AUDIO COVER SONG IDENTIFICATION: MIREX 2006-2007 RESULTS AND ANALYSES

J. Stephen Downie, Mert Bay, Andreas F. Ehmann, M. Cameron Jones
International Music Information Retrieval Systems Evaluation Laboratory
University of Illinois at Urbana-Champaign
{jdownie, mertbay, aehmann, mjones2}@uiuc.edu

ABSTRACT

This paper presents analyses of the 2006 and 2007 results of the Music Information Retrieval Evaluation eXchange (MIREX) Audio Cover Song Identification (ACS) tasks. The Music Information Retrieval Evaluation eXchange (MIREX) is a community-based endeavor to scientifically evaluate music information retrieval (MIR) algorithms and techniques. The ACS task was created to motivate MIR researchers to expand their notions of similarity beyond acoustic similarity to include the important idea that musical works retain their identity notwithstanding variations in style, genre, orchestration, rhythm or melodic ornamentation, etc. A series of statistical analyses were performed that indicate significant improvements in this domain have been made over the course of 2006-2007. Post-hoc analyses reveal distinct differences between individual systems and the effects of certain classes of queries on performance. This paper discusses some of the techniques that show promise in this research domain

1. INTRODUCTION

Founded in 2005, the annual Music Information Retrieval Evaluation eXchange (MIREX) is a community-based endeavor to scientifically evaluate music information retrieval (MIR) algorithms and techniques. Since its inception, over 300 music information retrieval (MIR) algorithms have been evaluated across 19 distinct tasks. These tasks were defined by community input and range from such low-level tasks such as Audio Onset Detection to higher-level tasks as Audio Music Similarity and Retrieval. More information about MIREX can be found at the MIREX wiki [8] where task descriptions and results are archived. This paper focuses on one specific task, namely Audio Cover Song Identification (ACS), which was first run in 2006 and repeated in 2007.

This paper is organized as follows: In Section 1.1, we discuss the motivation for conducting an ACS task. In Section 2, we introduce the task design and its evaluation dataset and the evaluation metrics used. In Section 3 we compare the results of the ACS 2006 and 2007 tasks. In Section 4, we focus on the ACS 2007 results and perform a set of statistical significance tests to investigate differences in system performance and the effects of the data on these performances. Section 5 summarizes what has been learned from the examining the different approaches to cover song identification. Section 6 contains the conclusion and future work.

1.1. Motivation

Aucouturier and Pachet's [1] seminal 2004 study identified the limitations of using audio-based timbral features to perform music similarity tasks. They performed more than one hundred machine learning experiments using spectral features and could only improve the performance 15% over a baseline. They called this problem the "glass ceiling". Whitman, et al. [11] investigated the "album effect", where they saw that the performances of artist identification systems were inflated by machine learning algorithms picking up on similarities in the production qualities of albums. The album effect has also been investigated by Kim, et al. [6]. Pampalk, et al. [9] addressed similar effects, where they evaluated genre classification systems on artist-filtered datasets and noted a marked reduction in performance.

The glass-ceiling, album and artist-filtering effects can also be seen throughout the MIREX 2005-2007 results. For example, comparing the best results for the Audio Genre Classification task of MIREX 2005, 82.34% (Bergstra, Casagrande and Eck), with the MIREX 2007 results, 68.29% (IMIRSEL (SVM)) [8] we see an apparent reduction in system performance across the two years. Similarly, the top Audio Artist Classification results for 2005, 72.45% (Mandel and Ellis) and 2007 48.14% (IMIRSEL (SVM)) also exhibit a seemingly large decline in performance. These performance drops can be partially explained by the fact that in 2005 there was no artist or album filtering of the test and training sets used these evaluations. In the 2007 Audio Genre Classification task, the data was filtered such that no track from the same artist could simultaneously exist in both the test and train sets in any cross-validation fold. Also, in the 2007 Audio Artist Identification task, the test collections did not include any track from the same album in test and training sets of any cross-validation folds.

The issues of an apparent glass ceiling, in conjunction with the absence of artist and album filtering in early MIR system evaluations overstating performance effectiveness,

indicated a need for the development and evaluation of methods using higher-order music descriptors in MIR similarity tasks. The ACS task was created to motivate MIR researchers to expand their notions of similarity beyond acoustic similarity to include the important idea that musical works retain their identity notwithstanding variations in style, genre, orchestration, rhythm or melodic ornamentation, etc. Because cover songs are known to span a range of styles, genres, and instrumentations, yet are often, in some sense, undeniably "similar," the evaluation of cover song identification performance can address the distinction between timbral similarity and "musical similarity". While identifying cover songs represents only a narrow scope of possible applications in regard to the use of higher-level features in MIR systems, it is an effective starting point in evaluating the usefulness of currently proposed "high-level musical descriptors", like those being investigated in [5][7].

2. ACS TASK DESCRIPTION

2.1. Evaluation Dataset

The ACS task dataset consists of 1000 tracks. Thirty different "cover song" groups each having 11 different versions for a total of 330 tracks are embedded within the ACS database. The original works in the cover song collection come from a variety of genres such as pop, rock, classical, baroque, folk, jazz, etc. Furthermore, within each group, the versions of each work similarly span a wide range of genres with different styles and orchestrations. The total length of the 330 cover songs is 21.2 hours with an average track length of 232 seconds (σ = 77 sec.). The remaining 670 tracks in the database are "noise". The noise tracks were chosen to be unrelated to any of the cover songs and their performing artists. The noise set also reflects a broad variety of genres and styles. The total length of the noise set is 45.8 hours with an average track length of 241 seconds (σ = 82 sec.). Thus the total length of the ACS dataset is 67 hours with an average track length of 242 seconds (σ = 72 sec.). Unlike many other MIREX tasks where 30 second clips were commonly used, the ACS task employed whole tracks to allow participants the opportunity to exploit the potentially important musical structure of the pieces.

All tracks in the dataset were encoded as 128 kbps MP3s and then decoded back to 22.05 kHz 16-bit WAV files using the LAME codec. Since the cover songs came from variety of sources with different encoding parameters, the MP3 encoding/decoding step was necessary to normalize the dataset to minimize the influence of coding effects on system performance.

2.2. Evaluation Methods and Metrics

The goal of the ACS task is to use each cover song track as a "seed/query" for identifying the 10 other versions of that piece in the dataset. All tracks in each cover song group are used as queries for a total of 330 queries. Since the underlying work of each individual cover song in a cover song group is known, the ground-truth for the ACS task is unambiguous and non-subjective. This distinguishes the ACS task from such other music similarity tasks as Audio Music Similarity, Audio Genre Classification, Audio Mood Classification, etc., which require the application of potentially subjective human judgments. The same dataset was used for both ACS 2006 and 2007. The identities of the pieces in the dataset have never been released to preclude the possibility of the *a priori* tuning of the submitted systems.

In ACS 2006, the overall evaluations were based on average performance and mean reciprocal rank (MRR). Average performance was defined as the mean number of covers identified within the top 10 returned items by the system. Rescaling the average performance score to the range of [0, 1] yields the precision at 10 (P@10) value for that system. Reciprocal rank was calculated as 1 over the rank of the first correctly identified cover song. In 2006 the systems only returned their top 10 candidates.

In ACS 2007, the participants introduced a new evaluation metric: mean average precision (MAP). For each query, average precision is calculated from the full returned list (i.e., 999 returned songs) as the average of precisions when the ranked list is cut off at each true item:

$$Ave.P = \frac{1}{10}\left(\sum_{r=1}^{999} p(r) \cdot I(r)\right) \quad (1)$$

where $p(r)$ is precision at rank r

$$p(r) = \sum_{j=1}^{r} \frac{I(j)}{r} \quad (2)$$

and $I(j)$ is a binary indicator function which is 1 if the j^{th} returned item in the list is a cover song, and 0 otherwise. The MAP is calculated as the mean of average precisions across all 330 queries. MAP is a commonly used metric in the text information retrieval domain [3]. Using MAP has the advantage of taking into account the whole returned list where correct items ranked closer to rank 1 receive the largest weights.

3. COMPARISON OF 2006-2007 RESULTS

Eight systems participated in ACS 2006 resulting in a task-wide P@10 of 0.08 (σ = 0.067), and a task-wide MRR of 0.19 (σ = 0.13). Table 1 (see Appendix A for legend) shows the P@10 values for each system. It is quite important to note that the systems labeled with '*' were not specifically designed for the ACS task. These systems, which were originally designed to participate in the Audio Music Similarity task, were graciously

volunteered to help the ACS organizers to determine whether standard music similarity algorithms would be adequate for the ACS task. Similarly, the task organizers included the IM system in the 2007 evaluations. The average P@10 of the top 4 (task-specific) systems in ACS 2006 was 0.13 ($\sigma = 0.073$). The average MRR of these 4 system was 0.28 ($\sigma = 0.14$).

Eight systems participated in ACS 2007 resulting in a task-wide MAP of 0.2062 ($\sigma = 0.1674$). The task-wide P@10 for ACS 2007 was 0.2057 ($\sigma = 0.1675$). The task-wide MRR was 0.38 ($\sigma = 0.27$). The average P@10 scores for the top 4 systems in 2007 were 0.34 with a standard deviation of 0.12.

Table 1 shows the P@10 values of the systems for both years. We can see a substantial improvement in ACS 2007 over 2006. The top 4 systems scores in 2007 are the same or better than the best score of 2006. After confirming that the top 4 mean P@10 values for both 2006 and 2007 are normally distributed using the Jarque-Bera (J-B) test [2] ($p < 0.05$), we ran a one-way ANOVA on the top 4 systems from each year to see if there is a statistically significant difference in performance between the years. The ANOVA indicated a significant difference between the P@10 means with $F(1,6) = 9.18$, $p = 0.023$. This result highlights that there has been a ~270% improvement of the top performing ACS systems in one year.

2006		2007	
DE	0.23	**SG**	0.50
KL1	0.11	**EC**	0.37
KL2	0.10	**JB**	0.26
CS	0.06	**JEC**	0.23
LR*	0.05	**KL1**	0.13
KWL*	0.04	**KL2**	0.09
TP*	0.04	**KP**	0.06
KWT*	0.03	**IM****	0.01

Table 1. Precision at 10 for ACS 2006 – 2007 results

It is also important to look at the non-task-specific systems. For example, the IM system which had the lowest score of all systems over both years was based on the same naïvely constructed spectral feature set (i.e., MFCC's, zero-crossing rates, spectral flux, spectral centroid, etc.) as the IM-SVM system that ranked amongst the top systems for the Audio Artist Identification, Audio Genre Classification and Audio Classical Composer Identification tasks in MIREX 2007. The weak performance of the non-task-specific systems strongly suggests that to capture the identity aspects of "music similarity," one should go beyond simple spectral features. Top-performing systems in ACS 2007 used higher-level features (e.g., rhythm, tonality, tonal sequences, etc.) so as to capture the musically important structures.

4. ANALYSIS OF 2007 RESULTS

In this section we will focus exclusively on the ACS 2007 results since the overall performance of the 2007 systems is significantly better than the 2006 group. Furthermore, in 2007 all but one system (i.e., IM) were specifically designed for the ACS task.

4.1. Descriptive Analyses of ACS 2007 Results

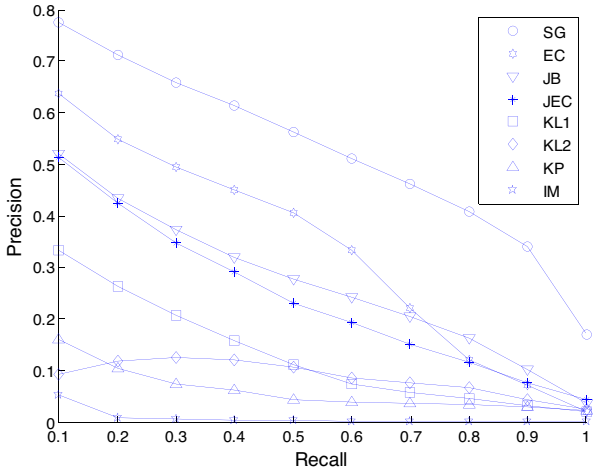

Figure 1. Precision-Recall curves for ACS 2007

The precision-recall graph in Figure 1 was generated by calculating the precision values at each recall level from 0.1 to 1 and averaging across all 330 queries for each system. Looking at the graph, we can see that SG and EC retrieved substantial numbers of relevant cover songs in early portions of their results list. In fact SG had 26 queries with perfect precision at recall equal to 1. SG had a further 10 queries where the entire relevant set was returned within the first 11 items. EC had 12 queries where 9 out of the 10 relevant cover songs were returned in the first 10 items. These results are quite remarkable because, given our dataset size (1000) and the number of relevant items per query (10), the probability of randomly returning the entire relevant set within the top 10 list once in 330 queries is only 1.26×10^{-21}! Also noteworthy, is the extraordinarily flat performance curve of the IM system.

Figure 2 shows the box-whisker plot of the distributions of MAP values for each system across the 30 cover song groups. The bottom, middle and top of each box represent the lower quartile, median and upper quartile values, respectively. The '+'s are the outliers for each distribution. The data were amalgamated with respect to their cover song groups because ACS task is primarily interested in system performance with regard to the identification of the 30 underlying works rather than the 330 individual pieces. Thus, a MAP score was calculated

for each cover song group as the mean of the average precision values for each group's 11 members.

In Figure 2, we see that there is a fair amount of variance across query groups with respect to system performance. This is especially noticeable in the top 4 performing systems (i.e., JB, JEC, EC, SG). This suggests that some query groups might have significant effects (positive or negative) on system performance.

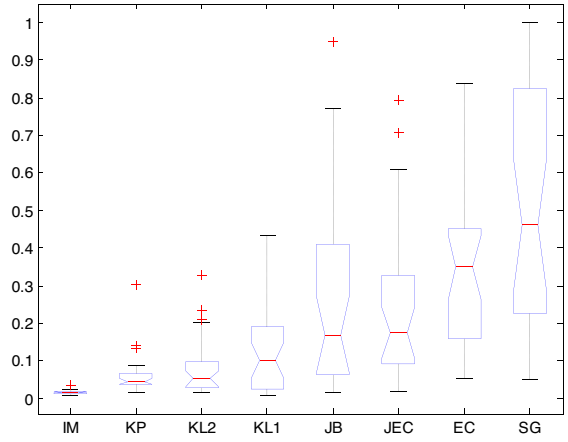

Figure 2. Box whisker plot of MAP distributions for each system across cover groups (query groups).

4.2. Inferential Significance Testing

In this section, we analyze the results in several different perspectives. We are interested in determining the answers to the following questions:

1. a) Is there any significant difference in performance means among the systems? and, b) If such differences exist, between which systems do they occur?
2. a) Is there any significance difference among performance means in query groups? and, b) If such differences exist, between which query groups do they occur?

Following the procedures outlined in [3], we determined whether the by-query-group MAP data discussed above were normally distributed across query groups using the J-B test. Most of those data did not conform to the normal distribution. However, after applying the arcsine square-root transformation:

$$y = arcsin(\sqrt{x}) \tag{3}$$

as recommended by [10], 7 out of 8 systems across query groups, and 28 out of 30 query groups across systems passed the J-B test ($p < 0.05$). Since the dataset is approximately normally distributed using arcsine square-root transformation, we selected a parametric test to investigate whether there are any significance differences among systems or query groups. Parametric tests are preferred, where appropriate, because they are more powerful than their non-parametric counterparts: they better detect differences that might be overlooked (i.e., they have lower Type II error rates).

A two-way ANOVA was chosen because it can provide answers to both of our system (Q.1) and our query group (Q.2) questions simultaneously. Table 2 shows the results of the two-way ANOVA on the transformed data. As one can see, there are significant differences among both systems and query groups.

Source	Sum Sq.	D.F.	Mean Sq.	F-stat	P
Systems	10.51	7	1.50	54.64	0.00
Query Groups	6.53	29	0.23	8.20	0.00
Error	5.58	203	0.027		
Total	22.62	239			

Table 2. Two-way ANOVA table

To further investigate the differences between individual system performance means (Q.1b), we performed the Tukey-Kramer Honestly Significantly Different (TK-HSD) analysis on the system data. TK-HSD was used because it can properly control the experiment-wise Type-I error rate unlike the commonly misused multiple *t*-tests [10]. Figure 3 shows the results of the TK-HSD on the transformed by-query-group MAP data for each system. The circled items refer to individual groupings based on the absence of significant differences within the grouping. The answer to Q.1b can be seen in Figure 3. It is evident that the SG system is significantly better than the other systems in this task. Also EC, JB and JEC have formed their own grouping. It is important to note that these results differ from those presented ACS 2007 results wiki [8] where the non-parametric Friedman's test was performed on the non-transformed data. The lower power of the Friedman's test appears to have missed the significantly better performance of the SG system.

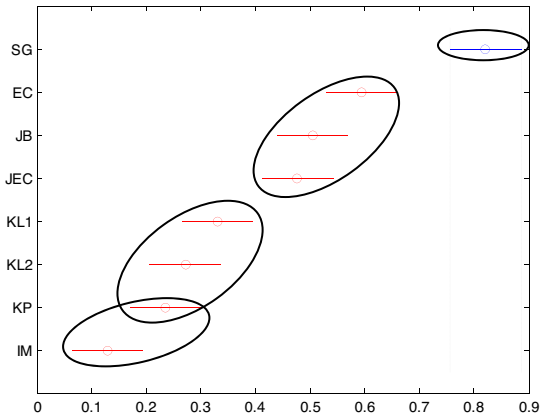

Figure 3. TK-HSD analysis on system effects based on the transformed by-query-group MAP data.

To answer Q.2b, We ran the TK-HSD test to determine where differences in performance occur with regard to the query groups. Table 3 displays the results. The first column represents the anonymized IDs of the query groups which are rank ordered with respect to the by-query-group MAP across the systems. The second column presents the by-query-group MAP. The shaded regions indicate sets of query groups which are not significantly different in how the algorithms performed. For example, the first column indicates that query groups 1 through 10 are not significantly different from one another.

Group no.	Avg. MAP
1	0.52
2	0.42
3	0.41
4	0.35
5	0.35
6	0.34
7	0.33
8	0.30
9	0.29
10	0.28
11	0.27
12	0.24
13	0.24
14	0.20
15	0.18
16	0.16
17	0.16
18	0.16
19	0.13
20	0.12
21	0.10
22	0.10
23	0.08
24	0.08
25	0.08
26	0.08
27	0.06
28	0.05
29	0.05
30	0.03

Table 3. Significance sets of cover song groups.

Table 3 clearly illustrates the wide range from 0.03 to 0.52 ($\sigma = 0.24$) of by-query-group MAP performances.

Since there is such a large discrepancy between the best query group and the worst, we investigated the two extremes to explain attempt to explain how systems behave in response to different query groups.

The best performing group (Group 1 in Table 3) is a late 17th-century canon with a regular structure and harmonic progression. All versions of Group 1 share the same rhythmic structure as they should, since canons are composed of replication of the rhythms and intervals of the same main theme. We surmise this makes it easier for algorithms to accurately compare the rhythmic and tonal structures of the songs. There is not much variance in orchestration or tempo in Group 1. This was one of the query groups, in which SG achieved its near-perfect MAP scores (0.998). The SG method uses sequences of tonal descriptors where songs are matched using dynamic programming for local alignment. The EC system also performed very well (MAP of 0.838) for this group. It uses correlation of beat-synchronous chroma features.

The worst performing group (Group 30 in Table 3) is an 18th century hymn set to its now-traditional 19th century melody. All the song versions in this cover group vary greatly in their harmonies, chord progressions, rhythms, tempi and dynamic ranges. The performances encompass many different styles such as country, blues, jazz, hip-hop, rock, etc. Group 30 songs exhibit a great deal of widely varying melodic ornamentation and several different languages. Virtually all task-specific systems use tempo and key-independent matching of the underlying tonal or harmonic structure of the pieces. Because the variations of the Group 30 songs contain a wide range of embellishments and changes to the harmonic structure, we believe the systems are sensitive to the varying nature of this group. SG scored a MAP of 0.052. EC scored the highest MAP of 0.06 for Group 30.

5. DISCUSSION

In the design of audio cover song algorithms, the top performing algorithms share a variety of attributes. A majority of the four top performing algorithms use chromagrams or pitch class profiles as the predominant feature representation, with methods for addressing possible changes in key and tempo in matching songs. Chromagrams represent the distribution of spectral energy quantized to the chromatic scale.

The EC algorithm addresses variations in tempo by performing a tempo induction stage and producing a beat-synchronous chromagram that contains a single chroma vector per beat, and uses cross correlation of the chromagrams for song matching. The SG system uses dynamic programming to align and match harmonic pitch class profiles. JEC also uses an underlying chromagram representation, but filters the logarithm of each of the 12 chroma channels into 25 logarithmically spaced bands. This 12×25 feature matrix captures the variation of each of the 12 chroma channel on scales of 1.5 to 60 seconds. Because a logarithm is used prior to filtering, large changes in tempo become apparent as simple shifts along the filter channel axis. Song matches are performed by calculating the Frobenius distance between feature matrices. JB also uses chromagrams, but these are used for performing HMM-based chord identification, with string alignment techniques being used to perform song matches on the chord transcription.

To address changes in key, the top performing algorithms perform circular shifts of their underlying representations to address possible transpositions. Therefore to calculate a possible song match, similarity scores are calculated multiple times for each transposition.

In contrast to the top performing algorithms, the worst performing algorithms across the two years are based predominantly on timbre features, which are highly effective for audio music similarity, genre identification, etc. However, for cover song identification, it is clear that

analysis of musical structure, and dealing with musical alterations to aspects such as key and tempo are necessary.

6. CONCLUSIONS AND FUTURE WORK

This paper presented an analysis of the evaluation of audio cover song identification systems. While the aim of identifying variations of musical pieces in audio collections is narrow in scope with regard to the overarching goals of MIR, it represents a necessary departure from a large portion of the MIR research done to date. In particular, we assert that cover song identification necessarily must explore "musical similarity" along structural dimensions, as opposed to those characterized merely by timbre. This is demonstrated by the poor performance of timbre-based audio similarity algorithms in identifying cover songs. However, we do not wish to imply that cover song identification is in some way superior to, or meant to serve as a replacement for related similarity and classification tasks (e.g. audio artist, audio genre, etc). Instead, it represents an interesting new direction of research because of its apparent need for analyzing underlying musical structure. The significant performance gains in a single year and the impressive performances of the top algorithms in 2007 suggest that some of the musical descriptors used by these algorithms are seemingly quite powerful. As discussed often in terms of the "glass ceiling" problem, it is our hope that such descriptors, in conjunction with all of the other research that has been carried out to date, can push the state of MIR research forward, and also allow musical similarity searches to be tailored along structural dimensions (e.g. "find me a song with a similar chord progression").

In the broader context of music similarity, an interesting future direction would be to test the ability of "cover song" systems to retrieve "relevant" songs (not necessarily cover versions) from an audio collection given a query. While we have seen algorithms intended to retrieve similar songs in the MIREX audio music similarity task performed poorly in cover song identification, it would be interesting to see if the reverse is true. That is, it could be beneficial to note whether these cover song systems, which rely more on matching tonal or chord sequences, would produce results that a human judge would deem "similar." This very topic is addressed by Ellis, Cotton and Mandel [4]. Although they found that traditional MFCC based approaches are superior to using only their beat synchronous chromagrams, the features used in cover song identification did perform significantly better than a random baseline.

7. ACKNOWLEDGEMENTS

MIREX has received considerable financial support from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF) under grants NSF IIS-0340597 and NSF IIS-0327371. Additionally, we would like to thank Professor David Dubin for his statistical advice.

8. REFERENCES

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

[2] Bera, A. K., Jarque, C. M., "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". *Economics Letters* 6 (3): 255–259. 1980.

[3] Di Nunzio, M. G., Ferro N., Mandl, T., and Peters, C. "CLEF 2007: Ad Hoc Track Overview", *In Nardi, A. and Peters, C., editors, Working Notes for the CLEF 2007 Workshop*, 2007

[4] Ellis, D. P. W., Cotton, C. V., Mandel, M. "Cross-Correlation of Beat-Synchronous Representations for Music Similarity", *Proc. ICASSP-08*, pp. 57-60, 2008.

[5] Gomez, E., "Tonal descriptions of music audio signals", *Ph.D Thesis*. Barcelona, 2006.

[6] Kim, Y. E., Williamson D. S. and Pilli S. "Towards quantifying the 'album effect' in artist identification," *Proc. ISMIR 2006*, pp. 393-394, 2006.

[7] Lidy, T., Rauber A., Pertusa A. and Iñesta, J. M. "Improving Genre Classification by Combination of Audio and Symbolic Descriptors Using a Transcription System", *Proc. ISMIR 2007*, 2007.

[8] MIREX Wiki. Available: http://www.music-ir.org/mirexwiki/.

[9] Pampalk, E., Flexer, A., Widmer, G. "Improvements of audio-based music similarity and genre classification," *Proc. ISMIR 2005*, pp. 260-263, 2005.

[10] Tague-Sutcliffe, J. and Blustein J. "The statistical analysis of the TREC-3 data", *Overview of the Third Text Retrieval Conference*, D. Harmon, Ed. (NIST, Gaithersburg, MD), pp. 385-398, 1995.

[11] Whitman B., Flake G. and Lawrence, S. "Artist detection in music with Minnowmatch", *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 559-568, 2001.

9. APPENDIX

CS	Christian Sailer and Karin Dressler
DE	Daniel P. W. Ellis
KL(1,2)	Kyogu Lee
KW(L,T)	Kris West (Likely), Kris West (Trans)
LR	Thomas Lidy and Andreas Rauber
TP	Tim Pohle
EC	Daniel P. W. Ellis, Courtenay V. Cotton
IM	IMIRSEL M2K
JB	Juan Bello
JEC	Jesper Højvang Jensen, Daniel P. W. Ellis, Mads G. Christensen, Søren Holdt
KP	Youngmoo E. Kim, Daniel Perelstein
SG	Joan Serrà, Emilia Gómez