A Music Genre Classifier Combining Timbre, Rhythm and Tempo Models

Franz de Leon[1,2], Kirk Martinez[1]
[1]Electronics and Computer Science, University of Southampton, United Kingdom
[2]Electrical and Electronics Engineering Institute, University of the Philippines
fadl1d09@ecs.soton.ac.uk

Abstract— **The changing music landscape demands new ways of searching, organizing and recommending music to consumers. Content-based music similarity estimation offers a robust solution using a set of audio features. In this paper, we describe the feature extractors to model timbre, rhythm and tempo. We discuss the corresponding feature similarity relations and how the distance measures are combined to quantify music similarity. The proposed system was submitted to 2011 Music Information Retrieval Evaluation eXchange (MIREX) Audio Music Similarity task for validation. Both objective and subjective tests show that the systems achieved an average genre classification of accuracy of 50% across ten genres. Furthermore, the genre classification confusion matrix revealed that the system works best on rap, hiphop and related types of music.**

Keywords- similarity estimation; music information retrieval

I. INTRODUCTION

Recent advances in technology and the Internet are changing the way music is distributed and accessed. The increased accessibility of music allows people to store thousands of music files in their portable media players, computing device, mobile phones, and other devices. The latest development called cloud storage enables online storage of digital music. This encouraged music distributors to adapt by establishing online channels such as Apple iTunes Store[1], Amazon MP3[2] and Napster[3]. In 2010, The International Federation of the Phonographic Industry reported that more than a quarter of the recorded music industry global revenues (29%) came from digital channels [1]. It is a market worth US$4.6 billion, an increase of 6% from 2009.

Given the large music collections available, there is a demand for new applications for searching, browsing, organizing, as well as recommending new music to users. A common method for searching for music is keyword search. People can enter the artist, track title or album name on a search engine and a list of nearest matching tracks are returned. However, this assumes that the user knows the specific keywords and that to some extent such tracks exist. It is also possible to perform a broader search by providing general keywords, e.g. genre, year, location, etc.

[1] http://www.apple.com/itunes/
[2] http://www.amazon.co.uk/MP3
[3] http://www.napster.com

There are other challenges that present new opportunities for research using large music collection for searching and retrieval of music related data. This research field is collectively called Music Information Retrieval (MIR). MIR is a multidisciplinary field that includes acoustics, psychoacoustics, signal processing, computer science, musicology, library science, informatics, and machine learning, etc. [2]. Its main goal is to provide a level of access to the world's vast music collection on a level at par, or exceeding, that being afforded by text-based search engines. This paper focuses on content-based methods for music similarity estimation. This involves extracting useful information or features from audio signals and developing a computational model for music similarity estimation.

The paper is organized as follows. The following section presents the related work. Section 3 details how the relevant audio features are extracted. Section 4 describes how the derived features are combined and used for audio music similarity estimation. Section 5 presents the validation results of the system using the 2011 Music Information Retrieval Evaluation eXchange (MIREX) data. Finally, the findings are summarized in Section 6.

II. RELATED WORK

Existing works on audio music similarity estimation focus on the estimation of similarity between one, or a combination of facets of music such as timbral and rhythmic information. Logan and Salomon were one of the first to publish a music similarity function based on audio content analysis [3]. The similarity function has been successfully applied to playlist generation, artist identification and genre classification of music. The method is based on the comparison of a 'signature' for each track using the Earth Mover's Distance (EMD), a mathematical measure of the difference between two distributions. For each track, the Mel-frequency Cepstral Coefficients (MFCCs) are computed. The signature is then formed using K-means clustering on the spectral features. Aucouturier and Pachet provided the groundwork for estimating music similarity using timbre [4]. A Gaussian mixture model (GMM) is trained on MFCC vectors from each song and is compared by sampling the distributions in order to estimate the timbral similarity between two songs. They also introduced the 'Aha' factor to correct unexpected associations.

In 2005, Mandel and Ellis presented a new system that uses single Gaussian distribution model to approximate timbre [5]. The first 20 MFCCs are calculated for a given audio frame. The mean and covariance matrix are computed for the resulting MFCC vectors. Thus, a song is represented by a 20x20 covariance matrix and a 20-dimensional mean vector. Similar to other timbre music similarity models, the temporal aspects are ignored. The similarity between two songs is then computed by Kullback-Leibler (KL) divergence. A symmetric KL divergence (SKLD) is derived to make it suitable as a distance function.

Pampalk combined Mandel and Ellis's timbre model with information from fluctuation patterns (FP) [6]. The fluctuation pattern describes the modulation of the loudness amplitudes per frequency bands. To some extent it can describe periodic beats. The resulting fluctuation pattern is a matrix with rows corresponding to frequency bands and columns corresponding to modulation frequencies (in the range of 0 to 10 Hz). The FP patterns are then summarized by computing the median of all FP matrices. The distance between FPs is computed by interpreting the FP matrix as high-dimensional vector and computing the Euclidean distance. The spectral similarity model is a single Gaussian with full covariance matrix. The distance between two Gaussians is computed using SKLD. Thus, the distance between two tracks is a weighted combination of the FP, spectral distances and derived features.

In our system, audio signals are modeled as long-term accumulative distribution of frame-based spectral features. This is also known as the "bag-of-frames" (BOF) approach wherein audio data are treated as a global distribution of frame occurrences. The disadvantage however is that temporal information is lost. There are several approaches to summarize the features. Tzanetakis, et al. used a single Gaussian with a diagonal covariance matrix [7]. Subsequent studies showed that using a single multivariate Gaussian with full covariance matrix can achieve the same level of performance [5][8]. In this work, we adopt this approach to benefit from reduced computational complexity compared to GMM.

III. Feature Extraction

This section describes the processes involved in deriving the features from audio signals. The features extracted from audio files are approximations of timbre, rhythm and tempo. The feature extraction and distance computation algorithms are implemented in MATLAB®.

A. Audio Preprocessing

Our system requires that the audio signals are sampled, or resampled, at F_s = 22050 Hz. This reduces the amount of data to be processed without compromising the salient features. The audio signals are then normalized such that they have maximum amplitude of one and have average value of zero. This removes DC component from the Fourier transform and also ensures that the amplitude of the transforms are of similar magnitude. Since the signals are almost of similar magnitudes after normalization, the *loudness* information is lost. However, this is acceptable since loudness is not as effective as other features for similarity functions [6].

B. Timbre

The timbre component is represented by the MFCCs [9]. The normalized audio signals signal is divided into frames with a window size and hop size of 512 samples (~23 msec.). The first 20 MFCC coefficients are derived but the *zeroth* coefficient is discarded. The resulting MFCC vectors are modeled as a single Gaussian distribution represented by its mean μ and covariance matrix Σ.

To enhance the timbre model, a number of coefficients are also derived. There are a number of simple features that can be computed from the spectrogram [10]. In this work, only the spectral flux is used.

1. *Spectral Flux* – defined as the squared difference between the normalized magnitudes of successive unfiltered spectral distributions. It measures the amount of local spectral change over a frame.

$$F_t = \sum_{n=1}^{N} (N_t(n) - N_{t-1}(n))^2 \quad (1)$$

   where $N_t(n)$ and $N_{t-1}(n)$ are the normalized magnitude spectrum of the Fourier transform at the current frame *t*, and the previous frame, *t-1*, respectively.

2. *Spectral Flux for Delta Spectrum* – given the unfiltered magnitude spectrum *M*, the delta spectrum *M'* is derived by getting the difference between successive frames. The spectral flux is computed from *M'* to determine the rate of change of the spectral flux.

Since the spectral flux coefficients are computed on the same time frames that MFCC used, the author proposes to append these coefficients on the MFCC matrix before taking the mean and covariance. In this way, the dynamic information is also preserved. This increases the dimensions of the mean vector and covariance matrix; hence it increases the computational complexity. Optimization is performed to find a balance between the number of dimensions and the performance of the algorithm. This is done by determining the best combination of MFCC values and spectral flux features that gives the highest genre classification accuracy. For example, in each time frame the last 19 MFCC values are appended with the spectral flux and spectral flux delta values. This results in a timbre model represented by 21x21 covariance matrix and a 21-dimensional mean vector.

C. Rhythm

The rhythm component is represented by the fluctuation patterns. Modulated sounds at low level modulation frequencies up to a modulation frequency of 20 Hz produce the hearing sensation of *fluctuation strength* [11]. The relationship between the fluctuation strength *F* and modulation frequency f_{mod} with masking depth ΔLs given by:

$$F \approx \frac{\Delta L}{\left(\frac{f_{\mathrm{mod}}}{4Hz}\right) + \left(\frac{4Hz}{f_{\mathrm{mod}}}\right)} \quad (2)$$

The sensation of fluctuation strength is most intense around 4Hz [11]. This is the basis of fluctuation pattern. The following steps describe the derivation of FPs:

1. Cut the spectrogram into short segments with window size of 3 secs. and hop size of 1.5 secs. The longer segment size is used, as compared to the size used for MFCC, to better capture the rhythm.
2. Map the 36 Mel-frequency bands into 12 such that the lower frequency bands are more preserved while the higher frequency bands are grouped.
3. For each segment and each frequency band, use FFT to compute the amplitude modulation frequencies of the loudness in the range of 0-10 Hz
4. Apply the weighting function using the model of perceived fluctuation strength (2).
5. Apply filters to reduce the influence of low frequencies and highlight the modulation frequency around 4 Hz.

The resulting FP is a matrix whose rows correspond to frequency bands and columns correspond to modulation frequencies (0 to 10 Hz). To summarize all FP patterns representing the segments of a music piece, the median of all FP's is computed. A single FP matrix represents a music file. The distance between pieces is computed by reshaping the FP matrix into a high-dimensional vector then solving for the Euclidean distance.

D. Tempo

Tempo is a fundamental property in music, particularly in western music. The tempo estimator used in this work is adapted from the work of [12]. It is simple and computationally efficient that can be easily integrated in our feature extraction system.

Global tempo estimation is done in two stages: onset envelope detection and tempo derivation. Given the Mel power spectrum, the first-order difference along time is calculated in each band. The negative values are set to zero, i.e. half-wave rectified, then the remaining positive differences are summed across all frequency bands. The resulting signal is then passed through a high-pass filter with a cut-off around 0.4 Hz to make it locally zero mean. The output is a one-dimensional onset strength envelope as a function of time that responds to proportional changes in energy summed across the 36 Mel frequency bands.

The second phase for the tempo estimator is to calculate the global tempo from the onset strength envelope. Given the onset strength envelope, autocorrelation will reveal any regular, periodic structure. For delays that align many peaks, a large correlation is observed. Human tempo perception is known to have a bias towards 120 bpm [13]. Hence, a perceptual weighting window is applied to the raw autocorrelation to de-emphasize periodicity peaks far from the bias. Finally, the scaled peaks are interpreted as indicative of the likelihood of a human choosing that period as the underlying tempo. The tempo period strength (TPS) is described by this equation.

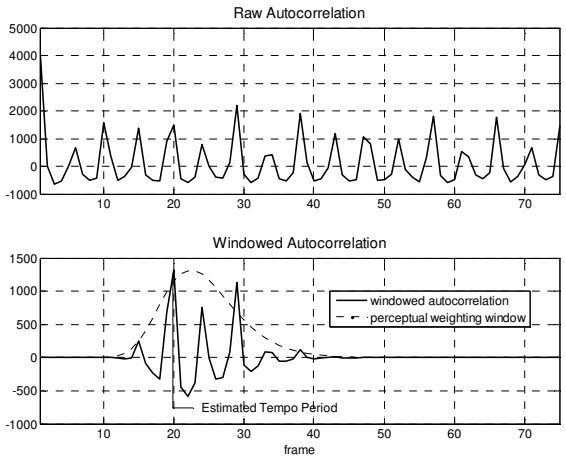

Figure 1. Tempo calculation. top: Raw autocorrelation of onset strength envelope; bottom: autocorrelation with perceptual weighting window applied and estimated tempo marked.

$$TPS(\tau) = W(\tau)\sum_{t} O(t)O(t-\tau) \tag{3}$$

where $O(t)$ is the onset strength envelope and $W(\tau)$ is a Gaussian weighting function on a log-time axis:

$$W(\tau) = \exp\left\{-\frac{1}{2}\left(\frac{\log_2 \tau/\tau_0}{\sigma_\tau}\right)^2\right\} \tag{4}$$

where τ_0 is the center of the tempo period bias, and σ_τ controls the width of the weighting curve. . The optimum τ_0 is 0.5 sec (corresponding to 120 bpm) and σ_τ of 1.4 octaves. The primary tempo is the delay τ for which the $TPS(\tau)$ is largest, see Fig. 1.

IV. Music Similarity Estimation

The timbre, rhythm and tempo distances between tracks are computed separately. A direct approach to combine timbral similarity with other features is to compute a weighted sum of the individual distances, see Fig. 2. Each distance component is normalized by removing the mean and dividing by the standard deviation of all the distances. The system is then optimized by determining the weights for each distance component that achieved the highest genre accuracy. It was determined the optimum weights for the timbre, rhythm, and tempo are w_{timbre}=0.95, w_{rhythm}=0.04, w_{tempo}=0.01, respectively. Symmetry is obtained by summing up the distances in both directions.

Using a single Gaussian with full covariance matrix to model timbre, the similarity can be computed using the Kullback-Leibler (KL) 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 [14]:

$$2KL(p\,\|\,q) = \log\frac{|\Sigma_q|}{|\Sigma_p|} + Tr\left(\Sigma_q^{-1}\Sigma_p\right) + (\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q) - d \tag{5}$$

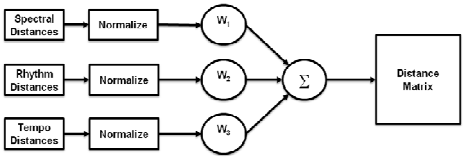

Figure 2. Block diagram of the proposed system for estimate music similarity.

where |Σ| denotes the determinant of the matrix Σ, *Tr( )* denotes the trace of the matrix. The resulting divergence is not symmetric. A symmetric version can be derived from (5). The symmetrised Kullback-Leibler divergence (SKLD) between two single Gaussian distributions $x_1 \sim N(\mu_1,\Sigma_1)$ and $x_2 \sim N(\mu_2,\Sigma_2)$ is defined as:

$$SKLD(x_1,x_2) = \frac{1}{2}KL(x_1,x_2) + \frac{1}{2}KL(x_2,x_1) \qquad (6)$$

The values obtained from SKLD have a large range, thus the *log* function is applied to (6). The Euclidean distance is used to compute the distance between rhythm patterns while the absolute distance is used for the tempo estimates.

V. EXPERIMENTS

Performance evaluation is a complex task that is often complicated with copyright issues. Researchers tend to evaluate their systems using their own database. This makes it hard to compare the systems on a level field. To solve this, the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) in the Graduate school of Library Information science at the University of Illinois conducts annual evaluations of many Music Information Retrieval (MIR) algorithms. Different MIR tasks work on a particular database and performance metrics. The annual evaluations are known as the Music Information Retrieval Evaluation eXchange (MIREX).

A. Setup

The author submitted the proposed system to MIREX 2011 audio music similarity task. The system was given 7000 songs chosen from IMIRSEL's "uspop", "uscrap" and "american" "classical" and "sundry" collections. The system then returned a 7000x7000 distance matrix. 100 songs were randomly selected from the 10 genre groups (10 per genre) as queries and the first 5 most highly ranked songs out of the 7000 were extracted for each query (after filtering out the query itself, returned results from the same artist were also omitted). Then, for each query, the returned results (candidates) from all participants were grouped and were evaluated by human graders using the Evalutron 6000 grading system [15]. Each individual query/candidate set was evaluated by a single grader. For each query/candidate pair, graders provided two scores to quantify similarity: a *fine* score from 0 (dissimilar) to 100 (exact), and a categorical broad score with 3 categories: Not Similar (0), Somewhat Similar (1), Very Similar (2).

TABLE I. COMPARISON OF SYSTEM PERFORMANCE

Algorithm	Genre Classification Accuracy	Average Fine Score	Average Broad Score
Proposed System	50.63%	50.49	1.09
Best Performing System	59.67%	58.64	1.31

B. Results

The objective results[4], based on the percentage of genre neighborhood clustering accuracy from indexed metadata, are tabulated on the second column of Table I. The results are artist filtered, that means an artist can only appear once in the training set or the testing set but not both. This process removes any biases of the algorithm towards a particular artist. The objective results are highly correlated with the human evaluation grades presented third and fourth columns of Table I. Both objective and subjective results show that returning the 5 closest songs to a given query, 50% of the candidate songs belong to the same genre. For comparison, the best performing system [16] in MIREX 2011 returns 60% of the candidate songs from the same genre.

To better understand the performance of the algorithm, the artist filtered genre confusion matrix for the proposed system is tabulated in Table II. The columns represent the true genres while the rows denote the predicted genres based on the genre of the candidate songs. Looking on the main diagonal of the matrix, it is observed that there is inconsistency in the accuracy. The algorithm performs effectively on raphiphop, metal, blues and country. Interestingly, it performed poorly on classical music. This is unexpected as previous runs using local database always had the highest accuracy with classical genre.

The western genres used in MIREX are not completely delineated. Hence, it is also observed from Table II that certain genres are often confused with other, but somehow similar genres. For example, metal is confused with rockroll, blues is confused with jazz, and so on. By clustering related genres, the resulting confusion matrix is shown in Table III. The average genre classification accuracy increases to 84%. Hence, there is a need to improve the performance of the algorithms in terms of differentiating songs from a cluster. This can be addressed by deriving additional features and improving the distance computation method. It is also worth considering the limitation of the "bag-of-frames" approach. The state-of-the-art uses larger time windows to preserve salient temporal information.

Among the three features used in the system, the highest weight is applied to the timbre component. This emphasizes the importance of timbre among other facets of music. Most audio signal processing techniques work on short frames or segments. Timbre encompasses all the spectral and rapid time-domain variability in the acoustic signal. Such information can be highly indicative of audio similarity as similar music may have similar instrumentation or orchestration. Other features, such as rhythm or tempo serve to complement timbre, and may even capture the mood of music. In addition, a more suitable model for the complex human judgment should be developed, rather than a simple sum of feature distances.

[4] http://www.music-ir.org/mirex/wiki/2011:MIREX2011_Results

TABLE II. ARTIST FILTERED GENRE CLASSIFICATION CONFUSION MATRIX

predicted\true	Metal	Blues	Baroque	Country	Rockroll	Jazz	Raphiphop	Edance	Classical	Romantic
Metal	0.64914	0.00543	0.00486	0.02714	0.24029	0.026	0.01771	0.12914	0.00029	0
Blues	0.00286	0.60657	0.002	0.03857	0.01057	0.20314	0.00486	0.00457	0.00857	0.00343
Baroque	0.00457	0.00571	0.41086	0.01143	0.01257	0.01743	0.004	0.00857	0.16371	0.15314
Country	0.05114	0.08657	0.04714	0.60743	0.26629	0.12	0.04629	0.098	0.00571	0.00371
Rockroll	0.23971	0.02457	0.02571	0.17171	0.37743	0.05486	0.03686	0.11629	0.00114	0.00429
Jazz	0.00571	0.18057	0.00571	0.05171	0.01686	0.46057	0.00914	0.03743	0.00686	0.00314
Raphiphop	0.01743	0.02543	0.00229	0.03429	0.03629	0.02057	0.818	0.214	0.00057	0.00086
Edance	0.01686	0.00314	0.00029	0.01686	0.01686	0.01714	0.06257	0.35686	0.00029	0.00086
Classical	0.00343	0.03543	0.28629	0.02229	0.00829	0.05086	0.00029	0.01086	0.30143	0.35
Romantic	0.00914	0.02514	0.21486	0.01857	0.01457	0.02943	0.00029	0.02429	0.51143	0.48057

TABLE III. ARTIST FILTERED, CLUSTERED GENRE CLASSIFICATION CONFUSION MATRIX

predicted\true	Metal	Blues	Baroque	Country	Rockroll	Jazz	Raphiphop	Edance	Classical	Romantic
metal,country, rockroll	0.93999	0.11657	0.07771	0.80628	0.88401	0.20086	0.10086	0.34343	0.00714	0.008
blues,jazz	0.00857	0.78714	0.00771	0.09028	0.02743	0.66371	0.014	0.042	0.01543	0.00657
baroque,classical, romantic	0.01714	0.06628	0.91201	0.05229	0.03543	0.09772	0.00458	0.04372	0.97657	0.98371
raphiphop,edance	0.03429	0.02857	0.00258	0.05115	0.05315	0.03771	0.88057	0.57086	0.00086	0.00172

VI. CONCLUSION

This paper presented a system for performing content-based music similarity estimation. The proposed system used features extracted from audio files to model timbre, rhythm and tempo. For the submitted algorithms to MIREX 2011 AMS task, both objective and subjective tests show that the systems achieved a genre classification of accuracy of 50%. By clustering related genres, the system's accuracy increases to 84%. The system can be further improved by considering other audio features and distance measures. This work serves to complement other approaches that may be limited by time and resources, such as manual annotation of music tracks. It recognizes that music similarity is very much dependent on user cultural background or preferences, and audio classification is best done by music experts.

ACKNOWLEDGMENT

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.

REFERENCES

[1] IFPI, "IFPI Digital Music Report 2011," http://www.ifpi.org/content/section_resources/dmr2011.html.

[2] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247-255, 2008.

[3] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, 2001, pp. 745-748.

[4] J.-julien Aucouturier and F. Pachet, "Music Similarity Measures : What's the Use?," in *3rd International Conference on Music Information Retrieval*, 2002, pp. 157-163.

[5] 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.

[6] E. Pampalk, "Computational Models of Music Similarity and their Application in Music Information Retrieval," Vienna University of Technology, 2006.

[7] G. Tzanetakis, G. Essl, and P. Cook, "Automatic Musical Genre Classification Of Audio Signals," in *Proceedings of ISMIR 2001: The International Conference on Music Information Retrieval and Related Activities*, 2001.

[8] E. Pampalk, "Audio-Based Music Similarity and Retrieval : Combining a Spectral Similarity Model with Information Extracted from Fluctuation Patterns," in *Submission to MIREX 2006*, 2006.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[10] 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.

[11] H. Fastl, "Psychoacoustics and Sound Quality," *Communication Acoustics*, 2005.

[12] D. P. W. Ellis, "Beat Tracking by Dynamic Programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51-60, Mar. 2007.

[13] M. F. McKinney and D. Moelants, "Ambiguity in Tempo Perception: What Draws Listeners to Different Metrical Levels?," *Music Perception*, vol. 24, no. 2, pp. 155-166, Dec. 2006.

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

[15] A. A. Gruzd, J. S. Downie, M. C. Jones, and J. H. Lee, "Evalutron 6000," in *Proceedings of the 2007 conference on Digital libraries - JCDL '07*, 2007, vol. 1, no. 217, p. 507.

[16] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees, "Using Block-Level Features For Genre Classification , Tag Classification and Music Similarity Estimation," in *Submission to Audio Music Similarity and Retrieval Task of MIREX 2011*, 2011, vol. 2, no. 1.