MIREX 2011 - AMS TASK: MULTI DESCRIPTOR BASED ALGORITHM

Simone Sammartino, Lorenzo J. Tardón, Isabel Barbancho, Cristina de la Bandera
Dept. Ingeniería de Comunicaciones, E.T.S. Ingeniería de Telecomunicación,
Universidad de Málaga, Campus Universitario de Teatinos s/n, 29071, Málaga, Spain
ssammartino@ic.uma.es

ABSTRACT

A method for the estimation of music similarity based on a combination of timbre, rhythm and tonal content of the songs, is presented in this report. The rhythm is estimated by evaluating the frequency distribution of the tempo pattern of the songs, while the tonal content is estimated by the calculation of an averaged chroma profile of the song. The algorithm is proposed to be submitted to the Audio Music Similarity task of MIREX 2011, in occasion of the 12th ISMIR Conference.

1. INTRODUCTION

In MIR community, many different approaches for automatic music recommendation are based on the retrieval of content-based descriptors that are able to estimate the audio similarity and, somehow, simulate the performance of the human brain with regard to the evaluation of music similarity.

Many classes of descriptors have been proposed for Audio Music Similarity (AMS). Among them, special mention must be given to the ones related to timbre, rhythm and tonal content. Logan and Salomon [10] and Foote [7] proposed the first examples of application of the Mel Frequency Cepstral Coefficients for the evaluation of music similarity and recommendation. The works by Foote et al. [6] and Phole et al. [14] describe the concept of the rhythmic similarity, highly exploited in AMS. Xiao ans Zhou. [17] propose the perceptual-based approach for music similarity, developing the use of the chromagram or chroma histogram proposed by Ellis and Poliner [4].

In this report, the combination of timbre, rhythmic and tonal analysis for music recommendation is proposed. The audio excerpts are analyzed and the MFCCs, the rhythmic pattern and the averaged chroma profile are extracted to be compared.

2. TIMBRE DESCRIPTOR

As early described by Foote [7] and Logan [11], the Mel Frequency Cepstral Coefficients are one of the most widely recognized spectral descriptors for music modeling and they have also been successfully employed in speech recognition tasks.

Depending on the methodology employed for the calculation of the MFCCs, some form of compression of the information is necessary in order to extract a compact representation of the cepstral behavior of the whole signal, to be used as a comparison mean among the different songs. Many authors proposed different solutions to this end. Pampalk [12], Foote [7], Aucouturier and Pachet [1] and Logan and Salomon [10] employ different approaches based on Gaussian Mixture Models, tree structured quantization, Monte Carlo distance and k-means method clustering, respectively.

2.1 The standardized variogram

The term 'variogram' stands for a statistical function describing the structured spatial/temporal evolution of a random field [16]. It is widely employed in Geostatistics for the so called Exploratory Spatial Data Analysis (ESDA), with the aim to describe the spatial autocorrelation of environmental variables. The Variogram can be employed in a unidimensional field as well, to study the time variability of an audio signal [9].

The variogram is defined as the semi-variance of the increment $[z_\alpha - z_{\alpha+h}]$, where z_α and $z_{\alpha+h}$ are two random variables z separated by the distance h. So, under the assumption of stationarity and ergodicity of the random variable, the experimental variogram (or semi-variance) can be defined as follows:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N_h} [z_\alpha - z_{\alpha+h}]^2 \qquad (1)$$

where N_h is the number of possible pairs of samples of the random process separated by distance h.

The experimental variogram can be fitted by a theoretical function, among a series of specific 'authorized' models [16]. The theoretical variogram is strongly related to the auto-covariance function of the increment: $Cov(h) \equiv Cov(z_\alpha, z_{\alpha+h})$. In particular, we can express the variogram in terms of the covariance function:

$$\gamma(h) = Cov(0) - Cov(h) \quad (2)$$

Hence, the typical shape of a variogram function (the theoretical model) fulfills:

- Its value at zero is zero:
  $\gamma(h = 0) = Cov(0) - Cov(h = 0) = 0$.
- It is a monotonically increasing function, because the corresponding covariance of the samples in a pair decreases with the distance.
- It tends asymptotically to the global variance of the random variable (its own autocovariance):
  $\gamma(h \gg 0) = Cov(0) - Cov(h \gg 0) \cong Cov(0)$.

In this work, the variogram is employed as a clustering tool for the MFCCs of an audio signal. In [15] more details on the ability of the variogram to represent the cepstral content of different audio sources are provided, as well as a more formal description of the variogram function.

The MFCCs matrices are computed for 12 DCT coefficients (from the 2nd to the 13th) and the temporal variability of each of the coefficients is described by means of the computation of the variogram function. In order to compress this information, the variogram is computed on a reduced number of distance lags (10), logarithmically distributed, and its values are normalized by the global variance (standardized variogram) [15].

The result is a compressed matrix of 12x10 elements (the song signature) that is conveniently reshaped in a vector of 120 elements with the aim of making simple the comparison of the cepstral signatures of the songs.

3. RHYTHMIC PATTERN

The rhythmic structure of a song plays a fundamental role in music similarity evaluation. If two songs do not share a similar tonal content (or they share it scarcely) but they have the same rhythmic pattern, they will be judged to be similar depending on the strength of the rhythm in the mood of the song.

Most of the works on music similarity based on rhythm, focus on the rhythmic spectrum. Foote and Uchihashi [5] propose the use of the beat spectrum that describes the rhythmic variations over the time. Peeters [13] describes the use of a spectral rhythmic pattern, estimated at the onset positions, which is employed for music genre classification.

In this work a very simple approach is adopted for the extraction of the spectrum of the inter-onset intervals. The onsets are detected on the envelope of the signal and the spectrum of the onset positions, weighted by the energy of the envelope, is extracted. In [2], a very similar approach has been employed for the estimation of the main tempo of the song, aimed to optimize the estimation of the tonal content of the audio signal. The onset positions are precisely located at the peaks of the envelope.

The algorithm employed in this process can be described in the following steps:

1. The signal is half-way rectified and low-pass filtered with a cut-off frequency of 100 Hz (the first subband of the filterbank employed by Dixon et al. [3]).
2. The envelope is computed using a low-pass filter on the transformed signal with cut-off frequency of 1 Hz.
3. The cumulative distribution function (cdf) of the differences of the zero crossing points of the first-order derivative of the envelope is computed.
4. The onset positions are defined as the values of the cdf exceeding the 25th percentile.
5. A triangular function is centered at each of the onset positions detected with height proportional to the local energy of the envelope.
6. The magnitude of the spectrum of this pseudo-rhythm signal is computed

The spectrum of the pseudo-rhythm signal built around the onset positions, defines the rhythm pattern of the signal, with clear peaks when the song shows a well defined bass line and a bunch of more representative peaks, in a more general case. Actually, the frequency distribution of the onsets is directly related to the tempo of the audio signal. In practice, the rhythmic spectrum represents the distribution of the contributions of each tempo values to the overall rhythmic structure of the song.

Using a Fourier transform of length 1024, and considering a fixed beat limit of 250 bpm, as the fastest possible tempo to measure, it is possible to reduce the spectrum to use in the comparison system to the bins under the frequency 4.16 Hz ($\approx$ 250 bpm), that is, to consider only the first 126 frequency bins (extended to 128 to round to a power of two).

4. TONE PROFILES

Another of the main factors that deeply affects the perceptual evaluation of music similarity is the tonal content of the audio.

Xiao and Zhou. [17] propose the use of chroma histogram as a descriptor for the melodic content of the song for music

similarity evaluation. They argue, as an example, that people tend to associate a high level of similarity to a pair of different versions (with different arrangement) of the same song. De la Bandera et al. [2] presented the use of the chroma profiles to estimate the contributions of tonalities and evaluate music similarity.

In this work, an average chroma profile is extracted from each music excerpt and it is employed as an additional descriptor to evaluate music similarity, together with timbre and rhythm.

The chroma is a 12-dimensional vector representing the overall contribution of each note to the tonal content of the song. In this work, the chroma vector is computed on the basis of the summation approach proposed by Fujishima [8].

Recalling the concept expressed in [2], in which a large temporal window was employed to guarantee the analysis of a complete melodic cell, the spectrogram of the audio signal is computed using windows of 1 second. In order to reduce the noisy contribution of low-amplitude samples of the spectrum, only the components exceeding the 25% of the maximum amplitude are taken into account.

After calculating the spectrogram, the magnitudes of the spectrum around each of the bins of the notes between C1 and B7 are summed, and an 84-dimensional amplitude vector is extracted:

$$H_t(k) = \sum_{i=f(k)-\Delta^-_{f(k)}}^{f(k)+\Delta^+_{f(k)}} M_t(i) \qquad (3)$$

where $H_t(k)$ stands for the k-th element of the 84-dimensional vector at time interval t, $f(k)$ is the pitch of the k-th note, with $k = 1, 2, \ldots, 84$, and M is the magnitude of the spectrum. The summation is calculated within the frequency range comprised between the left margin $f(k) - \Delta^-_{f(k)}$ and the right one $f(k) + \Delta^+_{f(k)}$. Note that $\Delta^-_{f(k)}$ and $\Delta^+_{f(k)}$ are different.

Now the amplitude vector is mapped into the twelve semitones. This process is done summing each tone magnitude over the seven octaves analyzed. The chroma vector is computed as follows:

$$C_t(k) = \sum_{i=1}^{7} H_t(k + (i-1) \cdot 12) \qquad (4)$$

where the chroma value $C_t(k)$ of the k-th semitone, with $k = 1, 2, \ldots, 12$, at time t, is the sum of the seven values of magnitude H_t of the seven octaves involved. The term i indexes the octaves.

The 12-dimensional mean chroma for the whole song is computed simply averaging all the chromas obtained for each temporal fragment. Before the summation, all the chromas are normalized such that the sum of all the elements of each profile is 1.

5. CALCULUS OF THE DISTANCE MATRIX

In this framework, the Euclidean distance (the two-norm of the difference) of the descriptor vectors is employed as similarity measure among the songs. The distance is weighted in the case of the variogram-based descriptor, in order to give more relevance to the first lag values, where most of the information on structural variability of the timbre can be found. The weights are computed as follows:

$$W(l) = \begin{cases} 20 & l = 1 \\ 11 - 10^{x/10} & l > 1 \end{cases} \qquad (5)$$

where l is the lag ordinal with $l = 1, 2, \ldots, 10$. The weights decrease logarithmically with the lags, with the exception of the first value (lag = 1) that is manually fixed to approximately twice the second value. Finally, the weights are normalized such that their values sum 1.

Both a full dense matrix and a sparse matrix of the 100 most similar elements for each song are returned.

6. REFERENCES

[1] J.-J. Aucouturier and Pachet F. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] Cristina de la Bandera, Simone Sammartino, Isabel Barbancho, and Lorenzo J. Tardón. Evaluation of music similarity based on tonal behavior. In *Proc. of the 7th Int. Symposium on Computer Music Modeling and Retrieval (CMMR 2010)*, pages 221–233, Málaga, Spain, 2010.

[3] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *Proc. of Int. Symposium on Music Information Retrieval (ISMIR 2003)*, pages 159–165, Baltimore, MD, October 26-30 2003. John Hopkins University.

[4] D. P. W. Ellis and G. E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, volume 4, 2007.

[5] J. Foote and S. Uchihashi. The beat spectrum: a new approach to rhythm analysis. In *Proc. of the IEEE International Conference on Multimedia and Expo, (ICME 2001)*, pages 881 – 884, 2001.

[6] Jonathan Foote, Matthew Cooper, and Unjung Nam. Audio retrieval by rhythmic similarity. In *Proc. of the International Conference on Music Information Retrieval*, pages 265–266, 2002.

[7] Jonathan T. Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, pages 138–147, 1997.

[8] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proc. of the International Computer Music Conference*, pages 464–467, 1999.

[9] A. Kacha, F. Grenez, J. Schoentgen, and K. Benmahammed. Dysphonic speech analysis using generalized variogram. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2005)*, volume 1, pages 917–920, 2005.

[10] B. Logan and A. Salomon. A music similarity function based on signal analysis. *Proc. of the IEEE International Conference on Multimedia and Expo (ICME 2001)*, pages 745–748, 2001.

[11] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. of Int. Symposium on Music Information Retrieval (ISMIR 2000)*, 2000.

[12] E. Pampalk. *Computational Models of Music Similarity and their Application to Music Information Retrieval*. PhD thesis, Vienna University of Technology, Vienna, March 2006.

[13] Geoffroy Peeters. Rhythm classification using spectral rhythm patterns. In *Proc. of Int. Symposium on Music Information Retrieval (ISMIR 2005)*, pages 644–647, 2005.

[14] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. On rhythm and general music similarity. In *Proc. of Int. Symposium on Music Information Retrieval (ISMIR 2009)*, pages 525–530, 2009.

[15] Simone Sammartino, Lorenzo J. Tardon, Cristina de la Bandera, Isabel Barbancho, and Ana M. Barbancho. The standardized variogram as a novel tool for music similarity evaluation. In *Proc. of Int. Symposium on Music Information Retrieval (ISMIR 2010)*, pages 559–564, 2010.

[16] Hans Wackernagel. *Multivariate Geostatistics: An Introduction With Applications*. Springer-Verlag Telos, January 1999.

[17] Linxing Xiao and Jie Zhou. Using chroma histogram to measure the perceptual similarity of music. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME 2008)*, pages 1317 –1320, 2008.