EVALUATING AUTOMATICALLY ESTIMATED CHORD SEQUENCES

Johan Pauwels and Geoffroy Peeters

STMS IRCAM-CNRS-UPMC
johan.pauwels@ircam.fr, geoffroy.peeters@ircam.fr

ABSTRACT

In this paper, we perform an in-depth evaluation of a large number of algorithms for chord estimation that have been submitted to the MIREX competitions in 2010, 2011 and 2012. Therefore we first present a rigorous scheme to describe evaluation methods in a sound, unambiguous way that extends previous work specifically to take into account the large variance in chord estimation vocabularies and to perform evaluations on select sets of chords. Then we take a look at the evaluation metrics used so far and propose some alternative ones. Finally, we use these different methods to get a deeper insight into the strengths of each of the competing algorithms and show that the choice of evaluation measure greatly influences the ranking.

Index Terms— music information retrieval, chord estimation, evaluation procedure, large scale evaluation

1. INTRODUCTION

Chord estimation has established itself as one of a growing number of tasks in the field of Music Information Retrieval (MIR). Especially since its addition to the yearly returning MIREX contest in 2008, the field has seen a steady flow of published papers [1, 2] and data sets [3, 4]. However, this increase in available data has not led to a proportional increase in generally applicable knowledge. Among the reasons for this, is the fact that not all automatic estimation algorithms generate the same vocabulary of chords and that the evaluation procedure in most papers is tailored to the used chord vocabulary, often described in an ad-hoc way that leaves some room for interpretation. Therefore it is hard to exactly reproduce the evaluation method and to compare results between papers. Furthermore, we feel that the richness of available data coming out of the MIREX contest has been under-exploited. It is unique in its comparison of a multitude of algorithms on the same data sets with the same evaluation methods, but only a minimal study of the results is being done each year. Therefore, we take a deeper look in this paper at the data of the 2010, 2011 and 2012 editions.

In order to make an evaluation reproducible, we need an unambiguous way to discuss metrics. The most thorough discussion of chord sequence evaluation up to now has been done by Harte [5]. He established a general framework to describe evaluation measures. Therefore he first studied the comparison of chord pairs and builds upon that to arrive at the comparison of two chord sequences. For the latter problem, he proposes the "dictionary-based recall evaluation" (DBRE) to study the performance for user-defined subsets of chords. Alternative methods to get a deeper insight into the performance of algorithms include the reporting of results per chord type [6] and of the amount of confusion with related chords [7, 8, 9]. The scheme we propose builds on the work by Harte. More specifically, we formulate a new framework to describe the evaluation of chord sequences, in which we reuse his work on the comparison of chord pairs. It is intended as a replacement to his DBRE, especially designed to deal with differences in estimation vocabulary between distinct algorithms.

As far as large-scale evaluations of chord estimation algorithms are concerned, the most known is of course the MIREX contest itself, but as previously said, its study of the results is rather limited. More extensive evaluations have been performed by [5, 8], but at the time of their evaluations, the vocabularies of the systems under test were mostly limited to major and minor triads. Our work will pay special attention to the assessment of more complex chords.

Next, we will detail our scheme to describe evaluation methods in Section 2. Then the measures used so far and some alternative ones will be formulated according to this framework in Section 3. They will afterwards be used to analyse the raw algorithmic output in Section 4. Finally, we draw some conclusions in Section 5.

2. EVALUATING: WHAT, WHERE AND HOW?

The objective of an evaluation procedure is to quantify the extent to which an estimated chord sequence resembles a certain reference sequence (preferably a manual annotation). This resemblance can be quantified according to a number of different definitions, such as resemblance in terms of chord segmentation [5, 6] or fragmentation and estimated vocabulary [8], but in this text we limit ourselves exclusively to similarity in harmonic content, mainly because the other ones have already been thoroughly treated by others. In order to evaluate this similarity in harmonic content, we compare the estimated chord sequence with the reference sequence on a pair-wise basis. In the literature, these pairs have been calculated in a frame-based (where both reference and estimation get discretised on a grid and every grid point gives a pair) or a segment-based fashion (where segment pairs of variable length are formed by joining chord boundaries of both sequences). We adhere to previous conclusions [5, 6] that state that the latter should be preferred because of its advantages related to rounding errors and speed. A clear evaluation strategy should then answer 3 questions about the comparison of these pairs of reference and estimated chords: "what", "where" and "how" do we evaluate? These questions and their possible answers will be further explained in the following sections.

2.1. What: handling different chord vocabularies

Typically, the reference chord sequence is annotated much richer than the estimated sequence, the former often having an unconstrained vocabulary only defined by the data itself while the latter has an algorithm dependent vocabulary defined before the analysis starts. An option to reduce the variability between the vocabularies of the reference and one or more estimated sequences, is to define a mapping between chords. In practice, it always is a mapping of chord types: the root is preserved and complex chord types get mapped to more simple chords (i.e. a reduction of the number of

This work was partly supported by the Quaero Program funded by Oseo French agency.

constituting chromas). The "what" question thus asks what exactly will be compared: full chord labels or their reduced versions.

Some mappings that we will use later on, are "triads" where chords get reduced to their triads or "tetrads & triads" where every chord gets mapped to its tetrad or in case that doesn't exist, to its triad. The no-chord symbol is mapped to itself for both. The most simple mapping is of course "none", leaving the chords as they are. In the remainder of this text, we will also use a toy example for illustrative purposes that consists of three mapping rules: 7→maj, maj→maj and min→min.

2.2. Where: selecting points of interest

The "where" question deals with which segments should be included in the evaluation. This should first and foremost only be determined by the reference sequence, in order to avoid making the evaluation depend on the estimated sequence. A mapping itself imposes implicit constraints on which pairs get evaluated. With each mapping $\mathcal{M}$ an input domain $\mathcal{C}_{MI}$ is associated for which the mapping is defined and an output domain $\mathcal{C}_{MO}$ which lists all possible chords that can be produced by the mapping. The mapping can then be notated as a surjection $\mathcal{M} : \mathcal{C}_{MI} \twoheadrightarrow \mathcal{C}_{MO}$. For our example mapping, $\mathcal{C}_{MI} = \{\mathrm{maj}, \mathrm{min}, 7\}$ and $\mathcal{C}_{MO} = \{\mathrm{maj}, \mathrm{min}\}$.

If the mapping is executed on a chord that is not part of $\mathcal{C}_{MI}$, its output is undefined. For instance, our example does not mention what to do with diminished chords, therefore the mapping of diminished chords is undefined. Consequently, if a chord in the reference annotation does not belong to $\mathcal{C}_{MI}$, then that segment should be skipped from evaluation. On the other hand, if a chord in the estimated sequence does not belong to $\mathcal{C}_{MI}$, the evaluation should fail and a different mapping should be used. Summarized, if we represent a chord of the reference vocabulary as $c_{ref} \in \mathcal{C}_{REF}$ and one of the estimation vocabulary as $c_{est} \in \mathcal{C}_{EST}$, then the segment pairs (c_{ref}, c_{est}) that will be evaluated are those where $c_{ref} \in \mathcal{C}_{MI}$. The actual labels that will be compared are $\mathcal{M}(c_{ref})$ and $\mathcal{M}(c_{est})$ where the mapping should be chosen such that $\mathcal{C}_{EST} \subseteq \mathcal{C}_{MI}$.

The input domain for our previously defined "triads" and "tetrads & triads" mappings is the same, namely the no-chord and all chords than can be mapped to a triad. We concretize this as all chords that contain a major or minor third or one of a major second or perfect fourth plus the no-chord. The "none" mapping obviously has the collection of all possible chords $\mathcal{C}_{ALL}$ as input domain.

The most common use-case is to evaluate on all chords (or at least all for which a mapping is defined), as this makes maximal use of the available data. In other cases however, it might be useful to restrict the evaluation to places where the reference chord belongs to a subset. This can lead to a better understanding of the performance of an algorithm on a certain category of chords, e.g. how well tetrads are recognized compared to triads. One way to do this would be to modify the mapping to map more input chords to "undefined" outputs, but that would mean that different mappings need to be used for reference and estimated chord sequences. Therefore we propose a more elegant solution where we keep the same mapping for both sequences, and the input domain of the mappings as general as possible. Instead, we introduce the additional requirement that evaluation only takes place when the reference chord c_{ref} belongs to a user-defined input limiting set $\mathcal{C}_{LI}$. In theory, this allows us to exactly specify the chords for evaluation, but for some cases this might be impractical. Suppose we want to have a more detailed look at the performance of chords that are mapped to a major triad. For our example mapping, this would mean setting $\mathcal{C}_{LI} = \{\mathrm{maj}, 7\}$, which is still workable, but for mapping functions with a larger input domain, this could mean that we need to list every possible combination of a major triad and a number of tensions (e.g. maj7, maj9, maj11, etc.). This quickly becomes impracticable. Therefore we introduce a similar mechanism to limit the chords at the mapped side by specifying an output limiting set $\mathcal{C}_{LO}$. Applied to our example, the large input limiting set $\mathcal{C}_{LI}$ can then simply be replaced by the singleton $\mathcal{C}_{LO} = \{\mathrm{maj}\}$.

Both limiting sets are optional, their default value is $\mathcal{C}_{ALL}$ such that they have no influence. Combining both limiting domains, this gives that segments pairs (c_{ref}, c_{est}) will only be evaluated when $c_{ref} \in \mathcal{C}_{LI} \cap \mathcal{C}_{MI}$ and $\mathcal{M}(c_{ref}) \in \mathcal{M}(\mathcal{C}_{LI} \cap \mathcal{C}_{MI}) \cap \mathcal{C}_{LO}$.

Taking the chord sequences in Figure 1, the initial number of evaluation segments is 6: (Bdim, Dmin), (Dmin, Dmin), (Dmin, Bmin), (G7, Bmin), (Cmaj, Bmin) and (Cmaj, Cmaj). When using our toy mapping, the first segment pair will be discarded because the mapping for Bdim is undefined. Additionally, setting $\mathcal{C}_{LO}$ to $\{\mathrm{maj}\}$ will further discard the second and third segment from the evaluation, because $\mathcal{M}(\mathrm{min}) = \mathrm{min}$ is not part of $\mathcal{C}_{LO}$. Only the last three reference-estimation pairs then remain and due to the mapping they will be evaluated as (Gmaj, Bmin), (Cmaj, Bmin) and (Cmaj, Cmaj).

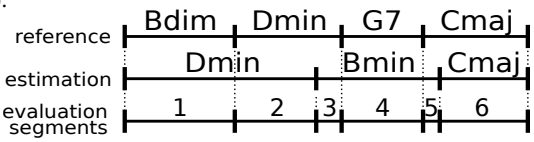

Fig. 1. Example chord sequences.

The introduction of the input limiting domain shows some similarity to Harte's "dictionary-based recall evaluation" (DBRE). The main difference lies in the way how chords that don't belong to a user-defined dictionary are handled. Segments with out-of-dictionary chords in the annotation are discarded from evaluation, as we propose, but segments with out-of-dictionary chords in the estimated sequence get a score of 0 by default. For our example in Figure 1 with dictionary $\{\mathrm{maj}, 7\}$, this would mean that the first 3 segments get discarded, but also that the (G7, Bmin) and (Cmaj, Bmin) segments get a score of 0 in all cases, whatever the evaluation measure might be. We argue that these 2 pairs should be evaluated as well, which can still lead to a score of 0, but does not have to. The resulting value will depend on the way the pairs are compared, where one can argue that on one hand, an estimation of Bmin for a G7 reference fails to find the correct root, but on the other hand, that all chromas of the estimated Bmin are part of the G7 chord.

2.3. How: assigning a score value

The last part of an evaluation scheme should answer the question of how to assign a value to the retained reference-estimation pairs. Therefore a scoring function $S : \mathcal{C}_{SR} \times \mathcal{C}_{SE} \to \mathbb{R}^+$ should be defined where $\mathcal{C}_{SR}$ and $\mathcal{C}_{SE}$ are the reference, respectively estimation, chord vocabularies of the scoring function. At the minimum, the scoring function should be able to handle all chords pairs coming out of the mapping subject to optional limiting. This means that $\mathcal{M}(\mathcal{C}_{MI} \cap \mathcal{C}_{LI}) \cap \mathcal{C}_{LO} \subseteq \mathcal{C}_{SR}$ for the reference sequence and $\mathcal{C}_{MO} \subseteq \mathcal{C}_{SE}$ for the estimation sequence need to hold, but it is advisable to make the scoring function input domains as large as possible. Nonetheless, the combination of limiting the evaluation to a subset and a mapping can make it easier to define the actual comparison by reducing the number of reference-estimation combinations it needs to be able to handle. For instance, in [10] the evaluation is only performed on segments annotated as triads in order to avoid the need to define a scoring function between triads and more complex chords, in particular between triads and tetrads that contain those triads.

This particular part of the evaluation scheme is thoroughly discussed in Harte's thesis [5], so we'll only briefly recap some major points. The measures can be divided into two categories: either

chords are evaluated as a whole, where the chromas are organised with respect to the root, or as a bag of chromas, where all constituting chromas are considered equal (ordered and unordered set comparisons in Harte's terms). A simple example of the first case assigns a score of 1 to an exact correspondence, taking enharmonic equivalence into account and ignoring the bass note, and 0 to other cases. This scoring function will be referred to as "exact" in the remainder of the text. For the bag of chromas approach, some examples of a scoring function are chroma based recall and precision. Returning to the (G7, Bmin) pair of the previous section, we see that its score would be 0 using the "exact" measure, but 0.75 and 1 for chroma based recall, respectively precision.

2.4. Summary

To summarize, an evaluation measure can be unambiguously described by the combination of a mapping function $\mathcal{M} : \mathcal{C}_{MI} \rightarrow \mathcal{C}_{MO}$, a scoring function $\mathcal{S} : \mathcal{C}_{SR} \times \mathcal{C}_{SE} \rightarrow \mathbb{R}^+$ and optionally input and output limiting domains $\mathcal{C}_{LI}$ and $\mathcal{C}_{LO}$. The average over multiple segments is computed by taking the average of all individual scores weighted by their segment lengths. Whether these segments come from the same or multiple songs is not important in this regard.

3. EVALUATION MEASURES

The results reported for MIREX editions 2010, 2011 and 2012 are limited to a single value per song. While the measures used for editions 2008 and 2009 are thoroughly discussed in Harte's thesis [5], the only information about the measure of the 2010 edition (which has been kept for the following editions) is its source code[1]. Therefore we start by describing it in terms of our newly defined scheme and comparing it with the previous evaluation versions. The advantages and drawbacks of these methods are then discussed, after which some alternative measures are proposed.

In contrast to the 2008 and 2009 versions, the 2010 edition doesn't use a mapping. This is of course the most simple solution, but has the disadvantage that the scoring function needs to be defined for every possible chord pair. The 2008 and 2009 versions followed an orthogonal approach, all chords are mapped to either major, minor or the no-chord. A particularity of both mappings, although they are not exactly equal, is that every possible chord is mapped to one of these three options, so there is no input for which the mapping is undefined. While it has the advantage that absolutely all available data is used for the evaluation, some of the mappings are rather stretched and cause a bias towards major chord types. A more detailed description of both mappings as well as their criticisms can be found in [5].

The scoring function from MIREX 2010 is of the bag of chromas category. First, reference and estimated chords (including their bass notes) are converted into a set of chromas. Then the cardinality of the intersection is taken, ignoring pitch spelling. If it is 3 or more, the segment receives a score of 1, otherwise its score is 0. For segments whose reference chord can be mapped to an augmented or a diminished triad, the threshold is lowered to 2 or more chromas. Of course, segments annotated with the no-chord, thus having an empty chroma set, are only assigned a score of 1 when the estimation also equals the no-chord symbol.

This method suffers from some important drawbacks. First of all, like all bag of chromas scoring functions, the root chroma looses its special position among other chromas: no distinction is made between chords that differ in root, but contain the same chromas. Secondly, there are no musicological underpinnings that warrant the special treatment of the augmented and diminished chords, nor can two-chroma chords ever be estimated correctly. Lastly, generating superfluous chromas in the estimated chord is not penalized at all. In comparison, the MIREX editions before 2010 used the "exact" scoring function.

In addition to the "Mirex2010" evaluation measure, we will use some other metrics. All have a scoring function that ignores the pitch spelling. The first two, named "Triads" and "Tetrads" use the aforementioned "triads" and "tetrads & triads" mapping. The mapped chords are then compared using the "exact" scoring function. From these, we derive two more measures that use a limiting set, in contrast to the first two. "TriadsInput" is equal to "Triads" with the addition of an input-limiting set that consists of all triads, including inversions. Next is "OnlyTetrads", similar to "Tetrads" but with an output limiting set containing all tetrads, but no triads. The following two metrics don't use a mapping, but have a simple scoring function. "Root" assigns a score of 1 only to chord pairs that have the same root and "Bass" does the same for bass notes. When a bass chroma is not explicitly given, we assume it is the root. The "Root" metric has also been used in MIREX 2008 and 2009 as an additional evaluation. Finally, we have two measures, "ChromaPrecision" and "ChromaRecall" that don't use a mapping and calculate precision and recall for the bags of chromas.

4. A NEW LOOK AT EXISTING DATA

Since the introduction of the NEMA (Networked Environment for Music Analysis) platform [23] for MIREX 2010, all algorithmic output and the used ground truth has been publicly available, offering us output from 42 systems for two data sets. The first one has been used in editions 2010, 2011 and 2012 and is named "Isophonics" [3]. It consists for the main part of Beatles songs (180) with some additional songs by Queen (19) and Zweieck (18). The second one has only been used in 2012 and is called "Billboard" [4]. It contains 197 songs that have charted in the Billboard 100 between 1958 and 1991. Unfortunately, only the annotations according to triads (without inversions) are publicly available for the latter set, therefore it is exclusively used in combination with the "Mirex2010", "Triads", "Root" and "ChromaRecall" metrics. For other mappings, it is impossible to verify whether any chromas that have been estimated in addition to the triad are correct and just haven't been annotated or whether they result from an incorrect estimation. Due to space constraints we have not printed all results here, but a selection of the most interesting can be found in Table 2 and Table 1 for the "Billboard" and "Isophonics" sets respectively. We retained the letter identifiers for the algorithms as used on the MIREX website[2], with the addition of the year in which they participated. The complete results for all the algorithms, as well as the algorithmic output used to calculate them, can be found on-line[3].

Comparing the two data sets, we see that the best results on the "Billboard" set are 10 % lower than on the "Isophonics" set, and this for all evaluation measures. This can be explained by the fact the "Isophonics" set has been publicly available before the contest, while the "Billboard" set has been secret. It is therefore likely that algorithms are optimised for the former.

The "Triads" scores are consistently lower than the "Mirex2010"[4] ones, but the extent to which they are lesser depends on the algorithm. For example, the algorithm with the best "Mirex2010" score,

[1] `http://nemadiy.googlecode.com/svn/`

[2] `http://www.music-ir.org/mirex/wiki/MIREX_HOME`

[3] `https://github.com/jpauwels/mirex-tools`

[4] Our reported results for the "Mirex2010" score differ somewhat from the "Weighted average overlap ratio" score on the MIREX site because of differences in the evaluation of no-chords and in the chord parsing

Algorithm	Mirex2010	Triads	Tetrads	TriadsInput	TetradsOnly	Root	Bass	ChromaRecall	ChromaPrecision
CWB1 ('10) [2]	77.60	76.54	65.95	78.21	0.00	80.28	77.68	82.10	85.60
EW4 ('10) [11]	77.79	77.13	66.79	79.29	0.00	80.99	79.39	81.97	85.33
KO1 ('10) [12]	76.95	76.53	65.08	77.30	0.00	79.65	78.62	80.49	83.99
MD1 ('10) [6]	77.96	76.42	65.56	78.98	17.66	79.69	76.83	82.91	84.49
MM1 ('10) [13]	77.42	74.46	53.39	76.81	36.32	79.54	78.88	84.05	80.47
OFG1 ('10) [14]	73.07	71.80	62.47	74.37	0.00	75.76	75.21	79.34	82.69
UUOS1 ('10/'11) [15]	77.09	75.99	65.93	78.41	0.00	79.34	77.88	81.04	84.46
CB2 ('11) [16]	79.29	78.08	67.55	80.15	0.00	81.65	79.93	82.95	86.47
KO1 ('11/'12) [17]	81.40	**80.69**	**73.88**	**82.13**	**52.65**	**82.92**	82.06	86.13	86.61
NM1 ('11) [18]	80.35	79.54	68.51	81.34	0.00	82.92	**82.42**	83.48	**86.97**
NMSD3 ('11) [18]	81.40	79.01	68.36	81.17	46.12	81.42	81.01	86.30	85.67
CCSS1 ('12) [19]	77.14	76.18	65.23	77.35	0.00	80.08	78.47	80.83	84.37
NG1 ('12) [20]	73.32	72.13	62.61	74.38	0.00	75.78	75.50	79.11	82.49
NMSD1 ('12) [18]	**82.15**	70.50	63.88	76.00	19.24	72.80	81.93	**87.09**	86.04
PMP1 ('12) [10]	74.76	73.68	64.00	76.19	0.00	76.80	75.57	80.07	83.43

Table 1. Evaluation scores for the "Isophonics" set.

Algorithm	Mirex2010	Triads	Root	ChromaRecall
CCSS1 ('12) [19]	66.21	65.90	71.33	75.97
DMW1 ('12) [21]	62.65	62.33	69.71	74.81
KO1 ('11/'12) [17]	69.80	**69.31**	**73.91**	79.03
NG1 ('12) [20]	62.49	62.27	66.99	74.56
NMSD4 ('12) [22]	**72.51**	60.45	63.72	**81.40**
PMP1 ('12) [10]	67.29	67.01	70.96	77.10

Table 2. Evaluation scores for the "Billboard" set.

"NMSD1/4 ('12)", decreases the most in the ranking. This is because the "Mirex2010" score does not require the root to be correctly estimated to get a good score, whereas the "Triads" score does. This explanation is corroborated by the "Root" scores. Consequently, we conclude that this algorithm is very capable of retrieving chromas in a chord, but does not manage to identify the root very well.

Overall, the score lowers even more for "Tetrads", but here too some distinctions can be made between algorithms. For most of them, the relative decrease is almost 15 %, but there are outliers in both directions. The relative decrease for "MM1 ('10)" is almost 30 %, significantly worse than when no tetrads would be estimated at all, while the decrease for "KO1 ('11/'12)" and "NMSD1 ('12)" is less than 10 %. The picture gets clearer when we only look at chords that are mapped to tetrads (the "TetradsOnly" score). Most algorithms that decrease 15 % in score actually can not generate triads at all. The 15 % just corresponds to the proportion of the annotations that are mapped to a tetrad. Only 5 algorithms do generate tetrads, but their estimation is lower than for exact triads ("TriadsInput" score) in all cases, ranging between 18 % and 53 % instead of 77–82 %. However, the performance on the tetrads themselves does not entirely reflect the ranking of "Tetrads" (in this case, "MM1 ('10)" is significantly better than "NMSD1 ('12)"). This brings us to the conclusion that due to the relatively small proportion of tetrads to triads, attention should be paid as to not overestimate triads as tetrads before any possible gain due to the increased chord vocabulary can be expected. It should be noted that the frequency of occurrence of tetrads in relation to triads is strongly dependent on genre, for example tetrads are more used in jazz than in rock music, so the results are likely to change for other data sets, and as a result, different algorithms can have different "preferences" for certain genres.

So far, we have ignored all bass notes during evaluation. There are however algorithms that have extended their vocabulary to be able to generate chord inversions. The measure "Bass" is specifically designed to evaluate this. It is somewhat surprising that the two best performing algorithms, "NM1 ('11)" and "KO1 ('11/'12)", can not generate inversions, so by default they always estimate root position. The explanation is that the ratio of inversions to ground positions is rather small, just like the tetrads/triads ratio. So similarly to the latter, the benefit of possibly estimating the right inversion is not enough to outweigh the option to wrongly estimate the bass on a root position, at least not with the current implementations.

As can be seen with the two previous metrics, it is obvious that a strongly skewed chord frequency makes it hard to evaluate improvements on less frequently occurring chords. A way to deal with this imbalance is to weigh all chords after mapping $\mathcal{C}_{MO}$ equally instead of according to duration. The drawback is that the mapping needs to be chosen in function of the estimation algorithm such that there are no mapped chords that cannot be generated by the algorithm (i.e. $\mathcal{M}(\mathcal{C}_{EST}) = \mathcal{C}_{MO}$), as that would strongly influence the average. Consequently, it is useful for per algorithm evaluations, but not for a large scale one, as that would necessitate a mapping that produces the lowest common estimation vocabulary. In our case, it would be a mapping with $\mathcal{C}_{MO} = \{\mathrm{maj}, \mathrm{min}\}$ only, thereby defeating the purpose of a more in-depth evaluation.

Lastly, we take a look at the bag of chromas evaluation measures. "ChromaRecall" is similar to the "Mirex2010" metric, but without threshold that binarises the score. Therefore the results are consistently higher. The best performing algorithms are those that can generate tetrads, the others are penalised because they simply cannot generate enough chromas at times. In combination with the "ChromaPrecision" measure, we can conclude that of those 5 tetrad generating algorithms, "MM1 ('10)" clearly has a tendency to overestimate triads as tetrads, the others are better balanced.

5. CONCLUSIONS AND FURTHER WORK

We started this paper by presenting a scheme to describe chord evaluation measures, extending the work of Harte [5]. Then the measures as used in the MIREX competitions from 2008 till 2012 have been described according to the proposed scheme as well as some alternative ones. Finally, we used these metrics to analyse the raw data coming from the competition. This showed us that the effects of a change in evaluation procedure is not equal for all algorithms: some maintain a consistent ranking, while others only excel in one particular measure. In general, we can conclude that extending the vocabulary of algorithms towards tetrads and inversions does not necessarily result in a better estimation, although the state-of-the-art is able to estimate some tetrads correctly, albeit not as well as triads. Future algorithmic improvements are therefore advised to verify progress in that domain by using similar evaluation measures as presented here.

In the future, we'd like to carry out a similarly detailed evaluation of tetrads and inversions on the "Billboard" set if we can get access to more detailed annotations. We also hope to integrate this extended evaluation directly into the NEMA framework so that future MIREX editions can directly benefit from a more extensive evaluation.

6. REFERENCES

[1] Hélène Papadopoulos and Geoffroy Peeters, "Large-scale study of chord estimation algorithms based on chroma representation and HMM," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2007, pp. 53–60.

[2] Taemin Cho, Ron J. Weiss, and Juan P. Bello, "Exploring common variations in state of the art chord recognition systems," in *Proceedings of the Sound and Music Computing Conference*, 2010.

[3] Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Chris Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler, "OMRAS2 metadata project 2009," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009.

[4] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga, "An expert ground-truth set for audio chord recognition and music analysis," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011, pp. 633–638.

[5] Christopher Harte, *Towards automatic extraction of harmony information from music signals*, Ph.D. thesis, Queen Mary, University of London, August 2010.

[6] Matthias Mauch, *Automatic chord transcription from audio using computational models of musical context*, Ph.D. thesis, School of Electronic Engineering and Computer Science Queen Mary, University of London, March 2010.

[7] Hélène Papadopoulos, *Joint estimation of musical content information from an audio signals*, Ph.D. thesis, Université Pierre et Marie Curie, July 2010.

[8] Laurent Oudre, *Template-based chord recognition from audio signals*, Ph.D. thesis, Telécom ParisTech, November 2010.

[9] Maksim Khadkevich, *Music signal processing for automatic extraction of harmonic and rhythmic information*, Ph.D. thesis, University of Trento, 2011.

[10] Johan Pauwels, Jean-Pierre Martens, and Marc Leman, "Improving the key extraction accuracy of a simultaneous key and chord estimation system," in *Proceedings of the International Workshop on Advances in Music Information Research (AdMIRe)*, 2011.

[11] Adrian Weller, Daniel Ellis, and Tony Jebara, "Structured prediction models for chord transcription of music audio," in *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2009, pp. 590 – 595.

[12] Maksim Khadkevich and Maurizio Omologo, "Use of hidden Markov models and factored language models for automatic chord recognition," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 561–566.

[13] Matthias Mauch and Simon Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, 2010, pp. 135–140.

[14] Laurent Oudre, Cédric Févotte, and Yves Grenier, "Probabilistic template-based chord recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2249 – 2259, November 2011.

[15] Yushi Ueda, Yuki Uchiyama, Nobutaka Ono, and Shigeki Sagayama, "Mirex 2010: Joint recognition of key and chord from music audio signals using key-modulaton HMMs," in *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*, 2010.

[16] Taemin Cho and Juan Pablo Bello, "A feature smoothing method for chord recognition using recurrence plots," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011, pp. 651–656.

[17] Maksim Khadkevich and Maurizio Omologo, "Time-frequency reassigned features for automatic chord recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, December 2011, pp. 181 – 184.

[18] Yizhao Ni, Matt McVicar, Raúl Santos-Rodríguez, and Tijl De Bie, "An end-to-end machine learning system for harmonic analysis of music," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 6, pp. 1771–1783, August 2012.

[19] Ruofeng Chen, Weibin Shen, Ajay Srinivasamurthy, and Parag Chordia, "Chord recognition using duration-explicit hidden Markov models," in *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, 2012, pp. 445 – 450.

[20] Nikolay Glazyrin, "Audio chord estimation using chroma reduced spectrogram and self-similarity," in *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*, 2012.

[21] W. Bas de Haas, José Pedro Magalhães, and Frans Wiering, "Improving audio chord transcription by exploiting harmonic and metric knowledge," in *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, 2012, pp. 295 – 300.

[22] Yizhao Ni, Matt McVicar, Raúl Santos-Rodríguez, and Tijl De Bie, "Using hyper-genre training to explore genre information for automatic chord estimation," in *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, 2012, pp. 109 – 114.

[23] Kris West, Amit Kumar, Andrew Shirk, Guojun Zhu, J. Stephen Downie, Andreas Ehmann, and Mert Bay, "The networked environment for music analysis (NEMA)," in *IEEE Fourth International Workshop on Scientific Workflows (SWF)*, 2010, pp. 314 – 317.