Visually-based Audio Texture Segmentation For Audio Scene Analysis

R. GHOZI, O. FRAJ, M. JAÏDANE

Unité Signaux et Systèmes (U2S), Ecole Nationale d'Ingénieurs de Tunis, Tunisia

ABSTRACT

In an analogy with image texture segmentation in visual scene analysis, this paper describes a method for segmenting sound textures in a stream of audio. In particular, we propose a visual scheme to partition an audio stream signal into pieces of audio textures. This visual representation is based on the inter-similarity matrix of the MFCC feature in the signal frames. Classical image enhancement such as binarization and median filtering are applied to the inter-similarity matrix in-order to partition the matrix into homogenous regions. A novelty test operator is then used to localize the boundaries of the image regions, which correspond to audio textures boundaries, signalling thereby a change of audio scene. The perceptual and computational advantages of this visually-based audio texture segmentation are illustrated using a wide range of sound textures of varying degree of complexity.

Key-words: audio texture – inter-similarity matrix audio/image segmentation – image enhancement - Novelty test - MFCC.

1. INTRODUCTION

With the increasing multimedia applications and the rising volume of audio data, the question of automatic audio scene analysis and thereby interpretation becomes of great importance in many settings [1]-[2]. For instance, the sound of traffic jam is directly linked to city environment. The sound of a crowd is typically associated with sport event if that crowd is cheering, a sitcom setting if the crowd is laughing, or could even describe the atmosphere of a marketplace or a shopping center. Similar scene analysis could be achieved in association with sound textures such as fire cracks and rain drops among others. We believe that the concept of audio texture, though relatively recent in audio media, is a key ingredient in most audio productions such as film and video scenes.

Examples of audio textures range from rainfall, crowd sound to fire-cracks. Natural textured sounds often present key information on the audio scene, which is regarded as background "noise" in classical speech/music processing. Just as visual texture segmentation (see for example [3]-[4]) presented critical clues to visual scenes analysis, we believe that the recently introduced audio class of textures has a similar role in analyzing and interpreting audio scenes. Therefore the ability to segment an audio stream composed of a sequence of textured sounds is as important to signal processing as visual texture segmentation has been to image processing.

Over the past decades, researchers have investigated various methods for audio segmentation and classification for a variety of applications, least of which speaker identification, speech/music segmentation and audio stream analysis for indexing and retrieval needs [5]-[8]. On the other hand the literature on audio texture is relatively recent and has focused mainly on different techniques for generating such audio signals [9]-[10].

There is an inherent difficulty in segmenting two textured sounds by examining their time waveform or their spectrograms because of the inherent similarities, imposed by the long-range stationnarity property of these types of signals. In this work we propose a visually-based characterization approach where we can take advantage of the available image processing tools in-order to best detect the presence of two or more audio textures.

The paper is organized as follows: Section 2 presents the visually-based characterization of a given audio texture and describes its perceptual and computational advantages. Section 3 details the audio texture segmentation steps based on the enhanced inter-similarity image. Section 4 illustrates the obtained results and discusses ways of improvement.

2. IMAGE-BASED AUDIO TEXTURE CHARATERIZATION

In an analogy to visual textures, an audio texture is regarded as a class of sounds, characterized by a repetition of structural building elements, called sound grains. Natural audio textures often display randomness in their time appearance and relative ordering, while preserving certain essential temporal coherence [10]-[11], and [16].

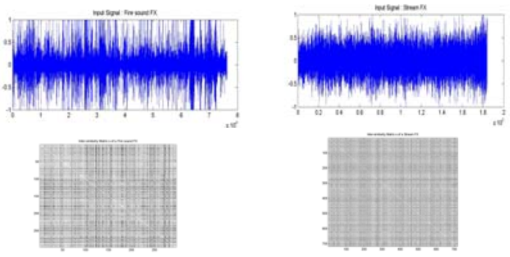
Having a visual representation of the audio texture and thereby the audio scene, where we can use image processing techniques, is the main idea behind our approach for partitioning the audio stream. However, one should note that the matrix representation of the inter-similarity among an audio signal features is not new and several works in the literature have relied on this technique to analyse the constituent's structures of the classical (speech and/or music) audio signal [10]-[12].

2.1 Visual representation of an audio texture

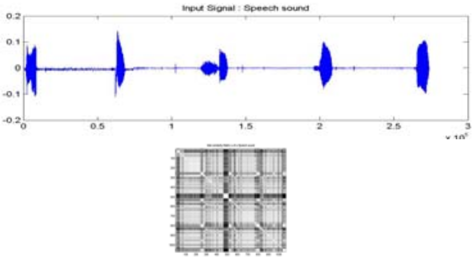
We based our analysis of the audio signals on the classical Mel-Frequency Cepstral Coefficients (MFCC) representation, due to its compact and faithful representation of the signal time information [13]. The pre-processing stage therefore consists of first partitioning the audio signal into frames of 256 samples and then each frame is represented via the first 13 MFCC coefficients. Let C_i denote the row feature vector of frame *i*. One of the simplest operations one can perform is to capture the degree of similarities between any two frames, using the classical vector product; Let S denote the matrix whose elements are given by:

$$S(i, j) = C_i . C_j^{\mathrm{T}} \qquad (1)$$

The result is a symmetric square matrix, which will fasten visibly the computation time in the following processing stages. It was shown in [13] that a visual display of this measurement, treating the entries as "grey- levels", yields a highly homogenous image characterization of S when the signal is textured, and highly non-homogenous for non-textured audio such as speech (see Figure 1).



(a) (b)



(c)

Figure 1: Samples of sounds and their similarity matrices: (a) fire-cracks, (b) water stream, and (c) speech. All sounds were sampled at 44.1 kHz and of 5 second duration each.

In this (image) grey level representation, dark pixels correspond to high similarity, while light pixels represent low level of similarity.

2.2 Multiple texture characterization

This visual representation of audio allows an easy way to detect the presence of two or more different textures. In fact, an audio stream playing a sequence of two successive audio textures yields a similarity matrix with a well structured image: two homogenous blocks on the diagonal, reflecting the stationary behaviour of each texture, and a third block which measures the cross-correlation between the textures, as shown in figure 2.

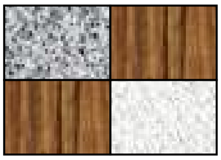


Figure 2: Generic block structure of the similarity matrix of a sequence of two audio textures of the equal duration.

It is clear that the image characterization of audio texture allows not only for a quick visual test of the degree of "textured-ness" of the signal, but it also opens the door to the set of image processing tools available for image enhancement and region detection, as will be illustrated below.

3. AUDIO TEXTURE SEGMENTATION

Once the inter-similarity image is obtained, the audio texture segmentation process requires two stages: the first stage makes use of basic image enhancement operations and the second performs a region detection operation, which marks the time frame boundaries of the constituent audio textures.

3.1 Similarity matrix image enhancement

The simple structure of the similarity matrix offers an easy processing procedure via classical image enhancement techniques. In our work we have performed standard image binarization (via mean-based thresholding), followed by a median filtering to reduce the speckle-like resulting noise [15]. Resulting image yields a clean view of the matrix making the next stage of segmentation process computationally faster.

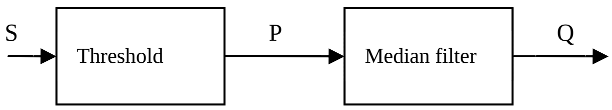


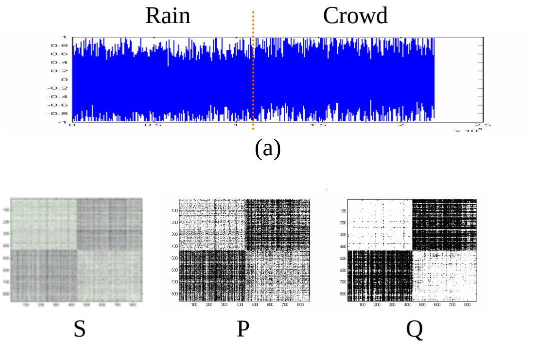
Figure 3: Image enhancement of the inter-similarity matrix *S*.

$$P(i,j) = \begin{cases} 1 & \text{if } S(i,j) < m \\ 0 & \text{if } S(i,j) > m \end{cases} \qquad (2)$$

Where *m* is the mean value of the matrix *S*,

$$Q(i,j) = median \begin{Bmatrix} P(i-k, j-l), \\ (k,l) \in W \end{Bmatrix} \qquad (3)$$

Where *W* is window of odd size (in our case size 5x5). The algorithm for median filtering requires arranging the pixel values in the window in an increasing or decreasing order and picking the middle value [15]. Figure 4 illustrates this image enhancement process through an example.



(b)

Figure 4: Enhancement of the similarity matrix S. (a): rain-crowd signal, sampled at 44.1 khz for a 2.5 sec equal duration of each texture. and results obtained in (b) based on the steps of figure 3.

3.2 Novelty score and texture boundaries

The purpose of this visually-based method is to localize time boundary to segment audio scenes (sound textures). In this sense, a region detection procedure is adopted; in particular, a gradient operator is applied over the diagonal block of the clean binary view of S. This one-pass operation produces a novelty score signal, which identifies the time frame at which the new texture is introduced, yielding thereby a set of segmented audio textures. The novelty score test is obtained by correlating the matrix S with a kernel *K* that captures the similarity/dissimilarity of a given frame *i* with its surrounding frames:

$$N(i) = \sum_m \sum_n K(m,n)\, S(i+m, i+n) \qquad (4)$$

We have used the following 2x2 and 4x4 kernel operators [15]:

$$K_2 = \frac{1}{4}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad K_4 = \frac{1}{16}\begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \qquad (5)$$

The kernel operator *K* acts as an "edge detector" across the region of application. In the audio context, this kernel is applied on the diagonal of the enhanced matrix in order to detect changes among neighbouring time frames of the signal.

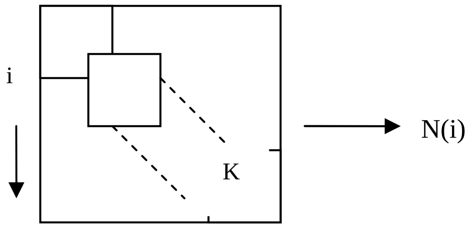
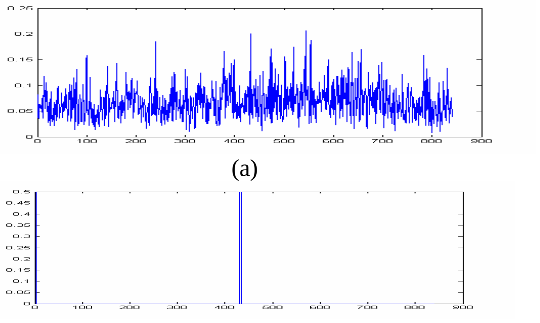


Figure 5: Novelty test operation applied to the similarity matrix.

The perceptual and computational gains in applying a gradient operator through the enhanced image of S reduce the novelty test to the following simple computation for the classical 2x2 kernel:

$$N(i) = \frac{1}{2}(1 - Q(i, i+1)). \qquad (6)$$

The advantages of using the enhanced matrix are far greater than that of the original matrix. In fact, the problem of threshold selection is greatly simplified with the cleaned version S. The novelty is marked by a binary signal that the occurrence locates the time frame at which the second texture is introduced; Figure 7 shows the results obtained for the rain-crowd sound of figure 5.



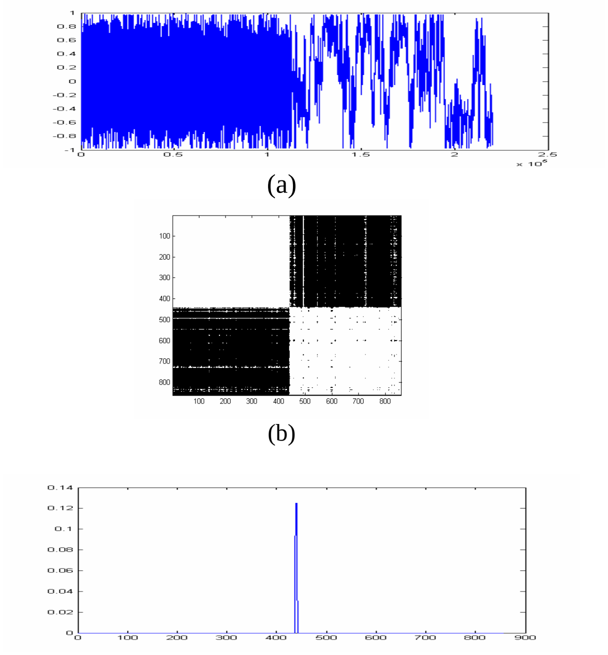
(b)

Figure 6: Novelty test results for the example of rain-crowd shown in figure 5 applied (a) to S, (b) to Q after cleaning (median filtering applied 3 times).

It is clear from figure 6 (a) that applying a Novelty score test on the original matrix S without any processing produces a signal rich of details, the level of which depends on the size of the kernel K. The difficulty in analyzing this kind of result resides in finding the proper level of thresholding capable of providing the desired segmentation. Figure 6 (b), on the other hand, shows the results of applying the Novelty score test on the enhanced image Q. It is clear that the problem of threshold selection is largely avoided, since most of the local variations within an audio texture have been cleaned during image enhancement.

4. RESULTS AND DISCUSSION

The proposed segmentation scheme was applied to many types of audio textures, including noise, which we consider a ‘primitive’ form of texture. The following figures show the results obtained. The audio textures sound files can be found in [16]. In all simulations, the audio signals were sampled at a frequency of 44.1 kHz. The total duration of each signal was 5 seconds (i.e., 220500 samples), The first half of the signal was taken from first texture, while the second half originated from the second texture.

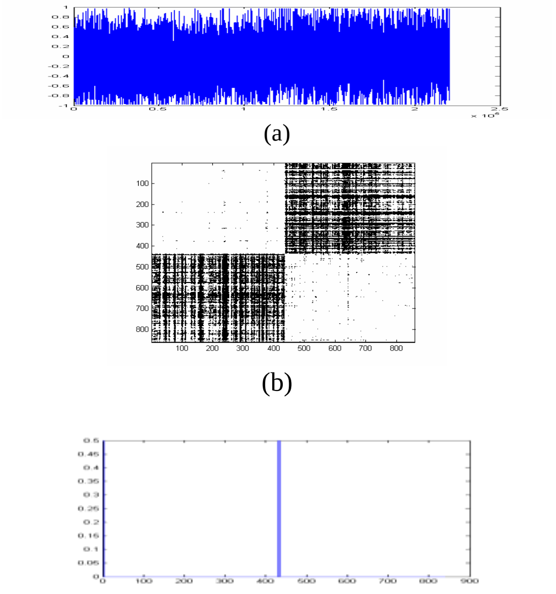


(a)

(b)

(c)

Figure 7: Segmentation results of white noise and Brownian noise. (a) Time signal, (b) enhanced similarity matrix, (c) Novelty score test obtained using a kernel of size 2 applied diagonally.

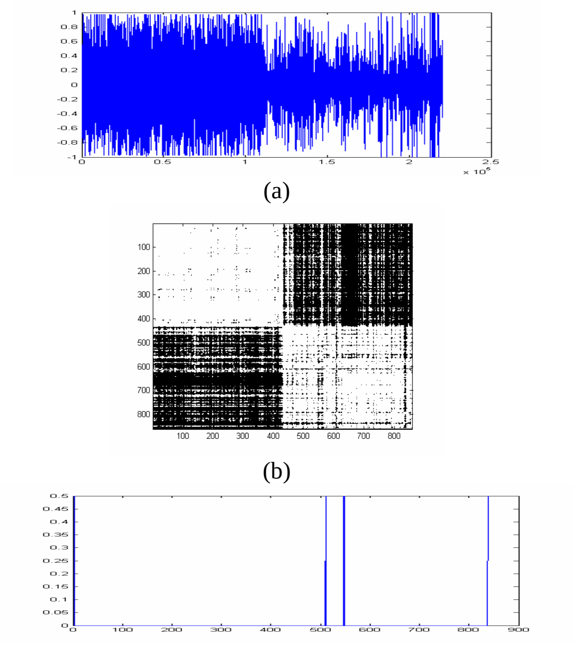


(a)

(b)

(c)

Figure 8: Segmentation results of the sound of rain followed by the sound of a crowd cheering. (a) Time signal, (b) enhanced similarity matrix with median filtering applied 3 times, (c) Novelty score test obtained using a kernel of size 4 applied diagonally.



(a)

(b)

(c)

Figure 9: Segmentation results of a stream of water followed by the sound of fire. (a) Time signal, (b) enhanced similarity matrix with median filtering applied 3 times, (c) Novelty score test obtained using a kernel of size 2 applied diagonally.

Discussion

It is clear from the above results that the proposed method offers relatively accurate segmentation results when the audio stream is composed of noise signals or textured ones as can be predicted from the clean visual feature matrix. However, we have noted that the proposed approach has less desirable performance when applied to more natural texture signals. In fact in those cases higher degree of "complexity" in the audio textures requires more advanced level of image processing. In order to correctly segment the signal, one may need to apply the median filtering multiple times, and to perform the Novelty score test horizontally, instead of diagonally, using larger size operator. The effects of processing parameters such as the number of MFCC coefficients and the size of the frame have been tried and we believe that the 'standard' optimal choice of such parameters can be adjusted for even faster time analysis of textured audio.

The time precision of this segmentation approach is at the frame level since all operations are based on the inter-similarity between time frames of the signal. Accurate results were found in segmenting simple textures (see figures 7 and 8). Less precision was noticed when the image enhancement results were not good (exp. Figure 9).

5. CONCLUSION

We have attempted to understand audio scene changes purely from an audio texture perspective. In particular, we described an efficient and fast method for detecting and localizing at the frame level the coming/entry of a new audio texture. It is up to the human user to judge which of the changes actually correspond to a semantic scene change. It would be interesting to associate our visual segmentation test with semantic boundaries. Such extension would require a correlation between our algorithm and human user revelations/reactions. We also believe that intelligent audio scene understanding would need to integrate existing speech/music analysis in addition to textured sounds in a given environment.

REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis: Perceptual Organization of Sound*, the MIT Press, Cambridge, Massachusetts, 1994.

[2]H. Sundaram and S.F. Chang. "Audio Scene Segmentation using Multiple Models, Features and Time Scales". *In IEEE ICASSP*, Istanbul, Turkey, June 2000.

[3] A. Lorette, X. Descombes, J. Zerubia, "Texture Analysis through a Markovian Modelling and Fuzzy Classification: Application to Urban Area Extraction from Satellite Images", International Journal of Computer Vision, Vol. 36, no. 3, pp. 219-234, 2000.

[4]M. Celenk, Z. Qiang, and D. Chelberg, "Equal-intensity Map Texture Modeling for Natural Scene Segmentation ", Proc. Fifth IEEE Southeast Symp.on Image Analysis and Interpretation, pp. 219-223, April 2002.

[5] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette and P. Depalle. Feature extraction and temporal segmentation of acoustic signals. Proceedings of the International Computer Music Conference (ICMC), Ann Arbor, MI, USA, pp. 199-202. September 1998.

[6] K. Hyoung-Gook, N. Moreau, and T. Sikora, "Audio Classification based on MPEG-7 Spectral Basis Representation", IEEE Trans. On Circuits and Systems for Vedio Technology, Vol. 14, issue: 5, pp. 716-725, May 2004.

[7] Wellhausen and H. Crysandt, "Temporal Audio Segmentation Using MPEG-7 Descriptors." *Proc. SPIE, Vol. 5021,* Santa Clara (CA) 2000.

[8]B. S. Ong and P. Herrara, "Semantic Segmentation of Music Audio Contents", *Proceedings of International Computer Music Conference 2005; Barcelona*

[9] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Synthesis of Sound Textures by Learning and Resampling of Wavelet Trees", Proc. ICMC 1999.

[10]L. Lu, L. Wenyin, and H. J. Zhang, "Audio Textures: Theory and applications", IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 2, pp. 156-167, March 2004.

[11] A.S. Arnaud and K. Popat, "Analysis and Synthesis of Sound Textures", *Computational Auditory Scene Analysis*, D.F. Rasenthal, G. Horoshi, and G. Akuno, editors, Lawrence Erlbaum Association, New Jersey 1998.

[12]J. Foote, "Automatic Audio Segmentation using a Measure of Audio Novelty", *proc. of IEEE Intl. Conf. on Multimedia and Expo*, **1**, pp. 452-455, 2000.

[13]R. Ghozi, W. El-Euch and M. Jaidane, "Two-dimensional Characterization of Audio Textures", 3rd Int. Symp. Video Comm. (ISIVC), Hammamet, Tunisia, 2006.

[14] P. Brodatz, Textures: A Photographic Album for Artists and Designers, Dover Publications, New York, 1966.

[15]A .K Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, New Jersy, 1989.

[16] http://research.microsoft.com/~llu/AudioTextures/