EVENT SYNCHRONOUS WAVELET TRANSFORM APPROACH TO THE EXTRACTION OF MUSICAL THUMBNAILS

Gianpaolo Evangelista and Sergio Cavaliere

[ACEL Group, Dept. of Physical Sciences]
Federico II Univ. of Naples, Naples, Italy
{gianpaolo.evangelista | sergio.cavaliere}@na.infn.it

ABSTRACT

Fast browsing of digital collections of music would largely benefit from the availability of representative audio excerpts of the pieces. Similar to their visual counterparts found in digital photo albums, music thumbnails should offer a comprehensive listening experience while requiring a limited storage space or communication data rate.

The approach to the generation of musical thumbnails of music proposed in this paper is based on an application of the Pitch-Synchronous Wavelet Transform, where the “pitch” is tuned to the elementary measure of the piece. The music thumbnail is encoded by the low rate coefficient sequence pertaining to the scaling residue of the transform. The scaling component represents the pseudo-periodic trend of the piece over several measures. Due to pseudo-periodicity, the time duration of the thumbnail can be arbitrarily extended in listening with no audible artifacts.

1. INTRODUCTION

In this paper we present an extended application of the Pitch-Synchronous Wavelet Transform, introduced by one of the authors [1][2], which concerns the production of musical thumbnails. We define an Event-Synchronous Wavelet Transform, whose components represent averages and differences of large segments of audio signals. When applied to the comparison of signal segments of duration equal to a musically relevant period – for example measure, beat or even phrase – it produces musically relevant components that can be used in order to enhance the variations of the piece or to produce average themes.

The problem of both representing and identifying musical pieces by means of thumbnails has recently gained the attention of several researchers (see, for example [3], [4], [5], [6], [7], [8]). Our approach aims at obtaining a representative fingerprint of a musical piece – or of part of it – by means of a peculiar average performed over several measures. In some cases a piece can be well represented by extracting a signal segment whose duration is a single measure. However, the choice of the proper measure is very critical so that when the selection has to be performed by automatic means, chances there are that the extracted measure will not be very significant. The strategy of averaging over several measures makes our thumbnails more robust with respect to arbitrary selection. Furthermore, the thumbnail includes several features of the original piece, which are played “in parallel” in a single measure. This usually provides a comprehensive flavor of the sonorities present in the piece in the shortest amount of time, a characteristic that is essential in fast browsing of musical pieces in a database by means of acoustic feedback. Moreover, the particular organization of the algorithm in an exact transform allows for progressive downloading of the signal, starting from the thumbnail, which is then refined by adding the complementary details as fluctuations or small scale components. This is the exact counterpart of what is currently available for progressive downloading of images. The thumbnail provides the flavor, which can be enriched in content, if desired, by the detail components to form the original piece. A further direction of this work, which will be developed in the near future, aims at automatically classifying and recognizing musical pieces from their low-rate thumbnails. The thumbnails extracted with the methods described in this paper are mostly oriented to perceptual recognition of the piece as opposed to audio fingerprinting [9], where the problem is that of retrieving a musical piece from a short fragment.

2. THE EVENT-SYNCHRONOUS WAVELET TRANSFORM

The thumbnail extraction algorithm developed in this paper is based on the Event-Synchronous Wavelet Transform (ESWT). This representation is an offspring of the Pitch-Synchronous Wavelet Transform (PSWT), which was applied to the separation of the harmonic resonance from the excitation noise, transients and wave shape fluctuations in pseudo-periodic (voiced) sounds of musical instruments [1][2]. The Wavelet Transform (WT) is a well-known tool for the multiresolution representation of signals and images. In the finite representation, the signal is decomposed in...
a low-pass trend (scaling component) plus fluctuations (details) at several scales. This is obtained by implementing a multirate average and difference scheme on adjacent signal samples. The PSWT is a vector generalization of the WT in which signal segments of one period length are averaged and differentiated sample by sample. The block diagram of the PSWT is shown in Figure 1. There, the sequence \(P(k)\) represents the local period expressed in number of samples. By means of a demultiplexer, whose number of output channels is selected by \(P(k)\), the signal is decomposed into synchronous frames, each of one period length. Each frame is stored in a vector \(v(k)\), whose components correspond to the samples in one period of the signal. As functions of the time index \(k\), these components are individually wavelet transformed in order to obtain the PS wavelet coefficients. The transform can be inverted by applying the inverse wavelet transform to each vector component and by multiplexing the result. The sequence of pitch periods \(P(k)\) is generally time varying, allowing the transform to tune to the local pitch even when the latter is a function of time. When \(P(k)\) is constant the pitch-synchronous wavelets are comb sequences, characterized by regularly spaced peaks in the frequency domain, as shown in Figure 2. The peaks of the scaling function \(\Phi_s(\omega)\) are tuned to the harmonics of the pseudo-periodic signal, while the peaks of the wavelet functions \(\Psi_s(\omega)\), \(k = 1, 2, ..., n\), form sidebands of the harmonics. Given a set of discrete-time wavelets

\[
\tilde{\psi}_{s,m}(k) = \tilde{\psi}_{s,0}(k - 2^m m), \quad m = 0, 1, ..., \quad s = 1, 2, ..., n \quad (1)
\]

and associated scaling sequences

\[
\tilde{\phi}_{n,m}(k) = \tilde{\phi}_{n,0}(k - 2^n m), \quad m = 0, 1, ... \quad (2)
\]

the PS wavelets and scaling sequences are, respectively, defined as follows:

\[
\psi_{s,m,q}(r) = \sum_k \delta(r - q - P(k)) \tilde{\psi}_{s,m}(k) \chi_q(k) \quad (3)
\]

\[
\phi_{n,m,q}(r) = \sum_k \delta(r - q - P(k)) \tilde{\phi}_{n,m}(k) \chi_q(k) \quad (4)
\]

where

\[
\delta(k) = \begin{cases} 
1 & \text{if } k = 0 \\
0 & \text{otherwise}
\end{cases} \quad (5)
\]

and, in terms of the pitch period sequence \(P(k)\), we have:

\[
\chi_q(k) = \begin{cases} 
1 & q = 0, 1, ..., P(k) - 1 \\
0 & \text{otherwise}
\end{cases} \quad (6)
\]

The scaling component in the PSWT representation, which is obtained by projecting the signal over the space spanned by suitably translated versions of the scaling function, represents the periodic trend of the signal. The projections of the signal over the PS wavelet subspaces represent fluctuations from the periodic behavior at several scales. It must be pointed out that the scaling projection is encoded in the scaling coefficient at a rate that is \(2^n\) slower than the original signal sampling rate.

The characteristic function \(\chi_q(k)\) in (3) and (4) allows us to deal with time-varying pitch. This method extends the signal as zero outside the local period. Each change of pitch is regarded as a transition, mostly shown in the wavelet components. It must be pointed out that the period extension is arbitrary and it does not influence the completeness of the representation; other options are available [1].
Figure 3: Event Synchronous analysis of drum pattern hidden in running water noise: (a) mixed signal, (b) water noise extracted by means of sum of wavelet projections and (c) drum pattern extracted as scaling projection.

Figure 4: ESWT analysis of excerpt from Mark Isham’s “Many Chinas”: (a) original signal; (b) sum of wavelet components; (c) scaling component.

3. EXTRACTION OF MUSIC THUMBNAILS BY MEANS OF THE ESWT

The ESWT can be directly applied to the problem of extracting meaningful music thumbnails from recorded pieces. Since the scaling component of the transform represents an average over musically significant time intervals (e.g., beat or measure), its characteristics are “musically” periodic. For this reason, in a large number of pieces a good representative element is given by the central portion of the ESWT scaling component of duration equal to one or more musical periods. The duration of the thumbnail can be periodically extended in order to provide a longer listening experience, which perceptually helps to better identify the piece at no extra data exchange costs. Due to pseudo-periodicity, the juxtaposition of identical “periods” does not introduce listening artifacts. Moreover, since the transform performs averages over $2^n M$ musical periods, where $n$ is the number of scale levels and $M$ is the length in samples of the impulse responses of the filters employed to compute the transform, then the obtained average spans several musical periods. For example, if the chosen musical period is the measure, then a 3 scale level transform based on 11 samples impulse responses yields an average over 88 measures! The ESWT scaling component is robust with respect to transitory variations of the musical content. Even replacing an entire measure with silence, as we did in a few experiments, does not affect the scaling component, which tends to replace the missing part with the average period.

An example of ESWT analysis of the piece “Many Chinas” by Mark Isham (Vapor Drawings, Windham Hill Records) is reported in Figure 4. The central period ($\approx 0.8$ sec.) of the scaling components in Figure 4(c) yields the music thumbnail shown in Figure 5. Here again the components were amplitude scaled to equal level. The sound of the obtained thumbnail contains most of the accompanying atmosphere and is devoid of variations, which allows one to closely identify the piece.

Tuning the transform to the given piece presents a two-fold
problem. The first issue is both aesthetical and interpretative and concerns the choice of the suitable musical period. While for a large class of music pieces the choice of the period equal to
the measure is particularly relevant, there are many exceptions in which other quantities such as beat or phrase length intervals are more interesting and produce more meaningful thumbnails. An approach to the automatic extraction of the suitable period could exploit the similarity analysis concepts described in [10], [11], [5]
[8], where the piece is first scanned in order to detect similar parts and then similar parts are averaged via ESWT in order to obtain the thumbnail.

The second issue concerns the development of a reliable event
detection algorithm that allows one to synchronize the ESWT with
the chosen musical period throughout the piece. It should be noted
that the ESWT allows for time-varying tracking of musical period
in order to synchronize, e.g., with variable tempo. In our simple
examples we employed an estimate based on the deterministic au-
tocorrelation of the signal. In order to enhance the estimate, the
signal is bandpass filtered before computing the autocorrelation.
The result is shown in Figure 6. The position of the leftmost max
at non-zero lag provides the relevant estimate of the large-scale
periodicity of the piece. It can be noted that the autocorrelation
function peaks at points coinciding with the tempo. However, the
largest peak corresponds to a 4/4 measure lap, which allows us
to synchronize the ESWT to this interval. With this choice, the
music excerpt of Figure 4(a) yields an ESWT thumbnail capturing
the self-similarity of the piece. This is apparent if one compares
Figure 7 with Figure 8, where Foote’s similarity measure based
on normalized MFCC scalar products are reported, respectively,
for the original piece and for the ESWT scaling component from
which the thumbnail was extracted.

It must be pointed out that the autocorrelation based measure
tracking algorithm works well with musical pieces with rhythms
cadenced by drums, claps or periodic musical textures. In more
critical situations one should resort to more refined note onset, beat
and measure estimation methods, as found in [12], [13], [14].

4. RESULTS AND CONCLUSIONS

In this paper we introduced an algorithm for extracting thumb-
nails of a musical piece based on the suitably defined Event-
Synchronous Wavelet Transform. The obtained thumbnails cor-
respond to averages of the musical content over several musical
periods (measures). These averages can be computed by means of
a comb filter (obtained by periodization of a low-pass filter) whose
spacing is tuned to the relevant musical period. A property that
supports the use of wavelet transforms rather than generic comb
filters is the possibility to implement progressive download from
the thumbnail to the entire piece.

The results illustrated in the figures correspond to a sin-
gle piece. However, we tested the algorithm for a variety of
pieces in different genres. We remarked that our thumbnail ex-
traction method works especially well with music cadenced by
stable rythmic patterns and it works worse with pieces consist-
ing of solo music themes. Examples are available at the site
http://acel.na.infn.it/thumbnails. More accurate
psychoacoustic tests to be performed both on musicians and on
wider audience are on the way.

A further direction of our work is to explore the use of the
ESWT thumbnails for the classification of musical pieces, as ap-
plied to the search in large databases. In this context, it will be
relevant to study the influence on recognition of the wavelet com-
ponents containing the details.

5. REFERENCES

[1] G. Evangelista, “Pitch Synchronous Wavelet Representa-
tions of Speech and Music Signals,” IEEE Trans. on Sig-
nal Processing, vol. 41, no. 12, pp. 3313–3330, Dec. 1993,
special issue on Wavelets and Signal Processing.

Figure 7: Similarity map of M. Isham’s piece showing regularity over measure.

Figure 8: Similarity map of ESWT scaling component of M. Isham’s piece, showing highly increased regularity over measure.


