TOWARDS EFFICIENT MUSIC GENRE CLASSIFICATION USING FASTMAP

Franz de Leon, Kirk Martinez
WAIS Group, Electronics and Computer Science
University of Southampton
Southampton, UK
{fadl1d09,km}@ecs.soton.ac.uk

ABSTRACT

Automatic genre classification aims to correctly categorize an unknown recording with a music genre. Recent studies use the Kullback-Leibler (KL) divergence to estimate music similarity then perform classification using *k*-nearest neighbours (*k*-NN). However, this approach is not practical for large databases. We propose an efficient genre classifier that addresses the scalability problem. It uses a combination of modified FastMap algorithm and KL divergence to return the nearest neighbours then use 1-NN for classification. Our experiments showed that high accuracies are obtained while performing classification in less than 1/20 second per track.

1. INTRODUCTION

Digital technology and the Internet have changed the music industry landscape. Millions of tracks are available through online channels such as Apple iTunes and Amazon MP3. Given the large music collections available, there is a need for new applications for browsing, organising, and discovering music for consumers. The research field of Music Information Retrieval (MIR) aims to address these challenges by using content-based techniques for performing tasks such as audio music similarity estimation and classification.

A particular aspect of music track classification is genre classification. The problem is to correctly categorize an unknown recording of a song with a music genre. Labels can be hierarchically organized in the collection of genres and subgenres. These labels are used to enhance the musical file with a semantic metadata or to organize a music collection. At present, genre classification is still biased towards Western music. Thus, genre labels are the ones commonly used in Western music stores.

There are several approaches to perform automatic genre classification. One method is to use Kullback-Leibler (KL) divergence to estimate timbre similarity then use *k*-nearest neighbours to perform classification [1][2][3]. This method has been effective as seen in the annual Music Information Retrieval Evaluation eXchange (MIREX) runs but it suffers from scalability problems. This is due to the KL divergence properties that limit its applicability to large scale databases: 1) it is computationally expensive, 2) the divergence is not a metric, and 3) it is vulnerable to issues associated with high dimensionality.

In this paper, we present a *k*-NN genre classifier that addresses the limitations of the KL divergence and the inherent scalability problem. We use an adaptation of the filter-and-refine indexing method that was used for fast music similarity search [4]. Given an unlabeled track, the *filter* step uses a modified FastMap [5] algorithm to quickly generate its nearest neighbours among the training set. Thus, the KL divergence does not have to be computed over the whole database reducing the total classification time. From the nearest neighbour results, the *refine* step is performed by applying the KL divergence as music similarity measure on Gaussian timbre models. The divergence values are rescaled to make them metric. The metric values form a distance vector that is normalized such that other features can be added to enhance the similarity measure. Finally, the genre of the nearest track is used to label the unknown track.

2. RELATED WORK

The study by Tzanetakis and Cook was among the first to introduce the problem of automatic music genre classification [6]. The classifiers used to evaluate these feature sets include single Gaussian models, Gaussian mixture models and a *k*-nearest neighbour classifier. They achieved classification accuracies that are comparable to the results from human musical genre classification.

The features used for genre classification are usually correlated with the ones used in music similarity estimation. Most studies use content descriptors related to timbre as the algorithms should be able to classify short excerpts of an audio recording. In a study by Gjerdingen and Perrott, it was found that humans can perform genre classification in as short as ¼ second [7]. It was argued that timbre encompasses all the spectral and rapid time-domain variability in the acoustic signal. Such information can be highly indicative of particular genres. Other features, such as melody or rhythm cannot be derived for such short audio clips. In contrast, audio fingerprinting or cover song identification works on longer samples that enable them to derive other features than timbre. For classification, the Gaussian mixture model has been classically used [8] but support vector machines (SVM) are increasingly becoming popular [1][9].

The timbral texture of a song can be modelled by deriving its Mel-frequency cepstral coefficient (MFCC) vectors. The vectors can be summarized as a single multivariate Gaussian with full covariance matrix. In this way, the closed form solution of the KL divergence is used to compute the similarity between two music models. To scale this approach to millions of tracks, a filter-and-refine method is proposed in [4] to speed up audio similarity queries that use the KL divergence as similarity measure. The method in [4] uses modified FastMap algorithm to map the Gaussian timbre models to *k*-dimensional vectors. The whole collection in the vector space is filtered to return a number of possible nearest neighbours. The result is then refined by computing the exact KL divergence. They reported that their system is able to process similarity queries on a 2.5 million songs database in less than a second.

3. FEATURE EXTRACTION

Feature extraction is the process of deriving a compact numerical representation to characterize a segment of audio. Our system extracts two features to model timbre, namely MFCC and ΔMFCCs.

3.1. Mel-Frequency Cepstral Coefficients

The MFCCs are the result of a cosine transform of the real logarithm of the short-term magnitude spectrum after it has been passed through a Mel-frequency scale filter bank. The Mel-frequency scale filters are intended to approximate the distribution of the ear's critical bandwidths with frequency, using filters placed roughly linearly at low frequencies and logarithmically at higher frequencies. The important aspects of the human auditory system which MFCCs model are: (1) the non-linear frequency resolution using the Mel frequency scale, (2) the non-linear perception of loudness using decibels, and to some extent (3) the perception of the spectral shape after using a Discrete Cosine Transform.

3.2. ΔMFCCs

In the field of speech recognition, MFCCs can be greatly enhanced by adding time derivatives to the basic static parameters [10]. In the same manner, these features may be used to enhance timbre model of a music track. The delta coefficients are computed using the following formula:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{1}$$

where d_t is a delta coefficient at time t, c is the cepstral coefficient, computed using a time window Θ.

3.3. Summarizing the Audio Features

The features derived from each audio track must be summarized efficiently and consider the similarity computation method that will be performed. In this paper, the single Gaussian model with full covariance approach is implemented to benefit from reduced computational complexity compared to the Gaussian Mixture Models. A single multivariate Gaussian probability density function is defined as:

$$\mathrm{N}(x \mid \mu, \Sigma) = \left(\frac{1}{2\pi}\right)^{1/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{2}$$

where x is the observation (n-dimensional feature vector), μ is the mean, Σ is an n x n covariance matrix. We also apply this to model the delta coefficients. Thus, the timbre for each audio file is represented by two single multivariate Gaussian models (for MFCCs and delta coefficients).

3.4. Mapping the Derived Features to Euclidean space

To accelerate the genre classification process, the MFCC and delta coefficient vectors are mapped to k-dimensional Euclidean space using a modified FastMap algorithm. The parameter k is arbitrary, with higher values leading to a more accurate mapping.

The modified FastMap algorithm starts by choosing two pivot objects. To determine the pivot objects, choose an arbitrary object and let it be the second pivot object O_b. Compute the distance to all other objects using the symmetrized KL (SKL) divergence. The KL divergence between two single Gaussians $p(x)=\mathrm{N}(x;\mu_p,\Sigma_p)$ and $q(x)=\mathrm{N}(x;\mu_q,\Sigma_q)$ is given by [11]:

$$\begin{aligned} 2KL(p \,\|\, q) &= \log\frac{|\Sigma_q|}{|\Sigma_p|} + Tr\left(\Sigma_q^{-1}\Sigma_p\right) + \\ &+ (\mu_p - \mu_q)^T \Sigma_q^{-1}(\mu_p - \mu_q) - d \end{aligned} \tag{3}$$

where $|\Sigma_n|$ and $Tr\,(\Sigma_n)$ denote the determinant and trace of matrix Σ_n, $n=\{p,q\}$, respectively. The SKL divergence are transformed into an exact metric with the function $T{:}SKL \rightarrow [\log(1+SKL)]^{1/2}$ [12]. Sort the divergences then select the median object as the first pivot object O_a. Similarly, update the second pivot object O_b by selecting the median object after computing all the distances from O_a.

For each object O_i, compute its projection $x_m(O_i)$ on the imaginary line (O_a,O_b) at m^{th} dimension.

$$x_m(O_i) = \frac{D(O_i,O_a)^2 + D(O_a,O_b)^2 - D(O_i,O_b)^2}{2D(O_a,O_b)} \quad m = 1 \ldots k \tag{4}$$

where $D(\ )$ is the transformed SKL divergence. Next, consider the projections of the objects on a hyperplane perpendicular to the line (O_a,O_b). The squared Euclidean distance $D'(\ )$ between the projections O_i' and O_j' can be computed as:

$$(D'(O_i', O_j'))^2 = (D(O_i,O_j))^2 - (x_i - x_j)^2 \quad i, j = 1 \ldots L \tag{5}$$

The algorithm is run recursively until the set dimension is reached. The output is an L x k projection matrix X where the i^{th} row is the image of the i^{th} object.

In summary, the feature extraction process derives the following for every song: 1) the means of the MFCCs and delta coefficient vectors, 2) their corresponding covariance and inverse covariance matrices, and 3) the projections of the MFCCs and delta coefficient vector means to a k-dimension Euclidean space.

4. GENRE CLASSIFICATION

The use of audio signals for similarity estimation is justified by an observation that sound signals of music belonging to the same genre share certain characteristics. These may include the instrumentation, rhythmic patterns and pitch distributions [13]. Accordingly, an unlabeled track can be tagged with its nearest neighbours from the training set.

The genre classification starts by filtering the training set to return a number of possible nearest neighbours to an untagged track. This is done by computing the squared Euclidean distances on the mapped vectors. This process is much faster, even with high values of k, than performing a linear scan over the training set using SKL. The result is refined by computing the transformed SKL on the candidate subset. Two distance values are produced (from MFCCs and ΔMFCCs) then combined to return the *true* nearest neighbours. Finally, the genre of the nearest track is used to label the untagged track (1-NN).

4.1. Setup

Two datasets were used in our experiments. The first dataset includes the training and testing sets for the ISMIR 2004 genre classification contest [14]. The second dataset was the GTZAN genre collection [6]. Ten-fold cross validation experiments were performed with the GTZAN dataset to avoid overfitting. The genre distributions for the two datasets are listed in Table 2.

For each track, a 30-second clip was selected from the middle. Then each audio signal is segmented into 23 ms non-overlapping windows from which MFCCs and ΔMFCCs were computed. Classification accuracies were derived from the confusion matrices. We investigated the effects of varying different parameters such as the number of Euclidean dimensions *k*, the filter size *R*, filter criteria, and distance weights.

4.2. Results

The initial experiments established the baseline system. This was done by performing genre classification using a full linear scan with SKL divergence over the training set. We tried several weights for the distances from the two features. The genre classification accuracies are tabulated in Table 1.

Table 1: *Genre classification accuracy of the baseline system using different weights on MFCC and ΔMFCC.*

Features	GTZAN	ISMIR2004	Mean
MFCC	0.7220	0.7305	0.7263
0.4MFCC+0.6ΔMFCC	0.7936	0.7318	**0.7627**
0.5MFCC+0.5ΔMFCC	0.7830	0.7307	0.7569
0.6MFCC+0.4ΔMFCC	0.7888	0.7348	0.7618

Results show that there is an improvement in the accuracies in combining the MFCCs with the delta coefficients. Based on the average performance, the best accuracies were obtained when the delta coefficients were given more weight than MFCC (0.4MFCC+0.6ΔMFCC). Hence, we use this as our baseline system. We also note that it takes around 0.59 seconds to perform genre classification per track. This information will be used to benchmark the proposed system using FastMap.

The proposed system uses the modified FastMap algorithm to filter the training set and return an approximate nearest neighbor subset. In our implementation, we use the MFCC and ΔMFCC distances as criteria. Figure 1 shows an example of the average distribution of the training songs after using Euclidean distance between an unclassified song and the training songs. The grey bars indicate the songs with the true genre of the query song. The grey histogram is skewed to the right that implies there is a high probability that the songs will be correctly classified. For a given filter size *R*, we return the nearest *R/2* objects based on the MFCC and ΔMFCC distances. The objects are combined to form the *R* nearest subset. Duplications are possible so the subset is further reduced by taking only the unique items.

We then refine the result and return the exact distance measures between the untagged track and the candidate subset. The genre of the closest track is used to label the unknown track. Figure 2 shows the performance of the proposed system using different *k* Euclidean dimensions and filter size. The filter size is expressed as a percentage of the number of items in the training set. The data presented are the average of the classification accuracies across all genres for a particular dataset.

Table 2: *Genre distribution of tracks for the ISMIR 2004 and GTZAN datasets.*

ISMIR 2004	
Training Set	
Songs	729
Genres	classical (320), electronic (115), jazz_blues (26), metal_punk (45), rock_pop (101), world (122)
Development Set	
Songs	729
Genres	classical (320), electronic (114), jazz_blues (26), metal_punk (45), rock_pop (102), world (122)
GTZAN	
Songs	1000
Genres	country (100), rock (100), reggae (100), blues (100), disco (100), hiphop (100), jazz (100), pop (100), classical (100), metal (100)

There is a more consistent pattern in the performance from the GTZAN dataset than the ISMIR 2004 dataset. This may be attributed to the uniform distribution of genre in the GTZAN dataset. In general, a larger filter size results in better accuracies. For the GTZAN dataset, a filter size of at least 7% produced accuracies in the vicinity, and sometimes even surpassing, that of the baseline system. This is not obvious for the ISMIR 2004 dataset but the average accuracies across *k* validate the observation.

It is expected that higher values for *k* used to map the timbre models lead to more accurate mapping in the Euclidean space. However, this did not translate to higher classification accuracies. On average, the best accuracies for the GTZAN and ISMIR 2004 datasets are obtained when *k*=60 and *k*=40, respectively. This shows the effect of the *curse of dimensionality* as higher dimensions lead to overfitting of the timbre models.

Suppose we choose a parameter combination of *k*=60 and filter size *R*=7% as the candidate system. The resulting accuracies for GTZAN and ISMIR 2004 datasets are 0.784 and 0.705, respectively. With this configuration, the performance of the candidate system is comparable to the baseline system. Moreover, it takes only 0.035 seconds to classify a track, or just 6% of the baseline system's classification time.

5. CONCLUSIONS

We have investigated an efficient method for automatic genre classification. The proposed system works on music tracks where timbre features are extracted, MFCC and ΔMFCC modelled as single multivariate Gaussian. To compute timbre similarity, the system uses a modified FastMap algorithm as filter and symmetrized KL divergence to refine the results. Performing genre classification using 1-NN showed that the accuracies obtained are comparable to the baseline system that uses SKL divergence only. Furthermore, it can handle genre classification in less than 1/20 of a second per track. Therefore, the proposed system has the potential to be used in very large databases.

In general, there is a trade-off between performance and computation complexity. More training data leads to a higher probability that a more similar track will be returned to a query. However, this translates to higher computational complexity. A compromise is to use high quality music tracks that can clearly delineate one genre from another. A further direction is to investigate other features and improve similarity estimation.

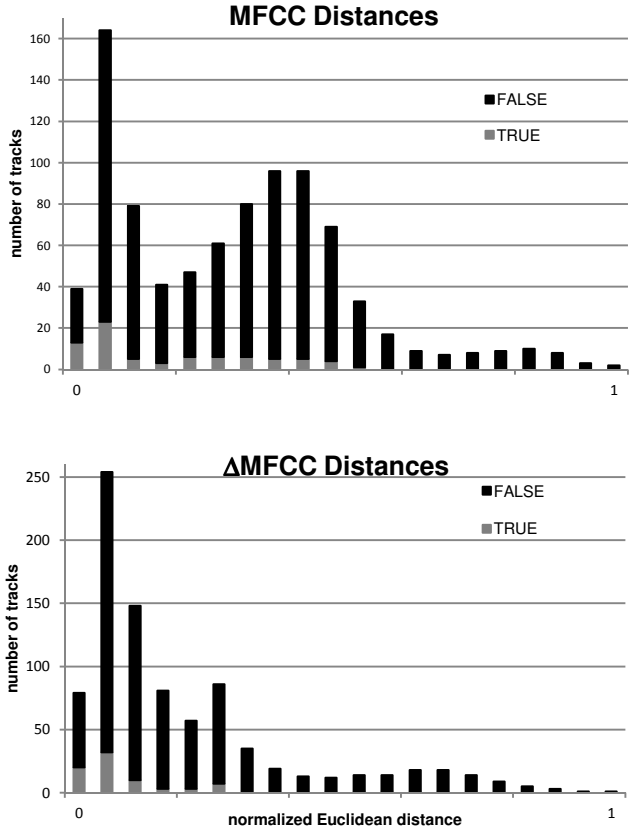

Figure 1: *Histogram of the average Euclidean distances between an untagged song and the training set (GTZAN, k=60). Grey bars indicate the songs true genre of the query song. The top panel corresponds to MFCC distances, bottom to ΔMFCC distances.*

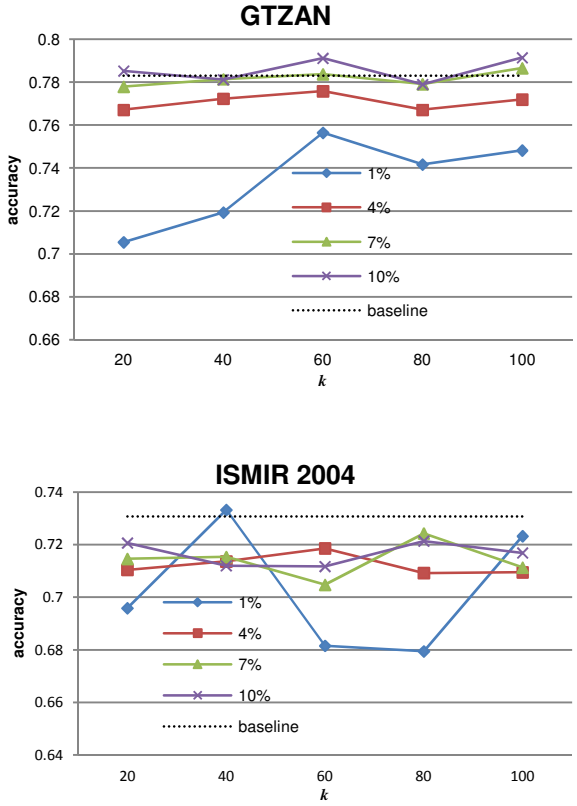

Figure 2: *Genre classification accuracy for GTZAN (top) and ISMIR 2004 (bottom) datasets as a function of k.* Each symbol corresponds to a filter size expressed as the percentage of the number of items in the training set.

6. ACKNOWLEDGMENTS

Mr. Franz de Leon is supported by the Engineering Research and Development for Technology Faculty Development Program of the University of the Philippines, and DOST.

7. REFERENCES

[1] M. I. Mandel and D. P. W. Ellis, "Song-level Features and Support Vector Machines for Music Classification," in *Submission to MIREX 2005*, 2005, pp. 594-599.

[2] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of Audio-Based Music Similarity and Genre Classification," in *Proceedings of ISMIR 2005 Sixth International Conference on Music Information Retrieval*, 2005.

[3] T. Pohle and D. Schnitzer, "Striving for an Improved Audio Similarity Measure," in *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, 2007, no. 1.

[4] D. Schnitzer, A. Flexer, and G. Widmer, "A Filter-and-Refine Method for Fast Similarity Search in Millions of Tracks," in *ISMIR 2009*, 2009, no. April, pp. 537-542.

[5] H. Faloutsos and K.-I. Lin, "FastMap: A fast algorithm for indexing, data-mining and visualization of tradi- tional and multimedia datasets," in *Proceedings of the 1995 ACMSIGMOD international conference on Man- agement of data*, 1995, pp. 163-174.

[6] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, Jul. 2002.

[7] R. O. Gjerdingen and D. Perrott, "Scanning the Dial: The Rapid Recognition of Music Genres," *Journal of New Music Research*, vol. 37, no. 2, pp. 93-100, Jun. 2008.

[8] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic Annotation and Retrieval of Music and Sound Effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467-476, Feb. 2008.

[9] A. Meng and J. Shawe-Taylor, "An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier," in *Proceedings of 6th International Society for Music Information Retrieval Conference*, 2005, pp. 604-609.

[10] J.-julien Aucouturier and F. Pachet, "Improving Timbre Similarity : How high's the sky?," *J. Negative Results Speech Audio Sci.*, vol. 1, 2004.

[11] W. Penny, "Kullback-Liebler Divergences of Normal, Gamma, Dirichlet and Wishart Densities," 2001.

[12] C. Charbuillet, G. Peeters, S. Barton, and V. Gouet-Brunet, "A fast algorithm for music search by similarity in large databases based on modified Symetrized Kullback Leibler Divergence," in *2010 International Workshop on Content Based Multimedia Indexing (CBMI)*, 2010, pp. 1-6.

[13] W. J. Dowling and D. L. Harwood, *Music Cognition*. New York: Academic Press, 1986, p. 258.

[14] "ISMIR 2004 Audio Description Contest-Genre/Artist ID Classification and Artist Similarity." [Online]. Available: http://ismir2004.ismir.net/genre_contest/index.htm.