TEMPORAL ENTROPY-BASED TEXTUREDNESS INDICATOR FOR AUDIO SIGNALS

Olfa Fraj, Raja Ghozi, and Mériem Jaïdane-Saïdane

Unité Signaux et Systèmes
École Nationale d'Ingénieurs de Tunis
Université de Tunis El Manar

ABSTRACT

In this paper, we present a temporal entropy-based indicator that reflects the texturedness level of a given audio signal. Inspired from an image homogeneity evaluation via a multidirectional cumulative entropy computation, we similarly propose an audio signal homogeneity analysis through a direct and reverse progressive auditory information content tracking. A [0 - 5] texturedness indicator is then constructed using auditory-inspired parameters, and is inherently associated with short listening time. Using this indicator, a new audio signals classification is proposed where speech signals are assigned low texturedness scores, and academic noise signals are assigned the higher range of texturedness scores. Classically known audio textures are assigned, in this scale, various intermediate to high texturedness scores.

Index Terms— Audio textures, texturedness indicator, short listening time, direct and reverse progressive auditory information, cumulative entropy.

1. INTRODUCTION

Recently, the concept of audio texture has emerged as a key audio analysis feature [1-2]. It has been mainly proposed as a tool in multimedia audio background synthesis [3-6], audio stream segmentation and restoration, and environmental soundscape analysis [7-9]. However, signal processing studies on the relationship between the concept of audio texture and the large field of audio features extraction remain not well established. Indeed, audio signals are usually recognized and categorized into the large classes of speech, music, and noise. In this work, we argue how an arbitrary audio signal can be analyzed over a new continuum of audio signal "texturedness" content level. For that, we propose an indicator that evaluates the texturedness degree of audio signals inherently related to its information content dynamics. This indicator was constructed using a cumulative entropy-based technique, originally intended for image homogeneity analysis and texture discrimination, which we here adapted to audio signals. The originality of this work is that it offers a novel audio feature that assigns a given audio signal a "texturedness level" over a quantified 0 to 5 scale. This audio texturedness indicator is inherently dependant of several auditory parameters: the listening time duration, the basic analysis frame size, and the feature used in capturing the information content of the audio signal.

This paper is organized as follows: in section 2, we present the entropy-based homogeneity tracking method for image and we show how we adapt it for audio signals. Section 3 presents the design steps of the proposed audio texturedness indicator based on the direct and reverse temporal entropy analysis of section 2. The relevance of each of the design parameters in the proposed audio texturedness indicator is also discussed. In section 4, the texturedness indicator results are presented and discussed. Finally, Section 5 discusses the indicator performance results and presents some perspectives of this work.

This work is a part of a Tunisian-French inter-disciplinary project (CMCU- PHC UTIQUE project N 23242VJ)

2. CUMULATIVE DIRECT AND REVERSE ENTROPY FOR AUDIO SIGNAL HOMOGENEITY TRACKING

In contrast with audio signal processing, image processing has witnessed many advances in texture-based image analysis, where the task of texture discrimination was based on several features, such as local energy, contrast pixel correlation (co-occurrence matrices) and entropy (see [10-11]).

2.1. Multidirectional cumulative entropy tracking for image texture homogeneity assessment

A four-directional entropy tracking in image similarities and structures identifying corresponds to a quick image scanning, capturing therefore the image homogeneity degree. In Fig1, we illustrate this entropy tracking process applied to two different images : (a) sand image [12], which is highly textured, and (b) cameraman image which is non-textured. This process consists of a cumulative computing of the spatial entropy as follows:

- from each corner of the image an ($N \times N$) analysis window

is used to progressively scan the image. At each step i, the grey level entropy is evaluated through the analysis window that increases in size ($i.(N \times N)$) until covering totally the image quarter from the considered corner to the image center (see fig 1 (a)).
- using the obtained entropy values from each direction/corner, four cumulative entropy curves, denoted H_c, are obtained (see fig 1 (c) and (d)).
One can notice that the less textured the image is, the further apart are the resulting four entropy curves as illustrated in fig 1 (c) and (d). Therefore, a similarity analysis between the multidirectional cumulative entropy plots could be used as a basis for examining the image level of homogeneity and therefore its "texturedness" level.
Several computational and perceptually inspired parameters are considered in this texturedness level evaluation; in particular the basic analysis window size with respect to the size of the overall image, the increasing analysis window at each step, the choice of the underlying analysis feature, and the similarity measure used to assess the obtained homogeneity evaluation along the different directions.

2.2. Direct and reverse cumulative entropy evaluation in audio signals

In an analogous manner, the audio signal homogeneity content can be examined via a similar progressive entropy tracking in a direct and reverse directions (see fig 2). In fact, when listening to an audio signal in "normal" (forward) direction, then listening to the same signal in a reverse time (backward) direction, one can notice that the more the signal is "textured", the less differences are perceived by the human ear. Whereas when the sound contains perceptible changes in its informational content over time, the resulting direct and reverse listening are quite different, which is for instance the case of speech signals.
This direct and reverse listening to the audio signal has some auditory attention basis as well. It is worth noting, that several analysis studies in cognitive restoration of reversed speech showed that the intelligibility is not greatly affected as long as the basic frame size does not exceed $50ms$ (see for example [13-14]).
In addition to the proposed bidirectional audio analysis, we consider some computational and auditory parameters such as the observation time corresponding to the listening time and the analysis frame size. As analysis feature we propose the use of temporal entropy measure which is a mathematic quantity that captures all the audio amplitudes statistics.
Let $x(t)$ be the audio signal observed over a time duration T_{obs}, and its corresponding reverse signal denoted by $x_r(t)$, where:

$$x_r(t) = x(T_{obs} - t), t \in [0, T_{obs}]. \quad (1)$$

The audio signal $x(t)$ is sampled at T_s sampling rate and contains N samples. Once the analysis frame length T_f is fixed,

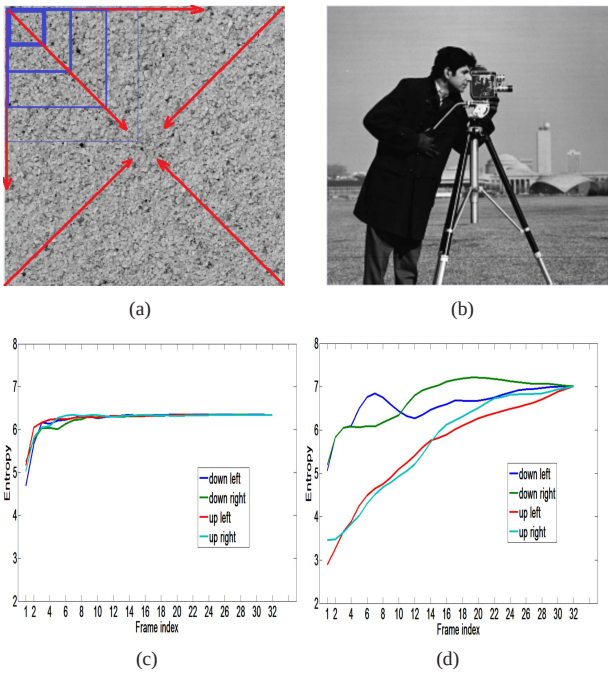

Fig. 1. Multidirectional cumulative entropy plots (H_c) of (a) textured image of sand, and (b) non textured image of cameraman. In (c) H_c of the image (a), and in (d) H_c of image (b). Image size 256×256, basic analysis window (frame) size 8×8.

a progressive framing of length iT_f is applied on x and x_r (as shown in figure 2), where $i \in \{1, 2, ...n_f\}$ is the step index, and $n_f = T_{obs}/T_f$ represents the number of frames which is also equal to the total number of framing steps.
At the i_{th} step, a temporal entropy measure $H(i)$ is computed by:

$$H(i) = \sum_{j=1}^{M} \overline{h}(i,j) \log(\frac{1}{\overline{h}(i,j)}) \quad (2)$$

This temporal entropy formulation is based on the normalized histogram of the signal amplitudes $\overline{h}(i,j)$ which we consider here as an approximation of amplitudes probability density. $\overline{h}(i,j)$ is given by:

$$\overline{h}(i,j) = \frac{h(i,j)}{i.N_f} \quad (3)$$

where $h(i,j)$ is the updated amplitudes histogram computed at the i^{th} analysis step, and $j \in \{1, 2, 3, ..., M\}$ refers to the level index in the M quantification levels of the signal amplitudes. In equation 3, $h(i,j)$ is normalized by the total number of signal samples at the step i; and N_f refers to the number of audio samples in T_f.
When applied to x and x_r, this proposed temporal entropy measure allows to construct their cumulative entropy curves denoted respectively H_d and H_r curves. Figure 3 shows the

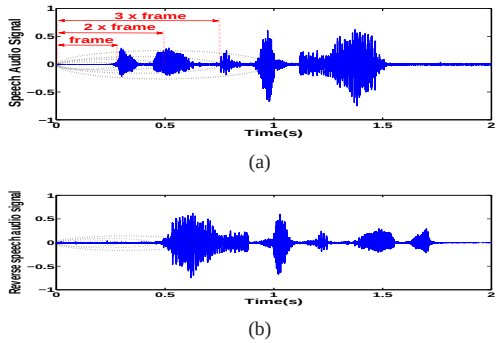

Fig. 2. Audio signal framing for cumulative entropy computation of (a) direct audio signal $x(t)$, and (b) its reverse version $x_r(t)$. Speech signal with 2s duration, 16kHz sampling frequency, increasing frame size analysis.

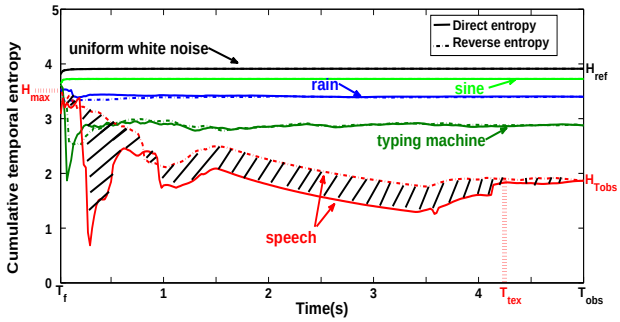

Fig. 3. Importance of the surface area S (hatched) between the cumulative direct (solid line) and reverse (dotted line) entropy curves, and the texturedness indicator design parameters: T_f, T_{obs}, H_{ref}, H_{max}, $H_{T_{obs}}$, and T_{tex}. Five audio examples with different texturedness properties: uniform white noise (black), sine wave(green), soft rain sound (blue), typing machine sound (dark green), and leadership speech (red). Signals duration $T_{obs} = 5s$, frame size $T_f = 20ms$, number of amplitudes levels for entropy computation $M = 50$.

bidirectional cumulative entropy results for five audio signals with different texturedness levels. We notice that for the uniform white noise and the rain signals, H_d and H_r plots are very close to each other. However for less textured audio signals, the curves are distant, which could clearly be seen for the considered speech signal.

The obtained surface area between the direct and reverse entropy curves differs widely depending on the nature of the signal. This surface, designated by S, allows to visually distinguish two very separable classes: highly textured audio and highly non textured audio. This surface area S will then be used as a key parameter in the evaluation of audio signals texturedness degree.

We also noticed that the final entropy value denoted $H_{T_{obs}}$, which expresses the overall signal entropy value, usually tends to decrease from a maximum value obtained for the uniform noise to lower values for highly non-stationary audio signals, such as speech.

3. THE AUDIO TEXTUREDNESS INDICATOR DESIGN

Our approach relies on a graphical representation of the main parameters which we represent through different surface areas of the cumulative temporal entropy tracking plots (see figure 3). The proposed indicator I_{tex} is given by:

$$I_{tex} = K.\rho_{ref}[1 - \rho_{\tau}.(\frac{S}{S_{max}})], \quad (4)$$

where :

- S is the main parameter, defined as the area delimited by H_d and H_r curves, and is given by:

$$S = T_{obs}.\sum_{i=1}^{n_f} |H_d(i) - H_r(i)|, \quad (5)$$

The surface S is first normalized by $S_{max} = T_{obs}.H_{max}$ which represents the extreme maximum value that S could reach during T_{obs}; $H_{max} = \max_{i\in\{1,2,..n_f\}}(H_d(i), H_r(i))$ is the maximum value reached by H_d or H_r.

-ρ_{τ} is a normalizing area ratio given by

$$\rho_{\tau} = \frac{S_{tex}}{S_{max}} \quad (6)$$

where $S_{tex} = T_{tex}.H_{max}$. The parameter T_{tex} is a "texturedness time" that expresses the time duration after which H_d and H_r curves merge to the same final value, within a margin θ such that $\frac{S_{tex}}{S_{max}} \leq \theta$.

- ρ_{ref} is a normalizing area ratio related to the "final" surface to that of the reference signal, given by:

$$\rho_{ref} = \frac{S_{T_{obs}}}{S_{ref}} \quad (7)$$

where $S_{T_{obs}} = H_{T_{obs}}.T_{obs}$ and $S_{ref} = H_{ref}.T_{obs}$.

H_{ref} is the entropy value of the selected reference audio signal adopted in this study which is the uniform white noise. When the signal amplitudes are quantized into M levels, the reference entropy is then $H_{ref} = \log M$. Entropy is maximal for such signal since all of its quantized amplitudes values are equally present resulting in a maximum of uncertainty and supposed to induce a high surprise effect.

- K is a scaling constant which we assigned the value 5 for a perceptual texturedness level indicator scaling in $[0 - 5]$ range, similar to PESQ (Perceptual Evaluation of Speech Quality), usually adopted in survey responses ranges [15], and the subjective Likert scale for attitudes measure [16].

Relevance of I_{tex} terms : discussion

- $\frac{S}{S_{max}}$ normalizes the surface S to the the maximum surface value obtained for the most non textured signal having $S = S_{max}$. It partially classifies all signals sharing the same H_{max} and compares them to the same reference of highly non textured signal. In case of highly textured audio signal $S \to 0$ then $\frac{S}{S_{max}} \to 0$, whereas for the highly non textured signal $S \to S_{max}$ resulting in $\frac{S}{S_{max}} \to 1$.

-ρ_τ allows the comparison between two audio signals that differ only by their T_{tex} values. Using T_{tex} the ρ_τ ratio compares audio signals "texturedness speed" for a given T_{obs}. Therefore a small T_{tex} indicates that the audio signal has quickly became textured, which is the case of highly textured signal, whereas a high value of T_{tex} indicates that the signal is rather less textured. For highly non textured signal $T_{tex} \to T_{obs}$, then $\rho_\tau \to 1$, whereas for a highly textured signal $T_{tex} \to 0$ then $\rho_\tau \to 0$.

- ρ_{ref} performs a comparison with a reference signal which is the most textured audio signal having the highest entropy value H_{ref}. This ratio compares the stabilization final entropy value $H_{T_{obs}}$ to that of the reference signal. In case of highly non textured signal $H_{T_{obs}} \to 0$, then $\rho_{ref} \to 0$, whereas in case of a highly textured signal $H_{T_{obs}} \to H_{ref}$ resulting in $\rho_{ref} \to 1$.

Special cases:

- The uniform (white) noise: is the reference audio signal for the texturedness indicator which has the highest and invariant entropy value. For this audio signal surface $S = 0$ then $\frac{S}{S_{max}} = 0$, and $T_{tex} = 0$ then $\rho_\tau = 0$. In addition $\rho_{ref} = 1$ resulting in $I_{tex} = 5$.

- A highly non textured audio: speech with non decreasing temporal entropy variations which is a non stationary signal presents large entropy variability captured by a large surface S. Theoretically for an extremely high non textured audio signal $S = S_{max}$ and $T_{tex} = T_{obs}$ then $\rho_\tau = 1$, resulting in $I_{tex} = 0$.

4. AUDIO SIGNALS: A CONTINUUM OVER A "TEXTUREDNESS SCALE"

The developed I_{tex} indicator has been tested on a large set of audio signals including academic signals (uniform white noise, pink noise, sine...), natural sounds (rain, fire crackling...), mechanical sounds (bells, train horn...), human originating audio signals (group applause, crowd cheering, motivational speeches...), rhythmic sounds (music, singer without music), and campus restaurant sound ambiances recorded at different moments of the day for different attendees numbers [17]. The behavior of the texturedness indicator I_{tex} for two different observation time durations, $T_{obs} = 3s$ and $T_{obs} = 6s$, is graphically represented in figure 4.

As expected, the academic noise signals obtained high texturedness scores and then are found on the top of the texturedness scale, whereas speech class had low texturedness scores and is then the least textured class. Classically known audio textures had, in their turn, high to middle texturedness degrees (see for example rain and group applause). For different short listening time, textured and highly textured audio signals preserve approximately the same texturedness value whereas non textured audio signals such as speech present different values due to their non stationary properties.

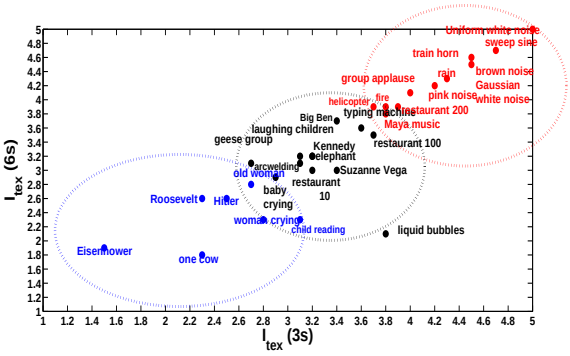

Fig. 4. A preliminary I_{tex} indicator classification into three large overlapping groups using K-means clustering algorithm. $T_{obs} = 3s$ and $T_{obs} = 6s$, frame size $20ms$, number of bins $M = 50$.

A robustness study has also been performed through small variations on its auditory parameters T_{obs} and T_f has shown good performances [18].

In the time range of few seconds (3s to 6s), which is linked to the human short term auditory memory capacity [19-20], the classification ability of I_{tex}, as a new audio feature, formally offers the class of relative audio texture signals a distinguished position on this 0 to 5 scale that lies between basic noise and speech/music. Therefore this new audio texturedness feature is in alinement with human "fuzzy" yet natural classification abilities.

5. CONCLUSION

In this paper, we have proposed an auditory-inspired audio texturedness indicator which evaluates the texturedness degree of a given audio signal on a 0 to 5 scale. This novel indicator is based on cumulative entropy tracking in the direct and reverse listening directions of the signal. Preliminary results on a large set of audio signals containing various audio categories were presented. These results showed a new audio classification scale for the audio continuum in which academic noise are on the top, then we find intermediate sounds with relative texturedness degrees. Low texturedness values were obtained in speech signals cases.

In support of this indicator, a subjective study was initiated to assess some perceptual parameters through the evaluation of the texturedness level in audio signals related to the listeners "spontaneous" scoring.

6. REFERENCES

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

[2] L. Lu, L. Wenyin, and H. J. Zhang, "*Audio Textures: Theory and Applications*", IEEE Transactions on Speech and Audio Processing, vol. 12, Issue 2, pp.156-167, 2004.

[3] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D.Lischinski, and M. Werman, "*Synthesis of Sound Textures by Learning and Resampling of Wavelet Trees*", Proceedings International Computer Music Conference, Beijing, China, 1999.

[4] A. Misra, P. R. Cook, and G. Wang , "*A New Paradigm For Sound Design*", Proceedings of the 9th Int. Conference on Digital Audio Effects, Montreal, Canada, September 18-20, 2006.

[5] G.Strobel, "Parametric Sound Texture Generator", master thesis, Austria, January 2007.

[6] D. Schwarz, "*State of The Art in Sound Texture Synthesis*", proceedings of International Conference on Digital Audio Effects, Paris, France, September 2011.

[7] R. Ghozi, W. El-Euch and M. Jaïdane, "*Two-dimensional Characterization of Audio Textures*", 3rd International Symposium in Video Communication, Hammamet, Tunisia, 2006.

[8] R. Ghozi, O. Fraj, and M. Jaïdane, "*Visually-based Audio Texture Segmentation for Audio Scene Analysis*", Proceedings of the 15-th European Signal Processing Conference, pp. 1531-1535, September 2007.

[9] R. Ghozi, O. Fraj, F. Hussein, and M. Jaïdane, "*Urban Soundscape Complexity Characterization via Audio-Visual Textures*", International Sound Act conference and Workshop, Aarhus, Denmark, 2010.

[10] M. Tuceryan, and A. K. Jain, "*The Handbook of Pattern Recognition and Computer Vision*", (2nd Edition), by C. H. Chen, L. F. Pau, P. S. P. Wang (eds.), pp. 207-248, World Scientific Publishing Co., 1998.

[11] F. Tupin, M. Sigelle, and H. Maître, "*Definition of a Spatial Entropy and Its Use for Texture Descrimination*", IEEE International Conference on Image Processing, Canada, September 2000.

[12] P. Brodatz, Texture: A Photographic Album for Artists and Designers, Reinhold, New York, 1968.

[13] K. Saberi, and D. R. Perrott, "*Cognitive restoration of reversed speech*", Nature (London), 398-760, 1999.

[14] S. Greenberg, and T. Arai, "*The relation between speech intelligibility and the complex modulation spectrum*", in 7th International Conference on Speech Communication and Technology, Scandinavia, pp. 473 476, 2001.

[15] UIT-T Rec. P.800.1, "*Mean Opinion Score (MOS) terminology*", 2003.

[16] R. Likert, "*A Technique for the Measurement of Attitudes*", Archives of Psychology 140: 1-55, 1932.

[17] R. Ghozi, O. Fraj, M. Jaïdane, and Mohsen Bel Haj Salem, "*Parametric Auditory Complexity Interpretation of Sound Ambiances in Confined Public Spaces*", transmitted to AES journal, (under revisions).

[18] O. Fraj,R. Ghozi, and M. Jaïdane, "*Audio texturedness indicator robustness study*", Internal report, Unité Signaux et Systèmes, March 2014.

[19] N. Cowan, "*The magical number 4 in short-term memory: A reconsideration of mental storage capacity*", Behavioral and Brain Sciences, vol 24, pp.87-185, 2000.

[20] A. Baddeley, " *Working memory : looking back and looking forward*", Nature review, Neuroscience, Vol 4, pp.829-839, October 2003.