MIREX 2011 - AMS TASK: MFCC/VARIOGRAM 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 the use of the standardized variogram as clustering algorithm for Mel Frequency Cepstral Coefficients, is detailed in this report. The standardized variogram is used for the compression of the information of MFCCs. The algorithm is 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). Logan and Salomon [4] and Foote [2] proposed the first examples of application of the Mel Frequency Cepstral Coefficients for the evaluation of music similarity and recommendation.

In the framework of the MIREX 2011 the method described in [7] is proposed for the AMS task.

2. TIMBRE DESCRIPTOR

As early described by Foote [2] and Logan [5], 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 [6], Foote [2], Aucouturier and Pachet [1] and Logan and Salomon [4] 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 [8]. 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 [3].

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:

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

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 [8]. The theoretical variogram is strongly related with 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:

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

Hence, the typical shape of a variogram function 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 Figure 1, a typical experimental variogram of an audio signal is shown. The shape of the variogram well sumarizes the structural variability of a signal.

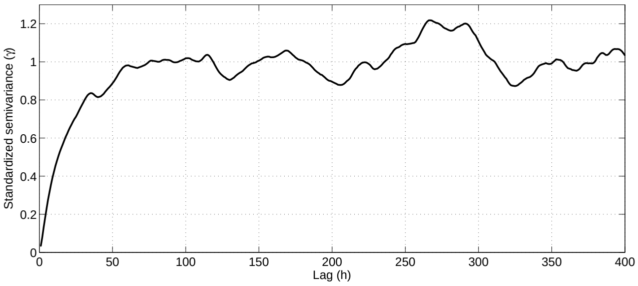

Figure 1. An example of a typical experimental standardized variogram of an audio signal.

In this algorithm, the variogram is employed as a clustering tool for the MFCCs of an audio signal. In [7] 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) [7].

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. 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 defined by the following expression:

$$W(l) = \begin{cases} 20 & l = 1 \\ 11 - 10^{x/10} & l > 1 \end{cases} \quad (3)$$

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 (see Figure 2). Finally, the weights are normalized so that their values sum 1.

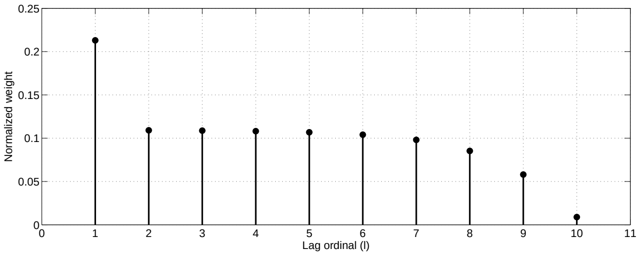

Figure 2. The weights applied to the Euclidean distance for the variogram-based descriptor. The values are normalized so that their values sum 1.

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

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

[3] 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.

[4] 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.

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

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

[7] 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.

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