CROSS TASK STUDY ON MIREX RECENT RESULTS: AN INDEX FOR EVOLUTION MEASUREMENT AND SOME STAGNATION HYPOTHESES

Ricardo Scholz **Geber Ramalho** **Giordano Cabral**

Universidade Federal de Pernambuco
Recife, PE, Brasil

reps@cin.ufpe.br glr@cin.ufpe.br grec@cin.ufpe.br

ABSTRACT

In the last 20 years, Music Information Retrieval (MIR) has been an expanding research field, and the MIREX competition has become the main evaluation venue in MIR field. Analyzing recent results for various tasks of MIREX (MIR Evaluation eXchange), we observed that the evolution of task solutions follows two different patterns: for some tasks, the results apparently hit stagnation, whereas for others, they seem getting better over time. In this paper, (a) we compile the MIREX results of the last 6 years, (b) we propose a configurable quantitative index for evolution trend measurement of MIREX tasks, and (c) we discuss possible explanations or hypotheses for the stagnation phenomena hitting some of them. This paper hopes to incite a debate in the MIR research community about the progress in the field and how to adequately measure evolution trends.

1. INTRODUCTION

In the last 20 years, mainly due to growth of audio data available in the Internet, Music Information Retrieval (MIR) has been an expanding field of research. It encompasses various problems or tasks, whose solutions have impact in music market. Since 2005, the MIR Evaluation eXchange (MIREX) [7] is the main evaluation "arena" in MIR field, proposing datasets, tasks and metrics to compare MIR solutions. A shallow analysis of its results shows they are continuously evolving for some tasks, whereas they seem stagnated for other ones.

There are several MIR and MIREX meta-analysis papers [6][7][23][24]. However, to our knowledge, a transversal study over stagnation of results on MIR tasks is lacking, as well as an index for evolution trend measurement. Also, stagnation phenomenon on many of these tasks is not yet being deeply discussed by the community.

Both the existence of common reasons and task specific reasons for stagnation on MIR tasks are very probable. Therefore, a deep study of stagnation phenomena is task-dependent, and demands the analysis of techniques, datasets and metrics used in recent years. Then, it is out of the scope of this paper to perform a deep analysis on the reasons of stagnation for each one of the MIREX tasks. This paper intends, instead, to provoke researchers involved with MIREX tasks (stagnated or not) to test some general hypotheses we suggest, and to propose their own task specific hypotheses.

Understanding of stagnation phenomena may be improved by objective evolution trends measurement. Comparing evolution trends between different datasets or metrics, for a given task, possibly help to identify how metrics and datasets bias observable results, or how each sub-problem of the task is more or less developed. In addition, evolution trends comparison between different tasks provide an overall picture of evolution in MIR research, drawing attention to what kind of methods and strategies are being used on developing tasks that could be adapted for stagnated ones.

This paper presents an accurate empirical analysis of MIREX recent results. It also proposes a configurable quantitative index for evolution trends measurement. Finally, it raises some hypotheses and questions that could possibly explain stagnation phenomena and/or hopefully help MIR research community to exchange more information about it in order to move forward.

Section 2 presents method used to analyze data. We explain and formalize a configurable index for evolution measurement on Section 3. Section 4 raises hypotheses about possible causes of stagnation. Section 5 draws some general conclusions on the performed analysis. Finally, future works are listed on Section 6.

2. METHOD

MIREX is the MIR competition that became the main evaluation venue in MIR field. It has been running since 2005. According to MIREX 2015 final results' poster, 107 researchers from 64 teams participated in the last edition and submitted algorithms for 21 active tasks, resulting in 402 runs, over 47 different datasets [11].

MIREX contributions to the MIR community are evident. Influential MIR researchers have identified four key contributions of MIREX: "training and induction into MIR", "dissemination of new research", "dissemination of data" and "benchmarking and evaluation" [7].

In order to evaluate research progress in MIR tasks, we could have tried to compare results published in recent years. However, we decided to focus analysis in MIREX because it can be more systematic, since: (1) MIREX tasks are well defined, and (2) submissions from different years to a given task/subtask run over the same datasets and (3) the results are evaluated using the same metrics. We do acknowledge the limitations of this methodological choice, since not all MIR algorithms developed have been evaluated in MIREX competition. But, for the sake of comparison precision and extensiveness, it seemed to be the best choice.

A timeline of tasks, subtasks and datasets used for each task or subtask was constructed with data collected from MIREX results between 2010 and 2015 [11]. In order to analyze tendencies, it is necessary to consider a relevant time frame, as well as to guarantee comparisons over time are consistent. As inclusion criterion, only tasks for which there was at least one dataset used for at least five editions since 2010 were admitted. The rationale of this choice is that observing a unique dataset ensures consistency and comparability of results, whereas considering at least five editions provides a reliable time window for trend analysis. Though, from all 28 tasks proposed between the first edition in 2005 and the last in 2015, 4 tasks were discontinued until 2008, other 6 tasks were considered very recent (started in 2013 or later), whereas the remaining 18 tasks were analyzed in this study, including 3 active tasks in 2014 which did not run in 2015.

We assumed datasets and methods for metrics computation did not change, except when explicitly documented on the task's MIREX official wiki or results' pages [11]. Among the remaining candidates, one dataset for each task or subtask was chosen to collect data for analysis. When more than one dataset was available, older datasets were preferred, to allow future researches to extend this work by comparing backwards. For Audio Genre Classification task, two datasets were equally older: Mixed Set and Latin. Mixed Set was then chosen, as a more generic set tends to provide a more realistic picture of the state of the art.

Among 18 analyzed tasks, 3 tasks presented more than one subtask. "MF0 Estimation & Tracking" is divided into "MF0 Estimation" and "Note Tracking". Actually, we believe that they could be two different tasks themselves, due to the different nature of their objectives. Then, both subtasks were analyzed. For "Query-by-tapping" (QBT), two subtasks are available: "QBT with symbolic input" (subtask 1) and "QBT with wave input" (subtask 2). We analyzed subtask 1, since onset files allow participants to concentrate on similarity matching, which is the main objective of the task, instead of onset detection. Finally, "Query-by-Singing/ Humming" (QBSH) presented two subtasks: "Classic QBSH evaluation" and "Variants QBSH evaluation". Classic evaluation (subtask 1) was chosen, since the variants evaluation adds constraints to the original problem – for instance, considering queries as variants of "ground-truth" midi.

Each task has several metrics computed. As our analysis needs to rank results, for the sake of comparison, one metric for each task or subtask was chosen. As this analysis aims to understand evolution of the state of the art on each task, more general metrics were assumed to provide a more realistic picture of each task's performance. Then, metrics often used in MIREX Overall Results Posters [11] and metrics measuring overall performance were chosen, at the expense of those measuring a given characteristic of the algorithms. For instance, F-Measure was preferred when tasks also compute Precision and Recall, as Precision and Recall compute specific performances whereas F-Measure relates to both Precision and Recall.

Considering the chosen metrics, top results for each task were analyzed. We then noticed that two groups emerged: "tasks presenting stagnated results" and "tasks presenting evolving results". The first group included tasks which presented no significant improvement on results in the last years of the competition. And the second group included tasks whose results' evolution is noticeable in the last six years. Of course, there is a high level of subjectivity on deciding when a given task is evolving (and at which pace), or stagnated. For a systematic analysis, a quantitative index for evolution measurement is necessary.

3. AN INDEX FOR EVOLUTION MEASUREMENT

To perform our study, we needed a quantitative index for measuring results' evolution trends, in order to distinguish stagnated results from evolving ones. In this section we introduce what we called "Weighted Evolution Measurement Index" (WEMI), and we discuss its semantics.

Measuring stagnation phenomena by just looking to evolution graphs has limitations, as similar graphs may

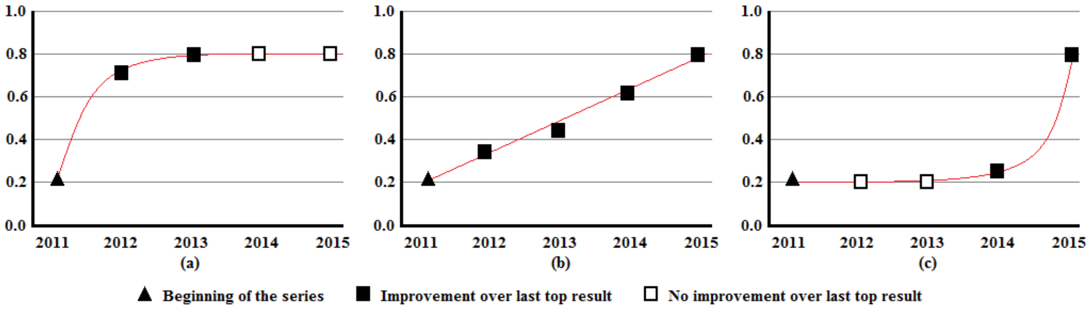

Figure 1. Examples of different evolution trends: (a) stagnation trend; (b) continuous evolution trend; and (c) recovery from recent stagnation trend.

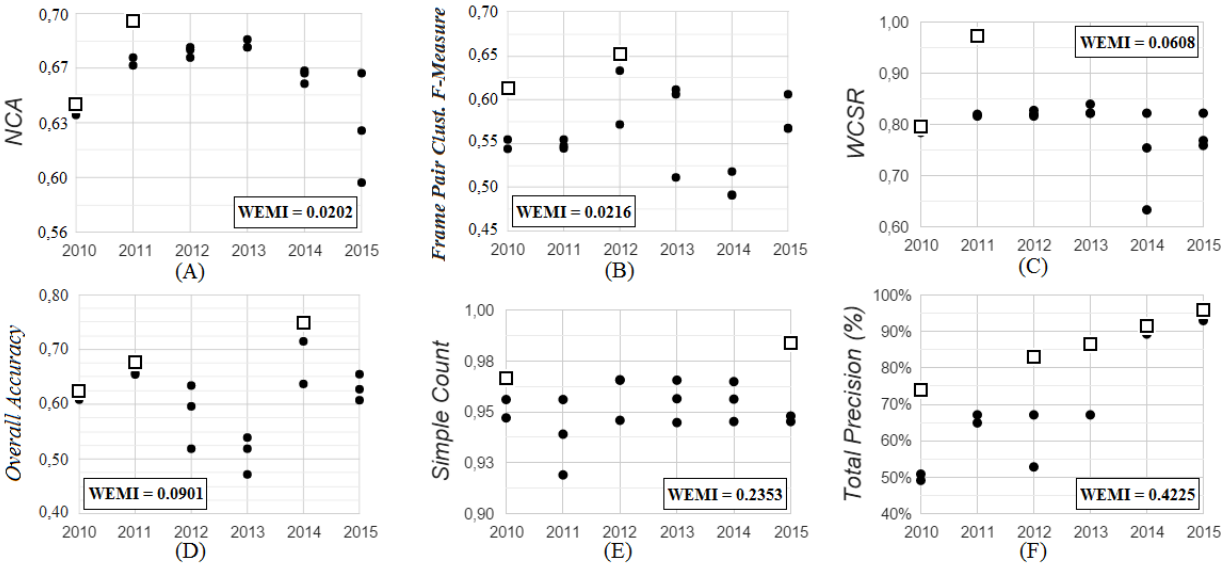

Figure 2. Some tasks results evolution plots (3 top results per year) and respective WEMI values (*w=0.6* and *c=0.0713*): (A) “Audio Music Mood Classification”; (B) “Music Structure Segmentation”; (C) “Audio Chord Estimation”; (D) “Audio Melody Extraction”; (E) Query-by-Singing/Humming”; and (F) “Score Following”; top historical results (squares) were used for trend analysis.

be hard to distinguish. For a state of the art analysis of trends, we must be able to objectively differentiate *continuous* from *intermittent* evolution, as well as measure *how* evolution occurred over time and consistently compare evolution of different tasks. For instance, consider Figure 1, which shows different hypothetical evolution scenarios. In all cases, results evolve from 0.2 to 0.8, so the first and the last result of all series coincide (overall error drop was exactly the same). However, the way evolution occurred is different in each case, so evolution trends are not the same. First series (a) shows a clear *stagnation trend*, as no recent improvement occurred after a huge improvement in the past. Second series (b) shows a *continuous evolution* of results, with small improvements every year. Finally, third series (c) shows a huge *recovery from a recent stagnation period*, as recent improvements occurred after many years without any improvement. Therefore, it is interesting that an index for evolution trend measurement can be properly balanced to differentiate these scenarios. In addition, such an index must be consistent in scenarios of complete stagnation (i.e., no evolution since the beginning of the series).

Considering that we are interested in state of the art evolution, it does make sense to discard results which did not overcome the top result achieved so far. Then, the proposed index considers evolution as a monotonically increasing function. Figure 2 shows various examples of actual top results per year, and results selected for trend analysis.

The index we propose considers a series of results from year *i* to year *f* (in this study, *i = 2010* and *f = 2015*). According to the chosen metric and state of development of each task, bias may occur if we observe top results directly. In order to avoid it, we consider relative error drop rate from one year to the next.

Error, in a given year *y*, such that $i \leq y \leq f$, is defined as:

$$e_y = 1 - max(r_i, r_{i+1}, \dots, r_y) \qquad (1)$$

Where r_j is the top result achieved in year *j*. The error drop rates are then computed for each pair of successive years (*y-1* and *y*, such that $i+1 \leq y \leq f$), as:

$$\Delta e_y = 1 - \left({}^{e_y}\!/_{e_{y-1}}\right) \qquad (2)$$

Recent evolution is reinforced by higher weights of error drop rates for recent years, so that recent improvements tend to push WEMI up more than results achieved many years ago, even if the error drop rate in both cases was the same.

Continuous evolution is reinforced with a direct proportionality between WEMI and the number of actual improvements within the time frame. This way, WEMI tends to be higher when continuous evolutions are achieved each year, in comparison with the situation in which the same overall evolution is achieved from one year to the next, at once. Then, WEMI is defined as:

$$WEMI = \left(\frac{\sum_{y=i+1}^{f} w^{f-y+1} \times \Delta e_y}{\sum_{y=i+1}^{f} w^{f-y+1}}\right) + c\frac{q}{f-i} \qquad (3)$$

The number of improvements over previous top result between *i+1* and *f* (i.e., the number of times Δe_j is larger

Task	Dataset Used	Metric Observed	YHTR	*q*	OEDR	WEMI
Audio Music Sim. and Retr.	Default	Av. Fine Score Human Eval.	2011	1	0.04	0.0188
Audio Music Mood Classif.	MIREX 2007	Normalized Class. Accuracy	2011	1	0.14	0.0202
Music Structure Segment.	MIREX 2009	Frame Pair Clust. F-Measure	2012	1	0.10	0.0216
Audio Tag Classification	Maj/Min Tag	Tag Classification Accuracy	2011	1	0.11	0.0254
MFFE&T – MF0 Estimat.	MIREX 2009	Chroma Precision	2011	1	0.36	0.0324
Audio Class. Comp. Ident.	MIREX 2009	Normalized Class. Accuracy	2011	1	0.39	0.0341
Symbolic Melodic Simil.	Essen Col.	"Fine" score[1]	2013	2	0.12	0.0398
Audio Key Detection	MIREX 2005	Weighted Key Score	2013	1	0.26	0.0540
Audio Chord Estimation	MIREX 2009[2]	Weigh. Chord Symbol Recall	2011	1	0.87	0.0608
Classic Query-by-Tapping	Roger Jang	Simple Count	2012	1	0.29	0.0738
Audio Onset Detection	MIREX 2005	Average F-Measure	2013	3	0.40	0.0801
Audio Genre Classification	Mixed Popular[3]	Normalized Class. Accuracy	2014	2	0.37	0.0829
Audio Melody Extraction	MIREX 2005	Overall Accuracy	2014	2	0.33	0.0901
Audio Tempo Estimation	MIREX 2006	Average P-Score	2015	2	0.18	0.1014
Audio Beat Tracking	MCK	F-Measure	2015	4	0.20	0.1038
MFFE&T – Note Tracking	MIREX 2009	Average F-Measure[4]	2014	3	0.58	0.1731
Query-by-Singing/Humm.	Roger Jang	Simple Count	2015	1	0.51	0.2353
Score Following[5]	Not identified[6]	Total Precision	2015	4	0.83	0.4225
Audio Cover Song Identif.	Mixed Collec.	Total num. of cov. id. in top 10[7]	2013	1	N/A	N/A

[1] Sum of fine-grained human similarity decisions. | [2] Major/minor triads classification. | [3] Also known as US Pop Music. | [4] For onset only over *chroma*. | [5] Also known as "Real Time Audio to Score Alignment". | [6] MIREX result pages mentions 3 datasets, but we could not identify which one was considered for the results in provided tables. | [7] Metric "mean number of covers identified in top 10 (average performance)" would be preferable, but is not available for all years.

Table 1. Analyzed tasks' general information; WEMI computed for w = 0.6 and c = 0.0713; YHTR stands for "Year of Historical Top Result"; OEDR stands for "Overall Error Drop Rate", computed as $OEDR = 1 - (e_{2015}/e_{2010})$.

than zero, for $i+1 \leq j \leq f$) is called q (if no improvement occurred, WEMI must be zero). Two configurable constants, w and c, are defined such that $0 < w \leq 1$ and $c > 0$. Clearly, the closer w is to zero, the greatest the weight of *recent improvements* on final index, whereas the closer it is to *1*, more equalized weights are used. Also, the closer c is to zero, the lowest the weight of *constant evolution* on final WEMI value. In this study, we computed WEMI for a variety of w and c values (results are available at *https://goo.gl/bxwrDy*). Balancing w and c depends mainly on the importance one gives to *recent* against *continuous* evolution. Therefore, we believe a discussion on MIR community about this trade off would lead to more appropriate balancing of the index for MIREX tasks, considering the goal of identifying bottlenecks of evolution and/or evaluation.

The "Audio Cover Song Identification" task could not have WEMI computed, as expected "total number of covers identified in top 10" (T10) is not available. However, results almost doubled T10 from 908 in 2010 to 1714 in 2013, regardless of the absence of improvements in other MIREX editions.

A total of 18 tasks were analyzed, with "MF0 Estimation & Tracking" comprising two subtasks. This resulted in 19 task/subtasks. Table 1 shows a summary of the analysis, with examples of output for $w = 0.6$ and $c = 0.0713$ (the average weighted sum of error drop rates of all tasks, considering $w = 0.6$).

WEMI is a first proposal and a provocation for a broader discussion about evolution measurement indexes for MIR tasks, especially on MIREX competition. Objective and early identification of stagnation trends may raise earlier discussions in the community about the appropriateness of metrics, datasets or current methods for a given task, probably helping to shorten future stagnation periods or to improve current metrics and/or datasets. Evidently, computing WEMI for a single metric of a task may be misleading. On the other hand, computing it for many metrics of a task will probably lead to a greater understanding of specific bottlenecks in task evaluation and/or evolution.

4. SOME STAGNATION HYPOTHESES

Stagnation on most MIR task results is already acknowledged by MIR community, as in the case of singer identification in polyphonic audio [13], music transcription [2], emotion and genre classification [14][19], music similarity [8][18], and so on. In spite of this acknowledgement, there is not much discussion about possible hypothesis which could explain the phenomena. Sturm [22], among others [23][24], have recently raised questions about the experimental validity in MIR evaluation, stating reliable evaluation remains neglected in MIR research. Even though, data put together in this research inspire some questions. This paper intends to provoke this kind of questions, and its explanation hypotheses.

In order to encourage this discussion, identifying whether MIREX manages to satisfactorily measure improvements of performance for its various tasks is necessary. If this is true, why so many of them are stagnated? Temporary solution stagnation phenomena are a normal stage in scientific development. However, some mechanisms could be employed to shorten them.

As we said before, since the explanations for stagnation may be task-dependent, it is difficult to provide general explanations for stagnation, and then hints on how to overcome it. Nevertheless, from some discussion on the literature of specific tasks, coupled with our own experience in MIR, we have formulated two hypotheses that may possibly help researchers to move forward. These hypotheses are not meant to be correct, but rather to start a transversal discussion among stagnated tasks.

The first hypothesis is: MIR approaches should perhaps be more musical knowledge-intensive. According to Downie [6], in 2008, community members were becoming aware of the limitations of MIR "generic approaches", i.e. the application of information retrieval solutions for music, without relying on musically meaningful features. However, since then, most of works in MIREX seems to still rely on more generic IR techniques than on an in-depth use of specific music knowledge. It is true that embedding music expert knowledge pawn generality of the approaches, but perhaps this could be path to move away from stagnation.

Let's take "Audio Chord Estimation" as an example. Chord estimation is apparently stuck into a kind of glass ceiling. Very often, approaches are agnostic, neglecting contextual information or musical knowledge after feature selection. The most successful approaches in MIREX often use probabilistic machine learning techniques, mostly through neural networks, as HMM, MLN or Bayesian [3][17][20][25]. A few approaches make use of specialist knowledge, applying it on the lower levels of symbolic information, in order to improve feature vector quality [4][12]. A deeper study on "Audio Chord Estimation" state of the art, performed by McVicar et al. [16], observed advances in feature extraction and modeling stage, as well as expert musical knowledge use for model training, but no musical knowledge use for post-processing, for instance.

Musical creation process is essentially artistic. Then, perplexity of harmonic sequences in real world tends to be high, implying less predictability. In fact, one cannot talk about a correct or wrong chord sequence, as in most classification problems. A composer not only is free to create novel chord sequences, but he or she tends to look for them. Therefore, purely probabilistic approaches are limited by the predictability of analyzed corpus, meaning that uncommon (artistically novel) chord sequences may be misrecognized. In addition, other variables may interfere in harmonic sequences, such as genre (jazz harmony differs strongly from rock harmony) or style (a given musician tends to prefer some chord sequences).

There are evidences that musical knowledge can improve chord estimation [21][1][5][15]. Therefore, we believe that improvements can be achieved using musical knowledge on higher levels of information and contextual information to decide what chord is represented by a given feature vector. By higher levels of information we basically mean musical theory applied over symbolic information. For instance, the use of functional harmonic analysis, which has been proved to add relevant information to chord sequences [21][5], to chose, among candidate chords, the ones which lead to more meaningful chord sequences, even when their feature vector are not the first options provided by a feature vector based classifier.

Music structure information has also been shown to add relevant contextual information for chord estimation [15]. For instance, the classification of "easy" chords first and the use of this information to help classification of "harder" ones, according to the harmonic meaning of the sequence they would lead to, or using harmonically similar pieces already classified of the same song (sometimes with better conditions to feature extraction and classification, such as less noise, transients, arpeggios or ornamentation) may lead to improvement on current results.

The second hypothesis is: the number of techniques employed by the MIREX community is perhaps too limited. It is very difficult to prove that a particular technique is not used by the community, as failed attempts are rarely published. But observing recent ISMIR publications, we noticed that each task presents a small set of often used techniques. For instance, chord estimation tends to rely mostly on HMM, but also on MLN or Bayesian networks, for classification.

To reinforce our hypothesis, we analyze the impact of a specific technique in MIREX results, showing how the use of a new technique can affect results. Chosen technique was Deep Learning, which dates back to the Neo-cognition introduced by K. Fukushima in 1980 [9], but only a few years ago have been found promising for MIR.

In 2012, Humphrey warned about the lack of deep learning approaches in MIR research [10]. Analyzing the top results from 2010 to 2015, among the 19 tasks/subtasks considered in this work, we can notice that, in the last 3 years, 11 tasks had their results improved, but only 3 of the improvements came from approaches using deep learning techniques, according to the technical reports submitted to MIREX. This may suggest that (1) deep learning is not being extensively explored yet in MIR and (2) if deep learning could improve results in three of the tasks, it is fair to consider there are possibly other techniques, yet not explored, with similar potential.

Regarding the first assertion, it might be due to the lack of enough labeled data, meaning that some tasks are not even eligible for Deep Learning yet. In this case, it would be fair to consider the creation of new datasets or enlargement of existing ones as a possible path to overcome stagnation on these tasks.

Of course, other hypotheses could be deeper investigated. For instance, the lowest WEMI values (even for several different w and c values) belong to tasks which use human generated ground truth data. Further investigation of this relation could lead to relevant information. Unsuitability or limitation of the datasets and metrics of the stagnated tasks are worth to investigate. However many of these hypotheses are task dependent and such an investigation would be better performed by specialists on each task.

5. CONCLUSIONS

Trying to understand recent practical advances in MIR research, a compilation of the last 6 years of MIREX results for 18 tasks (one of them comprised of two sub-tasks) was performed. Aiming to encourage discussion on how to measure progress in MIR, we propose a configurable quantitative index of improvement, the "Weighted Evolution Measurement Index" (WEMI), in order to objectively measure trends on each task, in a comparable way, reinforcing *recent* and *continuous* advances. We believe such an index may help understanding bottlenecks of evolution or measurement issues, by comparing different datasets and metrics for a given task (intra-task analysis), as well as helping to overcome stagnation, by task comparisons, observing whether methods and strategies of evolving tasks are being applied to stagnated ones (inter-task analysis). The index can be balanced, according to the community's understanding of what is most relevant: *continuous* or *recent* evolution. Also, we raise hypotheses and questions about stagnation affecting many of MIR tasks and we point some possible insights on this matter. We believe that a deeper discussion in MIR community about stagnation phenomena affecting many of MIR tasks may help to find general mechanisms or strategies which will allow overcoming it, as well as improving MIREX interest and relevance.

6. FUTURE WORK

It would be possible to obtain a more detailed overview of MIREX tasks' trends with a number of additional information, for instance: (1) comparing WEMI computed for other metrics or other datasets, in a given task, will probably help understanding if metrics or datasets are biasing observable evolution trends at first sight; and/or (2) a deeper study on discrepant results (for instance, "Classical Composer Identification, 2011" and "Chord Estimation, 2011", as seen in Figure 2) in order to identify overfitting or other distortions of top result will certainly improve accuracy. Another interesting improvement would be adaptation of WEMI so that total weights sum is normalized by the time window, as this would allow more consistent comparisons between time windows of different lengths, if this makes sense in a given context. Finally, a formal evaluation of the index is still missing.

7. REFERENCES

[1] Bello, J. P.; Pickens, J. 2005. A Robust Mid-level Representation for Harmonic Content in Music Signals. *Proc. of the 6th International Society for Music Information Retrieval Conference* (London, United Kingdom, September 11 – 15), ISMIR'05. 304-311.

[2] Benetos, E., et al. 2013. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41, 3 (July 2013), 407-434. DOI=10.1007/s10844-013-0258-3.

[3] Chen, R.; Shen, W.; Srinivasamurthy, A.; Chordia, P. 2012. Chord recognition using duration-explicit hidden Markov models. *Proc. of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal, October 8 – 12, 2012), ISMIR'12. 445-450.

[4] Cho, T.; Bello, J. P. 2011. A feature smoothing method for chord recognition using recurrence plots. *Proc. of the 12th International Society for Music Information Retrieval Conference* (Miami, USA, October 24 – 28, 2011), ISMIR'11. 651-656.

[5] De Haas, W. B.; Rodrigues Magalhães, J. P. and Wiering, F. 2012. Improving Audio Chord Transcription by Exploiting Harmonic and Metric Knowledge. *Proc. of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal, October 8 – 12, 2012), ISMIR'12. 295-300.

[6] Downie, J. 2008. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29, 4, 247-255. DOI=10.1250/ast.29.247.

[7] Downie, J., et al. 2014. Ten years of MIREX: reflections, challenges and opportunities. In *Proc. of the 15th International Society for Music Information*

Retrieval Conference (Taipei, Taiwan, October 27 – 31, 2014). ISMIR'14. 657-662.

[8] Flexer, A. 2014. On Inter-rater Agreement in Audio Music Similarity. In *Proc. of the 15th International Society for Music Information Retrieval Conference* (Taipei, Taiwan, October 27 – 31, 2014), ISMIR'14. 245-250.

[9] Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, 4, 193–202. DOI = 10.1007/bf00344251.

[10] Humphrey, E. J.; Bello, J. P.; Lecun, Y. 2012. Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. *Proc. of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal, October 8 – 12, 2012), ISMIR'12. 403-408.

[11] ISMIRSEL. 2016, January 8. *MIREX Home [Online]*. Available: http://www.music-ir.org/mirex/wiki/MIREX_ HOME.

[12] Khadkevich, M.; Omologo, M. 2011. Time-frequency reassigned features for automatic chord recognition. *Proc. of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing* (Prague, Czech Republic, May 22 – 27, 2011), ICASSP'11. 181-184.

[13] Lagrange, M., Ozerov, A., and Vincent, E.. 2012. Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning. *Proc. of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal, October 8 – 12, 2012), ISMIR'12. 595-600.

[14] Lidy, T., et al. Improving Genre Classification by Combination of Audio and Symbolic Descriptors Using a Transcription System. *Proc. of the 8th International Society for Music Information Retrieval Conference* (Viena, Austria, September 23 – 27, 2007), ISMIR'07. 61-66.

[15] Mauch, M.; Noland, K. and Dixon, S. 2009. Using Musical Structure to Enhance Automatic Chord Transcription. *Proc. of the 10th International Society for Music Information Retrieval Conference* (Kobe, Japan, October 26 – 30, 2009), ISMIR'10. 231-236.

[16] McVicar, M.; Santos-Rodríguez, R.; Ni, Y. and De Bie, T. 2014. Automatic Chord Estimation from Audio: A Review of the State of the Art. *Proc. of IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22, 2, 556-575. DOI = 10.1109/TASLP.2013.229458.

[17] Ni, Y.; McVicar, M.; Santos-Rodríguez, R.; De Bie, T. 2012. An end-to-end machine learning system for harmonic analysis of music. *Proc. of the IEEE Transactions on Audio, Speech and Language Processing*, 20, 6, 1771-1783. DOI = 10.1109/TASL.2012.2188516.

[18] Pachet, F. and Aucouturier, J. 2004. Improving timbre similarity: How high is the sky?. *Journal of Negative Results in Speech and Audio Sciences*, 1, 1, 1-13.

[19] Panda, R., et al. 2013. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. *Proc. of the 10th International Symp. on Computer Music Multidisciplinary Research* (Marseille, France, October 15 – 18, 2013), CMMR'13. 570-582.

[20] Papadopoulos, H.; Tzanetakis, G. 2012. Modeling chord and key structure with Markov logic. *Proc. of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal, October 8 – 12, 2012), ISMIR'12. 127-132.

[21] Scholz, R.; Vincent, E.; Bimbot, F. 2009. Robust modeling of musical chord sequences using probabilistic N-grams. *Proc. of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing* (Taipei, Taiwan, April 19 – 24, 2009), ICASSP'09. 53–56.

[22] Sturm, B. L. A Simple Method to Determine if a Music Information Retrieval System is a "Horse". 2014. In *IEEE Transactions on Multimedia*, 16, 6 (July 2014), 1636-1644. DOI = 10.1109/TMM.2014.2330697.

[23] Urbano, J. 2011. Information Retrieval Meta-evaluation: Challenges and Opportunities in the Music Domain. In *Proc. of the 12th International Society for Music Information Retrieval Conference* (Miami, USA, October 24 – 28, 2011), ISMIR'11. 609-614.

[24] Urbano, J., Schedl, M. and Serra, X. 2013. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 41, 3 (December 2013), 345-369. DOI= 10.1007/s10844-013-0249-4.

[25] Yoshii, K.; Mauch, M.; Goto, M. 2011. A unified probabilistic model of note combinations and chord progressions. *Workshop Book of Neural Information Processing Systems, 4th International Workshop on Machine Learning and Music: Learning from Musical Structure* (Sierra Nevada, USA, December 17, 2011), MML'11. 46.