COMPARING
SOUNDS AND THE ORGANISATION OF THEIR ONSETS IN TIME

Tim Pohle[1], **Dominik Schnitzer**[1,2]
[1)]Dept. of Computational Perception
Johannes Kepler University, Linz, Austria
[2)]Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria

ABSTRACT

This abstract describes the algorithm we submitted to the MIREX 2009 Audio Similarity (AMS) Task. The algorithm has two main components. One component aims to describe the similarity of the "timbre" or "sound" that appears in the two tracks to compare, and the second component focuses on comparing the periodicities and frequencies of the onsets appearing in the two tracks. Both components are weighted equally (i.e., $1:1$). The algorithm is a variant of the algorithm described in [1], and ranked first in the MIREX 2009 AMS task.

1 ALGORITHM DESCRIPTION

This abstract contains a superficial description of the algorithm components. For more information, the reader is referred to [1]. The algorithm has two major components which are weighted equally (i.e., $1:1$), a *rhythm* component and a *"timbral"* component.

1.1 Rhythm Component

The rhythm component is based on a modification of the Fluctuation Patterns [2]. Calculation of the rhythm component includes the following steps:

- The audio excerpt is transformed into a *cent/sone* like representation. Sone values s are estimated from the amplitudes a by $s = 2^{log_{10} a}$ (cf. [3]).
- An onset estimation is performed, and the number of frequency bands is reduced.
- For each frequency band, periodicity estimation is done on segments of 2.63 sec length. Periodicities are scaled to assign each metrical level the same number of bins (assuming only meters of two).

The matrix resulting for each segment is transformed by applying a 2D cosine transform. Coefficients 0 and 1 are kept in the frequency dimension, and coefficients $0..17$ are kept in the periodicity dimension. These values are stacked to form a 36 dimensional vector for each segment. The rhythm feature data for a track is the mean and full covariance matrix of these vectors over all segments.

The rhythm component distance of two songs is estimated by calculating (cf. [4, 5])

$$D(\mathcal{N}_1, \mathcal{N}_2) = H(\mathcal{N}_3) - \frac{H(\mathcal{N}_1) + H(\mathcal{N}_2)}{2} \qquad (1)$$

where H denotes the entropy, and $\mathcal{N}_3$ results from merging $\mathcal{N}_1$ and $\mathcal{N}_2$. We use the square root of D. A way to merge two Gaussians into one is given in [6], setting the weights of $\mathcal{N}_1$ and $\mathcal{N}_2$ to 0.5 each it follows:

$$\begin{aligned} \mu_3 &= 0.5\mu_1 + 0.5\mu_2 \\ \Sigma_3 &= 0.5\Sigma_1 + 0.5\Sigma_2 + 0.5\mu_1\mu_1' + 0.5\mu_2\mu_2' - \mu_3\mu_3' \end{aligned}$$

The entropy H of a single Gaussian can be computed by (e.g., [5])

$$H(\mathcal{N}) = \frac{1}{2} \log \left((2\pi e)^d \, |\Sigma| \right) \qquad (2)$$

where d is the number of dimensions, and $|\Sigma|$ denotes the determinant of covariance matrix Σ.

1.2 "Timbre" Component

The "timbre" component consists of the well-known MFCCs [7] (coefficients $0..15$), Spectral Contrast Feature [8] using the "2N" method [9], and for each frame, two feature values estimating the amount of harmonic and percussive elements in the current audio frame (cf. [10]). Feature values are represented by a single Gaussian, which are also compared by calculating the square root of (1).

1.3 Distance Computation

Rhythm and "timbre" distances are calculated separately. Before they are combined, each of the two distance measures is normalized by mean removal and division by standard deviation (based on a track's distance to all other tracks in the music collection). Symmetry is re-created by subsequently summing up the distances in both directions for each pair of tracks (cf. [11]).

Team ID	*Team Members*
ANO	Anonymous
BF	Benjamin Fields
BSWH	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, Perfecto Herrera
CL	Chuan Cao, Ming Li
GT	George Tzanetakis
LR	Thomas Lidy, Andreas Rauber
ME	François Maillet, Douglas Eck
PS	Tim Pohle, Dominik Schnitzer
SH	Stephan Hübler

Table 1. Team IDs.

2 DISCUSSION OF RESULTS

Overall, 15 algorithms were submitted. Participating teams are listed in Table 1. For 100 query tracks, each algorithm was used to retrieve five tracks that are most similar to the query track according to the respective algorithm. Human graders assessed the actual similarity of the retrieved tracks to the repective query track. A detailed description of the experimental setup is given on *http://www.music-ir.org/mirex/2007/index.php/Audio_Music_Similarity_and_Retrieval*. The discussions and figures in this section are based on the results given on *http://www.music-ir.org/mirex/2009/index.php/Audio_Music_Similarity_and_Retrieval_Results*.

Our submission (denoted PS2) ranked first, while our re-submitted algorithm from 2007 (denoted PS1, which is a modification of the G1C algorithm [2])) ranked second. In Figure 1 results of the Friedman test are given. This test is based on the rank of an algorithm after sorting the algorithms according to the average fine score for a given query song. Thus, the theoretical optimum value is equal to the number of algorithms.

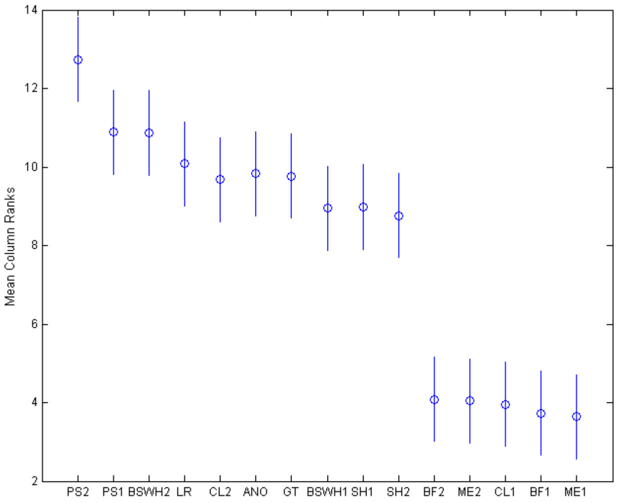

Figure 1. Friedman test based on human-assigned fine scores.

PS2 has an average rank of 12.7 (of the theoretical maximum value of 15.0), while the mean ranks of PS1 and BSWH2 are 10.89 and 10.87, respectively. No significant difference between these three algorithms is measured by the Friedman test. However, a significant difference between PS2 and the other 12 submitted algorithms is measured.

While the Friedman test measures the relative performance of algorithms, absolute values are given by the histogram of the human-assigned broad scores in Figure 2. Overall, more than half (56.4%) of the songs retrieved by the PS2 algorithm were rated "very similar". For PS1, this figure is 45.2%. Correspondingly, PS2 produces about one third (36.3%) less "outliers" (i.e., songs that were rated "not similar") than the next best algorithm in this respect (BSWH2), and 38.9% less outliers in comparison to PS1.

2.1 Comparison to 2007

Comparability of results obtained in the MIREX 2007 AMS task and the MIREX 2009 AMS task is supported in two ways. First, a similar experimental setting was used (same music collection, same number of queries, same number of candidate tracks per query). Second, the algorithm that ranked first in the 2007 AMS task was re-submitted unaltered (denoted PS and PS1 in 2007 and 2009, respectively). Table 2 lists the number of candidates retrieved by this algorithm that were rated very similar, somewhat similar, and not similar in the 2007 and 2009 AMS tasks.

	VS	SS	NS
2007	221	177	102
2009	226	179	95

Table 2. Comparison of results of PS07 algorithm in the AMS Tasks held in 2007 and 2009. Number of retrieved songs rated "Very Similar" (VS), "Somewhat Similar" (SS), and "Not Similar" (NS).

It can be seen that the absolute numbers are quite comparable, in spite of different query songs, different sets of graders, and different sets of participating algorithms. We see this as an indication that there is a good comparability of results between the AMS tasks held in 2007 and 2009.

2.2 Runtime

Runtimes of the algorithms are shown in Figure 3. It can be seen that the runtime of our submission was neither particularly slow nor fast in comparison to the other submissions.

2.3 Outlook

PS2 has two components. The first is the "timbre" component, that is intended to calculate the overall "sound similarity" of two tracks. The second is the "rhythm" component, that aims at determining rhythm similarity, i.e., aspects of how the onsets of sounds are organized in time. This may evoke the definition of music as "organized sound" given by Edgar Varese (e.g. [12]). Seen in this light, the question may arise in how far the used algorithm

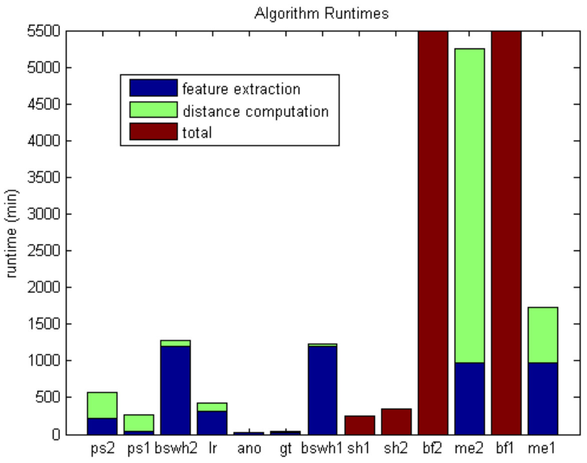

Figure 3. Algorithm runtimes. Same ordering of algorithms as in Figure 1. CL1 and CL2 were left out, as no runtimes are given. Values were cut off at 5500 min, for BF1 and BF2, a runtime of 10 days and 0 minutes is reported. When no separate runtimes for feature extraction and distance computation (blue and green, repsectively) are reported, the overall runtime is shown (red).

concept is suited to actually compare "music", and not just "timbre". Regardless of the answer to this question, the techniques used in our algorithm might also be of use in the context of other algorithm concepts. For example, one could think of using the Gaussian representation of rhythm in a similar way as [13] in a classifier-based similarity measure such as BSWH1.

3 ACKNOWLEDGMENTS

This work is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung under project number L511-N15.

4 REFERENCES

[1] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer, "On rhythm and general music similarity," in *Proceedings of the* 10^{th} *International Conference on Music Information Retrieval (ISMIR'09)*, 2009.

[2] E. Pampalk, *Computational Models of Music Similarity and their Application in Music Information Retrieval*, Docteral dissertation, Vienna University of Technology, Austria, March 2006.

[3] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics*, Springer Series in Information Sciences. Springer, third edition edition, 2007.

[4] Jianhua Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, pp. 145–151, 1991.

[5] Marco Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe Hanebeck, "On entropy approximation for gaussian mixture random vectors," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2008.

[6] Jinwen Ma and Qicai He, "A Dynamic Merge-or-Split Learning Algorithm on Gaussian Mixture for Automated Model Selection," in *Proceedings of 6th International Conference on Intelligent Data Engineering and Automated Learning - IDEAL*, July 6–8 2005, pp. 203–210.

[7] Beth Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the First International Symposium on Music Information Retrieval (ISMIR)*, Plymouth, Massachusetts, oct 2000.

[8] Dan-Ning Jiang Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

[9] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[10] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. International Conference on Music Information Retrieval (ISMIR'08)*, 2008.

[11] Tim Pohle and Dominik Schnitzer, "Striving for an Improved Audio Similarity Measure," in *4th Annual Music Information Retrieval Evaluation Exchange*, 2007.

[12] Bill Hammel, "An essay on patterns in musical composition transformations, mathematical groups, and the nature of musical substance," http://graham.main.nc.us/~bhammel/MUSIC/compose.html, 2000.

[13] Michael Mandel and Dan Ellis, "Song-level features and support vector machines for music classification.," in *Proc. ISMIR*, 2005, pp. 594–599.

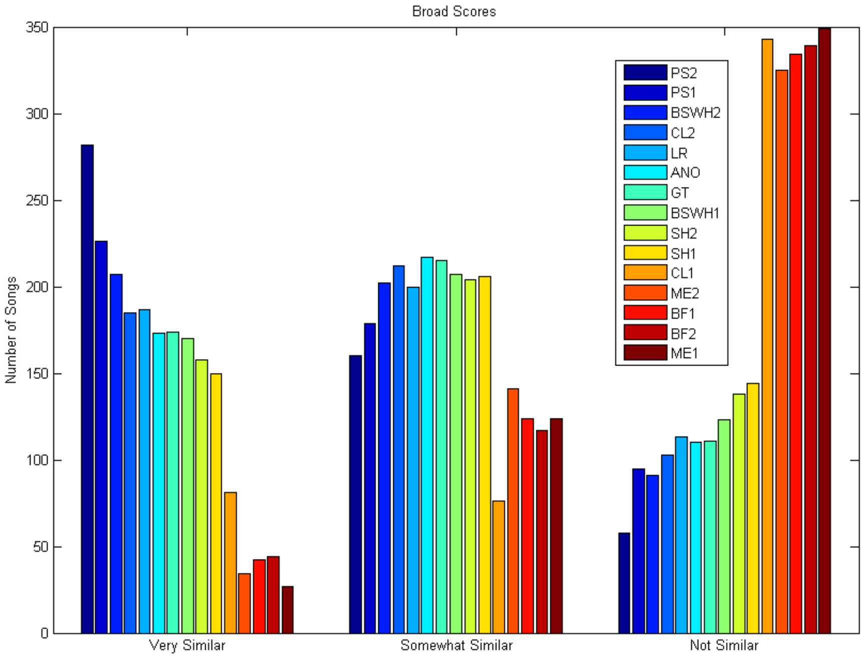

Figure 2. Histogram of human-assigned broad scores. Each submitted algorithm was used to retrieve 500 tracks. Our submission PS2 ranked first. Compared to the second-ranked algorithm PS1, 24.8% more tracks retrieved by PS2 were rated as "very similar", and 38.9% less tracks were rated "not similar".