Nearest-Neighbor Artist Identification

Beth Logan
Hewlett Packard Labs
One Cambridge Center
Cambridge MA USA
btlogan@ieee.org

ABSTRACT

We propose and investigate the performance of a simple artist identification system. Our approach learns a set of signatures for songs by known artists and then given an new song chooses the nearest neighbor from amongst these to identify the artist. The song signatures consist of clusters of MFCC frames hence artist similarity is based on timbral properties. On a 75 artist task, our algorithm has 17.0% accuracy. On the MIREX task our overall performance is 26.0%. This is substantially less than the best performance for this task of 72.5%. We thus conclude that while our approach is better than random, it does not compare well with other more complex systems. The extra modeling in such systems is thus justified.

Keywords: artist identification, music evaluations.

1 INTRODUCTION

The growth of online music in recent years has motivated the development of algorithms to automatically organize large audio repositories. Artist identification from audio properties is one such task that has attracted interest. While in many situations it may not be necessary to automatically identify the artist as it will be included as metadata with the song, this task is still interesting as it has a very well defined ground truth. It is thus easy to compare different feature extraction and modeling algorithms for this task. This could lead to new insights for other music retrieval tasks for which the ground truth is less easy to determine.

There has been considerable prior work on artist identification, often couched in terms of "singer identification" (e.g. Whitman et al., 2001; Kim and Whitman, 2002; Berenzweig et al., 2002; Maddage et al., 2004). Typical techniques focus on sung sections of the music and thus first identify the sections of music in which the artist is singing and then focus on modeling these sections. The most complex of these approaches (Maddage et al., 2004) models both the voiced and instrumental sections separately then makes a decision on the artist based on a linear combination of both these models. Although several techniques report accuracies near 90%, such results are obtained on small databases with only around 10 artists to be identified. When more artists are added the accuracy drops dramatically.

For example, the 2004 MIREX task included an artist classification track. This looked at artist classification of 30 and 40 artists. The best results for this task were 34% for the 30 artist task and 28% for the 40 artist task.

In this paper, we take a very simplistic approach to artist identification. Our technique is based in the idea that many popular artists are renowned for being derivative. For example, it is common to speak of a band that has a particular "sound". This is particularly true of artists with short-lived careers, although undoubtedly there are characteristics such as voice and typical instrumentation which are particular to any artist and are likely to identify their work.

Given a set of songs by known artists, we analyze these songs to extract a signature describing the main sounds present. We then identify the artist for an unlabeled song by comparing its signature to those of the known artists and assigning the artist label of the nearest neighbor.

Part of our motivation is to see how well such a simple approach performs on a large dataset against the more complex algorithms that will no doubt be part of this year's MIREX artist detection track. Since no previously published complex algorithms were demonstrated to perform well on a 100 artist task (either because they did not have good performance or because the authors never performed the experiment) we hope to learn whether the extra work extracting and modeling the sung sections is beneficial.

2 Methods

To analyze the audio, we use a previously published song analysis technique originally developed for song similarity (Logan and Salomon, 2001). This approach has been shown to give 63% accuracy on an artist similarity task (Berenzweig et al., 2003) and is comparable to many other music similarity algorithms.

Our approach is as follows. We first divide the audio for each song into a series of overlapping frames. We then convert each frame to a set of Mel frequency cepstral coefficients (MFCCs) (Rabiner and Juang, 1993). These features are a compact representation of the amplitude spectrum, hence we are primarily using timbre to distinguish between artists.

Given the set of MFCCs for each song, we then cluster these frames into groups which are similar. We use the K-means clustering algorithm (Duda et al., 2000) to achieve this, although any clustering algorithm could be used. The set of clusters, characterized by the mean, covariance and weight of each cluster is then denoted the signature for the song. The process of obtaining the signature for each song is shown in Figure 1.

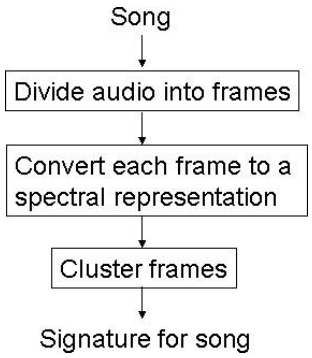

Figure 1: Process of obtaining the signature for each song

We compute the signature for each song in the training and test sets. We then use the Earth Mover's distance (EMD) (Rubner et al., 1998) to compute the distance between the signatures for each song in the test set and each song in the training set. We need to use the Earth Mover's distance as there is no closed form solution for computing the distance between two signatures.

The EMD calculates the minimum amount of 'work' required to transform one signature into the other. Let $P = \{(\mu_{p_1}, \Sigma_{p_1}, w_{p_1}), \ldots, (\mu_{p_m}, \Sigma_{p_m}, w_{p_m})\}$ be the first signature with m clusters where μ_{p_i} and Σ_{p_i} are the mean and covariance respectively of cluster p_i and w_{p_i} is the weight of that cluster. Similarly, let $Q = \{(\mu_{q_1}, \Sigma_{q_1}, w_{q_1}), \ldots, (\mu_{q_n}, \Sigma_{q_n}, w_{q_n})\}$ be the second signature. Let $d_{p_i q_j}$ be the distance between clusters p_i and q_j. In our work, we compute this using a symmetric form of the Kullback Leibler (KL) distance.

Define $f_{p_i q_j}$ as the 'flow' between p_i and q_j. This flow reflects the cost of moving probability mass (analogous to 'piles of earth') from one cluster to the other. We solve for all $f_{p_i q_j}$ that minimize the overall cost W defined by

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j} \quad (1)$$

subject to a series of constraints. That is, we seek the cheapest way to transform signature P to signature Q. This problem can be formulated as a linear programming task for which efficient solutions exist. Given all $f_{p_i q_j}$, the EMD is then calculated as

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{p_i q_j}}. \quad (2)$$

We use the EMD to compute the distance between each test song and all artists in the training set. The artist of the nearest neighbor in the training set according to the EMD between their signatures is the hypothesized artist for the test song.

3 Experiments

In this section we describe the results of experiments on our own in-house data and on the MIREX task.

3.1 Conversion to Song Signatures

For all experiments we convert each song in the training and test set to signatures as described in Section 2 above. We first convert each song to a series of MFCC features. We achieve this by first downsampling the audio to 16kHz mono and dividing this signal into frames of 25.6ms overlapped by 10ms. We then convert each frame to 40 Mel-spectral coefficients and take the logarithm and the discrete cosine transform to obtain 40 cepstral coefficients. Of these, only coefficients 1 to 19 are used with the 0th and higher coefficients being discarded.

Given the set of MFCCs for each song, we then construct a signature by clustering the vectors using K-means clustering. We use 16 clusters per song. These clusters are the signature for each song.

3.2 In-house Database

In order to gauge the efficacy of our algorithm and to test our implementation before submission to MIREX we first conducted experiments on an in-house dataset. The training set for this data contains 10 songs for each of 75 artists to give a total of 750 songs. The artists were chosen randomly from a much larger database containing many styles of music, although predominantly of the *Rock* style. We excluded music from the *Classical* style from this experiment. The test set contains one song for each artist in the training set. The training and test sets are disjoint although we did not attempt to ensure that the training and test sets did not contain songs from the same album (thus the "album effect" could mean the results on this database are an overestimate of performance).

3.3 Results

Table 1 shows the results of our experiments. The first line shows results for our in-house experiments. We see that our algorithm has an accuracy of 17.0%. While this is much worse than human performance, it is significantly better than random and a reasonable result for such a simplistic algorithm.

The second line of Table 1 shows the overall results for the MIREX artist identification task. This task identifies artists from the *Magnatune* and *USPop* databases. Each database contains 1158 training files. There are 642 and 653 test files for *Magnatune* and *USPop* respectively.

From Table 1 we see that on the MIREX task, our system has overall accuracy of 26.0%. On the third line of Table 1 we report results for the best entry to MIREX for artist identification (Mandel and Ellis). This system has an overall accuracy of 72.5%. It thus appears that there is merit in constructing models that analyze the various components of songs rather than just building a general simplistic model.

Table 1: Results on In-house and MIREX databases

Database	Accuracy (%)
In-house	17.0
MIREX	26.0
Best MIREX	72.5

3.4 Timing

We timed our algorithm on an Intel Xeon 2.4GHz CPU with 2.6G of RAM running V2.4.18 of GNU/Linux. Converting a 5 minute audio file at 44.1kHz to a song signature takes around 7.5s. Therefore, converting 1000 files on a single processor should take around 2 hours. The Earth Mover's distance scales linearly with the number of artists in the training database. For a 750 file database the computation takes about 1.5s. Therefore, to assign an artist to 100 songs given a training database of 750 artists would take several minutes. It should be noted that we made no attempt to optimize our algorithm for speed since it met the criteria for entry to MIREX.

4 Conclusions

We have presented a very simple artist identification system. Our system first learns a model for each training song for each known artist. Then for an unlabeled song, we choose the closest model to identify the artist. On an in-house 75 artist task, our algorithm had an accuracy of 17.0%. On the MIREX task, our algorithm had an overall accuracy of 26.0%. The best performing MIREX system had an overall accuracy of 72.5%. We thus conclude that while our simple approach is better than random, more complex approaches are justified and lead to significantly better performance. This conclusion would have been difficult to reach without the chance to evaluate algorithms on a common task.

ACKNOWLEDGEMENTS

We would like to express our extreme gratitude to the MIREX team for enabling this evaluation. Their work will undoubtably promote major progress in the field of MIR as researchers will finally be able to compare algorithms on non-trivial audio databases.

References

A. Berenzweig, D. P. W. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *AES 22nd International Conference on Virtual, Sythetic and Entertainment Audio*, 2002.

A. Berenzweig, B. Logan, D.P.W. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proceedings International Conference on Music Information Retrieval (ISMIR)*, pages 103–109, 2003.

R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, 2000.

Y. Kim and B. Whitman. Singer identification in popular music recordings using voice coded features. In *ISMIR*, 2002.

Beth Logan and Ariel Salomon. A music similarity function based on signal analysis. In *ICME 2001*, Tokyo, Japan, 2001.

N. C. Maddage, C. Xu, and Y. Wang. Singer identificaton based on vocal and instrumental models. In *17th International Conference on Pattern Recognition (ICPR)*, 2004.

L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice–Hall, 1993.

Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. Technical report, Stanford University, 1998.

B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with Minnowmatch. In *IEEE Workshop on Neural Networks for Signal Processing*, 2001.