SPEEDING UP MUSIC SIMILARITY

Elias Pampalk
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, 1010 Vienna, Austria

ABSTRACT

This paper describes (1) the submission to the ISMIR'04 genre classification contest and (2) the submission to the MIREX'05 (Music Information Retrieval eXchange) audio-based genre classification and artist identification tasks. The main difference between the submissions is the reduction of computation time in the order of magnitudes.

This paper concludes with a discussion of the relationship between genre classification and artist identification, the relationship between similarity and classification, and references to related MIREX'05 submissions.

1 IMPLEMENTATION OVERVIEW

Features are extracted from 22kHz mono wav input (2 minutes from the center of each piece are used for further analysis). For the 2004 submission these features are cluster models of MFCC spectra. The 2005 submission additionally uses fluctuation patterns and two descriptors derived from them: Gravity and Focus.

For each piece in the test set the distance to all pieces in the training set is computed. A *nearest neighbor classifier* is used. There is no training other than storing the features of the training data. Each piece in the test set is assigned the genre label (or artist's name) of the piece closest to it.

1.1 M2K Specific

The functions are implemented in Matlab 7 and submitted with an M2K wrapper. The 2004 submission requires the Netlab Toolbox and the signal processing toolbox. The 2005 submission does not require any additional toolboxes. The same functions are used for the genre classification and artist identification tasks.

1.2 Computation Time

The CPU times given in Table 1 are measured on a 1.3GHz Intel Centrino laptop. The 2004 submission does not fulfill the MIREX'05 time constraints (72 hours per task). For example, it takes 10 days to compute the (symmetric) distance matrix on a collection with 3000 pieces. The 2005 submission completes this in less than 4 hours.

	Feature Extraction (for each song)	Distance Computation (for each pair of songs)
2004	60 seconds	500 milliseconds
2005	3 seconds	3 milliseconds

Table 1: Approximate CPU times on a Centrino 1.3GHz.

2 SIMILARITY MEASURES

This section describes the algorithms and parameters used for both submissions. In terms of classification accuracy the 2005 submission generally performs equally or better than the 2004 submission depending on the music collection. For example, on the Magnatune collection there are no significant differences, while on two of the collections (DB-S and DB-L) used in [1] the performance is slightly better.

2.1 Preprocessing

Both submissions use two minutes from the center of each piece (22kHz, mono) for analysis. Both first compute MFCCs using 19 coefficients (after ignoring the first). The only difference is that for the 2004 submission the FFT window size is 512 with 50% overlap (hop size 256) while in 2005 the size is 1024 with no overlap (hop size 1024).

The exact window size does not have a critical impact on classification accuracies. The reason why the hop size is not larger for the 2005 submission (e.g. twice as large) is that the Mel spectrum is used for the fluctuation pattern computations. This requires a spectrogram without large gaps.

2.2 Submission 2004

The 2004 submission (which won the genre classification contest) implements the spectral similarity presented by Aucouturier and Pachet [2, 3]. The implementation is available in the MA toolbox [4].

2.2.1 Feature Extraction (Frame Clustering)

For the MFFC spectras of each song a GMM is trained (using the Netlab toolbox) with 30 centers and a diagonal covariance matrix. The GMM is initialized using k-means.

2.2.2 Distance Computation (Cluster Model Similarity)

Aucouturier and Pachet suggest to use Monte Carlo sampling to compare two songs. To compute the similarity of pieces A and B a sample from each is drawn, S^A and S^B respectively. A sample size of 2000 is used in the 2004 submission. The log-likelihood $L(S|M)$ that a sample S was generated by the model M is computed for each

MIREX 2005, 2nd Annual Music Information Retrieval Evaluation eXchange, September 11 – 14, London

piece/sample combination. The distance is computed as

$$\begin{aligned} d_{AB} &= L(S^A|M^B) + L(S^B|M^A) - \\ &\quad L(S^A|M^A) - L(S^B|M^B). \end{aligned} \tag{1}$$

The reason for subtracting the self-similarity is to normalize the results.

2.3 Submission 2005

The similarity measure is a combination of information from fluctuation patterns [5] and spectral similarity. The details of this combination and evaluation experiments[1] can be found in [1]. In particular, the combination is the sum of 65% spectral similarity combined with 15% fluctuation patterns, 5% Focus, and 15% Gravity. Prior to the linear combination the distances are variance normalized based on the distance matrices computed on DB-L [1].

The differences between the 2005 submission and the approach presented in [1] (which uses the code of the 2004 submission) are:

A. For the spectral similarity a different approach is used which combines ideas from Logan and Salomon [6] with ideas from Aucouturier and Pachet [2]. This approach is described below.

B. The Mel spectrogram (before DCT) is used instead of the sonogram for the computation of the fluctuation patterns. This cuts preprocessing time in half and does not seem to have a negative impact on the results.

C. Fewer frequency bands are used for the fluctuation patterns: only 12 instead of 20 are used. In particular, the width of higher frequency bands is increased. This results in 720 instead of 1200-dimensional patterns. For Gravity and Focus the exact number of frequency bands does not play a critical role.

D. Performance wise the 2005 submission is magnitudes faster while the classification accuracy is reduced only slightly.

2.3.1 *Fast Spectral Similarity*

As suggested in [6] k-means is used to cluster the MFCC frames. In addition, two clusters are automatically merged if they are very similar. In particular, first k-means is used to find 30 clusters. If the distance between two of these is below a (manually) defined threshold they are merged and k-means is used to find 29 cluster. This is repeated until all clusters have a minimum distance to each other. (Empty clusters are deleted.)

The maximum number of clusters per song is 30 and the minimum is 1. The threshold is set so that most songs have 30 clusters and only very few have less than 20. In practice it does not occur that a song only has 1 cluster (unless it contains only silence). This optimization can be very useful since the distance computation time depends quadratically on the number of clusters.

Unlike the approach suggested in [2] we draw no random samples from the cluster models. Instead the cluster centers are used as sample (as suggested in [6]). However, instead of using the Earth Mover's distance (as suggested in [6]), the probability for each point of this sample is computed (as suggested in [2]) by interpreting the cluster model as GMM. Since such a sample does not reflect the probability distribution (due to the different priors) the log-likelihood of each sample is weighted according to its prior before summarization:

$$L(S^A|M^B) = \sum_{i=1}^{k_A} P_i^A \log\left(\sum_{j=1}^{k_B} P_j^B N(S_i^A|M_j^B)\right), \tag{2}$$

where k_A is the number of centers in model M^A. P_i^A is the prior probability of center i. $N(S_i^A|M_j^B)$ is the probability that sample S_i^A (i.e. the mean of center i) was generated by cluster j from model M^B (assuming a Gaussian distribution and diagonal covariance). To compute the distances Equation 1 is used.

The genre classification performance of this fast spectral similarity is not as good as the 2004 submission. However, the effects are reduced after the combination with the information from the fluctuation patterns.

3 DISCUSSION

The following two subsections discus the relationship between the MIREX'05 genre classification and artist identification tasks, and how this similarity based approach relates to the classification task. The third subsection points out relationships to other submissions based on the abstracts submitted to MIREX'05.

3.1 Genre Classification and Artist Identification

An algorithm that performs well on artist identification might not perform well on genre classification. In particular, this can be the case if the algorithm focuses on production effects or a specific instrument (or voice) which distinguishes the artist (or even a specific album). That is, if the algorithm focuses on characteristics which a human listener would not consider relevant for defining a genre.

Genre classification is often evaluated on music collections where all pieces from an artist have the same genre label. In addition, usually no artist filter is used for cross evaluation. An artist filter ensures that all pieces from an artist are either in the test set or the training set. An algorithm that can identify an artist would also perform well on genre classification if no artist filter is used.

The parameters used for this submission have been optimized using an artist filter. That is, they are optimized for genre classification and not artist identification [1].

[1] As the Magnatune collection has played an important role in these experiments, overfitting could be an issue.

3.2 Similarity and Classification

A music similarity measure can be used to generate playlists, give recommendations, or visualize collections. A simple way to evaluate similarity is through genre classification. The assumption is that pieces from the same genre are similar to each other. A classifier used in the evaluation of similarity should not modify the similarity measure itself (e.g. by changing the weights depending on the training data). A straightforward choice is to use a nearest neighbor classifier.

The goal of the work in [1] is a similarity measure which does not need to be adapted to each collection it is applied to. Also this submission does not optimize the weights based on the training data. However, it is possible to do so. For example in [1], a set of parameters was found that yielded 41% classification accuracy on the DB-S collection, while the overall best (average performance on four collections) set of parameters only yields 38%.

3.3 Related MIREX'05 Submissions

This submission is very similar to Beth Logan's submission. It would be interesting to investigate how the spectral similarity based on the Earth Mover's distance [6] compares to the approach suggested in this paper without using the additional information from the fluctuation patterns.

Thomas Lidy and Andreas Rauber also use fluctuation patterns (referred to as rhythm patterns) and compute statistics from these. However, they do not use Focus (mean of the fluctuation pattern after normalizing the pattern so that the maximum value equals 1) or Gravity (center of gravity on the modulation frequency axis minus the theoretical center of gravity).

Most MIREX'05 submissions use MFCCs in some way or the other. Several submissions explicitly combine features related to timbre (such as spectral similarity) with complementary features related to rhythm or tempo (such as fluctuation patterns).

4 Analysis of the Results

The details of the results are available online from the MIREX webpage.[2] The similarity measure performed very well in terms of quality and computation time. The results for artist identification are given in Tables 2 and 3. The results for genre classification are given in Tables 4 and 5.

The intended application of the similarity measure is not genre classification. To compare the other submissions on the level of a similarity measure would require to run them with a nearest neighbor classifier. The only directly comparable submission is the one by Beth Logan in the artist identification task which uses a nearest neighbor classifier.

[2] http://www.music-ir.org/evaluation/mirex-results

			Norm.	Time	CPU
	Participant	Raw	Raw	[hh:mm]	Type
1	Bergstra et al. (1)	77.26	79.64	24:00	B
2	Mandel & Ellis	76.60	76.62	03:05	A
3	Bergstra et al. (2)	74.45	74.51	–	–
4	**Pampalk**	66.36	66.48	01:11	B
5	Tzanetakis	55.45	55.59	00:44	B
6	West & Lamere	53.43	53.48	07:38	B
7	Logan	37.07	37.10	–	–

Table 2: Artist identification results for the Magnatune collection. For training 1158 tracks were used and 642 for testing. CPU Type A is a system with WinXP, Intel P4 3.0GHz, and 3GB RAM. CPU Type B is a system with CentOS, Dual AMD Opteron 64 1.6GHz, and 4GB RAM.

			Norm.	Time	CPU
	Participant	Raw	Raw	[hh:mm]	Type
1	Mandel & Ellis	68.30	67.96	02:51	A
2	Bergstra et al. (1)	59.88	60.90	24:00	B
3	Bergstra et al. (2)	58.96	58.96	–	–
4	**Pampalk**	56.20	56.03	01:12	B
5	West & Lamere	41.04	41.00	07:28	B
6	Tzanetakis	28.64	28.48	00:41	B
7	Logan	14.83	14.76	–	–

Table 3: Artist identification results for the Magnatune collection. For training 1158 tracks were used and 653 for testing.

Acknowledgements

This research was supported by the EU project SIMAC (FP6-507142). OFAI is supported by the Austrian ministries BMBWK and BMVIT.

References

[1] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR)*, 2005.

[2] J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proc International Conference on Music Information Retrieval (ISMIR)*, 2002.

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

[4] E. Pampalk. A Matlab toolbox to compute music similarity from audio. In *Proc of International Conference on Music Information Retrieval (ISMIR)*, 2004.

[5] E. Pampalk. Islands of music: analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology, 2001.

[6] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proc IEEE International Conference on Multimedia and Expo*, 2001.

			Norm.		Norm.	Time	CPU
	Participant	Hierarch.	Hierarch.	Raw	Raw	[hh:mm]	Type
1	Bergstra et al. (2)	77.75	73.04	75.10	69.49	–	–
2	Begstra et al. (1)	77.25	72.13	74.71	68.73	06:30	B
3	Mandel & Ellis	71.96	69.63	67.65	63.99	02:25	A
4	West	71.67	68.33	68.43	63.87	12:02	B
5	Lidy & Rauber (RP+SSD)	71.08	70.90	67.65	66.85	01:46	B
6	Lidy & Rauber (RP+SSD+RH)	70.88	70.52	67.25	66.27	01:46	B
7	Lidy & Rauber (SSD+RH)	70.78	69.31	67.65	65.54	01:46	B
8	Scaringella	70.47	72.30	66.14	67.12	06:19	A
9	**Pampalk**	69.90	70.91	66.47	66.26	00:55	B
10	Ahrendt	64.61	61.40	60.98	57.15	01:22	B
11	Burred	59.22	61.96	54.12	55.68	03:28	B
12	Tzanetakis	58.14	53.47	55.49	50.39	00:22	B
13	Soares	55.29	60.73	49.41	53.54	06:38	A

Table 4: Genre classification results for the Magnatune collection. 1005 tracks were used for training, 510 tracks for testing, and about seven genres needed to be classified.

			Norm.	Time	CPU
	Participant	Raw	Raw	[hh:mm]	Type
1	Bergstra et al. (2)	86.92	82.91	–	–
2	Begstra et al. (1)	86.29	82.50	06:30	B
3	Mandel & Ellis	85.65	76.91	02:11	A
4	**Pampalk**	80.38	78.74	00:52	B
5	Lidy & Rauber (SSD+RH)	79.75	75.45	01:26	B
6	West	78.90	74.67	05:09	B
7	Lidy & Rauber (RP+SSD)	78.48	77.62	01:26	B
8	Ahrendt	78.48	73.23	02:42	B
9	Lidy & Rauber (RP+SSD+RH)	78.27	76.84	01:26	B
10	Scaringella	75.74	77.67	06:50	A
11	Soares	66.67	67.28	03:59	A
12	Tzanetakis	63.29	50.19	00:22	B
13	Burred	47.68	49.89	02:34	B
14	Chen & Gao	22.93	17.96	–	–

Table 5: Genre classification results for the USPOP’02 collection. 940 tracks were used for training, 474 tracks for testing, and about four genres needed to be classified.