Improvements of Audio-Based Music Similarity and Genre Classification

Elias Pampalk[1], **Arthur Flexer**[1,2], **Gerhard Widmer**[1,3]

[1]Austrian Research Institute for Artificial Intelligence (OFAI)

[2] Institute of Medical Cybernetics and Artificial Intelligence, Center for Brain Research, Medical University of Vienna

[3] Department of Computational Perception, Johannes Kepler University Linz, Austria

ABSTRACT

We combine spectral similarity with complementary information from fluctuation patterns including two new descriptors derived from them. The performance is evaluated in a series of genre classification experiments on four music collections. The main findings are:

1. Although the improvements are substantial on two of the four collections our experiments confirm earlier findings [1] that we are approaching limits of simple audio statistics.
2. Evaluating similarity through genre classification is biased by the collection (and genre taxonomy) used.
3. In a cross validation no pieces from the same artist should be in both training and test set.

COMBINED SIMILARITY

We linearly combine four distance measures. Prior to weighting the distances they are normalized such that the standard deviation of the individual distance matrices equals 1.

1. Spectral Similarity is related to timbre. However, characteristics such as attack or decay are not modeled. We use the approach suggested by Aucouturier and Pachet (AP) in [1].
2. Fluctuation Patterns (FPs) describe loudness fluctuations in frequency bands [2]. They complement spectral similarity.
3. Focus (FP.F) describes the distribution of energy in the FP. FP.F is low if the energy is focused in small regions, and high if the energy is spread out over the whole FP.
4. Gravity (FP.G) describes the center of gravity of the FP on the modulation frequency axis. A slow piece usually has a low value. However, the perception of tempo is not modeled.

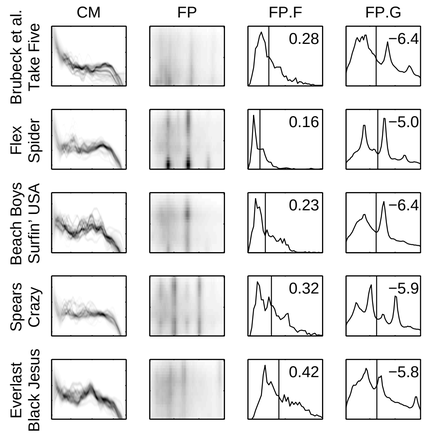

Figure 1: *Visualization of the features.*

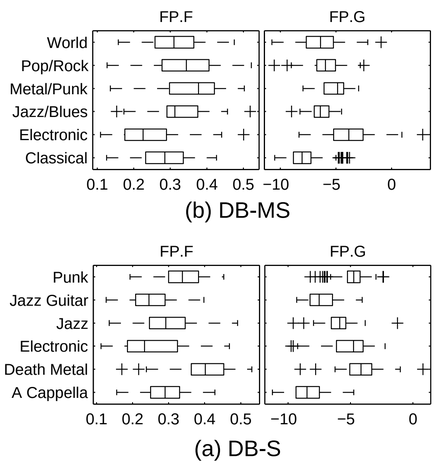

Figure 2: *Distribution of FP.F/FP.G per genre.*

DATA

			Tracks/Genre		
	Genres	Artists	Tracks	Min	Max
DB-S	16	63	100	4	8
DB-L	22	103	2522	45	259
DB-MS	6	128	729	26	320
DB-ML	10	147	3248	22	1277

Table 1: *Statistics of the four collections.*

GENRE CLASSIFICATION

We use a nearest neighbor classifier and leave-one-out cross evaluation. To justify this approach for the evaluation of similarity it is necessary to assume that pieces very similar to each other are in the same genre.

ARTIST FILTER

We apply a filter to ensure that all pieces from an artist are either in the training set or test set. The resulting performance is significantly worse. For example, on DB-MS (using AP) we obtain 79% accuracy without and only 64% with artist filter. On DB-L (using AP) we obtain 71% without and only 27% with filter.

COMBINING TWO

Combining AP with FP.F performs poorly, while combinations with FP.G perform surprisingly well (considering that FP.G is a simple scalar descriptor). The improvements on DB-ML and DB-MS are marginal. The smooth changes of the accuracy with respect to the mixing coefficient are an indicator that the approach is robust.

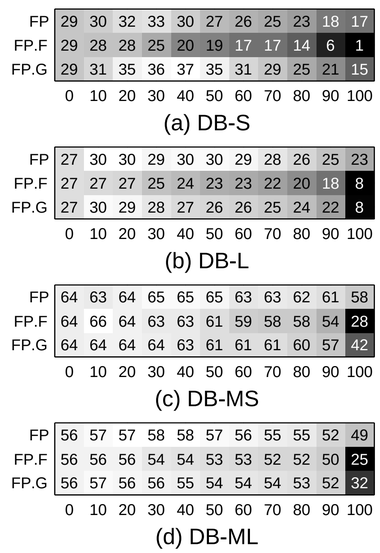

Figure 3: *Combining AP with one other.*

COMBINING ALL

There are a total of 270 possible combinations using a step size of 5 percent-points and limiting AP to a mixing coefficient between 100-50% and the other measures to 0-50%.

Analogously to the results of combining two, FP.F has the weakest performance here and the improvements for DB-MS and DB-ML are hardly significant.

There is a danger of overfitting. For example, for DB-S using as little AP as possible (highest values around 45-50%) and a lot of FP.G (highest values around 25-40%) gives the best results. On the other hand, for DB-MS the best classification accuracies are obtained using 90% AP and only 5% FP.G.

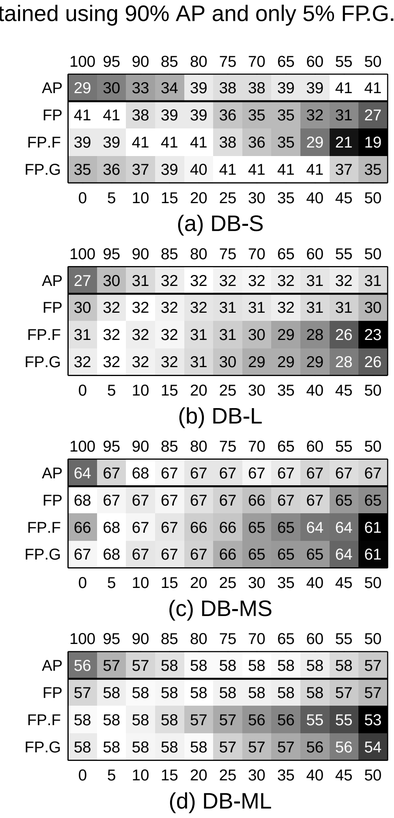

Figure 4: *Combining all measures. Classification accuracies are given in percent. Each table summarizes 270 experiments.*

CROSS COLLECTION

To avoid overfitting we compare the performance of combinations across collections. The worst combination (using 50% AP and 50% FP.F) is in average (across all collections) 15% below AP. The highest increase is in average 14% higher than AP.

	Weights				Accuracy			
Rank	AP	FP	F	G	S	L	MS	ML
1	65	15	5	15	38	32	67	58
2	65	10	10	15	38	31	67	57
3	70	10	5	15	38	31	67	58
248	100	0	0	0	29	27	64	56
270	50	0	50	0	19	23	61	53

Table 2: *Performance on all collections.*

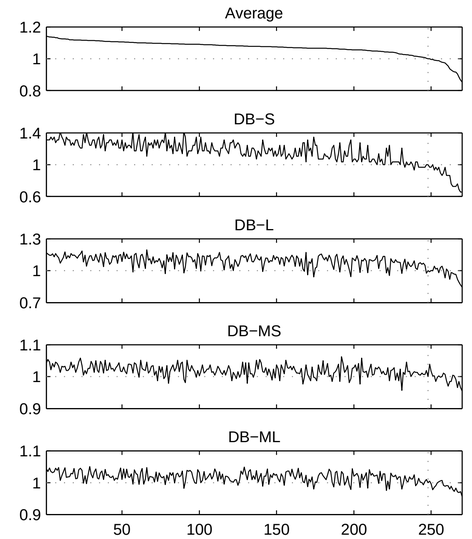

Figure 5: *Relative performance increase.*

Acknowledgments

This work was supported by the EU project FP6-IST-507142. OFAI is supported by the Austrian ministries BMBWK and BMVIT.

References

[1] 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.

[2] E. Pampalk. Islands of music: Analysis, organization, and visualization of music archives. MSc thesis, 2001.