# MIREX 2006
# Audio-Based Music Similarity and Retrieval Evaluation: Main Results

For a detailed analysis see: http://staff.aist.go.jp/elias.pampalk/papers/pam_mirex06.pdf

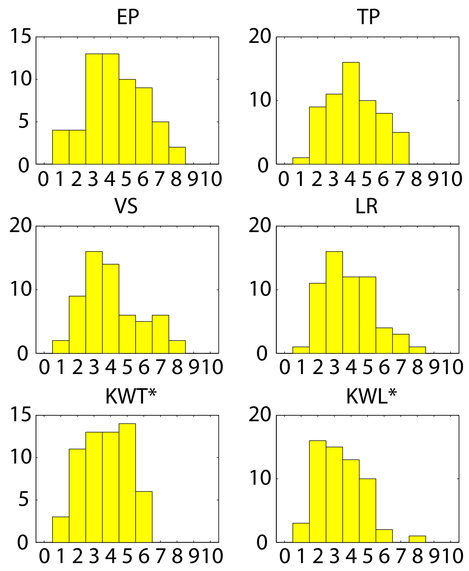
**Listening Test Conducted by IMIRSEL**
- o 60 query songs were randomly selected (query-by-example scenario).
- o For each submission 5 candidates were computed per query.
- o Each query/candiate pair was evaluated by 3 subjects.
- o Query/candidate similarities were rated on a scale from 0-10.
- o The subjects could re-adjust ratings as many times they wished.
- o Each query/candiate pair was evaluated in the context of all other candidates for the respective query.

**Participants**
- o Elias Pampalk (EP)
- o Tim Pohle (TP)
- o Victor Soares (VS)
- o Thomas Lidy & Andreas Rauber (LR)
- o Kris West (KWT* & KWL*)

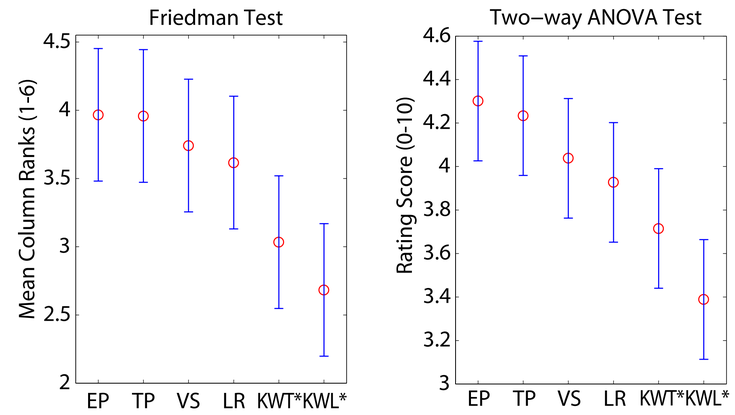
**Ranking Procedure**
- o Each query is treated as one observation.
- o For each query a score is defined as the mean of all 15 ratings (5 candidates x 3 subjects).
- o The Friedman test was chosen to test the significance of differences, because the distribution of the samples was unclear.

**Results**
- o The listening test did not show significant differences in quality.
- o In terms of computation time (for a 5000x5000 distance matrix):
  - LR is favorable with respect to the distance computation time, and
  - EP is slightly favorable with respect to overall time.
- o TP presented an interesting approach to overcome some "always similar" song problems for a large class of similarity measures (including EP's submission).

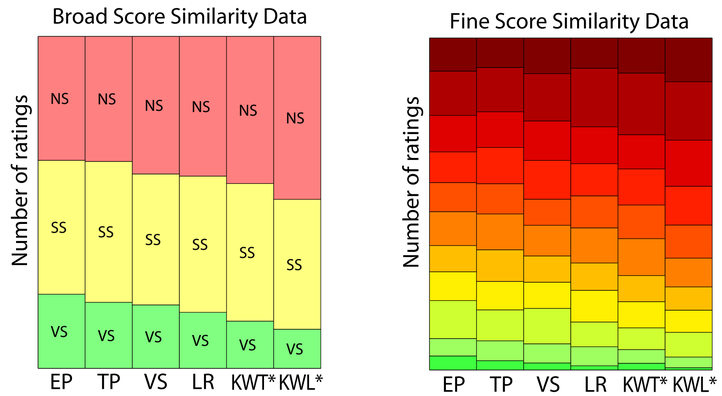
**Conclusions**
- o Overall all submission scored poorly (in average lower than 5 on the 0-10 scale).
- o The overall differences in quality are small (glass ceiling?).
- o The listening test set up was not sufficient to measure significant differences:
  - the number of queries was too low, and
  - the local context (30 candiates/query) was too big which might have been one of the reasons why the ratings were more inconsistent than in previous studies.



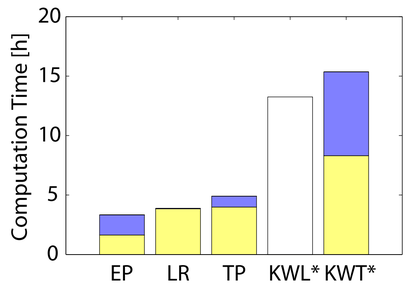

Histogram of the 60 scores per submission. If the distribution of the scores can be approximated with a normal distribution, then the two-way ANOVA test can be applied instead of the Friedman test.




Ranking of the submissions. Blue lines mark the significance boundaries for p=0.05 level. The ranking only changes marginally when using the two-way ANOVA test instead of the non-parametric Friedman test.




The left side shows the number of times per submission the songs were rated with each of the broad similarity scale categories. The categories on the broad scale are: not similar (NS), somewhat similar (SS), and very similar (VS). The right side uses data from the fine scale (0-10). The lowest block (green) corresponds to a rating of 10, the highest block (red) to 0. All ratings are rounded to the nearest whole number.




Computation times for some of the submissions. The lower part of each bar (yellow) is the feature extraction time (for 5000 songs). The upper part (blue) is the distance computation time for the complete distance matrix (which requires computing the distance of 12.5 million song pairs). For KWL* the times for the individual parts were not recorded. The VS submission was not able to compute the full distance matrix within a reasonable amount of time. The times were measured on a machine equipped with: Dual AMD Opteron 64, 1.6 GHz, 4 GB RAM, running Linux (CentOS).

Poster compiled by Elias Pampalk. Presented at MIREX, October 8-12, Victoria, Canada.