Evaluating the Automatic Chord Estimation and Alignments tasks needs using metrics from a code challenge

Valter Jorge da Silva[1], **Giordano Ribeiro Eulalio Cabral**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco
Av. Jorn. Aníbal Fernandes, s/n, – 50740-560 Recife, PE

vjs3@cin.ufpe.br, grec@cin.ufpe.br

Abstract. *Automatic Chord Estimation is a subject of Music Information Retrieval that tries to extract the chords of a song in a usable manner. In the last year, many researchers tried to overperform the quantitative metrics, but the results lack reproducibility by who needs them, musicians. In this article, we reviewed the state of the art of some of these areas and performed a code Challenge that was evaluated by some of the MIREX metrics and by musicians. Then, with these results, we evaluated the need for evolution on the Estimation task and the Alignment Task of the MIR area.*

1. Introduction

Within the Music Information Retrieval (MIR) process, various methods of extracting information from songs are available. One of the most discussed topics is chord identification in an automated way, with is called Automatic Chord Estimation (ACE) [1, 2]. In ACE it is proposed that a system can identify which chord is being played at a moment, given an input that can be a song or just an independent sound signal. The use of ACE could improve several aspects of the music area, such as teaching, accompaniment activities (tag-along), meta-information extraction (genre, duration, etc), automatic creation of chords sheets or scores, among others.

This area has been receiving innovative methods in recent years, but it is going through a period of stagnation. Even with new technologies emerging, such as the use of Deep Learning, the tools are created focusing on quantitative results for the problem, such as general accuracy, or how much more it estimate right. This situation creates a problem in which applications focus on making fewer mistakes, evolving in small percentage portions of the accuracy metric between one tool or another, but causing a gap in the qualitative results.

The qualitative results of a system with this proposal must be related to how much the output can be used for the final purpose of the area, being this use given by an automatic tool or by a musician. For example, a result of these systems can hit 90% of all samples, but the 10% that it misses results in a final result that is a non-reproducible input, as it is very segmented.

Another problem related to the quality of the results is the issue of aligning the chords with the original audio. In ACE tools, the output usually has a timestamp telling you when the chord change happened. However, these resulting times usually have a temporal discrepancy in relation to the original audio, necessitating the use of another technique called Chord Alignment.

Automatic Chord Alignment (ACA) is a topic in the MIR area in which techniques for performing the temporal alignment of chords are proposed [4, 5]. Given the audio of a song and the chords played in it, a Chord Alignment system is capable of synchronizing what is in the written language with the audio it represents.

In this article, a review was carried out on the state of the art of ACE and ACA tools. With this state of the art in hand, an ACE challenge was proposed and executed, to analyze the subjectivity of the results of the systems, looking for the qualitative gaps in which the ACA system could act, thus improving the results.

2. Review of the State of the Art

As an initial step, we decided to conduct a state-of-the-art review of the ACE and ACA themes. Fortunately, the MIR area has a few articles focused on compiling the state of the art for these subjects.

In Mcvicar's article [1], the authors summarize the technical concepts of music that are used in ACE systems, such as chords, notes, beat times, etc. In addition, they discuss the building of ACE systems, defining a common pipeline between the observed systems. Finally, they summarize 14 years of contribution to the area, demonstrating the results of the Music Information Retrieval Evaluation eXchange (MIREX) event, which takes place annually, and results of works important to the theme of ACE. This article is being considered as the basis for the ACE research, considering the summary of the last years in the area.

In conjunction with the work above, Pauwels et al also made a study of the state of the art for chord recognition systems [2]. In their article, the work from Fujishima [3] is cited as essential to the theme, due to the proposition of a Chromagram as a 12-dimensional array, which is the standard used today. The authors demonstrate 7 existing problems in ACE activities, showing the works carried out in the last 20 years that imply these problems. Finally, they demonstrate the change in methodology that has been taking place in the MIR area, which is unifying the themes to achieve better results, citing as a recurrent problem for this change the lack of data to carry out these activities.

For the ACA theme, we can cite Dannerberg [4] as the base work. In their work, the authors define the theme of Score Alignment, and its use by computers for

automatic score alignment. In their work, they define the term automatic music accompaniment, which we call tag-along in this work. Finally, they cite some tools that did this follow-up activity and cite the potential of its use in the music industry. Hanna et al [5] carried out experiments with the theme of Chord Alignment, in which they created a system that performed the alignment of scores considering their representation by chords.

3. Running the Challenge

To collect more data for the analysis of this work, a partnership was made with the AI startup Moises.AI, to carry out a challenge. The proposal was to carry out an event in which participants could build tools for the realization of Automatic Chord Estimation using Artificial Intelligence methods, audio, and chords sheets with music as system inputs.

The expectation was that at the end of the challenge, we would have enough data and projects to quantitatively analyze the results, that is, how well they hit the chord at the correct time. It would also be possible to carry out a qualitative analysis of the results, how much they can be used by musicians or other tools for their purpose.

In addition to the challenge proposal, an evolution of the Decibel tool [6] was also released. This software allows you to carry out ACE activities through multiple sources (Audio, Chords sheets, MIDI), using available meta-information to increase the final accuracy in terms of chord and timing change. However, this tool only uses the beat time to check if the change occurred, causing the loss of information between Beats. Another point of improvement in the tool is the chord recognition functionality, which allows only Major and Minor chords. The released version, entitled DECIBEL PLUS, proposed changes in its structure to allow the use of the trained network for songs that were not in its training dataset. Another allowed change was the optional use of MIDI, as the main objective of the proposed challenge is the use of ACE with chords sheets.

As final validation, all solutions have run using a dataset containing 30 songs with varying difficulties, available in table 2 at the end of the document. Dataset information was not made available to the participants until the end of the challenge, to avoid its use in training the solutions. The metrics used were the same as MIREX [7], which uses the mir.eval library, available in Python. The main metrics chosen were CSR (accuracy) and Segmentation, using the Sevenths method available in the library.

As output, systems should return a file with the extension ”LAB” in the format required by the mir.eval library mentioned above or in a JSON file containing the minimum information contained in the ”LAB”. The minimum information required was:

- The starting time the chord was detected;
- The end time at which the chord was detected, which must be equal to the start time of the next chord unless it is the last one in the song;
- The detected chord.

In figure 1 an example of the desired output is available. For cases where the system output occurred in JSON format, a parser would be executed to generate a file with this pattern.

Initial Time	End Time	Chord
0.000000	2.612267	N
2.612267	11.459070	E
11.459070	12.921927	A
12.921927	17.443474	E
17.443474	20.410362	B
20.410362	21.908049	E
21.908049	23.370907	E:7/3
23.370907	24.856984	A
24.856984	26.343061	A:min/b3

Figure 1: Example of output that the system should return, opened as CSV sheet

In addition to the projects being evaluated quantitatively through the metrics mentioned above, the results were evaluated by musicians, to assess how reproducible these results were.

The challenge received entries from 11 participating teams. Of this amount, 6 teams delivered projects, but only 3 returned results within the required standards and were evaluated. The results follow in the next section.

4. Results of the Challenge

After the delivery of the projects, several environments were created to run the systems. To maintain equality between them, all systems processed the same songs and should return output in the MIREX format [7] or in a JSON format that contains all the MIREX information.

As mentioned in the previous section, the mir.eval library, used in MIREX, was used here. For its use, the LAB of Ground Truth and the LAB returned by the system as output was needed for each song and participant. Below is the result of each of the projects evaluated in all stages.

4.1. Project 1: BARÕES

For project 1, entitled BARÕES, the participants used the software released for the challenge as a basis, with changes in its execution and data entry.

System navigation occurs using a WEB interface, in which the user provided the name of the song, not the audio. Then the system searches YouTube for songs similar to the term entered and then the user should choose the one that fits the search term. Finally, the system executed its algorithm on the YouTube music and returned to the user an interface that allowed the execution of the music together with the output, allowing tag-along activity and the correction of missing or wrong chords, to then provide the final output in ”LAB” format.

We verified that the Score metric was not effective, containing an average accuracy of 33% and an average segmentation of 62%, that is, 38% of the music was segmented. When analyzing the results with the tool, we verified that the accuracy was impacted due to differences in the audio used by the tool and in Ground Truth, which could be from a difference in silence at the beginning or end of the audio or different versions of the original music.

For the qualitative evaluation, the user's functionality to correct the system output in real-time, listening to the music in parallel, was well evaluated by the musicians, as this facilitates the reproduction of the result with its proper adjustments. So despite this returning a relatively high average targeting, it allowed real-time adjustments that would greatly reduce the segmentation value.

4.2. Project 2: RAGE

Project 2, entitled RAGE, used the software released for the challenge as a reference, with the addition of Sevenths chord recognition.

The system is executed in a semi-automated way. The user provides the audio as input and the system performs a search on chord sheets sites to collect the music's chords sheets, using the file name as the search point. With the audio and chord sheet available, it executes the base algorithm to return the "LAB".

This project returned an average accuracy close to 31% but had an Average Segmentation of 37%, that is, 63% of the results were segmented. When analyzing the project as a whole, we look at two possible causes for a drop in performance. The first was the addition of Sevenths chords without adding new songs in the training base that contained chords within the search spectrum, that is, new chords were added in the dictionary, but the training base continued to contain songs with only Major and Minor chords. Another problem that occurred was on the automation of the chord sheets capture because if the search returned any error, the algorithm simply did not process the song, returning 0% accuracy and 0% segmentation as result. In the scenario where the error runs were removed, the result would change to 63% average accuracy and 75% average segmentation.

For the qualitative analysis with the musicians, they realized that the results had a shift at the beginning of the song and that the tool did not capture certain chords, remaining a long time inside a chord that had already ended.

4.3. Project 3: Climber

Project 3, entitled CLIMBER used the Chordino [8] as a basis. Using the NLSS algorithm [9] from your library for chord extraction.

Its execution took place through an interface, in which the user informs the location where the songs are, chooses a song from the folder and he starts executing the algorithm. First, the algorithm runs Chordino on the audio, extracting the chords from it. Soon after, the software searches on chord sheets sites and uses these chords sheets to reinforce the chord sequence returned by the algorithm. Finally, the system returns the result in the required LAB format.

This system returned an average accuracy of 59% and an average segmentation of 78%, but it was the only one to recognize complex chords and accidents in music, performing very well on more complex songs and failing on simpler ones. An example that we can mention is the song "Three Little Birds", in which it hit 94% of the chords, having 98% of segmentation, that is, only 2% of the segmented song.

For the analysis of the musicians, they realized that the result of the system was satisfactory, especially for more complex songs. For these, they were able to follow the music with the result of the system. An example cited was the song "A Loba", which recognized complex chords and accidents during the beats, going beyond the Sevenths chords.

4.4. Compiled results of the challenge

After all the data was collected, the results of the challenge were revealed to the participants and the dataset songs titles were made also available for verification purposes. The final results of the challenge were available in table 1 below.

4.5. Final analysis and difficulties encountered

Analyzing the results, we noticed that the systems returned similar metrics, but they all had high segmentation levels. When we took the results of the systems and put them in one software that played this music with the chords together, we noticed three behaviors.

The first was that the result usually contained a slight temporal misalignment from when the chord was played and identified by the system. This in itself would impact the performance of the song in tag-along tasks without a prior review by the user, which only Project 1 provided.

The second point that impacted the quality of the results, linked to the first, is the misidentification of chords or simply their non-recognition. In some cases, the main chords were not recognized, but their components. This was due to the chord recognition form of each software, which uses a pattern of verification by sampling. An example observed is that systems based on DECIBEL [6] recognized chords only in beat time. In more complex songs, these systems simply did not recognize chord changes between beats times, and in cases where the change happened a little before the beat time, the chord recognized was wrong due to the chromagram recognized in the exact moment containing a chord in decay.

The third point was observed only in the readings of the chords sheets in a later analysis of the results, which is the difference in music notation used between some countries. The MIREX evaluation method used in the challenge takes into account the notation of HARTE [10], but the notation used by several musicians in Brazil,

Table 1: Final Result of the Challenge

Project	CSR	Segmentation	Musician Evaluation
1	0.3291	0.6183	Good
2	0.3161	0.3739	Not Good
3	0.5877	0.7776	Good

included in most chord sheets sites in the country, is the notation proposed by Chediak [11]. This resulted in an extra effort by the musicians to understand the difference in notation. Of the 3 projects, only the last allowed the return in both patterns.

In all cases, only projects 1 and 3 were considered suitable candidates for evolution, due to their correction functionality, and recognition of complex chords, respectively. But for its use to be viable to an ACA system, it would be necessary to evolve its algorithms.

5. Conclusion and Next Steps

As noted in the results of the experiments, only three systems managed to reach the final stage. Of these, two brought results that could be used for external tasks, such as use in music accompaniment software. Even though these two systems return reasonable results, it would be necessary to evolve them so that they can perform the chord alignment activity more concretely.

As the next steps, we expect to evolve the chosen systems to experiment with techniques that focus on chord alignment. We also expect that the combination of these techniques with those of ACE will be able to bring qualitative results that can be used by music reproduction systems or by musicians and students in the tag-along activity.

For this, it is also planned to evolve the system made available in the challenge or create new independent software for chord alignment.

References

[1] Matt Mcvicar, Yizhao Ni, and Tijl De Bie. Automatic Chord Estimation from Audio : A Review of the State of the Art. *IEEE Transactions on Audio, Speech and Language Processing*, 22(2):1–20, 2014.

[2] Johan Pauwels, Ken O'Hanlon, Emilia Gómez, and Mark B. Sandler. 20 Years of Automatic Chord Recognition From Audio. *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 54–63, 2019.

[3] T Fujishima. Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music, 1999.

[4] Roger B. Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM*, 49(8):38–43, 2006.

[5] Pierre Hanna, Matthias Robine, and Thomas Rocher. An alignment based system for chord sequence retrieval. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, (January 2009):101–104, 2009.

[6] Daphne Odekerken, Hendrik Vincent Koops, and Anja Volk. *DECIBEL: Improving Audio Chord Estimation for Popular Music by Alignment and Integration of Crowd-Sourced Symbolic Representations*. 2020.

[7] Music Information Retrieval Evaluation eXchange: MIREX. `https://www.music-ir.org/mirex/wiki/MIREX_HOME`. Accessed: 2021-07-15.

[8] Chordino and NNLS Chroma. `http://www.isophonics.net/nnls-chroma`. Accessed: 2021-07-15.

[9] Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult chords. *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, (1):135–140, 2010.

[10] Christopher Harte, Mark Sandler, Samer Abdallah, and Emilia Gómez. Symbolic representation of musical chords: A proposed syntax for text annotations. *ISMIR 2005 - 6th International Conference on Music Information Retrieval*, pages 66–71, 2005.

[11] Almir Chediak. *Dicionário de Acordes Cifrados*. Irmãos Vitale, 1984.

Table 2: Dataset Songs

ID	Artist	Song
1	Alcione	A Loba
2	Bob Dylan	Knockin on heavens door
3	Bruno e Marrone	Dormi na praça
4	Cassia Eller	Gatas Extraordinárias
5	Kid Abelha	Garotos
6	Luiz Gonzaga	Asa Branca
7	Lynyrd Skynyrd	Sweet home alabama
8	Marilia Mendonça	Alô porteiro
9	The Black Eyed Peas	I gotta feeling
10	Nando Reis	Por onde andei
11	Aline Barros	Digno é o Senhor
12	Barão Vermelho	Bete balanço
13	Bob Marley	Three Little Birds
14	Capital Inicial	Primeiros erros
15	Cassia Eller	Malandragem
16	Creedence Clearwater Revival	Have you ever seen the rain
17	Jorge e Mateus	Louca de saudade
18	Legião Urbana	Quase sem querer
19	Legião Urbana	Pais e filhos
20	Luiz Gonzaga	Derramaro o gai
21	Marilia Mendonça	Eu sei de cor
22	Maroon 5	Animals
23	Melim	Ouvi dizer
24	Melim	Meu abrigo
25	Nando Reis	Relicario
26	Pink Floyd	Wish you were here
27	Roupa Nova	Os corações não são iguais
28	The Beatles	Eleanor Rigby
29	Titãs	Porque eu sei que é amor
30	Victor e Leo	Borboletas