frame level where the chord estimation accuracy can be finally evaluated.

2.1 Audio chord estimation

To label a chroma vector with a chord triad, each chroma is compared to 24 different triadic chord templates, 12 for major and 12 for minor triads. In this study we use the simplest possible triadic templates, to avoid bias in the results due to over-fitting on the corpora. Two templates T are used, one for major and one for minor chords:

$$\text{Cmaj} = (1,0,0,0,1,0,0,1,0,0,0,0)$$
$$\text{Cmin} = (1,0,0,1,0,0,0,1,0,0,0,0)$$

Chord templates for all other chord roots are readily constructed by rotation of the C major or C minor templates.

The Non-Negative Least Squares (NNLS) chroma [9], widely popular in the recent chord estimation literature, provides a pitch class description of an audio frame in the form of a 12 dimensional vector. The influence of each pitch-class (C, C#, D, ...) is thus described by a real-valued number. This chroma vector is compared to the triadic chord templates: for each of the 24 chord templates, a scalar product with the chroma vector provides a correlation score. More precisely, the correlation C between a chord template T and a 12 dimensional chroma vector V is defined as:

$$C_{T,V} = \sum_{i=1}^{12} (T[i] \cdot V[i]) \quad (1)$$

The higher the correlation is for a particular triad according to Equation 1, the more likely according to the model the chord corresponding to the template is sounding for the considered observation. The 24 different triadic chords are thus ranked from higher correlation to lower correlation. The first best chord for the considered observation is the highest correlated chord with the observation, the second best chord is the second highest correlated chord, and so on.

2.2 The oracles

Oracles provide partial information about time segmentation. Here we consider oracles with three different levels of power, each revealing partial ground truth data to an audio chord estimation method. An oracle knows the exact triadic chord structure of songs, and reveals one of three types of information for every song:

- *beat oracle*: the onset time of every beat,
- *beat segmentation oracle*: the same information as above and additionally, if every new beat belongs to same chord as the beat before (thereby, it gives the number of beats in every chord),
- *chord oracle*: the onset time and duration of every chord.

Figure 1 illustrates the information revealed by an oracle. Each of the temporal segment defined by an oracle must then be labelled by a chord.

2.3 Segment labelling

At the frame level, the V in Equation 1 are the simply chroma vectors computed for each frame. When using the oracle, the different chroma vectors at the frame level must be combined to take into account the segmentation to form a *metachroma*, which is the global chroma vector for the considered time segment. This metachroma forms the V in Equation 1. The remainder of this section describes how the different metachromas are computed from the frame chroma vectors.

2.3.1 Chord

To label a chord, we consider all the frames located between the starting and ending times of the chord. If i is the first frame of the chord H and j is its last frame, we construct the metachroma V_H so that:

$$V_H[i] = \sum_{k=i}^{j} V_k[i]$$

with V_k the chroma vector describing the k-th frame. We then substitute V with V_H in Equation 1.

2.3.2 Beat

To label a beat, we consider all the frames located between the beat onset and the next beat onset. If i is the first frame of the beat B and j is its last frame, we construct the metachroma V_B so that:

$$V_B[i] = \sum_{k=i}^{j} V_k[i]$$

We then substitute V with V_B in Equation 1.

2.3.3 Beat segmentation

To label a beat segment, we consider all the beats of the chord. If i is the first beat belonging to the chord H, and j its last beat, we construct the metachroma V_S so that:

$$V_S[i] = \sum_{k=i}^{j} V_{B_k}[i]$$

with V_{B_k} the metachroma of the k-th beat of the H chord. We then substitute V with V_S in Equation 1.

2.3.4 Propagation to the frame level

Once every metachroma is labelled, the ranking of the 24 chords is propagated to every frame belonging to the considered time segment. This propagation permits the final evaluation of accuracy at the frame level. Figure 2 illustrates the propagation of the chord rankings at the frame level, using the chord oracle. This figure presents a chord lasting for five frames. The corresponding chromas are summed to form a metachroma which length equals five frame lengths. The ranking of candidates according to this metachroma is then propagated to each of the five frames. The baseline accuracy is obtained at the frame level, using the raw V_k chroma vectors.

2.4 Mapping to major and minor triads

In audio chord transcriptions, chords have a root note and a chord quality which typically belongs to a large dictionary [6]. In this paper, as with the triadic MIREX audio

Dataset	Beatles	Billboard
number of songs	180	649
total number of frames	∼620,000	∼3,000,000
total number of beats	∼52,000	-
chord changes per song	69	89

Table 1. Properties of the two datasets.

full chord	Billboard mapping	evaluation mapping
maj	maj	maj
min7	min7	min
aug	NA	N
maj6	maj	maj
7	7	maj
sus2	NA	N
5	5	N

Table 2. Examples of chord mappings to maj/min triads as proposed by the Billboard dataset.

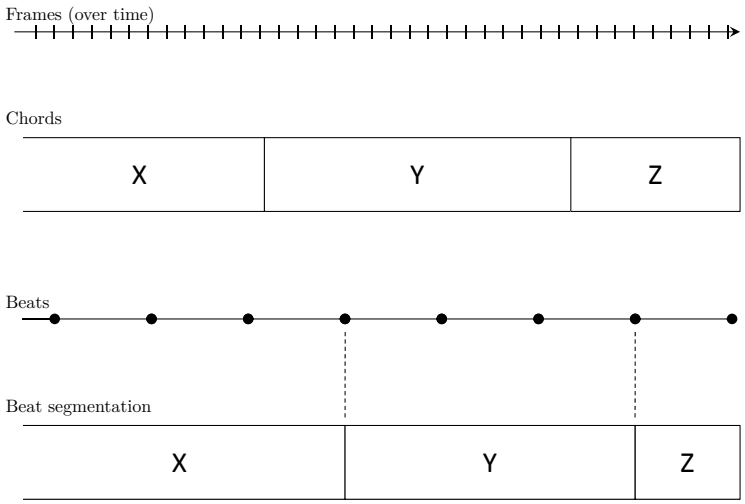

Figure 1. Different information revealed by a oracle: *chord oracle*, revealing the starting and ending times of the chords; *beat oracle*, revealing the beat onsets; and *beat segmentation oracle*, preserving the number of beats for each chord.

chord estimation evaluation [5], we only focus on the root note (C, C#, D, ..., B) and the mode (major or minor) of chords. All of the ground truth chords of the database have thus been mapped to major and minor triads using the mapping proposed by the Billboard annotations. Table 2 shows different examples of mappings according to the chord quality. When a chord cannot be mapped to a major or minor triad, the chord is not considered, and is subject to no evaluation. Silences and N-chords (part of a song in which no chord is played) are also ignored.

2.5 Recall score computation

The effect of the oracle on audio chord estimation is evaluated as follows. For each frame, we consider the random variable X as the rank of the correct chord in the frame rankings. Therefore $P(X = k)$ is the probability of finding the correct chord at rank k, and

$$r(k) = P(X \leq k) = \sum_{i=1}^{k} P(X = k)$$

is the cumulative probability of finding the correct chord within the top k triadic chords. The $r(k)$ recall score for a dataset is thus the number of frames for which the correct chord belongs to the top k correlated chords, divided by the number of total frames in the dataset.

2.6 Datasets

Two different audio chord transcription datasets, the Beatles discography and the Billboard Hot 100 dataset, were used to evaluate the effect of the oracle on audio chord estimation. These datasets are described in more detail below. For both datasets we apply the same settings for computing the chroma vectors:

- a frame length of 16384 samples ($\sim$ 0.37 sec),
- a hopsize of 2048 samples ($\sim$ 0.05 sec),
- a rolloff of 1%, as recommended for pop songs [1].

thereby mapping the standard Beatles dataset parameters to those used in the new Billboard dataset. The main properties of the two datasets are summarised in Table 1.

2.6.1 Beatles discography

The Beatles audio discography contains 180 songs with a 44kHz sampling rate. In this dataset, the average number of chord changes per song is 69, with an average of 7.7 distinct chords per song. Chord transcriptions were checked by Harte [6] and the MIR community, and are available online [2]. The corpus also includes the beat onsets within every song, for a total of 52,000 beats.

On the Beatles dataset, $\sim$ 41,000 out of $\sim$ 620,000 frames were discarded as silences or N-chords.

2.6.2 Billboard Hot 100

The Billboard Hot 100 is a weekly list of popular songs, ranked by radio airplay audience. The transcriptions of the chord progressions of 649 songs that appeared at some point in this list have recently been published [1,4]. In this dataset, there is an average of 96 chords changes per song, with an average of 11.8 distinct chords per song. No beat information is provided in this dataset.

On the Billboard dataset, $\sim$ 285,000 out of more than 3 million frames were discarded as silences or N-chords. Note that this represents a higher proportion of discarded chords than in the Beatles dataset, mainly because complex chords are more frequent in this dataset.

3. RESULTS AND DISCUSSION

This section presents the results obtained using the oracle approach outlined in Section 2, by simulation on the two audio chord transcription datasets.

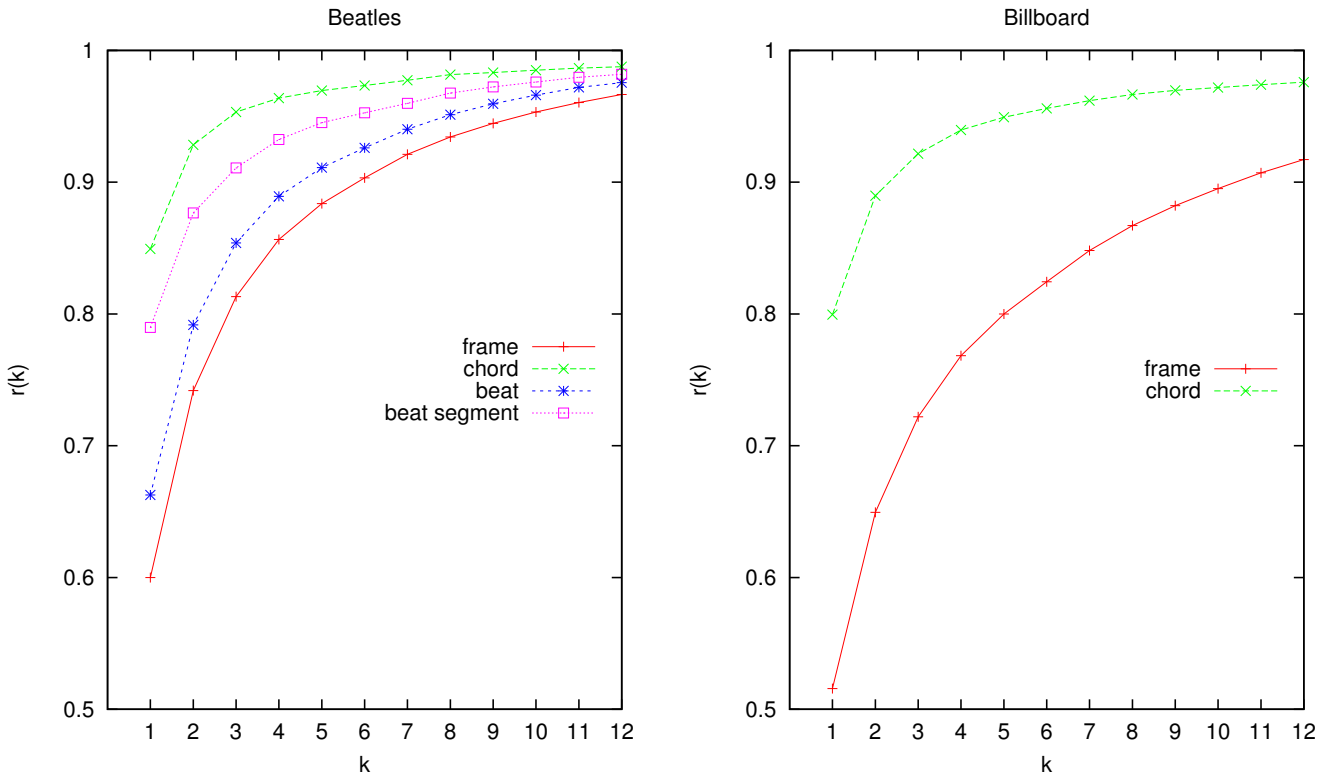

Figure 3. $r(k)$ recall scores on the Beatles dataset (left) and on the Billboard dataset (right).

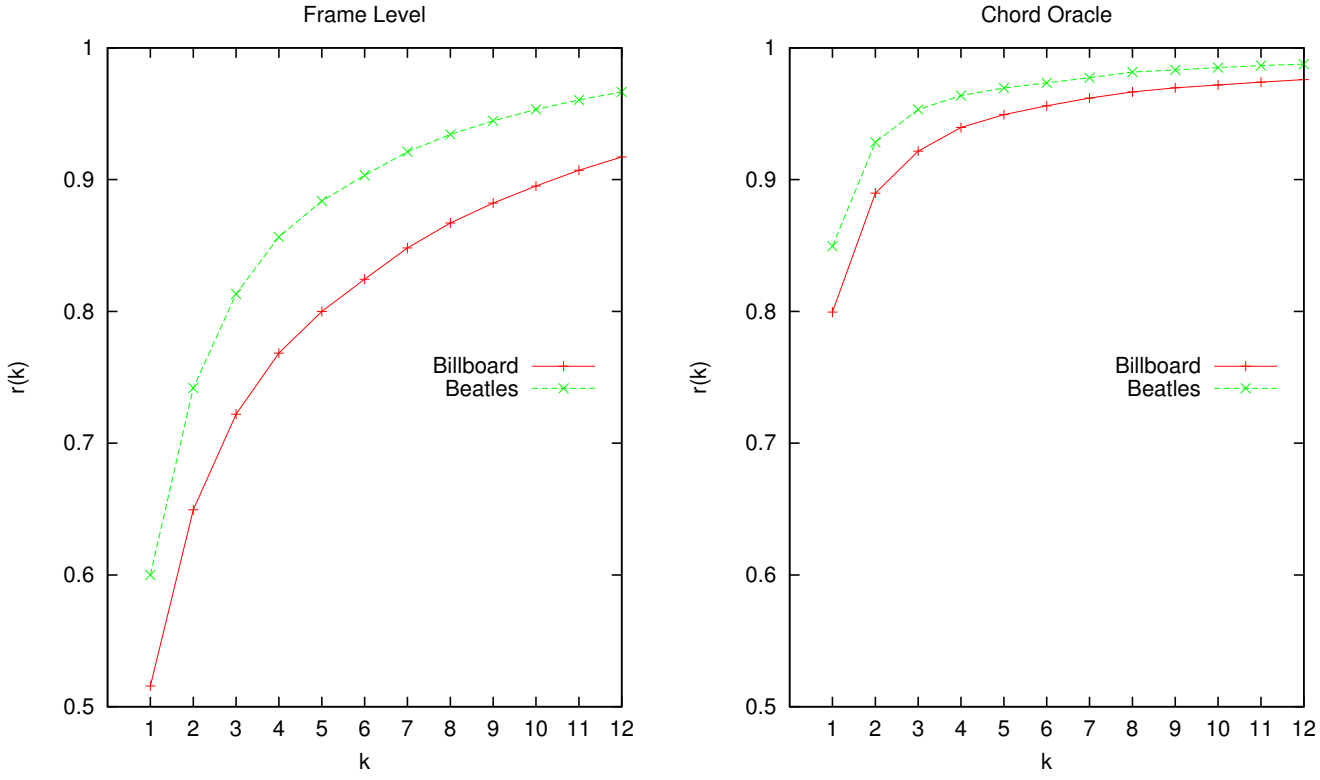

Figure 4. Comparison of the $r(k)$ recall scores on the frame (left) and chord (right) levels for the Beatles and Billboard datasets.

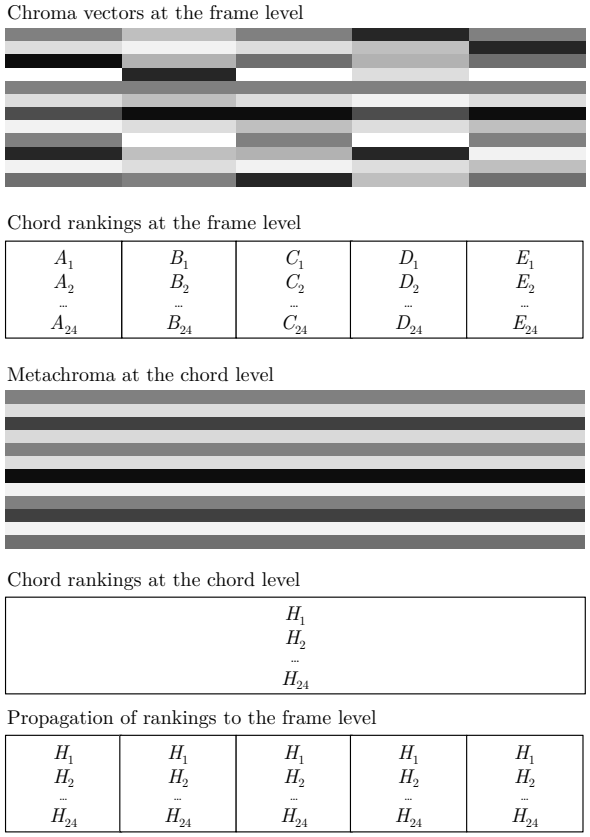

Figure 2. Illustration of the propagation of information provided by the oracle (here, the chord boundaries oracle) at the frame level, for a chord lasting for five frames.

Figure 3 shows the $r(k)$ recall scores on the frame level, for the three oracle segmentations presented in section 2. As the beat onsets were not provided in the Billboard dataset, it was not possible to compute the beat oracle nor the beat segmentation oracle scores for this dataset. Only the $r(k)$ for $k \leq 12$ were plotted on the figure, as $r(k)$ flatten out to 1 for $13 \leq k \leq 24$. The most notable results are presented in Table 3 and summarised as follows:

- *frame*: on both the Beatles and Billboard datasets, the $r(1)$ recall scores are lower than any $r(1)$ oracle score. The $r(3)$ scores are higher than $r(1)$ with the beat oracle, but lower than $r(1)$ with the chord and beat segmentation oracles.
- *chord oracle*: the most powerful oracle, with $r(1)$ scores higher than any other oracle, and $r(3)$ scores around 95% both on the Beatles and Billboard datasets,
- *beat oracle*: the least powerful oracle, but with $r(k)$ scores still notably higher than the frame level for $1 \leq k \leq 10$,
- *beat segmentation oracle*: less powerful than the chord oracle, but only for $r(k)$ scores with $k \leq 6$.

The $r(1)$ scores are especially important, because they represent what is immediately achievable in term of accuracy given the oracle. On the Beatles dataset, the $r(1)$ score on the frame level is 60%, while with the beat and beat segmentation oracles, it is 66% and 78%. The most notable jump is achieved with the chord oracle, which reaches 85%. This surprising result is confirmed on the Billboard dataset, where the $r(1)$ score jumps from 52% to 80%. This is to say that if the chord boundaries are known, a very naive chord estimation method (without any parameter or musical consideration of any kind) would probably outperform every automatic method, including the ones using pre-trained algorithms.

Also surprising is the jump between the $r(1)$ and $r(2)$ scores, notable on all four methods (frame, chord, beat and beat segmentation). This score increase is at least 8% (with the chord oracle on the Beatles dataset), and reaches more than 14% (with the beat oracle on the Beatles dataset). If we consider the difference between $r(1)$ and $r(3)$ scores, the difference becomes important, as presented in Table 3. The frame level jump (which is more than 20% on both datasets) can be explained by the relatively low $r(1)$ score. But with the chord oracle, the improvement from $r(1)$ to $r(3)$ is still more than 10% on both datasets, reaching even 95% on the Beatles dataset. These numbers show that we do not need to consider more than the top three chords, with the chord oracle, to reach a remarkably high potential accuracy in audio chord estimation.

Audio chord estimation systems, for example based on hidden Markov models, cannot profit directly from the oracle results as long as they work on the frame level. These systems must assign high probability to self transitions (remaining in the same chord), as it is likely that a chord lasts longer than a single frame. If they could be adapted to work at the chord level, the probability space would be divided between on chord changes, and results could then potentially reach the $r(3)$ recall scores.

4. CONCLUSION AND FUTURE WORK

This paper investigated the possible impact of time segmentation information for audio chord segmentation. By considering three oracles, we proposed a way to taking into account beats, beat segmentation and chord segmentation, by using a very simple average chroma computation. Results show that the information provided by those oracles could be highly beneficial to audio chord estimation methods.

A very naive method, using the chord oracle and without any chord transition consideration, post-smoothing algorithm or training process can outperform any existing method. This leads to the interesting problem of estimating the chord segmentation of the audio texture. The potential impact of segmentation on the audio chord estimation problem has not yet been fully considered by practical methods. Our results suggest an alternative view of audio chord estimation as first a segmentation problem, and then a labelling problem.

Another important remark is that it is possible for a lis-

Beatles				Billboard	
frame	chord	beat	beat seg.	frame	chord
$r(1)$=60	$r(1)$=**85**	$r(1)$=66	$r(1)$=79	$r(1)$=52	$r(1)$=**80**
$r(3)$=81	$r(3)$=95	$r(2)$=79	$r(2)$=91	$r(3)$=72	$r(3)$=92

Table 3. Notable recall scores (%) using oracles.

tener to detect some chord boundaries, as it may be easier for a listener to locate a chord change than to identify the chord label. This could lead to an informed chord estimation method, where the user provides the chord onsets (by indicating during listening when the chord changes) and a system could estimate in real time the label of the last chord. For the beat or beat segmentation oracles the $r(1)$ score may not be high enough to admit the above strategy, and to make use of the oracle a method must jointly estimate the number of beats in a segment in addition to the label of the segment.

This paper also compares the standard Beatles dataset for audio chord estimation with the new Billboard dataset. The Billboard dataset seems more difficult in terms of chord label estimation, and provides an interesting alternative to the Beatles corpus, especially with a larger number of songs along with a wider genre diversity. Both datasets validate the hypothesis that the oracles reveal valuable information, and that this information should be consider to improve audio chord estimation methods.

Future work will deal with ways to integrate of the oracle results into practical systems. Sophisticated statistical methods such as explicit duration models can jointly estimate the segmentation and labelling. Research on replacing the oracles with estimated information is also planned, to explore how well a fully automatic method can perform on audio chord estimation using estimated segmentation information.

5. ACKNOWLEDGMENTS

This research was supported by a grant *Modelos segmentales por detección de acordes de audio* from the Gobierno Vasco, Departamento de Industria, Innovacion, Comercio y Turismo (programme Saiotek 2010).

6. REFERENCES

[1] Billboard "Hot 100" website. `http://ddmal.music.mcgill.ca/billboard`. [Online; accessed April-14-2012].

[2] Isophonics website. `http://www.isophonics.net`. [Online; accessed April-14-2012].

[3] J. Bello and J. Pickens. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 304–311, London, UK, 2005.

[4] J. Burgoyne, J. Wild, and I. Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 633–638, Miami, USA, October 2011.

[5] J. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[6] C. Harte, M. Sandler, and A. Samer. Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proceedings of the 4th International Society for Music Information Retrieval (ISMIR)*, pages 66–71, 2005.

[7] M. Khadkevich and M. Omologo. Use of hidden Markov models and factored language models for automatic chord recognition. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 561–566, Kobe, Japan, 2009.

[8] K. Lee and M. 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):291–301, 2008.

[9] M. Mauch and S. Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 135–140, Utrecht, Netherlands, October 2010.

[10] M. Mauch and S. Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE Trans. on Audio, Speech and Language Processing*, pages 1280–1289, 2010.

[11] L. Oudre, Y. Grenier, and C. Févotte. Template-based chord recognition: influence of the chord types. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 153–158, Kobe, Japan, 2009.

[12] J. Pauwels, J. Martens, and M. Leman. The influence of chord duration modeling on chord and local key extraction. In *Proceedings of the International Conference on Machine Learning and Applications*, pages 136–141, Honolulu, Hawaii, USA, December 2011. IEEE.

[13] A. Sheh and D. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 185–191, Baltimore, MD, 2003.