Content-based music classification based on timbre similarity

Alexandre Savard
Schulich School of Music - McGill University
555 Sherbrooke St. West
Montreal, QC Canada H3A 1E3

Abstract— **Electronic music distribution is in need of robust music descriptors extracted automatically within the audio content. Such descriptors are essential in any tasks involving automatic classification engines. This paper will summarize an approach done by Aucouturrier and Pachet at the *Sony Computer Science Lab* within the *Cuidado* project. The algorithm that they have developed makes use of Mel-Frequency Cepstral Coefficients as well as Gaussian Mixture Models. Some other similar researches using mostly these two features will be also discussed.**

Index Terms— **Timbre similarity, Mel-Frequency Cepstral Coefficients, Gaussian Mixture Models**

I. Introduction

THE growing of large digital music database brought the needs for automatic classification engines based on the analysis of the audio signal. Users express the needs especially for automatic playlist generation. While music distributors see in these tools an application for music recommendation systems, which suggest to customers titles comparable to ones selected.

In the field of music features extraction and content-based music classification, timbre hasn't played an important role until very recent years. As an explanation, this exclusion is mostly attriutable to a lack of ground truths concerning timbre. Timbre is actually a very local attribute of sound and isn't meanungful to represent a whole musical piece. The concept of "global timbre" has been elaborated in order to characterize each title in a given database.

Unlike pitch or loudness, there is no evident acoustic features associated to timbre as a dimension of sound. Early researches on timbre demonstrated its multidimentional aspect (Grey 1975). Furthermore, not only timbre perception is related to various acoustic atributes but it seems to be influenced by additional cultural aspects. While problems for desinging an efficient algorithm to "quantize" timbre rise from the first assertion, this second one brings difficulties in the evaluation of the process when compared to human perception.

The litterature regarding specifically to timbre similarity in music is small. An important contribution to this research field is the work done by Aucouturier and pachet (Aucouturier & Pachet 2002,2004). It is mostly this approach that will be investigated in this short paper. Three main topics will be discussed: what to measure, how to measure it, and how to evaluate this measurement.

II. Timbre Definition

Before going foward in similarity measurement, we need to establish a valid timbre definition as a starting point for further explanation. We already presented timbre as a multidimentional feature of music and sound. In his works, John Grey points out three physical parameters related to timbre. A first one is the spectral flux that is the amount of change in sound components. An second important feature is the spectral gravity center that is related to the perception of brigthness. A last one is the attack time that is relevent from the temporal envelope of the sound.

Another attribute of sound is the sound component's harmonicity ratio (Plomp 1964). We can obviously notice that three parameters of timbre over four are related to the spectral envelope of sound. A good measurement should reflect those principal characterestics of timbre in regards of human perception. Timbre can't be completely defined using only those physical properties. However they provide significant informations that can be considered as sufficient for our purpose.

III. Timbre Measurement

Up to now, there exists no perceptually meaningful way to measure timbre. However, some attempts give us hope that this task is acheivable. An interesting avenue using Mel-Frequencies Cepstral Coeficient has been explored by different groups of researchers and presented good results (Logan & Salomon 2001, Liu & Hang 2002).

For each title in a database is associated a timbre descriptor. This descriptor is perceptually-relevent using the Mel scale for pitches. A set of coefficient is calculated to describe timbre. For that purpose, Mel-Frequecy Cepstrum Coeficients

(MFCC) given by the inverse Fourrier transform of the log-spectrum are used.

$$c_n = \frac{1}{2\pi} \int_{\omega=-\pi}^{\omega=\pi} \log(S(e^{i\omega}))e^{in\omega}\, d\omega$$

Higher coeficients represent the amount of fast temporal variations of the spectral envelope while lower coeficients are more related to slow spectral changes.

In Aucouturrier and Pachet's algorithms, a musical piece is divided in numerous of subsequent windows. A set of MFCCs is computed for each of these windows. They used sets of eight coeficients. This number goes up to 19 if we consider Logan and Solomon's researches. Such an amount of datas have to be stored efficiently in an appropriate structure. The most encountered one is the Gaussian Mixture Model (GMM).

Different approaches differ from each others in the way the segmentation of the musical work is done. Aucuturier and Pachet are using a fixed segment size of 50 ms. At the opposite, liu and hang chose a variable windowing that adapts itself to phoneme lengths of singing voice.

IV. Similarity Measurement

The similarity of two sets of MFCCs can be thought as inversely related to the distance between their GMMs. This distance is computed tacking 100 random samples from the GMM of a first song, and the probability that these samples can be obtained from a second song is calculated. Similarly, samples are taken from the GMM of the second song, and the probability process is repeated. It results a symmetric measure. The higher are these probabilities, the more similar are the songs.

V. Algorithm Evaluation

Some problems can arise when evaluating any classification algorithms. This situation is mainly due to subjective nature of human perception. In the case of timbre, an objective evaluation becomes difficult by the fact that usualy, metadatas don't include timbre descriptions. Also, the meaningfulness of "global timbre" associated to an entire piece of music is not widely recognized. A subjective evaluation actualy tends to demonstrate the effeciency of the algorithm. Aucouturier and Pachet claimed an effectiveness of 80 percent compared to a 15 percent correct match using objective genre classification.

VI. Conclusion

We had seen that an objective evaluation of an algorithm led to disapointing results concerning genre classification. On the other hand, a psychoacoustic survey demonstrated that the results wasn't that wrong and that effectively there could be a timbre similarity between a XXth century piano accompaigned pop song and a XIXth centure romantic lied. Aucouturier and Pachet concluded that there could be a ceiling in the performance of a classification engine exclusively based on timbre similarity.

References

[1] J. Aucouturier, F. Pachet, and Mark Sandler. 2004. The way it sounds: Timbre models for analysis and retrieval of music signals. *IEEE Transaction on multimedia.*

[2] J. Aucouturier, and F. Pachet. 2004. Improving timbre similarity : How high is the sky ?*Proceedings of the International Conference on Music Information Retrieval.*

[3] J. Aucouturier, and F. Pachet. 2002. Music similarity measures: What is the use ? *Proceedings of the International Conference on Music Information Retrieval.*

[4] C. Liu, and C. Huang. 2002. A singer identification technique for content-based classification of mp3 music object. *Proceeding of the Conference on Information and Knowledge Management.*

[5] B. Logan, and A. Salomon. 2001. A music similarity function based on signal analysis. *Proceeding of the International Conference on Multimedia and Expo.*

[6] J. Grey. 1975. An exploration of musical timbre. *Thesis presented at the Center of Computer Research in Music and Acoustics.*Stanford University, California, USA.

[7] R. Plomp. 1964. The Hear as a frequency analyzer. *Journal of the Acoustical Society of America*, 36, 1628–36.