Perceptual Validation of Chord Estimation Evaluation Standards

Jorge A. Cuarón Sánchez

MASTER THESIS UPF / 2014
Master in Sound and Music Computing

Master thesis supervisor:
Agustín Martorell Domínguez
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

Abstract

In the last years chord estimation has become important in Music Information Retrieval (MIR). The abundance of algorithms has driven a need for effective evaluation methods that are able to depict correctly their accuracy. In the last six years, one of the greatest efforts for this task has been done by MIREX, improving progressively with each yearly edition. To better portray the results, MIREX makes use of different chord vocabularies for this task. For the simpler chord vocabularies, more complex chords are mapped to simpler versions that contain shared notes. This simplification makes the assumption that the chords are very similar and can be considered as equivalent for said vocabulary. In this work, a test is proposed in order to validate that the mappings of more complex chords to simpler chords are perceptually relevant from a user centered perspective.

Contents

1 Introduction

Chord estimation algorithms have a wide variety of uses, ranging from direct transcription to being used as descriptors for tasks such as music similarity and segmentation. Since chord estimation algorithms can be a very meaningful source of information, it is of importance to have the most effective estimation algorithms. This brings the problem of evaluation.

To measure how good the estimations are, it is necessary to have reliable methods of evaluation. Nowadays, the most common method to evaluate these algorithms is by comparing their outputs against a ground truth and calculating different measures. Systematic evaluation for chord estimation is based on the labels that algorithms output to represent a chord. These labels are conventions that have an inherent musical meaning, but do not always represent the actual chord that was being played. This poses a *lost in translation* situation, where not all of the information is preserved.

1.1 Goal

Lately, one of the most important evaluation tasks for different Music Information Retrieval (MIR), including chord estimation, is done in the Music Information Retrieval Evaluation eXchange (MIREX). MIREX uses systematic evaluation to assess the performance of the different algorithms that are submitted yearly. Part of this evaluation for chord estimation is divided into 5 different categories, based on the complexity of the chord vocabulary for each category. For categories containing a simpler chord vocabulary, more complex chords are mapped to a chord considered as equivalent based on the notes they share. In this way, a 4 note chord will be mapped to a 3 note chord that contains the same notes, ignoring that extra note. In some cases this extra note gives a very characteristic sound to the chord, and it can sound significantly different than its 3 note simplification.

In these cases, if the evaluated chord shares the same basic triad as its 3 note simplification (e.g. C7 and Cmaj), the chord estimation will be considered as being correct against the ground truth. Perceptually, however, the chords might not be that highly related and may have a different effect on the listener. It is important to know this because when comparing algorithm performances, sometimes the simper vocabularies (major and minor chords) are the evaluation measures that are heeded first since they are more widely used. To compare algorithm performances MIREX sorts them based on the major-minor vocabulary, which is simplified to only contain two chord types as its name implies. This category makes use of the chord mappings previously described.

For this reason a perceptual evaluation test can be of great interest to validate whether the chord mappings made are perceptually relevant. Since most chord estimation end tasks are user centered, even if used as an intermediate step, it is pertinent to determine if all of these chord mapping assumptions are valid from this perspective.

Also, a brief review of some underlying concepts and previous work done is important to fully understand the aim of test to be done and the conclusions that can be reached.

2 Literature Review

2.1 Introduction

In this section a brief description of chord theory will be explained. After grasping the basic concepts of chord theory, an overview will be given for the most common chord estimation methods. This will be complemented by a short description of state of the art algorithms.

A summary of how algorithms have been evaluated will be recounted, as well as the problems that have risen, and how they have evolved in the last years. The section will especially focus on MIREX, since it has been the center for evaluation with standard rules for each edition. This has permitted objective comparisons between different algorithms for the same task.

2.2 Tonality and Chords

2.2.1 Definitions

In order to avoid ambiguity when describing some of the ideas that follow, the next concepts should be defined first:

Pitch is the fundamental concept, the basis of which all tonality and harmony is based on. Because of the lack of accurate definitions, many have been proposed such as *"..that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low"* by the American National Standards Institute[ANSI, 1973]. An expansion of the definition more related to MIR is given by Klapuri[Klapuri, 2006], stating that *"pitch is a perceptual attribute which allows the ordering of sounds on a frequency-related scale extending from low to high"*.

Furthermore, pitch sensation is subject to some perceptual attributes such as pitch circularity. This last concept, described by Shepard [Shepard, 1964], can be illustrated when dealing with chords. Although a chord can have different inversions as seen in Section 2.2.2, the chord is regarded as the same even when the notes composing it change in octave. This is true for functional harmony, which is the most common in Western music.

Chords are generally defined as pitch simultaneity, which can be a series of two or more notes grouped together. Benward & Saker[Benward and Saker, 2009] define a chord as *"a harmonic unit with at least three different tones sounding simultaneously"*. This definition can be a little weak and given to debate, since a dyad (two notes) could also be considered a chord (such as root and third, or root and fifth). Also, arpeggios can be generally recognized as chords even when they are not played simultaneously.

2.2.2 Chord Theory

Chords are built based on a musical scale, which in Western music corresponds to 7 notes. Chords built up from these scales will always be referred to according to their root as a *first degree* (I), *second degree* (II), up to the *seventh degree* (VII)[Shoenberg, 1969]. The most common chords are composed of three notes and denominated as triads. These most basic triads are given by the root, third and fifth intervals. When based on a major scale, these intervals will generate a sequence of chords (major, minor, minor, major, major, minor, diminished) for each scale starting on the root note, as seen in Figure 1. Chords can be extended by adding additional notes to create tetrads or more complex chords with even more notes.

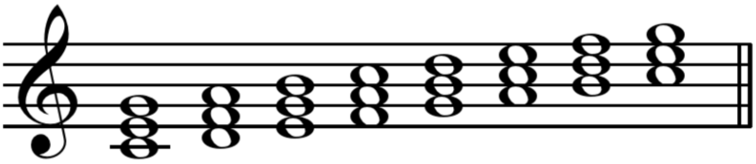


Figure 1: Triads for C major scale

These variations can become very large, and thus distanced from the original triad chords. When the chord possibilities grow it is important to label them correctly in order to avoid confusions and be as accurate as possible when representing them.

2.2.3 Notation

To label a chord correctly a type of notation that permits the representation of the chord without lending itself to confusion is desirable. Traditionally, a chord can be represented with standard musical notation as shown in Figure 1.

Apart from traditional musical notation, there are many ways in which chords can be represented. Benward & Saker[Benward and Saker, 2009] describe 3 of the most common methods for labeling chords: Roman numerals, macro analysis symbols, and popular music symbols. The roman numeral notation shows the chord as roman numerals, with upper case for major chords, lower case for minor chords, and a modifier for augmented and diminished chords. It has the advantage of showing the interval relationship with the root note based on a scale. This can also be a disadvantage, when the scale is unknown.

The macro analysis method serves the purpose of revealing harmonic gesture that might not be so evident with other types of notations. The system uses letters to denote the chords, distinguishing major chords with upper case, minor chords with lower case, and diminished and augmented chords by adding the ^{o} and $^+$ symbol. This notation is shown in Figure 2

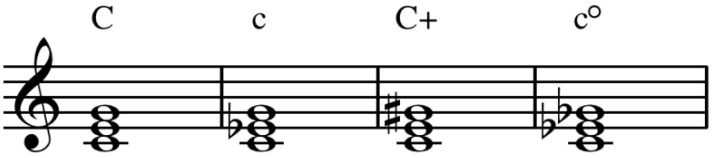


Figure 2: Macro Analysis notation for C major, C minor, C augmented, and C diminished.

The popular music symbols where developed as a short hand notation to aid musicians when performing. This is especially useful when improvising, since the notation will only mark the chords to be played without indicating the exact notes to play. Major chords are represented by a capital letter designating the root. Minor, diminished, and augmented chords correspond to the same letter with additional suffixes such as Mi, dim, and + added respectively. The bass can be denoted by a slash followed by the note. Figure 3 shows an example of this notation.

These type of traditional notations have served the purpose of simplifying for performers for years. However, a new problem with labeling chords has recently arisen when porting them to plain text used by computers. Having the notation in computers can serve a double purpose, being read by humans, or being parsed by computers to perform some operations or an algorithm. This problem has been tackled in the past few years, which resulted in a new labeling proposal as shown in [Harte et al., 2005] and [Harte, 2010].

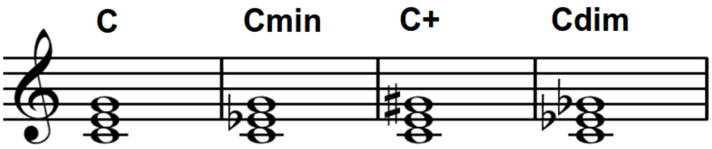


Figure 3: Popular Notation for for C major, C minor, C augmented, and C diminished.

In the proposed methods, it is noted that text annotations should comply certain characteristics. These include being context independent, unambiguous, flexible, intuitive for humans, and machine readable. Context independent means that a single chord must be able to be transcribed even without a context, which cannot be done when no key is given. Unambiguous refers to the notation being clear in every aspect. For example, in an ambiguous system the symbol b could refer to a flat modifier or to a B minor chord. A flexible system will allow for unconventional chords to be notated as well. Lastly, being human intuitive and machine parsable will make the notation easy to transcribe while allowing computer operations to be handled. The complete syntax as proposed in [Harte et al., 2005] can be seen in Table 1.

The proposed notation is consistent with the objectives and has been used for the MIREX evaluation since the year 2009. The notation becomes human readable by adding the *shorthand notation*, which is a predetermined type of popular notation. With this syntax a D minor seventh with an added 13th in second inversion could be written as:

```
D:min7(13)/5  =  D:(b3, 5, b7, 13)/5
```

Having an appropriate descriptive notation for chords is important to avoid as much information loss as possible. This is especially true when they are going to be used as intermediate steps for other tasks.

2.3 Chord Recognition in MIR

Chord recognition from audio is a task which has become increasingly popular since the beginning of the 21st century. There have been many advances, especially in the last few years. Detecting the harmonic structure can be very useful for many tasks, and chord recognition is a common method for doing so. Some of these tasks include segmenting, automatic transcription, structure analysis, tonal analysis, etc.

Table 1: Syntax for chord notation from [Harte et al., 2005]

```
<chord> ::= <pitchname> ":" <shorthand> ["("<ilist>")"]["/"<interval>]
          | <pitchname> ":" "("<ilist>")" ["/"<interval>]
          | <pitchname> ["/"<interval>]
          | "N"

<pitchname> ::= <natural> | <pitchname> <modifier>

<natural> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G"

<modifier> ::= "b" | "#"

<ilist> ::= ["*"] <interval> ["," <ilist>]

<interval> ::= <degree> | <modifier> <interval>

<degree> ::= <digit> | <digit> <degree> | <degree> "0"

<digit> ::= "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"

<shorthand> ::= "maj" | "min" | "dim" | "aug" | "maj7" | "min7" | "7"
              | "dim7" | "hdim7" | "minmaj7" | "maj6" | "min6" | "9"
              | "maj9" | "min9" | "sus2" | "sus4"
```

In general, chord recognition can be a difficult task by detecting the individual notes played by one or many instruments. A more effective approach, as introduced by Fujishima[Fujishima, 1999] suggests using Pitch Class Profiles (PCPs) as a tool. The underlying proposal has become the standard approach for chord recognition systems. As seen in Section 2.2.2, pitch has a chroma property, in which the same note in different octaves are perceived as similar. PCPs take advantage of this perceptual property of octave equivalence to weight the notes in all octaves and summarize them into a 12 bin vector.

Figure 4 shows a standard approach for chord detection algorithms, with common methods for each of the steps. These methods are explained subsequently.

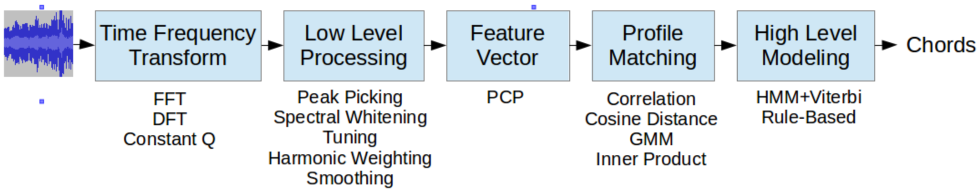


Figure 4: Chord detection flow diagram with different common methods for each step

2.3.1 Time Frequency Transforms

Discrete Fourier Transform (DFT) It is the most common time frequency transform, which is most regularly implemented through the Fast Fourier Transform algorithm (FFT) algorithm. It is normally preceded by a windowing and a zero padding function and it transforms the signal from time domain into frequency domain.

Constant Q Transform It is based on the Fourier Transform with a few differences. Based on the fact that human hearing perception is logarithmic, it is known that pitch sensibility on the lower range is more sensitive to frequency changes than in a higher range. In a DFT, the resolution is linear and changes depending on the window size used. The constant Q transform maintains a constant ratio of frequency resolution, which is normally around one quarter-tone. The result is a function in which the window length changes according to the frequency. For additional information refer to [Brown, 1991].

2.3.2 Low Level Processing

Peak Picking It refers to selecting peaks that are relevant for the signal. After a frequency domain transform, not all the information is relevant and therefore only peaks which convey information are chosen. The peaks can be chosen with a quadratic interpolation to improve resolution since a frame of N samples can only yield N frequency components. This technique is used in [Gómez, 2006] and shown in detail in [Serra, 1997].

Spectral Whitening It is a technique where the spectral peaks are normalized so that they become flat. It is called spectral whitening because white noise tends to have a flat spectrum. This process makes an algorithm more robust against timbre changes, since harmonics will end with a similar contribution to the spectrum.

Tuning Recorded songs are rarely tuned against 440 Hz. This process compensates for the differences when matching each frequency to its corresponding bin. This is done by estimating the frequency to which the song is tuned against and readjusting the mapping matrix to have the estimated frequency as a center. For more details refer to [Mauch, 2010].

Harmonic Weighting This process takes into account that any given frequency being mapped to its corresponding bin can be the harmonic of a lower F0 frequency of another bin. In this way, when the frequency is mapped against its corresponding bin it uses an exponential function to also weight some contribution to the other PCP bins whose frequency is a subharmonic of the frequency being mapped. This technique was introduced in [Gómez, 2006].

Smoothing One of the most common smoothing techniques is median smoothing. This technique uses a moving window of length n for each sample and calculates the median value. This type of filtering is useful for noise reduction, since it can remove spurious peaks in signals which are regularly caused by noise.

2.3.3 Feature Vector

Pitch Class Profiles Introduced by Fujishima [Fujishima, 1999], PCPs are a very effective way of conveying tonal information into a $n * 12$ bin vector, where each bin corresponds to a subdivision of $1/n$ of a semitone. Most commonly a PCP is a 12 bin vector with each of the bins corresponding to one semitone of the chromatic scale. The most basic way of calculating a PCP includes no low level processing and maps each frequency corresponding to a note in the chromatic scale over a defined frequency range. Same notes that span across different octaves are mapped to the same bin, adding more weight.

2.3.4 Similarity Measures

Correlation It can measure the correlation between a chord template and the obtained PCP. As proposed in [Gómez, 2006] it can be calculated by (1) where PCP and T are the PCP and chord template respectively, having expected values of μ_{PCP} and μ_T with a standard deviation of σ_{PCP} and σ_T.

$$\frac{E[(PCP - \mu_{PCP}) \cdot (T - \mu_T)]}{\sigma_{PCP} \cdot \sigma_T} \tag{1}$$

Cosine Distance It measures the cosine of the angle between two vectors. Since it is a measure of orientation, the result will be 1 when the vectors align perfectly, 0 when they are orthogonal with no apparent relationship, and -1 when the vectors are in opposite directions. It can be calculated by (2), where PCP is the PCP vector and T is the chord template vector.

$$cos(\theta) = \frac{PCP \cdot T}{\|PCP\| \, \|T\|} \tag{2}$$

Inner Product The inner product, or dot product is the multiplication of two vectors with a scalar result. When vectors are orthogonal the result will be 0. On the other hand when a vector V is multiplied by itself the result will be $\|V\|^2$.

Gaussian Mixture Models Single Gaussians or Gaussian mixture models can be used to model the PCP vector distribution for each chord. Gaussians are described by a mean vector μ_i and its covariance Σ_i. These metrics can yield a probability which can then be fed to a Hidden Markov Model or other probabilistic method to track a chord over time. A mixture model permits representing subpopulations of chord Gaussians which belong to the general population, thus allowing to characterize the same chord by different models.

2.3.5 High Level Modeling

Hidden Markov Models They can be used to track probability of change according to past states. The states in the HMM are the chords to be recognized, which could be as little as 24 for major and minor chords. Because HMM use past states, they can be very useful to predict chords according to the past states seen, which is the equivalent of predicting a chord based on the progression. Given the observations, the most likely chord is estimated having the chords as hidden variables and the spectral features as observed variables. This model has the advantage of incorporating musical training into the chord recognition system. One of the first implementations of this method for chord estimation is described in [Sheh and Ellis, 2003].

Rule-Based Models Custom rules can be applied to further refine the results of a chord estimation algorithm. Such rules can be very different from each other depending on the purpose intended. A simple example can be that given a chord estimation sequence on a short window based approach, replace any outliers with the mode of nearby chords. This could clean the sequence `Emin-Emin-Gmaj-Emin-Emin` where the `Gmaj` chord appears to be a misestimation, into the sequence `Emin-Emin-Emin-Emin-Emin`.

Most state of the art algorithms use a layered combination of these methods. Their differences in architecture can be rated when they are benchmarked in tests such as MIREX.

2.4 MIREX 2013 algorithms overview

Every year MIREX is held for different evaluation tasks. The audio chord detection task was introduced in MIREX in 2008. Although not all state of the art algorithms are sent to MIREX for evaluation, it does give a very good overview of what many state of the art algorithms are working on. Table 2 shows a summary of the methods used for all the algorithms that were submitted for evaluation in 2013, labeled as CB[Cho and Bello, 2013], CF[Cannam et al., 2013] & [Mauch, 2010], KO[Khadkevich and Omologo, 2011], NG[Glazyrin, 2013], NMSD[Ni et al., 2012], PP[Pauwels and Peeters, 2013b], SB[Steenbergen and Burgoyne, 2013]. Algorithms which were submitted twice as pre-trained and to be trained are merged into the same one.

Table 2: Overview of Audio Chord Recognition algorithms for MIREX 2013

Algorithm	**Time-Frequency Transform**	**Low Level Processing**	**Feature Vector**	**Profile Matching**	**High Level Modeling**
CB	Constant Q	Sub-band separation	6 sub-band chroma	GMM	Multi-Stream HMM
CF	Constant Q	Tuning, spectral whitening	Chroma Vector	Parametric expert function	HMM-Viterbi
KO	STFT	Tuning, TFR[1]	Harmonic Reassigned Chroma	GMM	Multi-Stream HMM
NG	Constant Q	Tuning, beat synchronization, harmonic weight, smoothing	Chroma Vector	Euclidian Distance Self Similarity Matrix	Rule-Based
NMSD	Constant Q	Tuning, HPSS[2], beat synchronization	Bass/Treble Loudness based Chroma	Single Gaussians	Harrmony Progression Analyser based on HMM
PP	Constant Q	Tuning, HPSS, beat synchronization	Loudness Based Chroma	GMM	HMM
SB	Constant Q	N/A	Chroma Vector	GMM	HMM Neural Network

[1] Time-Frequency Reassignment

[2] Harmonic/Percussive Sound Separation

The different methods used in each of the algorithms will result in different outputs in the estimation of the algorithms. It is desirable to have descriptive evaluation methods in order to compare properly their effectiveness.

2.5 Evaluation Methods for Chord Recognition

As with all tasks, it is important to have good evaluation methods for chord estimation systems. Good evaluation systems permit a better comparison of different approaches and to measure their performance accurately. Ideally, they should also give valuable information of whether the performance is good for the end use which most of the times relies in user oriented applications.

2.5.1 Common Measurements of Effectiveness in Information Retrieval

In information retrieval it is important to measure the effectiveness of a system. According to [van Rijsbergen, 1979], effectiveness *”is purely a measure of the ability of the system to satisfy the user in terms of the relevance of documents retrieved”*. Since chord detection algorithms are being dealt with, it can be said that an effective chord estimation algorithm is one that satisfies the user in terms of the relevance of chords estimated.

For this, it is useful to categorize the documents retrieved as a cross tabulation, or contingency table. From this table it is possible to derive some effectiveness measures to help us better understand the results. This is shown in Table 3, where A denotes relevant retrievals (chords correctly estimated), B denotes number of retrievals (number of chord estimations), and N denotes the total.

Table 3: Contingency Table

	Relevant	Non-Relevant	
Retrieved	$A \cap B$	$\overline{A} \cap B$	B
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$	$\overline{B}$
	A	$\overline{A}$	N

Precision It is the ratio of retrieved items that are relevant for a given query. This can also be interpreted as the fraction of relevant items out of all the items retrieved. Based on Table 3 it is defined in)3). A weakness of this measure when used alone is that it can be misleading when only few items are retrieved, for example, only when the confidence level is high.

$$\frac{A \cap B}{|B|} \tag{3}$$

Recall It is the the ratio of relevant items retrieved for a given query. This can also be interpreted as the fraction of relevant items retrieved out of all the relevant items. Based on Table 3 it is defined in (4). This measure when used by itself does not convey all the information needed. If too many retrievals are made (with many false positives), it can give a good score if most of the relevant items are fetched.

$$\frac{A \cap B}{|A|} \tag{4}$$

F-measure It is the harmonic mean of precision and recall. It is a better single metric when compared to precision and recall because both of them give different information that can complement each other when combined. If one of them excels more than the other, this metric will reflect it. It can be calculated by (5).

$$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

Precision, recall and F-measure can be very useful metrics in information retrieval systems, however they are not fully appropriate by themselves for a chord recognition task. Given that a chord estimation has to be time aligned, these metrics fail to display this side of the evaluation when the definition is taken as defined above.

2.5.2 Chord Estimation Evaluation in MIREX before 2012

Starting in 2008 MIREX has been evaluating audio chord estimation for all the algorithms submitted. The original proposal of metrics was done with a custom definition of precision and recall that takes into account that the task is time variable. The definitions from the 2008 contest are the following:

Recall *The number of time units where the chords have been correctly identified by the algorithm divided by the number of time units which contain detectable chords in the ground truth.*

Precision *The number of time units where the chords have been correctly identified by the algorithm divided by the total number of time units where the algorithm detected a chord event.*

Since the first edition the actual evaluation was done by measuring the overlap of the predicted chords against an annotated ground truth for each song in the dataset and then dividing it over the duration of the chord. Overlap can be calculated by (6). After, an average was made for all the songs to give the final score to each algorithm.

Another overlap metric was shown where the major and minor chords with the same root note were counted as a positive prediction when measuring the overlap. It is worth noting that the this chord evaluation was limited to only major and minor chords, truncating more complex chords into these categories (such as a Cmaj7 into a Cmaj). Also, an extra *chord N* is included to represent segments where no chord is played.

$$\text{Overlap} = \frac{\text{summed duration of correct chords}}{\text{total duration}} \tag{6}$$

When looking at the overlap formula, it can be seen that it is the same as the definition given for recall. Although the overlap measure can give a fair idea of the performance of an algorithm, it falls short when trying to make a more detailed analysis. It is possible to look at individual chord labels of each song to have a better notion of the performance, however this becomes unpractical.

2.5.3 Other Proposals

Mauch proposes in his Ph. D dissertation[Mauch, 2010] to also take into account chord segmentation for the evaluation. Correct alignment of chord estimations to the ground truth is something that has to be considered, since a quality chord estimation should be as close to the ground truth as possible. If there is an overlap score of 100% the alignment will also be 100% correct. However, in most cases it doesn't happen, and a segmentation metric can give more information regarding the performance of the algorithm.

The proposed metric is the *directional Hamming divergence*, which measures how much of the segments do not overlap with the maximum overlapping segments. It can be calculated by (7), where $B = B_i$ and $B^0 = B_i^0$ are the two segments.

$$h(B||B^0) = \sum_{i=1}^{N_B} (|B_i^0| - \max_j |B_i^0 \cap B_j|) \tag{7}$$

Based on this formula the segmentations $h(B||B^0)$ which shows B with respect to B^0 and $h(B^0||B)$ which shows B^0 with respect to B can be computed. Taking B^0 as the ground truth over-segmentation can be measured by calculating $h(B||B^0)$ and under-segmentation by calculating $h(B^0||B)$. Both cases yield useful information, and a smaller value means a better transcription. To consider both values, Mauch mentions that Chris Harte proposed in an online discussion to take 1 minus the maximum of the two values and normalizing by the duration of the song. This can be calculated by (8).

$$H(B||B^0) = 1 - \frac{1}{T} \max\{h(B||B^0), h(B^0||B)\} \in [0, 1] \tag{8}$$

Since the maximum value is being subtracted to 1 a high value indicates less segmentation, which depicts a more intuitive result.

2.6 Data Sets

The evaluation requires a ground truth to compare the results against. A data set can be generated using synthesized music, form where the ground truth is known. This method, nevertheless, can be ineffective because the timbre properties tend to differ a lot from real audio recordings. Since the synthesized sound is much cleaner, it is more likely that the chord estimation system will perform better. This type of evaluation is not fully representative of the end task which is audio chord recognition.

A method with more reliable results consists in having an annotated data set and evaluate the estimations by taking the annotations as the ground truth. The problem is the availability of such datasets, since it is very time consuming and takes a great effort to generate them. One of the most widely used datasets of chord annotations has been available since 2005, as presented in [Harte et al., 2005]. It consists of 180 songs from The Beatles out of 13 CDs.

These annotations were the first to become publicly available and are widely used. They also became the dataset for evaluation in MIREX. After this dataset was published in 2005, it remained as the only openly available until 2009 when additional annotations were made available in [Mauch et al., 2009]. This lack of diversity resulted in the algorithms becoming more and more tailored to this specific data set. Since many of the chord estimation systems involve machine learning, they became overfitted. The most recent and extensive dataset to be released is composed of more than 1000 songs taken form the *Billboard Hot 100* between 1958 and 1991.

2.6.1 The Beatles Dataset

This annotated dataset consists of 180 songs that are equivalent to 8 hours, 8 minutes and 53 seconds of total audio. The transcriptions were based on the harmonic analyses done by Pollack[Pollack, 2014]. It serves as a good data set for popular music, since The Beatles changed style gradually and have a certain diversity of style and sound. Earlier songs are of a more traditional rock-and-roll style, while the later songs tend to be more elaborate and in some cases orchestrated. It is described in detail in [Harte, 2010].

2.6.2 OMRAS2 Dataset

This dataset was presented in 2009 at the at the 11th Conference on Music Information Retrieval[Mauch et al., 2009], and it is comprised of four different annotation sets: beat and metric position, chords, key, and segmentation. The chord annotations include 14 pieces by Carole King, 20 by Queen, and 18 by Zweieck. Additionally, the data set contains the previously mentioned Beatles transcriptions. All the transcriptions were manually annotated at the Centre for Digital Music in Queen Mary University of London. No further information on the transcription process is given for this data set.

2.6.3 Billboard Hot 100 Dataset

Motivated by the lack of an extensive, reliable and more variated annotation dataset, Burgoyne et al[Burgoyne et al., 2011] underwent a project for creating a large one of more than 1000 songs, released in 2011. The songs were selected randomly from the *Billboard Hot 100* Chart, which is a weekly compilation that started in 1958 of the most popular singles in the United States. The songs selected range from 1958, when the chart started, to 1991. The end date was chosen mainly for two reasons, the first one being that starting from this date many Hip-Hop songs started to chart. These songs tend not to be as rich harmonically as other musical genres. The second reason is that after this date the chart was generated automatically with songs lingering for more time in the charts until a limit was imposed.

The transcription was done on plain text beginning by the song title, artist, meter, and key. The transcribers were given freedom to use the notation they found most natural, but afterwards all were converted to the format proposed in [Harte et al., 2005]. 17 persons were hired to do the transcriptions, they were either graduate students in musical performance, or professional jazz performers. Transcribers worked in pairs through a custom web interface, each doing their own annotations. A third transcriber would compare both versions and combine them into a final version, which was then time aligned and annotated with structural information.

The files are annotated in a custom format by bar with a resolution of up to eighth notes. The chords are generally written with the shorthand notation used in MIREX with a few additions. To make these transcriptions more useful, a parser for these files was developed by Haas and Burgoyne[Bas de Haas and Burgoyne, 2012]. This parser introduces the option of converting the format of the annotations into different formats. Some useful options include a mode for converting the files to the MIREX format, one with truncated shorthand notations, and the other displaying full note intervals. If a chord is unrecognized by the target vocabulary, the system introduces the X symbol.

2.6.4 Datasets used in MIREX

In 2012 MIREX introduced part of the Billboard dataset to the chord estimation task, giving room for comparison with the previous Beatles and OMRAS2 datasets. Pauwels and Peeters[Pauwels and Peeters, 2013a] conduct a good comparison of the datasets, in which they notice that the algorithms perform around 10% lower in the Billboard dataset than in the OMRAS2. This is reasonable, since the latter dataset has been available for a longer time making it more likely that the algorithms are optimized for the dataset.

In 2013, MIREX expanded the evaluation to include the segmentation metric proposed by Mauch and mentioned in Section 2.5.3. The segmentation evaluation takes into account over-segmentation, under-segmentation, as well as the harmonic mean between the two. This metric complements the existing overlap metric.

The evaluation was done with a more extended dataset. Taking advantage on the full release of the Billboard dataset, the evaluation was done separately on all the different available datasets: MIREX 2009, Billboard 2012, and Billboard 2013. This permits a better comparison between algorithms by showing the differences in performance.

The evaluation was also done for different chord dictionaries: root only, major and minor, major and minor plus inversions, sevenths, and sevenths plus inversions. Whenever the chord dictionary contained a simpler chord, the estimation was truncated to match the chord type. As mentioned in the MIREX webpage:[3] *"for instance, in the major and minor case, G:7(#9) is mapped to G:maj because the interval set of G:maj, {1,3,5}, is a subset of the interval set of the G:7(#9), {1,3,5,b7,#9}"*.

[3] `http://www.music-ir.org/mirex/wiki/2013:Audio_Chord_Estimation_Results_MIREX_2009`

3 Methodology

Given the current evaluation methods used for audio chord estimation, there is one major assumption done that is implicit in the methodology. This assumption is the mapping of more complex cords to its simplified versions, which considers that both chords are roughly equivalent. This makes sense from an analytic point of view, considering the fact that they share notes. However, since the chords are not identical, they might have meaningful perceptual differences. Because of this, it is important to validate whether the mappings to simplified chord vocabularies are perceptually relevant. To do this, a test with users can be done in which common misestimations are compared against the ground truth to determine if they are identified as being different. This test should be done in a relevant context, where the most frequent misestimations are the ones being tested.

3.1 Estimations Data Analysis

In order to determine the samples to be used in the test, it is necessary to know the behavior of the estimations when compared to the ground truth.

3.1.1 Chosen Dataset

For the latest audio chord estimation task in MIREX 2013, the algorithms were evaluated for three separate datasets labeled as MIREX 2009, Billboard 2012, and Billboard 2013. All of these datasets are comprised of collections described in Section 2.6. The MIREX 2009 dataset contains the Beatles, Queen, and Zweieck collections. The Billboard 2012 and 2013 datasets are both composed of songs from the Billboard dataset, with the difference that the 2012 dataset has its track-list publicly available, whereas the 2013 is still closed. MIREX released the output of all algorithms for each of these datasets, making its analysis possible.

From the 3 available datasets the Billboard 2012 was chosen for several reasons. Since the test to be done will have the need of playing an excerpt from the original audio as well as the algorithms' outputs, it is imperative to know to what song the outputs belong to. This prerequisite eliminated the possibility of using the Billboard 2013, since its track-list is not publicly available. From the two remaining datasets the Billboard 2012 posed as a better candidate since it has been available for less time, reducing the possibility of having the algorithms overfitted to it. As well, this dataset has more musical diversity which gives more variety to the chords estimated by the algorithms.

3.1.2 General Common Misestimations

For this dataset the output files consisted of 188 songs for each of the 12 algorithms, plus the ground truth for a total of 2444 output files. The output files contain two time stamps consisting of the start and end time for each chord, as well as the chord label with the notation proposed in [Harte et al., 2005] and mentioned in Section 2.2.3. These outputs were analyzed compared against the ground truth to generate some statistics concerning the most common misestimations. The purpose of having this data is to recognize appropriately the samples to be chosen for the test. They should represent the most common errors, as well as the most likely to happen under normal circumstances in songs similar to the dataset.

The first part of the data analysis consisted on a chord count of the estimations for all of the outputs (including the ground truth). This was done with a python script in which all the files were analyzed to look for all the different chords available and count them accordingly. The chord corpus of all the outputs is made up of 505 unique chord labels. Because the count was done on a label basis, a Db:maj chord would count as a different chord than C#:maj. The most common chords for the estimations on the Billboard 2012 dataset are shown in Table 4.

Table 4: Most common chords for all the outputs of the estimations on the Billboard 2012 dataset.

Chord	**Count**	**Percentage**	**Chord**	**Count**	**Percentage**
C:maj	28046	5.55%	Ab:maj	10052	1.99%
D:maj	27633	5.47%	B:maj	8510	1.68%
E:maj	27455	5.44%	G:7	8395	1.66%
G:maj	25997	5.15%	A:min	7485	1.48%
A:maj	24224	4.80%	D:min	6974	1.38%
N	19334	3.83%	E:7	6704	1.33%
Bb:maj	17852	3.53%	B:min	5589	1.11%
F:maj	17754	3.52%	Gb:maj	5438	1.08%
Eb:maj	10935	2.17%	E:min7	5342	1.06%
Db:maj	10254	2.03%	C:min7	5271	1.04%

Out of all these chords, a cross-check was made to see how many pairs of a Ground Truth Chord to an Estimated Chord existed. This count resulted in 15021 unique chord combinations (by comparing labels). It is worth mentioning that the count was made only on the intersections bigger than 75 ms, so it likely includes many chord pairs that exist due to the alignment not being exact between the ground truth and the estimations.

The chord combinations resulted in a very big number, so additional steps were taken to refine the results. First, the algorithm was tuned to convert each of the labels into a sequence of notes represented by its MIDI number and ordered from lowest to highest. The conversion was done following the same notation as proposed in [Harte et al., 2005] which is the standard for MIREX outputs. This step ensured that consonant chords with different names were treated as the same chord. To reduce these misalignment errors, only intersections larger than 500 ms were counted. This improved query resulted in 7488 different chord pairs.

As shown in in Figure 5, the misestimations that were counted were instersections bigger than 500 ms, which were then analyzed to see if they were correct or not. In this case only the Gmaj - G7 case would be counted, since it is the only intersection bigger than 500 ms were the estimation does not match the ground truth.

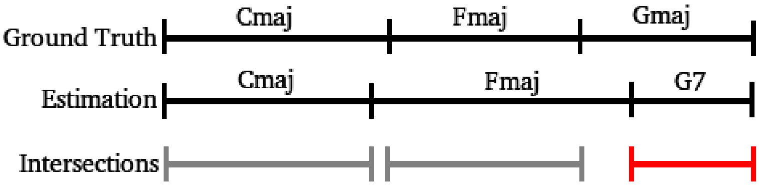


Figure 5: Only misestimations of intersections bigger than 500 ms were counted.

The percentage of these ground truth-estimation chord pairs that were correct was of 44.51%, compared to a 55.49% error rate. This metric is not as relevant as the Weighted Chord Symbol Recall used in the MIREX evaluation, however it does provide an additional perspective on the performance of the algorithms as an average. It can be seen that there are more misestimations than correct estimations for each one made with an overlap of at least 500 ms against the ground truth.

Out of these chord pairs obtained, the 100 most common errors were taken into account to group them into chord types and be able to get a broader picture of these errors. These results can be seen in Table 5. It can be noted that the most common error is estimating a dominant 7th (major minor 7th) chord as a major chord, and that a case for this happened for all 12 notes (which further emphathizes its commonality). In MIREX, when considering the major/minor vocabulary this would be taken as a valid chord estimation since the dominant 7th chord includes all the notes in the major triad.

Table 5: Most common errors by chord types

Ground Truth	**Estimation**	**Diff. Roots**	**Percentage**
7th	maj	12	21.76%
min7	min	12	14.35%
No Chord	Other Chord	6	5.95%
maj	maj7	6	5.63%
maj	5th displacement	5	5.62%
maj	4th displacement	5	4.36%
maj(9)	maj	5	4.35%
maj	7th	5	4.30%
maj7	maj	5	3.72%
min	maj	4	3.12%
Others	...	...	26.85%

From the results shown in Table 5 the most common chord types are described in Table 6. It can be seen that most of the common misestimations fall into a category where there is only a difference of having a missing note, or having one extra (such as 7th to maj). These chord estimations can be considered either correct or incorrect depending on the chord vocabulary used in MIREX. For simpler chord vocabularies the estimations would be considered correct.

Table 6: Most common chord types

Shorthand	**Chord Type**	**Example in C**
maj	Major	C, E, G
min	Minor	C, Eb, G
7th	Dominant 7th (Major minor 7th)	C, E, G, Bb
maj7	Major 7th	C, E, G, B
min7	Minor 7th	C, Eb, G, Bb
maj(9)	Major with added 9th	C, E, G, D

The previous analysis can give an overall picture of the nature of the algorithms, however not all of them behave in the same fashion. The differences in methodologies used in each of them will result in a different behavior that should also be considered.

3.1.3 Comparison Between the Different Algorithms

To gain a better insight on the differences between the algorithms submitted to MIREX 2013, a similar analysis to the one described in Section 3.1.2 was done on the outputs of each of the algorithms. The purpose of this analysis is to understand some of the contrasts of the algorithms against the general tendency of the group, and gather general statistics of the outputs, as well as describing the most common misestimations for each of them.

The first analysis, as shown in Table 7, displays a general overview of the nature of each of the algorithms. It helps to portrait if the algorithm tends to generate more pairs of misestimations, as well as how many misestimations it found. The misestimation pairs analyzed are of chord pairs with an overlap of more than 500 ms. Also, by knowing the average count, an idea of how many times these chord pairs were repeated for the entire dataset can be known.

Table 7: Algorithm Misestimation Comparison. Unique Cases are each of the misestimation pairs (ground truth to algorithm output); Total portrays the overall number of misestimations; Avg. Count shows how many times the misestimation pairs were repeated in average.

Algorithm	**Unique Cases**	**Total**	**Avg. Count**
CB3	1071	28056	26.20
CB4	1008	13868	13.76
CF2	1741	17860	10.26
KO1	916	15587	17.02
KO2	998	14466	14.49
NG1	694	18909	27.25
NG2	1066	21384	20.06
NMSD1	1144	15043	13.15
NMSD2	1081	15056	13.93
PP3	722	19417	26.89
PP4	1049	18294	17.44
SB8	559	38218	68.37

As a second analysis, the 10 most common chord misestimations are shown in Table 8 for each of the algorithms. This table permits seeing their behavior differences, by showing the most common misestimations by chord type. Similar misestimations may suggest that the algorithms have the same underlying principles in practice, therefore leading to the same patterns.

There are two main factors to be considered in these misestimations. The first one is a difference in chord type (such as Cmaj to Cmin) which were counted regardless of the root note. The second factor is a displacement of the root note by a certain interval, which is indicated by an asterisk (*) in the table. When a chord pair presents both, the chord type error is preceded by a " ->" sign. The top row pairs are the most common ones, while the bottom row pairs are the 10th most common to appear for the entire dataset.

Table 8: Most Common Misestimations for algorithms in MIREX 2013. The misestimations per algorithm are ordered from most common (top row) to least common (bottom row)

CB3	**CB4**	**CF2**	**KO1**	**KO2**	**NG1**
7-maj	7-maj	maj-maj7	7-maj	7-maj	7-maj
4th*	min7-min	maj-maj6	min7-min	min7-min	min7-min
min7-min	5th*	7-maj	4th*	5th*	4th*
5th*	4th*	min-min7	5th*	4th*	5th*
maj-7	maj-7	maj-7	maj7-maj	5th->maj/5-maj	3d-b*->min7-maj
5th->maj/5-maj	5th*->maj/5-maj	5th*	5th->maj/5-maj	maj-min	maj7-maj
maj-maj7	maj(9)-maj	4th*	min-maj	maj7-maj	min-maj
maj(9)-maj	2nd-b*	min7-min	3d-b->min7-maj	6th->maj-min	5th*->maj/5-maj
2nd-b*	min-min7	7-maj6	5-maj	min-min7	1/1-maj
min-min7	maj-min	min-maj	1/1-maj	maj(9)-maj	7th*

NG2	**NMSD1**	**NMSD2**	**PP3**	**PP4**	**SB8**
maj-maj7	7-maj	7-maj	7-maj	7-maj	5th*
6th*->maj-min7	Maj-maj7	4th*	min7-min	4th*	2nd*
7-maj	5th*	5th*	4th*	5th*	6th*
min-min7	4th*	maj-7	5th*	6th*->maj-min7	3d*
maj-7	maj-7	maj-maj7	maj7-maj	min7-min	7maj*
4th*	min7-min	min7-min	4th*->7-maj	maj7-maj	5th-a*
5th*	5th*->maj/5-maj	5th*->maj/5-maj	3d-b*->min7-maj	min-min7	4th*->maj/3-maj
6th*->7-min7	maj(9)-maj	min-min7	5th*->maj/5-maj	5th*->maj/5-maj	4th*
3d*->maj-min7	Min-min7	maj(9)-maj	1/1-maj	4th*->7-maj	7th*
2nd*->maj-min7	min-maj	4th*->maj-maj7	6th*->maj-min	1/1-maj	3d-b*

There are some observations that can be made from this data. From all the algorithms, CF2 and NG2 are the algorithms where the 4th and 5th displacement errors are not as common as in the rest. From the algorithm description found in Table 2 it can be seen that both are based on a chroma vector, where as the rest of the algorithms are based on a modified or custom feature vector. This may suggest that these type of chroma vectors are less prone to a 4th or 5th displacement error. It is also worth noting that 4th and 5th displacements can be considered as the same, since the output labels do not distinguish if the chord is ascending or descending. An ascending 5th will result in the same chord as a descending 4th, which on a label based notation as the one used is impossible to know.

Another of the observations that can be made is with respect to the SB8 algorithm. From Table 7 it can be seen that it is the one with the highest average count, as well as the one with the most misestimations and the least unique cases. Additionally, when looking at Table 8 it can be noted that all of the top 10 misestimations have the root displaced. It is the only algorithm which is based on Neural Networks, therefore hinting that the architecture might not be processing the input parameters correctly. Not having chord type errors can be considered somewhat strange, since different types tend to share more notes compared to a root displacement.

Some of these observations are not evident when looking at the data as a whole. Knowing how the data is distributed is essential in order to devise a good test.

3.2 Test Development

Much information can be gathered from analyzing the data taken from the outputs of the algorithms and comparing them against the ground truth. Some of the information that can be inferred was analyzed in Section 3.1 by bulking all the outputs from all the algorithms. This information, however, is analyzed systematically and when dealing with music it is reasonable to expect that not all users listen to music the same way, thus becoming subjective.

Based on the fact that most of the common errors would be considered as correct estimations in MIREX using the simpler chord vocabularies, it is reasonable to think that these errors are not so distant from the ground truth. However, from a user's point of view, these errors might not be musically aesthetic or they may sound somewhat off when compared against the original chords. On the other hand, since people are mostly used to hearing simple major and minor chords, these types of errors may sound more pleasant than the ground truth.

Based on these last statements, it is important to know how people perceive the errors from the algorithms' outputs. For this reason a test was devised to validate how users rate these common errors and how well they sound when compared against the original audio. To see how the chord estimation perceptions bear against the ground truth, it is necessary to include both of them and compare them against the original audio. A simple schema is shown in Figure 6. A reasonable hypothesis would be to expect that the ground truth files will always sound better than the chord estimations.

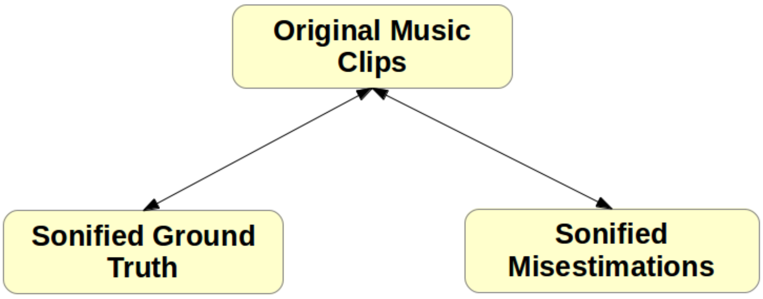


Figure 6: The test proposed compares the original audio to the sonified ground truth and estimations.

The test proposed consists in having 3 different audio excerpts. One corresponding to the original song, and the other two being the sonification of both the ground truth and the output of the estimation algorithm. These 3 excerpts correspond to the same segment of the song in order to have the same point of comparison. The segment is chosen according to be centered around the most common errors (misestimations), leaving 3 seconds before the chord and 3 after to give some context to the chord. The user interface has a timer indicating when the chord selected will appear in such a way that the user can know in which chord to focus and when. After listening to the excerpts the user needs to select which of them represents better the original song.

Since it is very likely that the performance in this task is subject to a number of variables from the user, some user data is gathered in order to arrive to better conclusions afterwards. This data includes knowing whether the subject has formal musical training, as well as if he/she has played an instrument. It is reasonable to think that trained musicians will discern better the differences in chords, and will probably tend to favor the ground truth excerpts.

It is expected that after a number of users complete the test there will be a tendency of leaning towards the ground truth in most cases. Depending on the chord type there might be stronger or milder differences in this tendency. Some of the chord types that have a stronger characteristic sound (such as a dominant 7th) will probably veer more against the ground truth, and this will suggest that the estimations in these cases are less permissive of errors from a perceptual point of view. On the other hand, if the tendency for that chord type is that there is no difference, or that the estimation sounds better than the ground truth, it will suggest that these type of errors are not so critical and are perceptually permissive.

The test was chosen to be developed by having separate audio excerpts. Another option would have been having the ground truth and estimation chords sonified on top of the original audio. One of the main reasons for choosing separate audio excerpts is that in this way, the user's memory is involved. When having the sonification playing over the song, it is easier to hear the differences especially when the chords are more dissonant. On the other hand, when the sonification is in separate files, the user will have to make memory of the excerpt. This in turn will be an indicator of his overall impression of the chord instead of his ability to notice the actual differences.

3.2.1 Sonification of Chord Estimation Outputs

One missing link between the information available and the test is having the output of the algorithms sonified. Since the outputs of the algorithms are written as time stamped labels, it is necessary to make some rules to sonify the labels and put them together to form a representation of the song. The most straightforward way to do this is converting the time stamped labels into MIDI chords, which can be sonified rather easily with one of the many tools available.

In order to make this first step of label to MIDI conversion, a python script was developed with the use of the MIDIUtil library [Conway Wirt,]. Firstly two dictionaries were made, one consisting of all the possible root notes mapped to its MIDI note equivalent. These mappings span across one octave, starting with A as MIDI note 57 (below middle C), and finishing in G#/Ab as MIDI note 68. The second dictionary consisted of all the possible chord types for the short hand notation mentioned in [Harte et al., 2005] mapped to each of the note intervals in semitones that compose each chord. In this way, the dictionary for a major chord is [0, 4, 7] with the intervals corresponding to the root note, major third, and fifth respectively. It is worth taking into consideration that the voice leadings are not sonified correctly, because this information is lost when transcribing the chords into labels.

Every output file was parsed to obtain the start and end time stamps

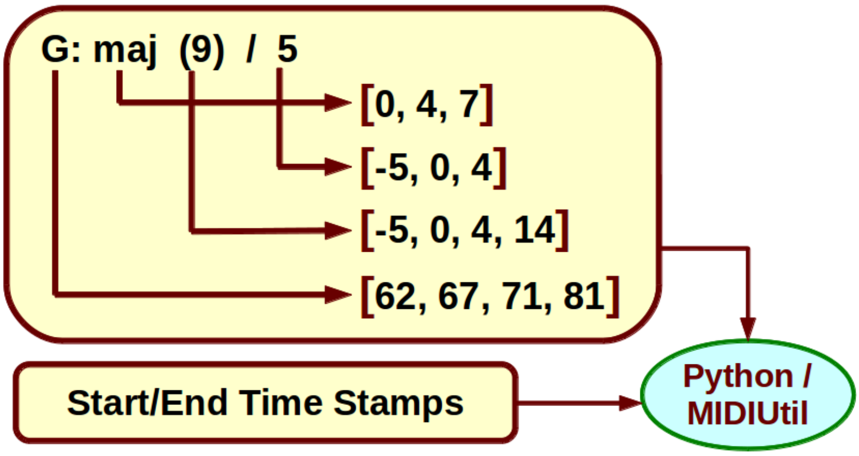


Figure 7: Label to MIDI chord. First the chord type template is retrieved. Then the bass note is modified or added. Consequently any additional notes are added to the chord. Finally the chord is offset by the root.

for each chord, as well as the label estimated. Each label was then parsed to separate it in up to four parts: the root note, short hand notation, bass note, and additional notes. To build up the MIDI chord, first the template for the short hand notation was retrieved. This template was then modified to include, where available, the bass note and the additional notes. The bass note was lowered one octave when it was already part of the chord template, or added it below the root note when it was not present. The additional notes were added above the root note depending on the intervals they corresponded to. This then leaves us with the template mapped, where the root note is mapped to 0 and the note intervals part from it. The MIDI number coinciding with the root note was then added to all the notes in the template, to end up with the complete chord. This process is shown in Figure 7

One of the main problems with this method (that is also inherent to representing a chord with a label) is that you cannot know much of the context of the chord. This means that it is unknown whether the chord is going up or down tonally. With the previously established parameters, an A chord will always sound lower than a G chord even when the progression is going up. This can pose a problem when doing the test because the users might perceive the progression differently. This does not have a big impact when comparing the ground truth and the estimations due to the fact that both are sonified under the same conditions. Because of this effect, a 4th and a 5th displacement should fall into the same category because a 4th sounding lower than the tonic will be a 5th perceptually and vice versa.

After this step the MIDI notes for the chord are defined. Taking its beginning and end time stamps it is then added to a MIDI track. The MIDIUtil library needs the time and duration for each note to be specified in bars. To simplify this a tempo of 120 bpm was chosen for all songs, and therefore the time stamps were multiplied by 2 to make the time in seconds match the bars to a tempo of 120 bpm. This process was repeated for each of the chords in the file appending each chord to the MIDI track. The end result is a translation of the output file to a complete MIDI track with all the chords in it at the specified time stamps.

A MIDI file was made for each estimation of all the algorithms, including the ground truth. The MIDI files were then sonified to synthesized piano chords by using the *Free MIDI to MP3 Converter* developed by *PolySoft Solutions*. This software allowed the processing of all the files to be done sequentially. The files were sonified entirely (as opposed to just sonifying the chosen excerpts) for two main reasons. When you skip to a section of a MIDI file where no chord is indicated it will sound as empty, because the chord that triggered the sound is not read. By sonifying before, it is possible to cut any segment of the audio and have the sound present even when the chord was triggered previously. Additionally, by sonifying all the files at once, the sample selection can be done quicker by just cutting the audio segments that are needed.

3.2.2 Excerpt selection

The excerpts were chosen in a pseudo-random manner taking into account the list of the most common errors. For each of the top 100 misestimated ground truth to estimation chord pairs, 10 occurrences were chosen randomly making an exception in cases where the chord would happen during the first or last 3 seconds of the song. The occurrences where looked up in all the algorithms compared against the ground truth. For each of these cases an excerpt was made by mapping to which algorithm and what song the chord misestimation belonged to. Reading the time stamps, the sonified audio corresponding to that estimation was cut leaving 3 seconds before and after the chord to give some musical context to the user. The algorithm also cut an audio excerpt for the ground truth corresponding to that algorithm to generate the matching pair to be used in the test.

After generating all of the ground truth and estimation excerpt pairs, a selection process was made. In order to have better results that focus only in the misestimated chord, the excerpts were listened at to determine which had fewer inconsistencies. An ideal pair of chords would be one that differs only in the misestimated chord, having the rest of the excerpt almost identical. This part is important because if it's not taken into account, other differences might bias the user towards not focusing on the actual chord. The selection was made focusing on the chord types that had the most misestimations, as shown in Table 5. The quantity of excerpts chosen does not reflect the ratio of appearances exactly, although excerpts with a bigger ratio were more common to appear. It is better to have a larger sample set for the different smaller cases as well, otherwise it would be difficult to reach convincing conclusions regarding them.

Once the best samples were selected, the excerpt from the original song was extracted with the ground truth time stamps. These stamps did not match entirely due to the fact that different versions of the song might have different timings. The timing for each chord was manually reviewed and noted down in order to have the exact moment at which the chord is sounding. This in turn enables for the test to let the user know when the chord will sound for maximum attention.

Listening to the sonified excerpts and comparing them with the original song, it was noted that in many cases the differences between the ground truth and estimation files seemed to broaden. It is probable for these differences to change the outcome of the test, as the user's attention might be deviated from the actual chord to be focused in. To further decrease the differences, the actual estimation excerpt was changed for a modified ground truth excerpt. This modification consisted in having the exact file as in the ground truth, but just changing the chord of interest to the misestimation from the algorithm. This ensured to get the lowest variability from file to file and thus excluding other factors in the audio that might veer the user's attention and perception. The general clip generation and selection process starting from the MIDI files is shown in Figure 8.

After all the selection filtering steps, the end result were 35 different excerpts with each having the sonification of the ground truth, the estimation, and the original song. This amounted to a total of 105 different audio files. These audio files will in turn be used in the test, which has to be delivered through an adequate platform.

3.2.3 Test Framework/Platform

It is important to deliver the test in an environment that suits its requirements and adapts to its needs and limitations. In this case it is not necessary to have an environment absolutely controlled by the test giver. Also, it does not require a prolonged attention span from the user because it will be di-

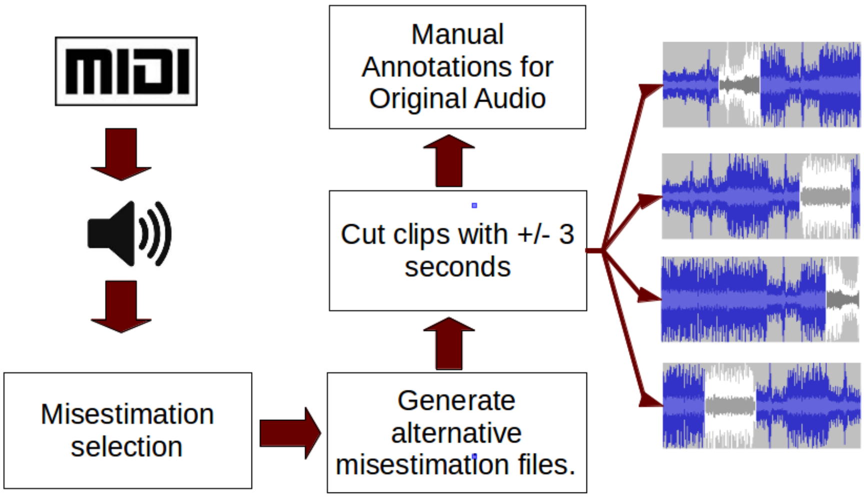

MIDI
Manual Annotations for Original Audio
Cut clips with +/- 3 seconds
Misestimation selection
Generate alternative misestimation files.

Figure 8: Clip Generation and Selection

vided into different questions corresponding to each of the excerpts to be rated. Not having these limitations gives some flexibility regarding the platform in which the test can be delivered. It is also desirable for the test to be reproduced easily in order to reach a greater amount of test subjects. With these premises in mind, an Internet based test appears to be the best option, since it is easily reproducible and permits a straightforward data collection.

To implement the test it is preferable to have as much control on its details as possible. A similar test was developed in [Hespanhol, 2013] with its code being kindly provided by the author. It is possible to adapt the test to meet the particular requirements by basing it in this template . The test works by having a front end in which the user can play audio and read information for a number of different questions and answer a form in which the answer will be recorded. This data is then passed to a PHP script that records the answers in the server.

After adapting the test for this particular situation, it was structured as shown in Figure 9 with the most importance given to two main phases. The first phase consisted in displaying the instructions and all the information that the user may need for the test. It also included a form for the user to fill his data in order to be able to further analyze the results and categorize them according to the users' profiles. The most relevant piece of information into which categorize the users is whether if they have musical training or play an instrument or not. The first case will mean they are more familiar with analyzing music and will therefore be more likely to differentiate the difference in the estimation of the algorithm against the ground truth.

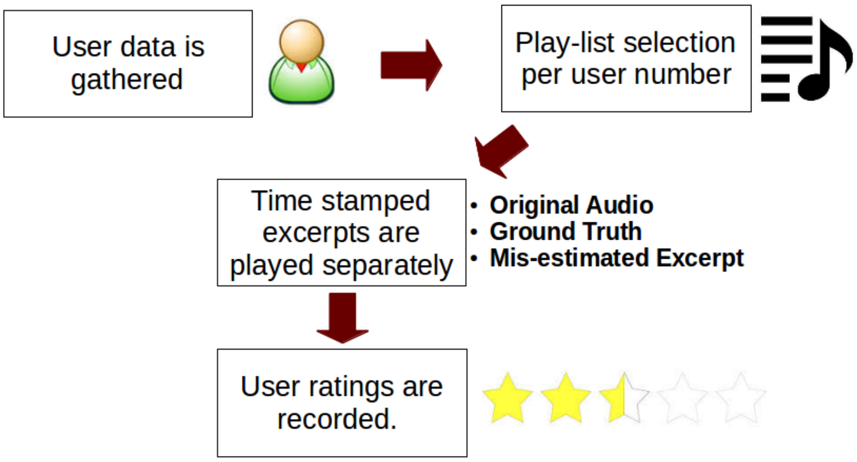
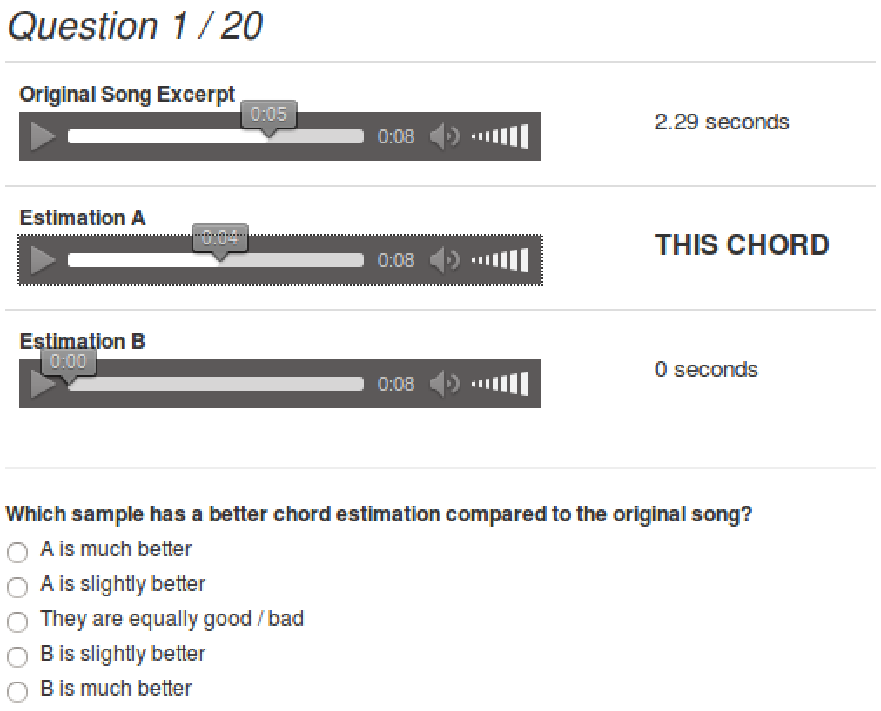

User data is gathered
Play-list selection per user number
Time stamped excerpts are played separately
• Original Audio
• Ground Truth
• Mis-estimated Excerpt
User ratings are recorded.

Figure 9: Test Description

The second phase consists in the actual test, where users will be hearing the samples through a HTML5 audio tag media player and evaluating them through a form for each question. A media player is included for each of the 3 files loaded. Each of the media players has a timer where the user can see when the chord to be compared will appear. The form includes a simple rating system where the user can say which of the samples has a better chord estimation when compared to the original song. The user is able to choose between 5 answers: whether sample A is much better, sample A is slightly better, bot samples are equally good/bad, sample B is much better, or sample B is slightly better. A screenshot of this interface is shown in Figure 10. These ratings were then recorded on a numeric scale from -2 to 2, where 0 is an equal perception of both algorithms.

As the test will not be in a controlled environment, it is not expected from the subjects to be concentrated for a long time. Also, by keeping it short it can be expected that more people will be willing to take it. Because of this, a test that lasts around 20 minutes appears to be ideal. Taking into account that most of the segments last around 7 - 10 seconds a user will need to listen to around 21 - 30 seconds per question and still have some time to re-listen to some of the samples when necessary as well as answer the question. With this parameters it is reasonable to think that the user can answer 20 questions in 20 minutes.

As mentioned in Section 3.2.2, there are 35 samples available to test the users with. Since all the available samples are not equally distributed



Question 1 / 20

Original Song Excerpt

2.29 seconds

Estimation A

THIS CHORD

Estimation B

0 seconds

Which sample has a better chord estimation compared to the original song?

○ A is much better
○ A is slightly better
○ They are equally good / bad
○ B is slightly better
○ B is much better

Figure 10: Online Test Front End

according to its family, there are more samples for some types and less for other. In order to get the families equally rated, three playlists of 20 clips were made. When there were less samples available per family, the playlists repeated the same samples. When a family had more samples to choose from, the playlists had different samples with a distribution by chord type as shown in Table 9. The three playlists were cycled by test taker number, meaning that the second playlist would be chosen after the first one, the third after the second, and the first after the third. For each playlist, the play order was randomized for each user.

The answers were recorded question by question by appending them to the user file. By the time the user finishes the questionnaire a file is stored with all the answers recorded as text. This text file contains a header with the user's information, followed by the answers and variables for each question. These consist on the number of question, the audio excerpt played, which excerpt contains the ground truth, and the rating that the user gave to the samples. The recorded answers were saved for each of the users to later analyze the complete results.

Table 9: Clip Distribution by Chord Type

Ground Truth	**Estimation**	**Clips**	**Per test**
Dominant 7th	Major	9	3
Minor 7th	Minor	5	3
Major	Major 7th	3	2
Major	4th / 5th Shift	2	2
Major Add 9	Major	4	2
Major	Dominant 7th	3	2
Major 7th	Major	3	2
Minor 7th	Major	3	2
Minor	Minor 7th	3	2
TOTAL		**35**	**20**

4 Results

After conducting the experiment the results were gathered from 32 participants out of a total of 43 tests taken. The rest were discarded mainly because they were incomplete. Out of these results some patterns in the users' answers are expected, pinpointing to tendencies in certain chord families which will lean towards either the ground truth or the chord misestimation as sounding better. The analysis will focus on central tendency statistical measures that will be taken in order to see towards where the users' perceptions are leaning.

Out of the 32 valid participants, the mean test taker would be 26.2 years old, male, with 6.75 years of musical training and 10.1 years playing an instrument. The sex distribution was 18.75% female to 82.25% male with ages ranged from 20 to 41 years old. The musical training extended from 0 to 16 years, while years playing an instrument varied from 0 to 20. From the users answering the test, it could be seen that the difference between the users with musical training and users that played an instrument was really small. The answers were correlated by 97.2%. This is due to the fact that most of the users that answered the test that played an instrument had musical training as well.

After all the results were gathered they were normalized on the same scale from -2 to 2, with the difference that a positive number means a preference for the ground truth, while a negative number to the misestimation. This was done to correct for the random factor used on the test, where the ground truth was chosen arbitrarily to be excerpt A or B by question.

As a first approach to analyze the data, a null hypothesis stating that the data comes from a normal distribution was tested. This was done for each of the chord families using SciPy's [Jones et al., 2001] normality test which is based on D'Agostino's and Pearson's [D'Agostino and Pearson, 1973] test for normality. As shown in Table 10, it can be seen that the p-values are very low, therefore rejecting the null hypothesis that the data comes from a normal distribution. Knowing that the data is not normally distributed will help analyze the data correctly.

Table 10: P-Values for Normality Test on Data

Chord Family	**P-Value**
4th/5th disp.	1.9150×10^{-5}
7 - Maj	7.0201×10^{-9}
Maj - 7	7.6578×10^{-6}
Maj - Maj7	2.5505×10^{-3}
Maj7 - Maj	0.0208
MajAdd9 - Maj	4.6434×10^{-3}
Min - Maj	0.0193
Min - Min7	0.0522
Min7 - Min	2.6682×10^{-5}

The results need to portray the main perception of the users to the test. This can be represented by using central tendency measures such as the mean, mode, and median values. Of these measures it is important to consider the possible limitations. The mean may not really represent well the central tendency because the test was based on a Likert scale. The scale is subjective because different users may have different score perception thresholds and one may therefore assign a "much better" rating more often that a "better" rating. The mode might be a more solid central tendency measure, showing what the most common answer was. The three of the measures compliment each other and permit a better analysis. Table 11 shows the central tendency for each of the chord families tested for all users.

To determine if the results are statistically significant, compared to a random flat response, a χ^2 test was made. This test was chosen because it is nonparametric, and as previously mentioned, the data obtained is not normally distributed. The p-value shown in the table tests the null hypothesis that the data has a flat distribution. Rejecting the null hypothesis suggests that there is a tendency of the data.

Additionally to the table, it is useful to represent the data graphically, as the tendencies can become more evident. Figure 11 and shows a box and whisker plot for all users. The division of all the data in to quartiles, as well

Table 11: Central Tendency of User Evaluation for all Users

Chords	**Count**	**Mean**	**Mode**	**Median**	**P-Value for** χ^2
4th/5th	64	0.4844	2	1	0.0191
7 - Maj	96	-0.3333	-2	-1	0.0493
Maj - 7	64	-0.0156	-1,0,1	0	0.8852
Maj - Maj7	64	-0.0625	-1	0	0.9098
Maj7 - Maj	64	0.4531	1	1	0.0218
MajAdd9 - Maj	64	-0.1719	-1	0	0.2656
Min - Maj	64	0.2500	1	0	0.1310
Min - Min7	64	-0.0313	0	0	0.8322
Min7 - Min	96	0.1563	1	0	0.1715

as showing the median (red line), mean (blue 'x'), and mode (blue dot) helps portrait to where the data is leaning. Also, the whiskers show the range of the data, which indicate if there are some answers not chosen by the users.

Table 12: Central tendency of user evaluation for users with more than 2 years playing an instrument

Chords	**Count**	**Mean**	**Mode**	**Median**	**P-Value for** χ^2
4th/5th disp.	48	0.4167	2	1	0.0762
7 - Maj	72	-0.2361	-2	0	0.2278
Maj - 7	48	-0.0625	-2	0	0.9693
Maj - Maj7	48	-0.1458	-2	0	0.8837
Maj7 - Maj	48	0.5208	1	1	0.0092
MajAdd9 - Maj	48	-0.1875	-1	0	0.4232
Min - Maj	48	0.1458	-1	0	0.5169
Min - Min7	48	0.1458	1	0	0.5509
Min7 - Min	72	0.2500	1	0.5	0.1273

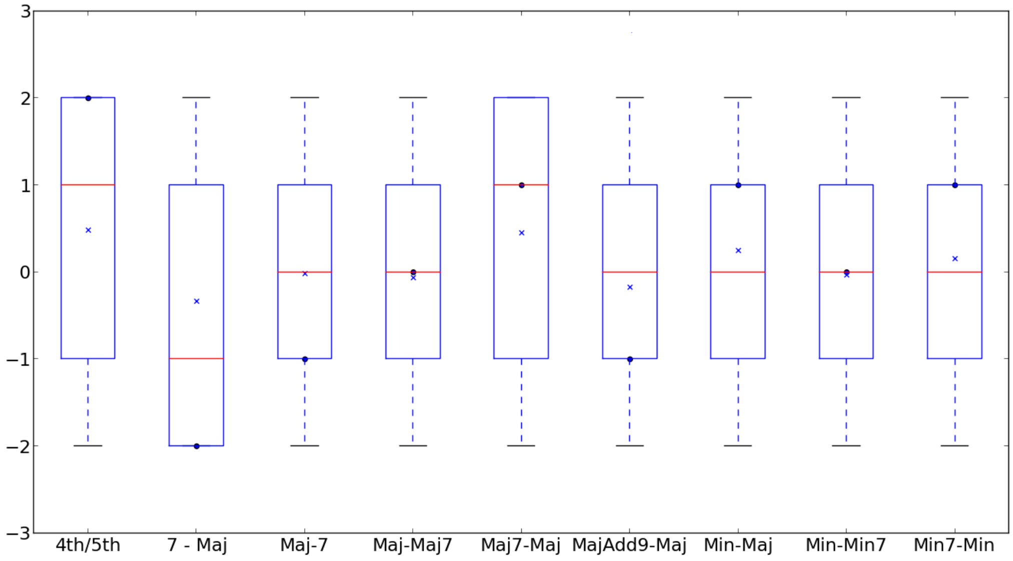


Figure 11: Box and Whisker Plot for all users. The 'x' represents the mean value and the dot shows the mode.

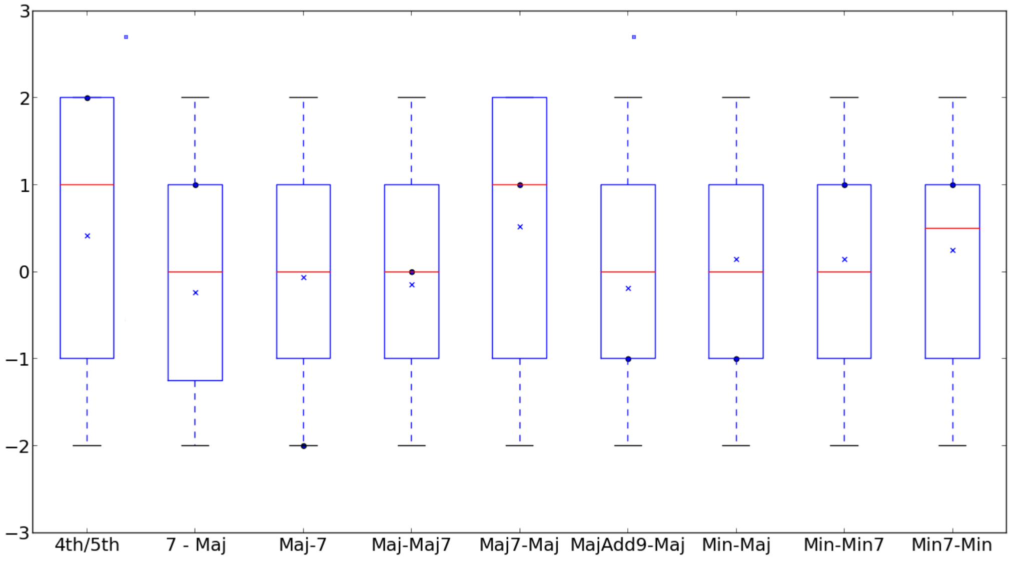


Figure 12: Box and Whisker Plot for users with more than 2 years playing an instrument. The 'x' represents the mean value and the dot shows the mode.

Table 13: Central tendency of user evaluation for users with less than 2 years playing an instrument

Chord	**Count**	**Mean**	**Mode**	**Median**	**P-Value** χ^2
4th/5th disp.	16	0.6875	2	1	0.2087
7 - Maj	24	-0.625	-2	-1	0.2325
Maj - 7	16	0.1250	-1	0	0.2626
Maj - Maj7	16	0.1875	-1	0	0.6005
Maj7 - Maj	16	0.2500	2	0	0.7128
MajAdd9 - Maj	16	-0.1250	-1	0	0.7128
Min - Maj	16	0.5625	1	1	0.0364
Min - Min7	16	-0.5625	-1	-1	0.3280
Min7 - Min	24	-0.1250	-1	-0.5	0.7664

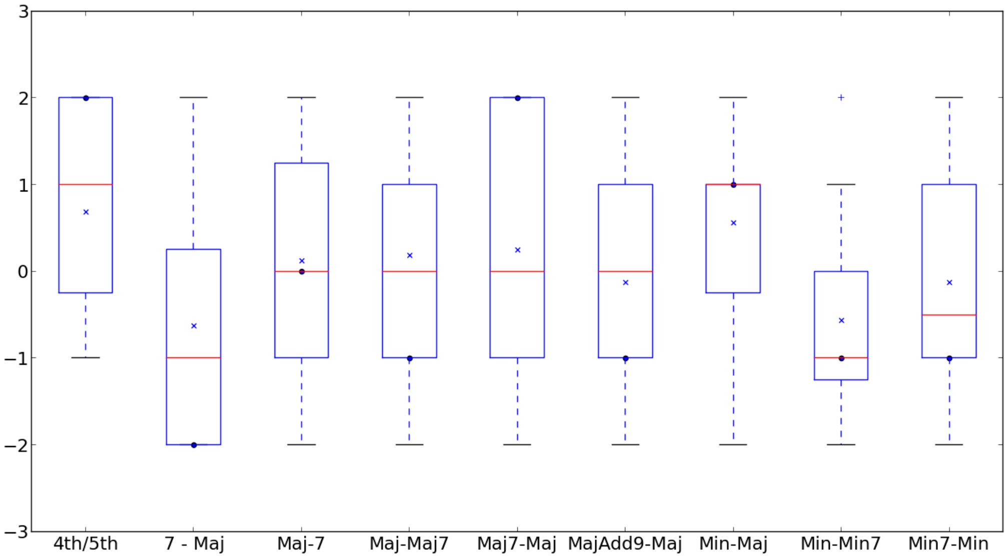


Figure 13: Box and Whisker Plot for users with less than 2 years playing an instrument. The 'x' represents the mean value and the dot shows the mode.

Just as Table 11 and Figure 11 show the central tendencies for all the users, Tables 12 and 13, and Figures 12 and 13 show the same data analysis done divided into two groups. The first group represents the part of the users who have played an instrument for more than two years. The second group represents the part of the users who have played an instrument two years or less. Two years of experience was chosen because it is reasonable to expect that these users will be familiarized with the sound of different chords. It is worth mentioning that the data is not equally divided because the majority of the people that took the test have played an instrument for more than two years.

After gathering the results and organizing the data into tables and graphs, there are some hypothesis that can be made from by analyzing them.

5 Conclusions

5.1 Results Analysis

From looking at Figure 11 it can be seen that there are 3 categories where data tends to group towards the edges. The first category is the 4th/5th, where the misestimation is displaced by a 4th or a 5th from the original root note. Both displacements are treated as the same because on a label based estimation it is not possible to know whether the chord sequence is ascendant or descendant. The data is clustered towards favoring the ground truth, and the mode indicates that the most common answer was to greatly favor the ground truth as a better estimation. When looking at both groups (with experience playing an instrument and without), it can be seen that they both follow the same tendency. For both groups the p-values are above .05, however when merged, the this value lowers to .019. This can be due to the number of samples, which makes the measure more reliable when there are more samples. These p-values however, do not necessarily mean that it has a high statistical significance.

When comparing a dominant 7th chord (ground truth) to a major chord, the users seem to prefer the major chord. At a first glance it may seem counter intuitive, since dominant 7th chords have a very characteristic sound and might therefore be easier to discriminate as sounding different. However, depending on the function that the chord serves in the sequence, the 7th chord may only be used as a *colorful* chord. In these cases, a reasonable hypothesis is to expect that the users developed a bias towards the major chord, which has a simpler sound and is used more commonly. When looking at the groups separately, it is interesting to see that users without instrument experience rated the 7th chord as sounding worse. This can be presumably because people who have played an instrument are more familiar with these types of chords. Also, the inverse case where the 7th chord is the misestimation does not seem to be correlated inversely. Although this inverse case does not have a high p-value, it also suggests that rating is dependent on the function of the chord.

The last prominent case is when a major 7th chord is misestimated to a major chord. The data seems to propose that the users favored the ground truth. Despite the fact that the mean is somewhat high when compared to the other families, the mode shows that the inclination is to only *slightly* favor it. It is interesting to see that the users with instrument experience clearly follow the same tendency to slightly favor the major 7th chord. Similarly to the previously mentioned dominant 7th chord, the opposite case where the dominant 7th is the misestimation does not seem to drift towards either side. It even seems to suggest, although with a high p-value, that the users are indifferent towards either chord.

Another interesting finding that is present in the people without instrument experience is that they seem to notice more the difference on minor to major chords. Although the sample number is low, the p-value is also low. This behavior is not seen on people with instrument experience, which seems odd considering that these are the most common chords in music and tend to be easily discernible for most musicians.

After examining the results for the different chord families investigated, it can be seen that not all cases are equal. There are some cases where the users do seem to have a bias towards either the ground truth or the estimation. These preferences are not inversely equal, which means that when the chord that was misestimated is the ground truth for another, and vice versa, the preferences are not preserved.

Two of the cases which showed an inclination from the users (7-Maj and Maj7-Maj) fall into a category where the ground truth is a 4 note chord, and the estimation is a 3 note chord. For these cases, when the algorithms are being evaluated in MIREX in simpler chord vocabularies (such as Major and Minor), they would be mapped to the same category. These mapping would treat the chord as being correctly estimated because they belong to the same chord vocabulary. This might not be the best approach and can be misleading in some cases when choosing a more accurate algorithm based on MIREX results. If the algorithm is to be used in a task in which user perception is relevant, then these type of chord mappings become significant as they will affect the score. This is an important fact to contemplate, especially considering that the major-minor vocabulary is the most relevant (as it is the category on which the score reports are sorted).

5.2 Contributions

The main contributions of this work can be summarized into the following points:

- Analysis of MIREX 2013 most common mis-estimations for Billboard 2012 Dataset.
- Automated tool for analyzing MIREX results.
- Sonification method for time stamped chord labels.
- User study of perceptual relevance.
- Analysis of results.

5.3 Future Work

As mentioned throughout the work, there are some aspects that could be improved on the previous work. Most of these involve doing a test of a similar nature with some improvements. One of the first improvements could be to extend the clips dataset to obtain a better distribution of chord types and specific harmonic functions. This way, the sample clips could be more reliable as having more variety would help generalize the cases better for the different chord families. Having less samples means that the answers are somewhat encased to those specific circumstances.

By testing other types of chords with more complex musical aspects such as a chord's role in tonality, it could be possible to obtain more specific results, as the chord's role can have a great impact on how the listener might rate a chord estimation. This can be seen with dominant 7th chords, as some times, such as in blues, they are played throughout the song adding a *flavor* without having a specific role in tonality.

Another improvement that could be made to a similar test is having a more homogeneous test population. For this study, most of the users that did the test had some musical training and instrument playing experience. Because of this, the comparison between the two groups could not be done with a great degree of trust, as one of the groups was significantly smaller than the other one.

Something that could help enhance the test is to consider the voice leadings of the chords. This could prove to be a time-consuming task, since the chord labels available from the ground-truth annotations do not inherently contain this information. Additional analysis on the desired excerpts would have to be made to build an appropriate dataset. The voice leadings should then be taken into account when sonifying the samples. With this change different results could be obtained for some of the chords. This would also permit the 4th and 5th displacements to be considered as separate because the sonification would be different.

Based on the results obtained in this work it would be reasonable to do another test with the mentioned improvements. This test should focus on the chord families where a tendency of user preference seems to be present. Two of the families which could the most promising to analyze are 7-Maj and 7maj-Maj. For these families, considering the chord's role in tonality could help clarify the cases in which these chords are noticed as different. An experiment could be set by obtaining samples with the different diatonic functions (tonic, supertonic, mediant, subdominant, dominant, submediant, or leading) for the same chord.

6 Appendix

6.1 Song Excerpts by Chords

Table 14: Songs from where excepts were chosen for each of the chord misestimations.

Chords	**Family**	**Artist & Song Name**
B:7-B:maj	7 - maj	The Dramatics - Hey You! Get Off My Mountain
B:7-B:maj	7 - maj	Johnny Horton - Johnny Reb
D:7-D:maj	7 - maj	Lesley Gore - You Don't Own Me
D:7-D:maj	7 - maj	The Everly Brothers - That's Old Fashioned (That's The Way Love Should Be)
D:7-D:maj	7 - maj	The Turtles - Happy Together
C:7-C:maj	7 - maj	Marty Robbins - I Walk Alone
C:7-C:maj	7 - maj	Johnny Tillotson - Talk Back Trembling Lips
Eb:7-Eb:maj	7 - maj	Chris Kenner - Land Of 1000 Dances
A:7-A:maj	7 - maj	Charlie Rich - Behind Closed Doors
E:min7-E:min	min7 - min	Elvis Presley - My Wish Came True
C#:min7-Db:min	min7 - min	David Crosby,Graham Nash - Carry Me
C:min7-C:min	min7 - min	Aretha Franklin - Oh Me Oh My (I'm A Fool For You Baby)
C:min7-C:min	min7 - min	Thelma Houston - Don't Leave Me This Way
C:min7-C:min	min7 - min	Dion - Love Came To Me
C:maj-C:maj7	Maj - maj7	Boston - Amanda
Bb:maj-Bb:maj7	Maj - maj7	Glen Campbell - Wichita Lineman
Bb:maj-Bb:maj7	Maj - maj7	Elvis Presley - For Ol' Times Sake
F:maj-C:maj	5th disp.	Thelma Houston - Don't Leave Me This Way
A:maj-D:maj	4th disp.	Wilson Pickett - In The Midnight Hour
Db:maj(9)-Db:maj	majAdd9 - maj	Simon & Garfunkel - The Sounds Of Silence
C:maj(9)-C:maj	majAdd9 - maj	Glen Campbell - Wichita Lineman
C:maj(9)-C:maj	majAdd9 - maj	The La's - There She Goes
G:maj(9)-G:maj	majAdd9 - maj	Prince (With Sheena Easton) - The Arms Of Orion
C:maj-C:7	Maj - 7	Cliff Richard - Devil Woman
C:maj-C:7	Maj - 7	The Beatles - A Hard Day's Night
C:maj-C:7	Maj - 7	Brenda Lee - All Alone Am I
D:maj7-D:maj	maj7 - maj	Little River Band - Happy Anniversary
D:maj7-D:maj	maj7 - maj	Minnie Riperton - Lovin' You
F:maj7-F:maj	maj7 - maj	Glen Campbell - Wichita Lineman
B:min-B:maj	min - maj	Duran Duran - Is There Something I Should Know
D:min-D:maj	min - maj	Duran Duran - I Don't Want Your Love
F:min-F:maj	min - maj	Pat Benatar - Heartbreaker
A:min-A:min7	min - min7	Boston - Amanda
A:min-A:min7	min - min7	Brenda Lee - All Alone Am I
E:min-E:min7	min - min7	Cliff Richard - We Don't Talk Anymore

References

[ANSI, 1973] ANSI (1973). *"American National Psychoacoustical Terminology"*. American Standards Association.

[Bas de Haas and Burgoyne, 2012] Bas de Haas, W. and Burgoyne, J. (2012). *"Parsing the Billboard Chord Transcriptions"*. University of Utrecht.

[Benward and Saker, 2009] Benward and Saker (2009). *"Music: In Theory and Practice"*, volume 1. McGraw Hill, 8 edition.

[Brown, 1991] Brown, J. C. (1991). "Calculation of a Constant Q Transform". *Journal of the Acoustic Society of America*, 89(1).

[Burgoyne et al., 2011] Burgoyne, J., Wild, J., and Fujinaga, I. (2011). "An Expert Ground-Truth Set for Audio Chord Recognition and Music Analysis". In *12th International Society for Music Information Retrieval Conference*, page 633.

[Cannam et al., 2013] Cannam, C., Mauch, M., Davies, M., Dixon, S., Landone, C., Noland, K., Levy, M., Zanoni, M., Stowell, D., and Figueira, L. A. (2013). *"MIREX 2013 ENTRY: VAMP Plugins from the Centre for Digital Music"*. MIREX.

[Cho and Bello, 2013] Cho, T. and Bello, J. P. (2013). *"MIREX 2013: Large Vocabulary Chord Recognition System using Multi-band Features and a Multi-stream HMM"*. MIREX.

[Conway Wirt,] Conway Wirt, M. "MIDIUtil Library". `http://www.emergentmusics.org/midiutil`. Retrieved in February 2014.

[D'Agostino and Pearson, 1973] D'Agostino, R. and Pearson, E. S. (1973). "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and b1". *Biometrika*, 60(3):pp. 613–622.

[Fujishima, 1999] Fujishima, T. (1999). "Real Time Chord Recognition of Musical Sound: a System Using Common Lisp Music". In *Proceedings of the International Computer Music Conference (ICMC)*, page 464.

[Glazyrin, 2013] Glazyrin, N. (2013). *"Audio Chord Estimation Using Chroma Reduced Spectrogram and Self-similarity"*. Ural Federal University.

[Gómez, 2006] Gómez, E. (2006). *"Tonal Description of Audio Music Signals"*. PhD thesis, Universitat Pompeu Fabra.

[Harte, 2010] Harte, C. (2010). *"Towards Automatic Extraction of Harmony Information from Music Signals"*. PhD thesis, Queen Mary, University of London.

[Harte et al., 2005] Harte, C., Sandler, M., Abdallah, S., and Gómez, E. (2005). "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations". In *Proceedings of the 6th International Conference on Music Information Retrieval*. ISMIR.

[Hespanhol, 2013] Hespanhol, N. (2013). *"Using Autotagging for Classification of Vocals in Musical Signals"*. Faculdade de Engenharia, Universidade do Porto.

[Jones et al., 2001] Jones, E., Oliphant, E., and P., P. (2001). Scipy: Open source scientific tools for python. `http://www.scipy.org`.

[Khadkevich and Omologo, 2011] Khadkevich, M. and Omologo, M. (2011). "Time-frequency Reassigned Features for Automatic Chord Recognition". In *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, page 181.

[Klapuri, 2006] Klapuri, A. (2006). Introduction to music transcription. In *Signal Processing Methods for Music Transcription*, pages 3–20. Springer US.

[Mauch, 2010] Mauch, M. (2010). *"Automatic Chord Transcription from Audio Using Computational Models of Musical Context"*. PhD thesis, Queen Mary, University of London.

[Mauch et al., 2009] Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S. Tidhar, D., and Sandler, M. (2009). "OMRAS2 Metadata Project 2009". In *Late breaking session at the 10th International Conference on Music Information Retrieval*. ISMIR.

[Ni et al., 2012] Ni, Y., Mcvicar, M., Santos-Rodriguez, R., and De Bie, T. (2012). "An End-to-end Machine Learning System for Harmonic Analysis of Music". *IEEE Trans. Audio, Speech, Lang. Process*, 20(6):201–213.

[Pauwels and Peeters, 2013a] Pauwels, J. and Peeters, G. (2013a). "Evaluating Automatically Generated Chord Sequences". In *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, page 749.

[Pauwels and Peeters, 2013b] Pauwels, J. and Peeters, G. (2013b). "Segmenting Music Through the Joint Estimation of Keys, Chords and Structural Boundaries". In *MM '13 Proceedings of the 21st ACM international conference on Multimedia*, page 741.

[Pollack, 2014] Pollack, A. W. (2014). "Notes on ... series". `http://www.icce.rug.nl/~soundscapes/DATABASES/AWP/awp-notes_on.shtml`. Retrieved in March 2014.

[Serra, 1997] Serra, X. (1997). *"Musical Sound Modeling with Sinusoids plus Noise"*, pages 91 – 122. Swets & Zeitlinger.

[Sheh and Ellis, 2003] Sheh, A. and Ellis, D. (2003). "Chord Segmentation and Recognition Using EM-Trained Hidden Markov Models". In *Proceedings of the 4th International Conference on Music Information Retrieval*, page 741.

[Shepard, 1964] Shepard, R. N. (1964). 'Circularity in Judgments of Relative Pitch". *The Journal of the Acoustic Society of America*, 36(12).

[Shoenberg, 1969] Shoenberg, A. (1969). *"Structural Functions of Harmony"*. Ernest Benn Limited, 2 edition.

[Steenbergen and Burgoyne, 2013] Steenbergen, N. and Burgoyne, J. A. (2013). *"Joint Optimization of a Hidden Markov Model - Neural Network Hybrid for Chord Estimation"*. MIREX.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *"Information Retrieval"*. London: Butterworths, 2 edition.