Audio Engineering Society

Convention Paper

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Maintaining sonic texture with time scale compression by a factor of 100 or more

Robert C. Maher

Electrical & Computer Engineering, Montana State University, Bozeman, MT 59717-3780 USA
rob.maher@montana.edu

ABSTRACT

Time lapse photography is a common technique to present a slowly evolving visual scene with an artificially rapid temporal scale. Events in the scene that unfold over minutes, hours, or days in real time can be viewed in a shorter video clip. Audio time scaling by a major compression factor can be considered the aural equivalent of time lapse video, but obtaining meaningful time-compressed audio requires interesting practical and conceptual challenges in order to retain the original sonic texture. This paper reviews a variety of existing techniques for compressing 24 hours of audio into just a few minutes of representative "time lapse" audio, and explores several useful modifications and challenges.

1. INTRODUCTION

The increasing availability of audio recording devices capable of capturing hours, days, or weeks of continuous audio data has opened many new possibilities for sonic analysis, especially in understanding the diurnal and seasonal changes in natural sound environments [1, 2]. Although digital storage of such recordings is now widely available at a steadily decreasing cost, the notion that a human listener would have the time—and aural concentration and stamina—to listen to weeks of continuous audio seems highly unlikely. Thus, such long-term recordings will require automated processing and effective time-scale compression in order to be used and studied.

Typical time scaling techniques fall into two categories: time domain techniques such as the method of Roucos and Wilgus [3], and frequency domain techniques such as the phase vocodor [4]. For time scale compression (i.e., output_length = input_length/Factor) up to a factor of two or so, techniques that separate the temporal envelope from the underlying periodic excitation signal and scale them separately work quite well for speech and other quasi-periodic signals.

However, for time compression factors of 20, 50, 100, or more, the vast majority of the original audio temporal information must be eliminated. While this major time compression will likely destroy the intelligibility of a continuous speech recording, the more common situation in recordings of natural sound environments is a plethora of overlapping sounds with distinct patterns in frequency and time that create an ensemble typically referred to as a *sonic texture*. For example, a recording from a wilderness area or a backcountry location in a national park might contain subtle overlapping *background* sounds such as running water, bird calls, wind, insects, amphibians, high-altitude passenger jets, etc., with an occasional *foreground* sound like footsteps or human conversation. These sounds would come and go throughout the hours of the recording.

The desirable features of a time-compressed version of the recording would be to retain the sequencing and relative prevalence of the various sound sources in the audio time lapse version. An individual listening to the time-compressed version should be able to hear the individual sound sources and judge the overall sonic texture as being consistent with the original real time recording.

This concept is depicted in Figure 1. The normal time axis is mapped linearly to the compressed time axis with slope 1/N (dashed line), where N is the time compression factor. However, rather than applying uniform time compression to the entire segment, the algorithm identifies portions of the input signal containing foreground sounds for reduced time compression (segments t_0-t_1, t_2-t_3, and t_4-t_5), portions with consistent background texture (segments 0-t_0 and t_5-T_S), and portions consistent with the adjacent textures and therefore redundant (segments t_1-t_2 and t_3-t_4). This results in a time warping function that is still monotonically increasing, but no longer a simple straight line mapping.

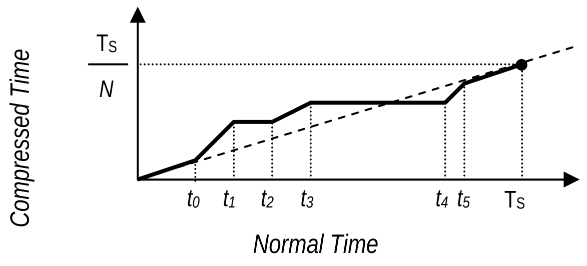

Figure 1: Time-compression by an overall factor N, with non-uniform warping to retain sound events.

Although the time warping function could conceivably be used to drive a conventional time-scale modification algorithm, the extreme time compression envisaged in this project (e.g., 60:1 or 100:1) would result in completely unnaturally compressed duration of individual sound events. Instead, a more holistic view and algorithm is needed.

There have been a variety of published papers in the area of major time-compression for audio. In addition to the techniques designed for modest compression factors of speech and music, there have been several notable efforts specifically dealing with sound textures. The contemporary composer R. Luke DuBois has coined the term "time lapse *phonography*" to describe his creative compositions involving sampled music segments [5]. Crockett and others have used the concepts of computational auditory scene analysis (CASA) to identify distinct sonic elements in a recording and to treat them individually [6]. In a complementary area, Frojd and Horner recently used the concepts of granular synthesis to create plausible sonic textures of auditory scenes such as applause and crowd noise, thereby synthesizing audio textures of arbitrary length [7].

One simple time-domain approach for extreme time scale compression is to do *block downsampling*. Short segments of the input signal are taken from widely spaced locations such that we select one segment out of N to produce the N:1 compression factor. Each short segment is windowed and then overlap-added to produce the time-compressed output sequence. The procedure begins by identifying a suitable block length, T, typically between a few hundred up to a few thousand milliseconds, which is chosen to be long enough to contain recognizable sounds, but short enough to allow suitable quality for the time lapse effect. The block downsampling method is shown in Figure 2.

The block downsampling approach is simple and straightforward to implement, but has the drawback that the arbitrary block segmentation does not take into account the details of the underlying signal itself, and therefore the sonic integrity of the output signal may be degraded. Nevertheless, it seems plausible that this drawback could be reduced by analyzing the input signal to select the location of the downsampling blocks to capture the individually recognizable foreground sounds, while still retaining the background sonic texture. This observation is the starting point for the work described in this paper.

The proposed method includes an automatic analysis/statistics stage and a synthesis/reconstruction stage. The automatic analysis stage determines a reasonably coarse description of the background sonic texture, and a reasonably fine segmentation of the foreground sonic events, which are relatively strong and presumably recognizable individual sounds.

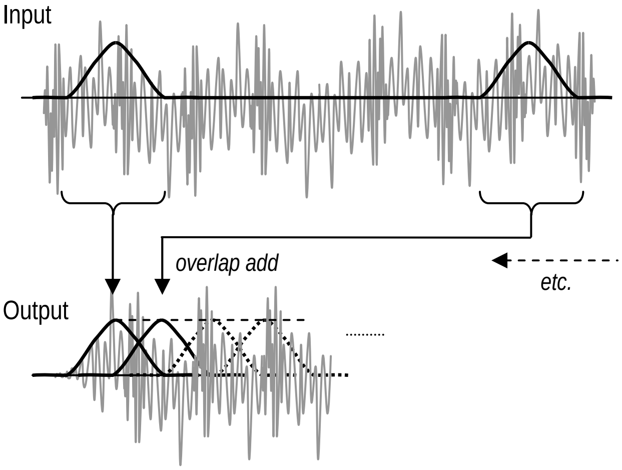

Figure 2: Simple time-domain block downsampling procedure.

2. SONIC TEXTURE: STATISTICS AND ANALYSIS

The goal of this work is to allow extreme time compression of audio recordings while maintaining sonic texture and distinctive foreground sounds. If we consider an example in which we seek a 60:1 compression, that is, keeping the equivalent of one minute out of every hour, we must blend or eliminate 98% of the temporal information while retaining the useful essence of the audio material. Thus, it is essential to have a practical and objective measure of the rather subjective concept of sonic texture.

For this investigation we start with a one second time interval and a 1/3rd octave frequency interval. Neither choice is necessarily optimum psychoacoustically, but the one second time interval and 1/3rd octave frequency interval are found to be sufficiently fine to capture the time-variant spectral character of the signal. Each one-second frame generates thirty 1/3rd octave band levels (20 Hz to 20 kHz range). The texture is defined by the time-variant fluctuation in the 1/3rd octave band levels.

The textural criterion that has been most useful has been to identify via automatic software analysis the frame-to-frame level differences in one or more bands that exceed a threshold, such as 1.5 dB. For each frame, the process looks at the next frame and at the subsequent five to ten frames to determine any repetitive fluctuations. The result is a map of the textural transitions.

As an example, consider the 1/3rd octave spectrogram of a 60 minute signal shown in Figure 3a. The primary background sonic texture of this signal includes a sustained low frequency component and a fluctuating mid-frequency component. There are also a few distinct events that are relatively short in duration, and some high frequency detail.

Applying the analysis algorithm with the frame-to-frame level difference criterion and a ten-frame median smoothing, the resulting background texture transition map is shown in Figure 3b. Note that the map indicates a few minutes of spectral continuity, punctuated by the distinct events and transitions. The goal is to retain the time segments with a high number of sonic texture transitions, at the expense of the segments with lower activity.

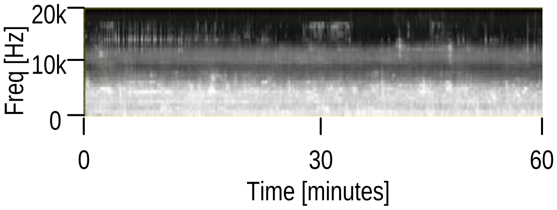

Figure 3a: Example 1/3rd octave spectrogram for 60 minute example signal.

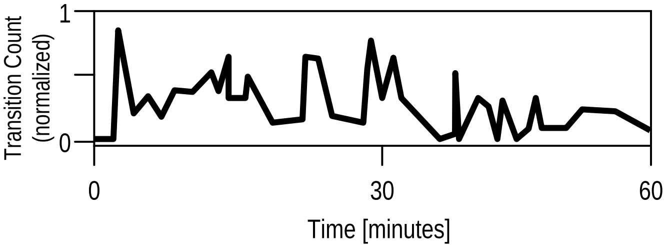

Figure 3b: Texture transition map analyzed from the spectrogram of Figure 3a.

Identifying the foreground sonic events is accomplished by the automatic software by looking longitudinally in groups of 1/3rd octave bands. If a particular band group shows an increase in level over one or two frames, the frame is flagged for event consideration. Applying this analysis to the example in Figure 3a produces the list of candidate foreground events shown in Figure 3c.

Once the entire length of the input audio file is processed, the rate of occurrence of the textural transitions and the number of foreground event flags is calculated. The idea is to look at the rate of occurrence of texture blocks and the timing of the event frames, and to attempt to maintain this rate after time compression.

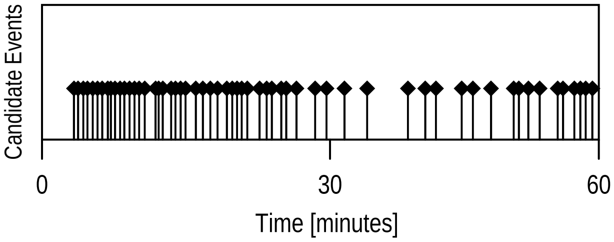

Figure 3c: Distinct sonic event map analyzed from the spectrogram of Figure 3a.

Reviewing the example of Figure 3a-c, the analysis indicates more than 50 distinct sonic events and at least a dozen texture transitions within the 60 minute signal segment. If the goal is a 60:1 time compression, i.e., on average each minute of the original recording becomes one second in the compressed version, the processing will focus on retaining the distinct events and the transitions, although it is clear that the duration of each event will need to be curtailed to fit everything into the time-compressed output.

3. LONG-TERM TEXTURE EXAMPLE

A more lengthy and complicated example will help illustrate the difficulties of the analysis procedure. A 24 hour segment of natural sound recorded at the Grant-Kohrs Ranch National Historic Site (Deer Lodge, Montana, USA) is shown in Figure 4, displayed as a 1/3rd octave spectrogram. Segment “00” in the top left of Figure 4 corresponds to the midnight – 1:00AM hour, segment “01” in the upper right corner corresponds to the 1:00AM – 2:00AM hour, and so forth. The brightness of the spectrogram details indicates the sound pressure level at that particular time and frequency range.

In order to allow a reasonable duration for audition, the 24 hours of audio ideally would be compressed into 12 minutes or so (120:1). This is reasonable for many of the hours in which the sonic texture is consistent and subtle, such as hours 00 through 05, but is difficult to handle for more lengthy sonic events, such as a passenger jet flyover during hour 12, and train whistles and track noise during hour 21 and during hour 23. Additional research is needed to determine the best way to time compress events like the flyover and the train sounds that last many minutes in normal time, but cannot occupy more than a few seconds in the compressed-time regime. This remains an open question.

4. SEGMENTATION AND ASSEMBLY

With 98% or more of the temporal audio frames discarded by extreme time scale compression, the segmentation and reassembly requires both elimination of redundant frames and seamless mixing and concatenation of the frames that are essential to maintain the sonic texture. In many examples the challenge is that the analysis detects more sonic events and texture transitions than can be accommodated in the time compressed output. Some priority ranking must be deduced or inferred to decide which events and transitions will be left out of the compressed result.

The current solution to this issue uses the windowed overlap-add to recreate the time compressed output signal. The assembly and reconstruction occurs in the time domain, with the distinct sonic event map guiding the selection of high-priority frames, and the texture transition map guiding the selection of background frames to fill in the required duration between the foreground sonic events.

5. CONCLUSION

Extreme time scale compression of audio material for time-lapse purposes has no naturally occurring counterpart, so establishing valid metrics for assessing quality and fidelity is largely a subjective process. Future work will be needed to quantify the degree to which the time compressed output signal retains the texture and sonic character of the input signal. In the meantime, the basic compression strategy and goals described in this paper have been found to be useful and practical.

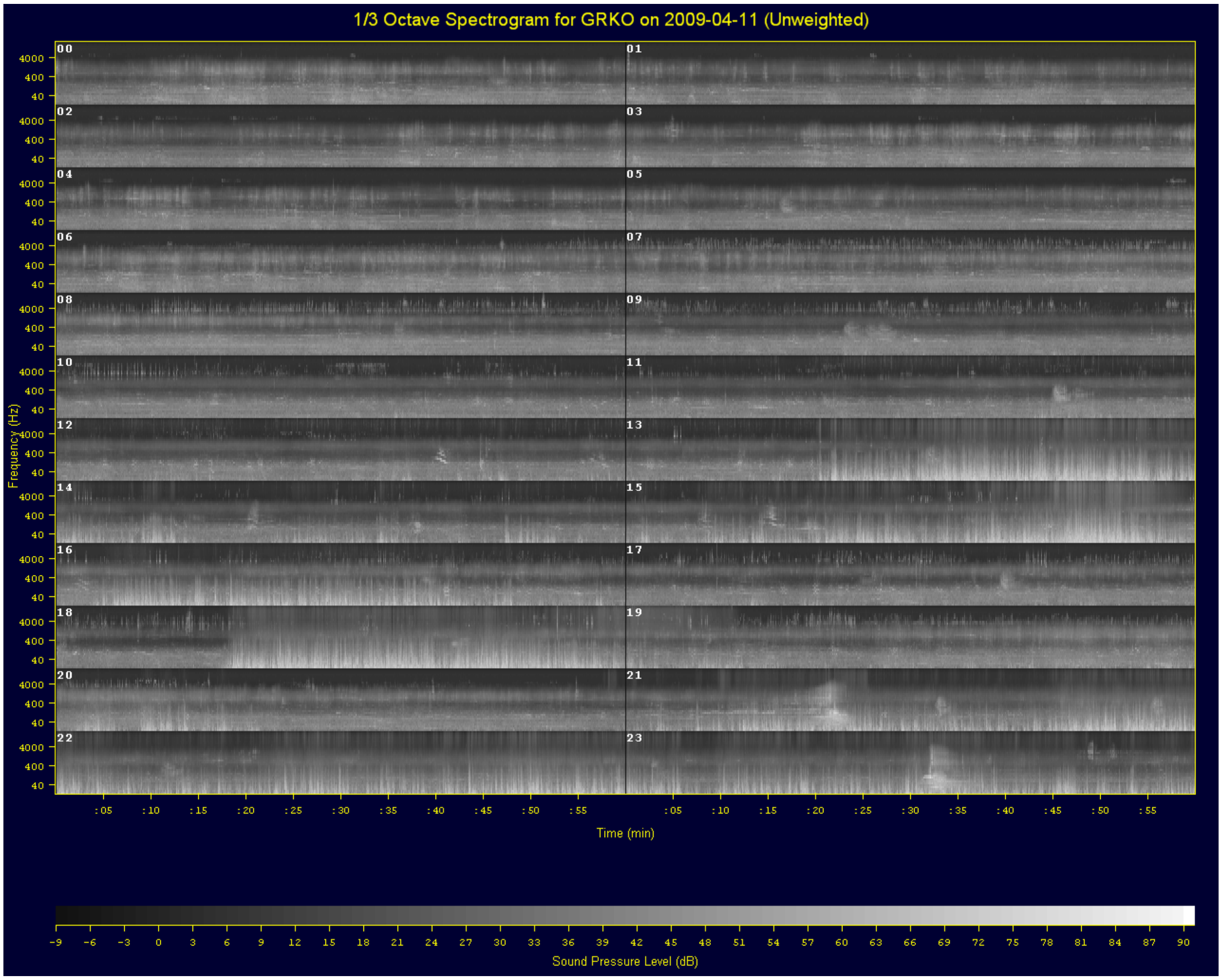

Figure 4: 24-hour spectrographic summary of the Grant-Kohrs Ranch monitor site on 2009 April 11.

6. ACKNOWLEDGEMENTS

This work was supported in part by the U.S. National Park Service through a contract with the Rocky Mountain Cooperative Ecosystem Studies Unit (RM-CESU). The author gratefully acknowledges the assistance of Chris Ford, Grant-Kohrs Ranch National Historic Site (Deer Lodge, Montana, USA), and Damon Joyce, National Park Service Natural Sounds Program (Fort Collins, Colorado, USA).

7. REFERENCES

[1] Krause, B., "Anatomy of the soundscape: evolving perspectives," J. Audio Eng. Soc., vol. 56, no. 1/2, pp. 73-80, Jan/Feb 2008.

[2] Maher, R., "Acoustics of national parks and historic sites: the 8,760 hour MP3 file," Proc. AES 127th Convention, preprint 7893, Oct 2009.

[3] Roucos S. and Wilgus A.M., “High quality time-scale modification for speech”, Proc. IEEE ICASSP, pp. 493-496, March 1985.

[4] Laroche, J., and Dolson, M., “Improved phase vocoder”, IEEE Trans. Speech and Audio Processing, vol 7, no. 3, pp. 323 –332, May 1999.

[5] DuBois, R. Luke, http://lukedubois.com/

[6] Crockett, B., "High quality multi-channel time-scaling and pitch-shifting using auditory scene analysis, Proc. AES 115th Convention, preprint 5948, Oct 2003.

[7] Frojd, M., and Horner, A., "Sound texture synthesis using an overlap–add/granular synthesis approach," J. Audio Eng. Soc., vol. 57, no. 1/2, pp. 29-37, Jan/Feb 2009.