Quantitative Comparisons into Content-Based Music Recognition with the Self Organising Map

Gavin Wood
University of York
York YO10 5DD
United Kingdom
gav@cs.york.ac.uk

Simon O'Keefe
University of York
York YO10 5DD
United Kingdom
sok@cs.york.ac.uk

Abstract

With so much modern music being so widely available both in electronic form and in more traditional physical formats, a great opportunity exists for the development of a general-purpose recognition and music classification system. We describe an ongoing investigation into the subject of musical recognition purely by the sonic content from a standard recording.

1 Previous Work

The self-organising map (SOM) is a neural method which may be used for dimensionality reduction of data. It can cope very well with high-dimensionality data, and is able to reduce a vector to a topologically-correct point on a (usually 2-dimensional) feature map. Because of the topology-aspect of the feature map, two input vectors that are similar will find their corresponding points in similar positions on the map. This aspect of the SOM is fundamentally important in the design of our recognition system.

In our previous work (Wood & O'Keefe, 2003), a simple system of utilising the spatial properties of the SOM was put forward as a benchmark recognition system. Each track is segmented (the exact number of segments being dependant upon length) and the segments are put through a particular audio feature extraction process. A SOM is trained upon a representative number of segments from varying tracks. This SOM is then used to translate each segment into a point on the feature map (this portion of the system is described in Rauber (Rauber & Frühwirth, 2001)). By combining [1] the segments' points from each track, an integer matrix may be formed. A technique termed as "bleeding" converts ('flattens') this integer matrix into a boolean matrix by making use of the topological nature of the feature map.

Tracks' similarity may then be measured by checking the similarity of the matrices; fast lookups may be done by correlation matrix memories. The other main design element was of extensibility, in as much as it would be a simple task to change the feature extraction techniques; simply feeding the SOM a different input vector.

The benchmark tests put forward examine how well the system is able to distinguish between music tracks that fall upon the same album and those that do not. These examinations take place on an archive of many contemporary music albums. The actual testing technique is simply to make the choice between two tracks, one of which appears on the same album as a third track. A random classifier would get the choice correct on average 50% of the time; the best system put forward made a correct choice around 80% of the time.

It must be noted that this is not meant to be a direct solution to a real-world problem; it is expected that, on average, tracks from the same album will be more perceptually similar that tracks from different albums. As such this is meant only to be an artificial, but objective, benchmark for recognition systems. It is anticipated that these objective experiments are a useful step towards the ultimate goal of a general purpose music recognition and classification system.

2 Techniques

Our previous work extracted only spectral features, we now test several signal-analysis approaches. We analyse the effect of using a rhythm spectrum as the feature extraction mechanism. This is calculated from the signal's self-similarity matrix in a technique put forward by Foote (Foote, 1999). The rhythm spectrum essentially gives the lag-correlation histogram of the audio. If there is a peak at X seconds, then the audio maintains a high similarity with itself at a period of X seconds. More peaks show more rhythm structures in the music; more distinguished peaks show more (sonically) pronounced repetition; peaks in the lower end of the histogram denote short-term percussive-based rhythmicity; in the higher end peaks may correspond to medium term rhythm structures such as verse-repetition or choruses.

Within this approach we vary the input data by conducting several psychoacoustic transformations, including Bark critical-band scaling (a technique to reduce the frequency spectrum to a small number of "critical" bands that we distinguish most fundamentally) and Sone loudness translation (a technique to scale the value of each frequency band to be both frequency-independent and have proportional loudness to that which we

[1] the combination takes place as a cumulative matrix addition

Table 1: Mean probability of correct decision, using optimal training parameters for the SOM

	Rhythm Spectrum (64D)	Rhythm Spectrum Feature Set (6D)
Basic acoustic spectra	0.70	0.67
Bark psychoacoustic spectra	0.66	0.63
Bark/Sone psychoacoustic spectra	0.68	0.70

perceive) (Cook, 1999). Evidence is found that psychoacoustic transformations that dramatically reduce the time to calculate structures such as the rhythm spectrum have little impact on recognition performance.

We also attempt to measure how well the system performs if the rhythm spectrum feature-vector used for training and retrieval with the SOM is compressed. To compress it, we extract specific statistical measurements from it, such as the distribution of the rhythm spectrum's intensities (regardless of actual frequency) and the distribution of the frequencies. It has been found that the beat spectrum can be compressed from a 64-band (dimension) vector to a six dimension feature vector with only a small loss of performance, with this loss taking place only in specific circumstances (as the results show). Indeed, using both Bark and Sone translations, as well as the rhythm spectrum feature extraction, gives as good a probability of successful choice as using basic acoustic spectra with the entire rhythm spectrum. This is useful since the former is far less computationally expensive both in analysis and training of the SOM, due to the downsizing of dimensionality by an order of magnitude.

3 Conclusion

The study shows that feature extraction techniques such as the rhythm spectrum can be used to extract aspects of a musical audio signal that facilitate recognition of sets of similar music. It also demonstrates a useful and compact representation of the rhythm spectrum which has evidence of performance as good as the 'raw' rhythm spectrum itself.

References

Cook, P. R. e. (1999). *Music, cognition and computerized sound: An introduction to psychoacoustics.* London, etc.,MIT.

Foote, J. (1999). Visualizing music and audio using selfsimilarity. In *Proceedings of ACM Multimedia, Nov '99, Orlando, Florida, USA. pp 77-80.*

Rauber, A., & Frühwirth, M. (2001). Automatically analyzing and organizing music archives. *Lecture Notes in Computer Science, 2163*, 402–??

Wood, G., & O'Keefe, S. (2003). A Quantitative Investigation into Content-Based Music Recognition with the Self-Organising Map. In *Submitted to 2003 IEEE International Workshop on Neural Networks for Signal Processing.* www-users.cs.york.ac.uk/˜gav/nnsp.ps.