SUBMISSION TO MIREX AMS TASK 2010

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

ABSTRACT

This preliminary abstract describes one of the algorithms we submitted to the MIREX 2010 Audio Similarity (AMS) Task. The algorithm is a modification of the algorithm we submitted to the MIREX 2009 AMS Task (abbreviated *PS09*). For comparison to the MIREX 2009 AMS Task, also PS09 is re-submitted, which ranked first in the 2009 AMS Task.

This abstract first briefly describes PS09, then the modifications in the present submission (*PSS10*) are given.

1 MIREX 2009 AMS TASK SUBMISSION (PS09)

This section describes our submission to last year's MIREX AMS Task (abbreviated *PS09*), which ranked first in 2009, and which we re-submitted for comparison purposes. The present submission described in this abstract is a modified version of this algorithm. The modifications are described in the next section.

This section contains a superficial description of the algorithm components of PS09, which is a variant of the algorithm described in [1]. 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, 12]).

2 MODIFICATIONS IN MIREX 2010 AMS TASK SUBMISSION (PSS10)

The optimisations for this year's submission had a focus on increasing precision (or genre classification accuracy) with respect to music genre labels, no focus was set on optimising rhythm similarity.

In this year's algorithm, an additional time-related component is used, denoted *vp*, which conceptually is somewhere between Onset Patterns (OPs, [1]) and Fluctuation Patterns. Instead of focussing on the onset parts as OPs, periodicities are estimated on the "full" signal without onset enhancement as in FPs. Features are represented as matrices of 12 frequency bands and 25 logarithmically scaled periodicity bands, without differently weighting different periodicities, and distances are compared by using minimum and maximum values (cf. [13], M_5):

$$D = 1 - \frac{\sum_{i=1}^{300} \min(f_i, g_i)}{\sum_{i=1}^{300} \max(f_i, g_i)} \quad (3)$$

where f and g are the two feature matrices of size $12 \times 25 = 300$ and i indicates the index of the matrix entry. Division by zero is avoided.

Distances of each component are adapted as described in Section 1.3, and combined with an empirically optimised weighting of 70 : 20 : 10 (timbre component : vp : onset coefficients), and the resulting distances are adapted again.

2.1 Some Genre Classification Results

Precision at	*1NN*	*3NN*	*5NN*
PS09	0.901	0.856	0.826
PSS10	0.841	0.794	0.756

Table 1. Average Precision on Ballroom collection.

Precision figures of the two submissions (PSS10 against PS09) on two music collections show that average precision increases on the ISMIR'04 genre classification contest training set [1] (729 tracks, no artist filter), while for the rhythm classes of the Ballroom collection [2] that had been used in the ISMIR'04 Rhythm Classification Contest [3] precision decreases.

Precision at	*1NN*	*3NN*	*5NN*
PS09	0.824	0.782	0.762
PSS10	0.851	0.807	0.786

Table 2. Average Precision on ISMIR'04 Genre Classification Contest collection, 30 sec excerpts.

[1] http://ismir2004.ismir.net/genre_contest/index.htm

[2] data from ballroomdancers.com

[3] http://mtg.upf.edu/ismir2004/contest/rhythmContest/

3 DISCUSSION OF RESULTS

(To be done.)

4 ACKNOWLEDGMENTS

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

5 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] T. Pohle, P. Knees, M. Schedl, and G. Widmer, "Automatically adapting the structure of audio similarity spaces," in *Proceedings of the* 1^{st} *Workshop on Learning the Semantics of Audio Signals (LSAS 2006),* 1^{st} *International Conference on Semantics and Digital Media Technology (SAMT 2006)*, 2006.

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

[13] Klaas Bosteels and Etienne E. Kerre, "Fuzzy Audio Similarity Measures Based on Spectrum Histograms and Fluctuation Patterns," in *Proc. 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*.