META-SONG EVALUATION FOR CHORD RECOGNITION

Yizhao Ni[1], **Matt Mcvicar**[1], **Raul Santos-Rodriguez**[2] **and Tijl De Bie**[1]
1. University of Bristol, U. K. 2. Universidad Carlos III de Madrid, Spain

Introduction. In recent years, audio *Chord Recognition* (CR) has become a very active field. The increasing popularity of Music Information Retrieval (MIR) with applications using mid-level tonal features has established chord recognition as an useful and challenging task. The annual MIREX (Music Information Retrieval Evaluation eXchange) competition has a task dedicated to chord recognition, where participants attempt to predict labels and boundaries for a song collection. In the most recent competitions, the dataset used consists of 217 songs from a collection of The Beatles, Queen and Zweieck albums for which the *ground truth annotations* are available [1]. Due to the limited amount of data, existing CR systems are usually trained & tested on the same songs, inevitably causing over-fitting on this dataset. Meanwhile, such evaluation is also heavily constrained by the simplicity of the data. For example, most of the songs in the dataset are from the Rock genre, implying that the performance might lack generalization to other genres.

To resolve these problems, the simplest but most costly solution would be to obtain more fully annotated data. Alternatively, we propose using a methodologically more challenging but cheaper and scalable approach: *meta-song evaluation*, which makes use of large and freely available online chord databases, such as *E-chords* [2] to help evaluate CR systems. The principle is to automatically generate high accurate but not perfect "pseudo chord annotations" for new songs, of which the chord sequences are available on these databases. The songs and the "pseudo annotations" are then used to estimate CR systems' performance via statistical theories. In our previous work [3] we have demonstrated a variety of models for generating such "pseudo annotations". This late breaking paper will show how to make use of these pseudo annotations to comprehensively evaluate performances of different CR systems [3].

[1] http://www.music-ir.org/mirex/wiki/2010:Audio_Chord_Estimation

[2] http://www.e-chords.com/

[3] The reader is referred to [6] for an extended version of this paper.

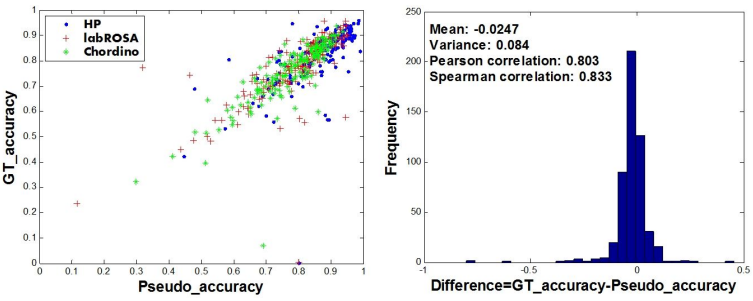

Figure 1. The relationship between the pseudo accuracies and the real GT accuracies on the 175 Beatles songs.

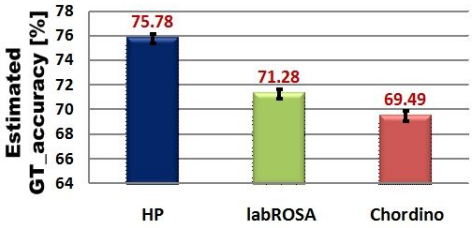

Figure 2. The estimated GT accuracies of the CR systems on 1840 songs. Error bars represent confidence intervals of performance within 95% confidence level.

Methodology. We use y_i^A and x_i^A to denote the *ground truth (GT) accuracy* and the *pseudo accuracy* (i.e. the accuracy of system's prediction compared to pseudo annotation) of CR system A's chord prediction for the i-th song. Then for each system we obtain two sets of data: a validation set $\{x_i^A, y_i^A\}_{i=1}^n$ and a test set $\{x_j^A\}_{j=n+1}^{n+m}$. Note that we only have ground truth annotations on the validation set and generally $m \gg n$. The CR system pool is denoted by $A \in \mathcal{A}$.

One observation from the validation set is that the pseudo accuracies are highly correlated with the GT accuracies (see Figure 1), as long as the pseudo annotations are accurate enough. In the ideal case, if all pseudo annotations are 100% accurate, the pseudo accuracies will converge to the GT accuracies. Inspired by this observation, we propose a linear regression framework to model the relationship between GT and pseudo accuracies on the validation set, which can then be applied to estimate GT accuracies on the test set.

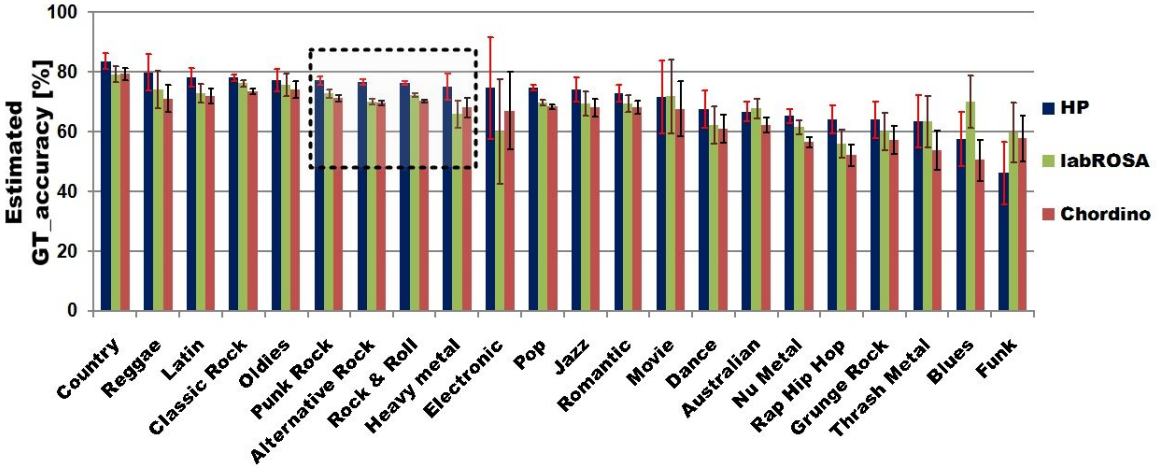

Figure 3. The estimated GT accuracies of the CR systems on each genre. See the caption of Figure 2.

Mathematically, we assume that (x_i^A, y_i^A) generated by CR system A are sampled *i.i.d.* from a Gaussian distribution

$$y_i^A = a_A x_i^A + b_A + \epsilon_i,\ 1 \le i \le n,\ \epsilon_i \sim \mathcal{N}(0, \sigma_A^2),\ \forall A \in \mathcal{A},$$

where the parameters $(\bar{a}_A, \bar{b}_A, \bar{\sigma}_A^2)$ can be estimated by the method of *least squares*. Therefore, in terms of the linear regression theory [4], a Gaussian distribution holds for all test examples $y_j^A \sim \bar{a}_A x_j^A + \bar{b}_A + \mathcal{N}\Big(0, \bar{\sigma}_A^2\big(1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}\big)\Big)$, with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^A$ and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^A - \bar{x})^2$.

Using the Gaussian properties, the test mean accuracy $\bar{y}^A = \frac{1}{m}\sum_{j=n+1}^{n+m} y_j^A$ follows $\bar{y}^A \sim \bar{a}_A \bar{x}^A + \bar{b}_A + \mathcal{N}(0, \hat{\sigma}_A^2)$, with $\bar{x}^A = \frac{1}{m}\sum_{j=n+1}^{n+m} x_j^A$ and $\hat{\sigma}_A^2 = \bar{\sigma}_A^2 \frac{\sum_{j=n+1}^{n+m} 1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}}{m^2}$. Via this distribution we can estimate the confidence interval of $\bar{y}^A$ with probability $1-\alpha$: $\bar{y}^A = \bar{x}^A + \bar{\mu} \pm Q(1-\alpha)\hat{\sigma}_A$, where Q denotes a normal quantile function $Q(p) = \inf\{y \in \mathbb{R} : p \le Pr(Y \le y)\}$. Meanwhile, we can also compare two CR systems A and B, by means of estimating the confidence interval of $\bar{y}^A - \bar{y}^B$ using $\bar{y}^A - \bar{y}^B = \bar{a}_A \bar{x}^A - \bar{a}_B \bar{x}^B + \bar{b}_A - \bar{b}_B \pm Q(1-\alpha)\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}$.

Experiments. The experiments involve three CR systems: the machine learning based HP [5] and labROSA [1] systems, which are trained on the MIREX dataset, and also Chordino [2], an expert knowledge based system. The validation set consists of 175 The Beatles' songs, of which we use the ground truth and pseudo annotations to train the relationship model. The test set consists of 1840 songs from a variety of genres, of which we can only derive pseudo annotations from E-chords. The objective is to estimate and compare the GT accuracies of the three systems on the test set, given their "pseudo accuracies" and the relationship model.

We first evaluated the systems on the whole dataset and the results are presented in Figure 2. We observed that the estimated GT accuracy of labROSA is slightly better than Chordino, implying a better generalization of ML-based systems over expert knowledge based ones. Meanwhile, HP achieves a large improvement over the other two systems, indicating its superiority. We then categorized the songs by their genres and estimated the performances of the systems on each genre. The results are illustrated on Figure 3. We observed that HP performs better on most of the genres, especially on Rock related genres. This conforms to the fact that HP is trained on songs mainly from the Rock genre.

1. REFERENCES

[1] D. Ellis and A. Weller. The 2010 LABROSA chord recognition system. In *Proc. of ISMIR (MIREX)*, 2010.

[2] M. Mauch and S. Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proc. of ISMIR*, 2010.

[3] M. McVicar, Y. Ni, R. Santos-Rodriguez, and T. De Bie. Using online chord databases to enhance chord recognition. *Journal of New Music Research*, 40(2), 2011.

[4] J. Neter, W. Wasserman, and M. H. Kutner. *Applied linear statistical models*. Irwin Press, Boston, 1990.

[5] Y. Ni, M. Mcvicar, R. Santos-Rodriguez, and T. Die Bie. An end-to-end machine learning system for harmonic analysis of music. In *http://arxiv.org/abs/1107.4969v1*, 2011.

[6] Y. Ni, M. Mcvicar, R. Santos-Rodriguez, and T. Die Bie. Meta-song evaluation for chord recognition. In *http://arxiv.org/abs/1109.0420*, 2011.