

EVENT SYNCHRONOUS MUSIC THUMBNAILS: EXPERIMENTS

Gianpaolo Evangelista

Federico II Univ. of Naples, Italy
Dept. of Physical Sciences

Sergio Cavaliere

Federico II Univ. of Naples, Italy
Dept. of Physical Sciences

ABSTRACT

The paper presents a recently introduced method for the generation of Music Thumbnails [1], an active field of research [4, 5, 6, 7, 8]. The method aims to the construction of short music pieces, listening to which one can acquire partial knowledge or, at least, perceive the general flavor of the music content of the whole piece. Therefore, the introduced method should have its relevance from a psycho-acoustical view-point, giving the opportunity of listening to a brief summary of the piece itself or of one of its parts. This will allow for enhanced navigation in large databases of music, where the thumbnail will give the flavor of the basic timbre features of the piece: only if it is of his/her interest the listener will proceed to acquire further acoustical details in order to confirm its choice. For this purpose, the method will be able to extract from a long piece of music common features that will be “condensed” in a shorter extract. Our method is synchronous with the basic phrase timing of the piece and it is therefore referred to as the *Event Synchronous Thumbnailing* (EST). Additionally, the method can be used as a means to navigate in large acoustical databases using thumbnailing for the purpose of collecting musical pieces having similar timbre flavors. The EST is therefore acting as a timbre signature to be searched for in the database.

1. INTRODUCTION

Searching for pieces in a music database requires the use of keys in order to properly index the acoustical and musical material. Feature such as timbre, spectral characteristics and rhythm, are particularly relevant, although features such as tempo, style are also important for complete recognition. Some of these features can be described in quantitative terms by means of the evolution of local features at different time scales, such as spectral centroid, spectral sharpness or, with the use of a much larger set of parameters, by means of the MFCC coefficients or of the coefficients of an LPC model. Generally, features are described by numerical parameters that are interesting for the automatic classification of the piece but fail to provide a specific listening experience. This experience, in turn, is really needed since it provides an acoustical overview of the music piece and could constitute the signature by which one can efficiently search for pieces. Our thumbnails are intended to provide an average listening experience of the

piece or of parts of it. The problem is that of estimating the mean characteristic of the music piece, which will best contribute to the construction of the summary. In a large number of pieces, according to genre, measure is an appropriate time scale at which pieces can be classified. In our point of view the average measure is evolving in time since it corresponds to a local average over a large scale, the size of which depends on the maximum scale level introduced in the method. Moreover, all local variations with respect to the average measure can be taken into account in the form of fluctuations at several scales with respect to the average measure (details). In the paradigm of an augmented recorder, our thumbnails could provide a summary of longer sound segments to be skipped or to be listened to. Furthermore, the details provide refinements available from larger to finer scale. This feature, built-in in the employed representation, is attractive in progressive download of musical pieces: from the thumbnail to the whole picture. Examples of the technique and of the produced sounds are at the url: <http://www.na.infn.it/mfa/acust/thumbnails>.

2. PITCH SYNCHRONOUS WAVELET TRANSFORM

The method takes its origin from a special type of wavelet transform previously introduced by one of the authors, namely the Pitch Synchronous Wavelet Transform [2, 3]. This transform is based on a synchronous representation of pseudo-periodic sounds. The mono-dimensional sound signal is transformed into a bi-dimensional signal where different periods of a pitched sound are stored in the rows of a matrix. The waveshape is possibly padded with constant values in order to conform to the maximum period of the signal. As a result, along the columns of the matrix we observe the slow variations due to the evolution of the sound period, in both shape and length. In the Pitch Synchronous Wavelet Transform an array of wavelet transforms, each operating on a single column of the matrix provides a representation in terms of an average or regularized period and of fluctuations from this period at several scales. The block diagram of a structure computing the Pitch-Synchronous Wavelet Transform is shown in Figure 1. The transform was proposed by the author as a signal representation useful for the analysis of pseudo-periodic signals [2, 3]. The feature of the transform of interest to us in the context of musical thumbnails is the ca-

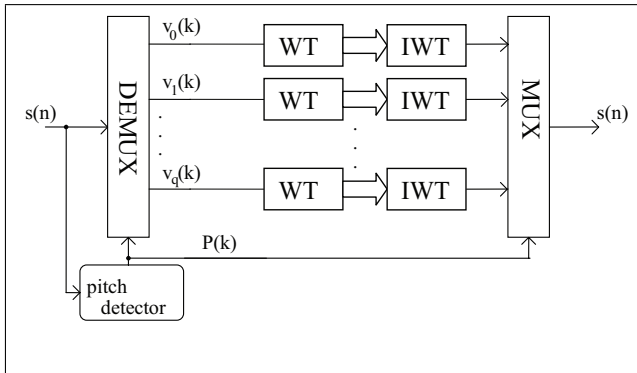


Figure 1. Block diagram of the Pitch-Synchronous Wavelet Transform and its inverse.

ability of separating the periodic part of the signal from its inter-period variations. In an isolated pitched sound these fluctuations are mostly due to the excitation noise (e.g. in the violin bow-string interaction) and/or to expressive local dynamics such as tremolo or vibrato. The Pitch-Synchronous wavelets adapted to a constant pitch signal are comb like. The average period waveshape is obtained by means of projection over the subspace generated by the scaling function. In the frequency domain this is given by the bottom diagram in Figure 2. When properly tuned, the scaling component represents most of the energy of the harmonics of the pseudo-periodic signal. The wavelet components are characterized by side-bands of the harmonics that become narrower and closer to the harmonics as the scale level increases.

The ensemble of fluctuations can be summed together in order to form the total noisy component of the signal. In turn, due to the completeness of the representation, the sum of the total noisy component with the scaling component retrieves the original signal. This provides us with a method for extracting, e.g., the blow noises in a trumpet sound from the remaining regular harmonic part, as shown in Figure 3.

3. EVENT SYNCHRONOUS WAVELET TRANSFORM

3.1. Event Synchronous Signal Representation

The approach of the pitch synchronous representation may be used at a different time scale. In order to reveal large scale periodicities, rather than representing the signal itself we may represent a local rms power of the signal. If the piece under examination exhibits periodic features at this time scale then the rms signal will have the same period. This is shown in Figure 4 where we show the surface of local rms values of the signal, where one row corresponds to the time interval of a measure. The periodicity due to the rhythmic structure can be distinctly observed in the form of vertical lines in the pattern. In Figure 5 we may check that the chosen time base is the proper one. Submultiples of the base time also show a clear underlying regular structure made up of periodic event sequences,

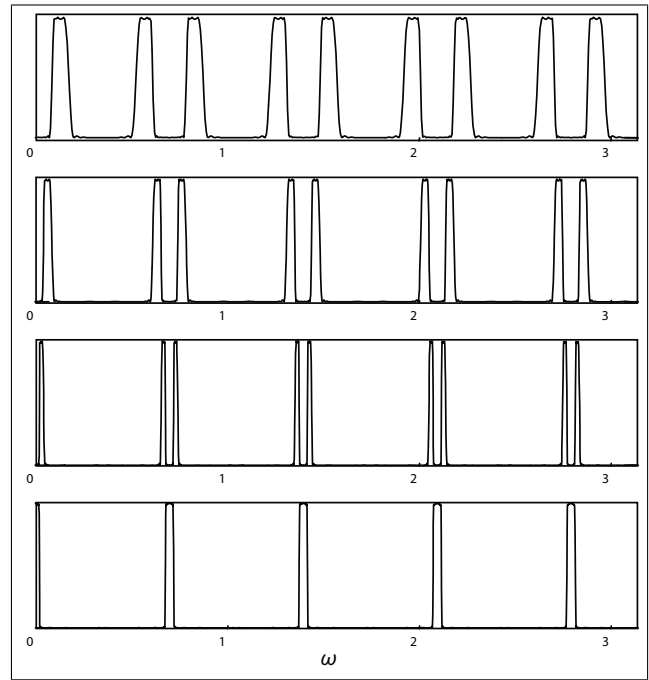


Figure 2. Comb-like frequency domain structure of the Pitch-Synchronous wavelets and scaling functions.

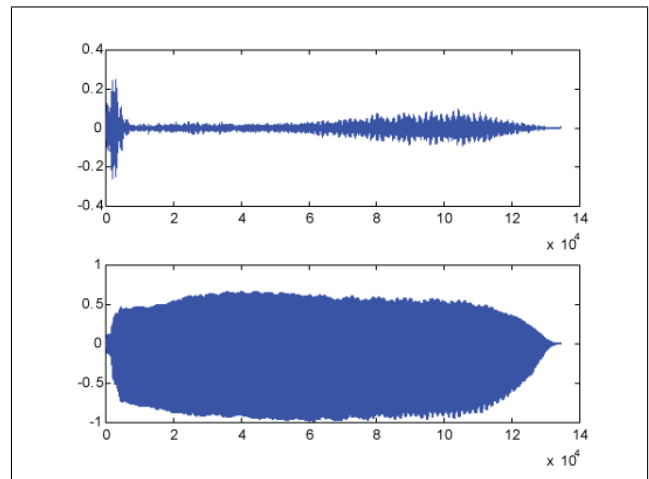


Figure 3. Trumpet tone split into its harmonic and transient parts.

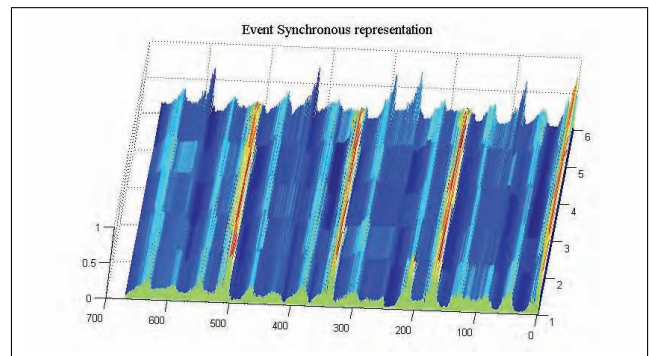


Figure 4. Event synchronous representation of a music piece: rms values are represented instead of the samples of the signal.

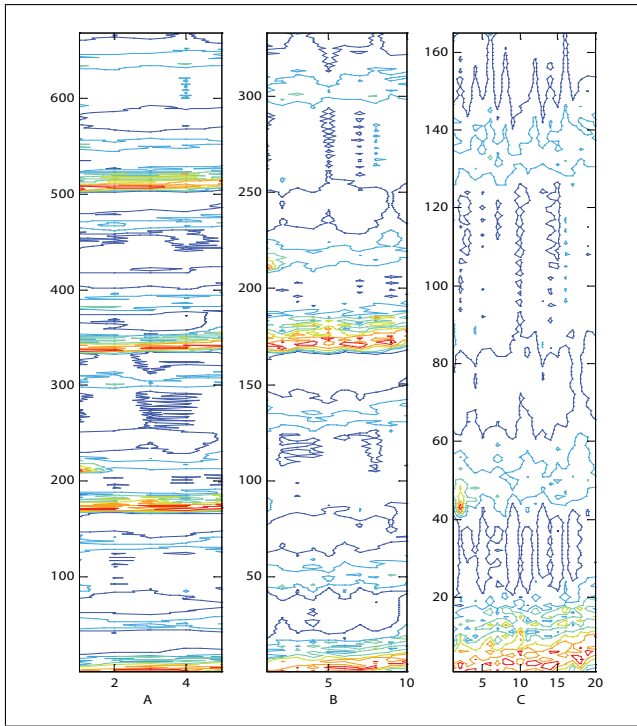


Figure 5. Event synchronous representation of a music piece obtained at the chosen scale (A), and submultiples of it: (B) half scale and (C) one quarter scale.

thus confirming the choice of the time base. The result obtained by choosing an unsuited time scale is shown in Figure 6. Alternately, Foote's similarity maps [9] can be used to detect and extract the base time of regular rhythmic and timbral pattern. Starting from the MFCC representation of a piece of music, these maps graphically represent values of the similarity between all short segments of the piece. Similarity of segments is obtained as the dot product of their MFCC vectors (see Figure 8).

As a preliminary stage, given a segment of music showing common features and a fixed rhythmic pattern we will identify its periodicity, or the time interval constituting the basic "measure" of the piece. A vast literature may be found on this aspect, offering very advanced automatic techniques for the purpose of identifying the periodicities of a musical piece [10, 11].

3.2. Event Synchronous Wavelet Transform

The starting idea for our realization is the use of a variant of the Pitch-Synchronous Wavelet Transform (PSWT) tuned to time intervals longer than the pitch period. If the chosen interval, say a beat or a measure or an entire musical phrase, is the fundamental period of the piece then the transform will be able to split the signal into two parts: a part containing features common to all beats or measures and a part mostly containing the variations from measure to measure or from beat to beat. The first part will eventually be able to summarize the piece and to reflect its basic rhythmic and timbral evolution structure; it will be

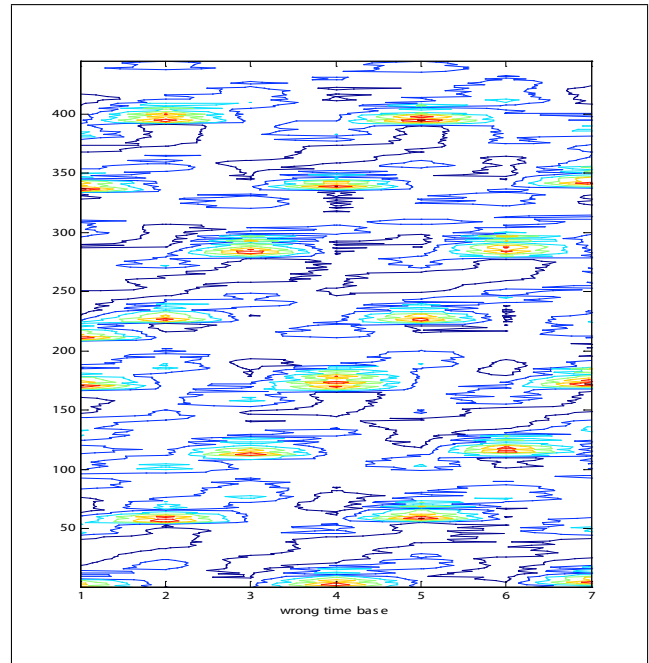


Figure 6. Event synchronous representation of a music piece: at a wrong time scale it shows no evidence of periodicity.

the thumbnail of the piece at least for the part under consideration, i.e. a local thumbnail. The piece will eventually evolve in some manner; the next part will be examined in the same way and split with the same technique. Therefore, given a segment of music we will first identify its periodicity, i.e. the time interval constituting the basic "measure" of the piece and then analyze it using a large-scale PSWT. The representation involving the PSWT with the pitch-detection module replaced by an event detection module is called the Event-Synchronous Wavelet Transform (ESWT). The result of the separation is shown in Figure 7, where the starting signal represents the superposition of a simple drum pattern with an uncoherent signal from running water in a river. The figure shows how the transform is able to extract the underlying common structure from period to period, thus capturing the essential texture of the sound signal. The scaling component of the representation shows a large degree of regularity, as it can be checked from the autosimilarity matrix [9] in Figure 9. This figure must be compared to the source similarity matrix shown in Figure 8: a much higher degree of regularity results from the separation. In fact, variations from beat to beat have been discarded from the regular part and collected in a separate sound signal.

3.3. Features of the Event Synchronous Wavelet Transform

The signal pair obtained by means of the ESWT has the property that by adding the signals in the pair we may rebuild the source signal exactly. This useful property is inherited from the completeness of the wavelet transform

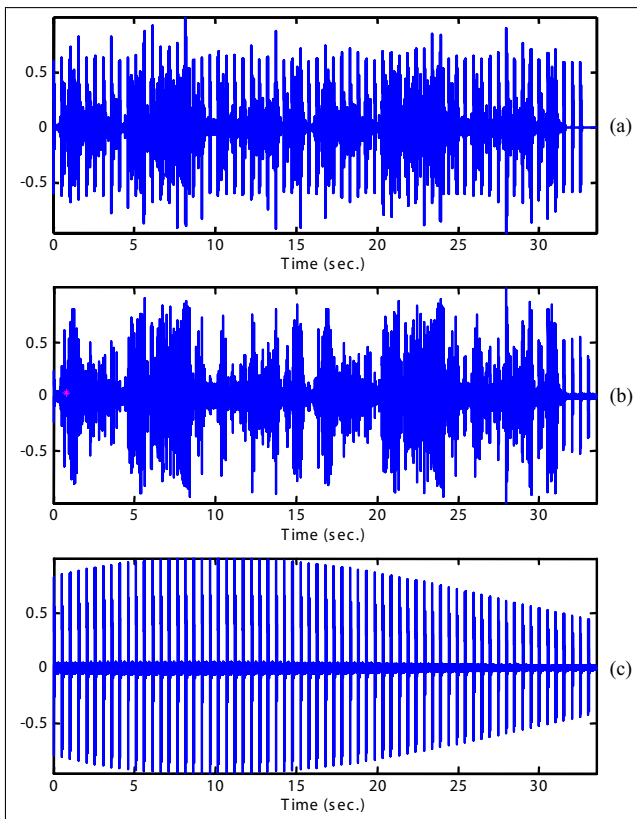


Figure 7. Event Synchronous Wavelet Transform: (a) source signal, (b) sum of Wavelet components (c) scaling residue.

itself. Other properties are inherited from the underlying dyadic wavelet transform, including the power of two under-sampling property from level to level at different detail definition. This means that the scaling residue has a sampling rate scaled by 2^{-N} where N is the number of wavelet scales used. Therefore this residue, when encoded by means of the coefficients of the scaling component, has a highly reduced rate.

The perfect reconstruction feature, again inherited from the Wavelet Transform, may be exploited in order to allow progressive loading of sound signals in order to detail them at successive stages, similar to what is usually applied to images when they are loaded from the internet and displayed at increasing resolution.

4. THUMBNAILS

4.1. Event Synchronous Thumbnails

What is exploited in order to build a suitable thumbnail is the property that the two parts in which the signal is decomposed have different meanings: the regular part, i.e. the scaling residue, collects features common from period to period and, in fact, inherits from the piece its basic underlying structure. The other part, on the contrary, collects the inter-period variations, which are in some way the details of the piece. The first part, the periodic component, will eventually be able to summarize the piece and reflects

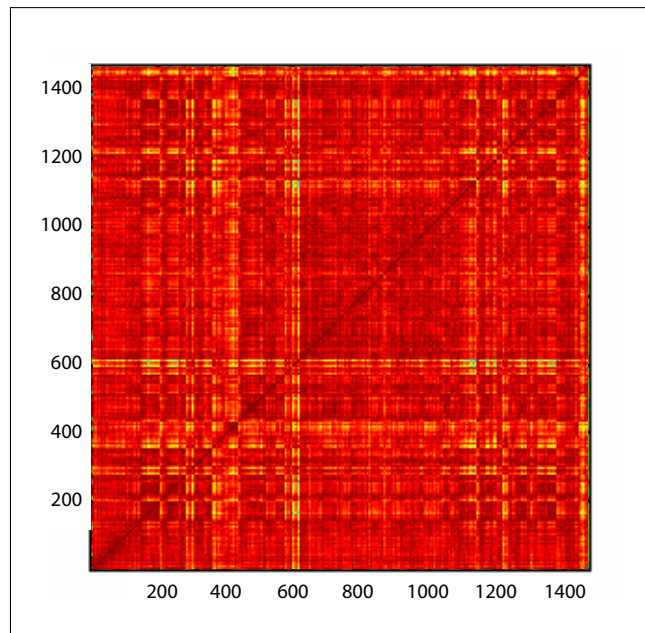


Figure 8. Autosimilarity matrix for the piece Many Chinas (by Mark Isham).

its basic structure. It will therefore be chosen as a thumbnail of the piece for the segment under consideration. We will denote it as the “periodic thumbnail” since it shows a marked periodic and repetitive structure. This signal has the same duration as the original piece. As already stated, the scaling residue requires a much reduced sampling rate. This undersampling works as a further means to reduce drastically the size of the proposed thumbnails simply and efficiently encoded by the wavelet coefficients of the scaling residue.

The next step is to select a single period from this periodic thumbnail, which will be used as the representative member of the whole periodic thumbnail; one period in the middle is usually well suited to the purpose, but other ones may work equally well. Figure 9 shows how the regular part, the periodic thumbnail, has greatly increased the autosimilarity of the signal, as opposed to the more complex starting structure in Figure 8. In what follows we will provide some statistical results supporting this hypothesis. The main results are anyway to be evaluated at the listening experience: essentially our method should have a musical and acoustical impact, as it may be seen from the examples at the url <http://www.na.infn.it/mfa/acust/thumbnails>. The piece will eventually evolve in some manner: the next sections of the piece will have different timbre and rhythmic structure and will undergo the same analysis and decomposition process. The collection of thumbnails obtained in this way will be a thorough summary of the piece well suited to give a short overview of timbres, musical atmosphere and rhythm of the entire piece.

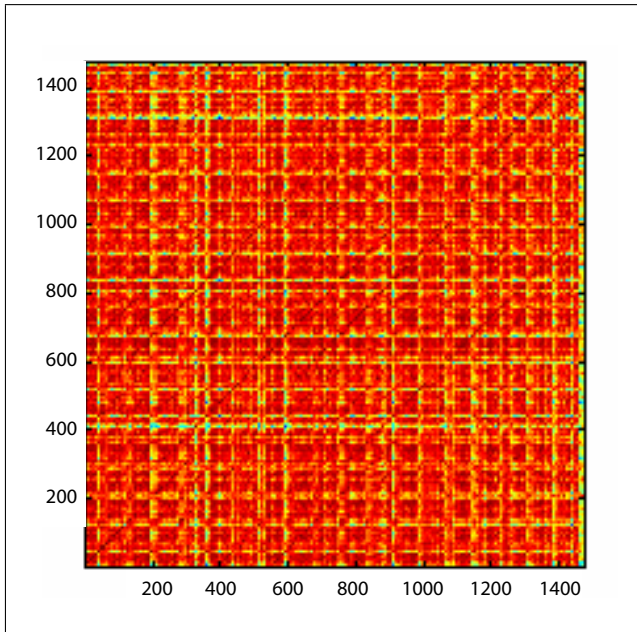


Figure 9. Autosimilarity matrix for the event synchronous periodic thumbnail extracted from the piece *Many Chinas* (by Mark Isham).

4.2. Features of the Event Synchronous Thumbnails

As already pointed out, the proposed thumbnails should have a real impact in listening. However, a statistical characterization may be given in some respects. We may check that the parameters describing the periodic thumbnail, i.e. the “regular” part of the signal, are more uniform than the parameters describing the original piece. We may, for example describe the sounds by means of the MFCC coefficients, obtained at a suitable time scale, or by means of the LPC coefficients, as it is common practice in the literature for recognition and identification purposes. In the case of MFCC coefficients, Figure 10 reporting the first 13 MFCC coefficients of a piece and of its periodic thumbnail shows that the variance of these coefficients is significantly decreased in the “regularized” component of the signal. Additionally, higher order moments of the parameter distribution can be similarly evaluated. We can also resort to a more detailed similarity analysis. As an example, the commercial *findsounds* program reveals a high degree of similarity between the original piece and its associated thumbnail. Another evidence can be given by means of clustering experiments. These experiments, although carried on a very reduced set of four songs, clearly show that by representing the songs in the form of their MFCC coefficients in a reduced dimensionality space results in sufficient clustering. The related thumbnails are well coupled to their source pieces in that their clusters are included in the clusters of the song from which they were extracted. Inter-cluster distances easily allow for perfect identification, as shown in Table 1.

We must again point out that the analysis techniques described in the above may just give a support to the

	Song 1	Song 2	Song 3	Song 4
Th.1	0.0529	0.9564	0.4910	0.3088
Th.2	0.9581	0.0603	0.8478	0.6598
Th.3	0.5375	0.8977	0.0093	0.3063
Th.4	0.3737	0.7282	0.2758	0.0351

Table 1. Cluster distances

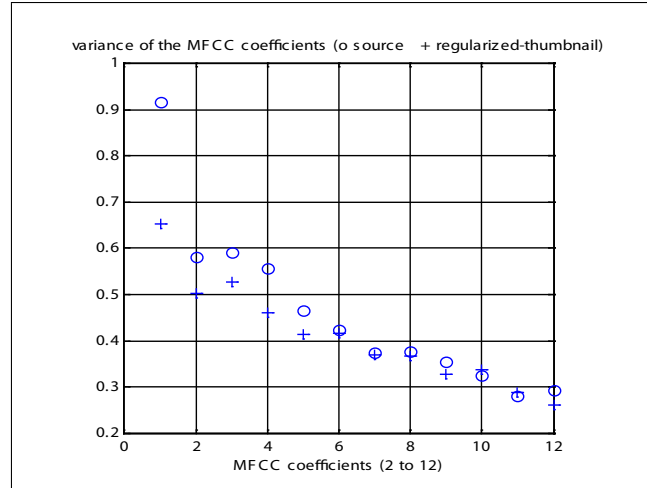


Figure 10. Variances of the MFCC for both the source and the ES thumbnail.

psycho-acoustical evidence of the effectiveness of the technique which, as already recalled, has mainly an impact as listening experience.

5. REFERENCES

- [1] G. Evangelista and S. Cavaliere, “Event Synchronous Wavelet Transform approach to the extraction of musical thumbnails,” *Proc. of DAFx’05*, Madrid, Spain, 2005, pp. 232-235.
- [2] G. Evangelista, “Pitch Synchronous Wavelet Representations of Speech and Music Signals,” *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3313-3330, Dec. 1993, special issue on Wavelets and Signal Processing.
- [3] G. Evangelista, “Comb and Multiplexed Wavelet Transforms and Their Applications to Signal Processing,” *IEEE Trans. on Signal Processing*, vol. 42, no. 2, pp. 292-303, Feb. 1994.
- [4] M. A. Bartsch and G. H. Wakefield, “Audio Thumbnailing of Popular Music Using Chroma-Based Representations,” *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 96-104, Feb. 2005.
- [5] C. Xu, N. C. Maddage, and X. Shao, “Automatic Music Classification and Summariza-

- tion,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 441-450, May 2005.
- [6] G. Peeters, A. La Burthe, and X. Rodet, “Toward Automatic Music Audio Summary Generation from Signal Analysis,” *Proc. 3-rd Int. Symp. on Musical Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 94-100.
- [7] J.-J. Aucouturier and M. Sandler, “Finding repeating patterns in acoustic musical signals: applications for audio thumbnailing,” *Proc. Audio Engineering Society 22nd Int. Conf. on Virtual, Synthetic and Entertainment Audio (AES22)*, Espoo, Finland, June 15-17, 2002, pp. 412-421.
- [8] M. A. Bartsch and G. H. Wakefield, “To chorus: Using chroma-based representations for audio thumbnailing,” *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’01)*, New Paltz, NY, 2001, pp. 15-18.
- [9] M. Cooper and J. Foote, “Automatic Music Summarization via Similarity Analysis,” *Proc. 3-rd Int. Symp. on Musical Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp.81-85.
- [10] E. Scheirer, “Tempo and Beat Analysis of Acoustic Musical Signals,” *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588-601, 1998.
- [11] A. Klapuri, A. Eronen, and J. Astola, “Analysis of the Meter of Acoustic Musical Signals,” *IEEE Trans. Speech and Audio Processing*, in press (2005).