High-Performance Tools to Generate and Visualize a Sonic Time-Lapse

Diego Espejo
Instituto de Acústica
Universidad Austral de Chile
diego.espejoa@gmail.com

Víctor Poblete
Instituto de Acústica
Universidad Austral de Chile
vpoblete@uach.cl

Pablo Huijse
Instituto de Informática
Universidad Austral de Chile
phuijse@inf.uach.cl

Felipe Otondo
Instituto de Acústica
Universidad Austral de Chile
felipe.otondo@uach.cl

ABSTRACT

*This paper describes an algorithm to generate a Sound-Time Lapse (STL), i.e., audio capsules or acoustic summaries that aim to reliably represent continuous and long audio signals recorded in wetlands from the Chilean city of Valdivia. The algorithm's input is a long duration audio signal ($>$ 20 hours), from which representative chunks are extracted and merged with minimum audible distortion to produce a brief audio file (*4-6 *minutes). The algorithm is a flexible and adaptable tool, its input parameters can be adjusted by the user to highlight specific portions of the original long recordings. Modern libraries focusing on high performance computing were considered for the implementation of the algorithm. The project aims to highlight the importance of the natural heritage of Valdivia's wetlands and bring them closer to the general public. Additionally, we recognize the opportunity to use STLs in scientific works so as to improve our understanding on biological diversity present in the natural sound composition of wetlands.*

1. INTRODUCTION

Located in the city of Valdivia in the south of Chile, wetlands are known as transition areas between earth and aquatic systems [1]. Wetlands' vast biological diversity provides the city's wildlife with water and shelter [2]. These wildlife environments are characterized by partial flood, where vegetation acts as a natural filter, purifying the water from toxic waste such as pesticides [3]. A large diversity of sounds are produced within the wetlands borders. These natural acoustic sources help us understand fragile structures of our planet. The life of amphibians such as frogs and toads depends on wetlands environments. The features of these animals allow them to live in earth and aquatic environments, they cannot travel long distances and find in these contexts the perfect places for mating [4]. During spring, there are common night singing birds that make use of complex sounds to attract their partners, defend their territories and also to alert others, communicate, and receive information [5]. The motivation of this research is to present to a wider audience in an open and creative fashion the various kinds of sonic interaction occurring in these kinds of wildlife environments. With this motivation in mind high-performance tools were designed and implemented to generate *Sound-Time Lapses* in order to make these interactions visible. The research presented here has two main aims. The first goal was to generate sonic capsules extracted from long wetlands' recordings that are both brief and rich in terms of bioacoustic content. The second goal was to generate high quality spectrograms of the long-duration field recordings ($>$ 24 hours), and also the sonic capsules. Future developments of the project will consider using the spectrograms as the basis for modeling the acoustic diversity of wetlands [1]. Specially tailored spectrograms and their use as a graphic visualization tool could help users interested in exploring morphological similarities and differences between the long-duration original spectrogram and the summarized time-lapse audio versions. Additionally, spectrograms and their visual interactive nature could also help to select, listen to, and label specific wetland acoustic events, such as rain, birds, or frog sounds.

2. RELATED CONCEPTS

Visible features that could be extracted from a landscape are varied: surface dimensions, geographical shape, fauna, flora, various kinds of morphology and human activity. These features are just a part of the existing characteristics of landscapes. Beyond what we can see, there are relevant complementary characteristics to a landscape, those are the audible features that are perceived by the hearing community, for instance, periodicities, spectral, and temporal variabilities. The concept of soundscape is used to describe the relation between a landscape, the composition of its sounds and the hearing community [6]. Wetlands sound variabilities change in spatial and temporal scales. Also, soundscape varies throughout the day and during each season. Extracting and identifying these variabilities in wetlands is vital if we aim to characterize and model such soundscape. In order to share this extensive wetland sound information, we require computational tools or platforms that facilitate reaching this goal.

2.1 Wetland signal spectrogram

Due to the nature of the wetland, the energy of the various oscillatory components exhibits a dynamic behavior. To

estimating time-varying spectra, we subdivided the entire signals into overlapping windowed signals called frames, and estimated the spectrum locally for each frame using short time Fourier transform based method, assuming the underlying process is quasi-stationary, that is, the local spectrum changes slowly with time. Thereby, the spectrogram was obtained by employing sliding windows with overlap in order to capture non-stationarity. The window length, the window function, window overlap and number of frequency bins in each window, were carefully chosen. An inherent characteristic of the wetlands is that the sound classes recorded come from different sources, such as animal vocalizations (*e.g.* dog vocalizations greater than 100 Hz), vehicle motors, rain, and wind, and exhibit certain shapes or marks when viewed as a spectrogram. Such sounds are characterized by broadband spectra. The energy is distributed at many different frequencies, with several sharp peak frequencies, transitions among harmonic sounds, and exhibits high frequency modulation throughout the entire signal. In order to get a desired time and frequency resolution, and bandwidth of the spectrogram, they were made using sampling rate $sr = 44.1$ kHz, then by the Nyquist-Shannon sampling theorem, signals will can contain frequency content up to 22050 Hz, a window length $L = 512$ samples (11.6 ms), a Hanning window (it has good frequency resolution and reduced spectral leakage), with a window overlap 50%, a fast Fourier transformation of $N = 512$-point, yielding frequency bin resolution of 43 Hz/bin ($\Delta f = (sr/2)/N$) and time resolution of 2.3 ms ($\Delta t = N/(sr/2)$).

2.2 Sound-Time Lapse (STL) design and optimization

The aim of constructing an algorithm that produces an STL, which originally receives a long duration audio signal (>24 hours), is generating a shorter new audio signal as output. This output is an audio capsule of the original signal, for instance, a 4 to 6-minute version. Thus, a sonic capsule signal provides a rough audible description of the recorded soundscape with minimum audible distortion, allowing us to listen to a short summary of the acoustic events that take place at a particular location. The STL design must reach a sonic overlap that is automatic and chronological, generating a sequence of short duration audio signals, which we call chunks. Additionally, the dynamic transition between two consecutive overlapped chunks should be as perceptually smooth as possible. In order to make these transitions subtle and gradual, different types of dynamic crossfade windows of various lengths are tested. This requires the use of a period of time in which the algorithm creates a gradual crossfade between the first chunk and the second chunk. For example, if we determine a crossfade window length of 2 seconds, the algorithm should attenuate the last 2 seconds of the first chunk, while the 2 first seconds of the second chunk should gradually appear. For this reason, the dynamic crossfade window shape is a highly relevant parameter in this task. In order to perceive a smooth aural mixing between two consecutive chunk transitions, we impose the following condition in the STL optimization: that during the transition the RMS level (in dB) is preserved between both chunks. This would increase the probability of avoiding perceptual audible variations between consecutive chunks.

2.3 Sound-Time Lapse duration

In the algorithm, n consecutive chunks of a signal are modeled, as composed by four parts: a Fade-in, a Constant, a Fade-out, and a Crossfade as shown in Fig. 1 in the case of n=2. In order to create the final STL audio file, the algorithm collects n chunks periodically each of length L_{Ch}, defined as $L_{Ch} = (L_C + 2L_W)$, in which L_C is the constant part length (Constant) and L_W is the Crossfade window length, or equivalently, the Fade length. If the starting point of chunk recordings given by P_0 is used as a variable together with the total original wetland signal length given by L_T and the chunk recording period such as ΔP, we will obtain the total number of fragments that STL will generate:

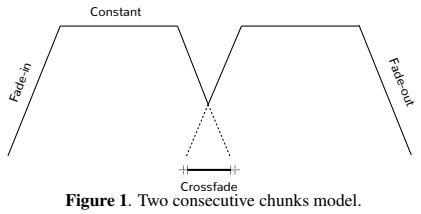

Figure 1. Two consecutive chunks model.

$$n = \frac{L_T - P_0}{L_{Ch} + \Delta P} \tag{1}$$

The initial point for the i-th chunk has the general formula:

$$P_i = P_0 + (i-1)(L_{Ch} + \Delta P) \tag{2}$$

Once we know the n value, empirically, Eq. (1) allows us to estimate the duration in seconds of the sonic capsule, named L_{STL}, where L_{Win} is the starting point of it, L_C is the constant length, L_W is the Crossfade window and L_{Wout} is the Fade-out length:

$$L_{STL} = L_{Win} + n \cdot L_C + (n-1) \cdot L_W + L_{Wout} \tag{3}$$

3. STL ALGORITHM

In practice, and due to the way in which wetlands recordings take place, the resulting STL audio file is fragmented in K consecutive smaller audio files. Hence, the number of fragments is:

$$n = \sum_{i=1}^{K} \frac{L_T(i) - P_0}{L_{Ch} + \Delta P} \tag{4}$$

where $L_{T(i)}$ is the total temporal length of the i-th fragment.

3.1 Input Parameters

As exposed in Fig. 2, in order to characterize a wetland through an STL, a linear system is assumed which requires the recorded audio signal and a set of parameters as input values. The first one is the starting point or P_0 (in minutes). The second one is the chunk duration or L_{Ch} (in seconds). The third one is the chunk recording period or ΔP (in minutes). The fourth one is the Crossfade duration o L_w (in seconds). To select an appropriate duration of L_W, according to the above definition, it must not be longer than $L_{Ch}/2$. Moreover, the fifth one is the crossfade window

shape or F, which defines the kind of the crossfade function. Although there are numerous options of crossfade shapes to overcome the problem of transitions between two concatenated segments, in this work we put special emphasis on the use of linear, exponential or logarithmic. Of course, the human auditory system performs operations far more complicated than detecting sudden changes in energy, but these three were chosen to be implemented based on their performance with respect to maintain constant the perceptual energy during the transitions. Nevertheless, we are using these shapes as inspiration for the STL method, rather than implementing a complete model for all numerous options of shapes.

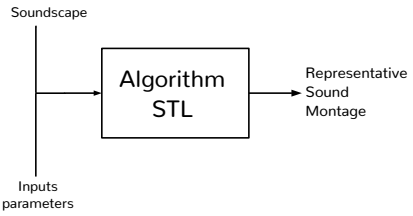

Figure 2. Construction of a representative STL version.

3.2 Algorithm Structure

Once the input parameters are defined, we can observe that Figure 3 describes a transformation sequence: $T_i\,|_{i=1}^{6}$ that linearly modifies X(t) signal of the soundscape. The Y(t) output represents the resultant STL audio signal. A description of each transformation is the following:

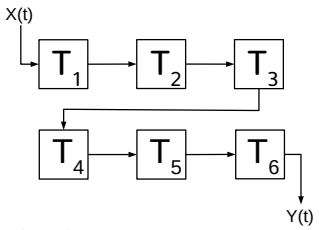

Figure 3. Signal transformation sequence.

As shown in Figure 3, T_1 represents the reading and preliminary data collection stages. The K fragmented consecutive audio files, provide the information related to the number of channels in which the audio was recorded, they also provide the value of each $L_{T(i)}$. T_2 transforms the temporal information from seconds to samples. A sample frequency of 44100 samples per second and a quantization rate of 16 bits were used in the standard WAV format. T_3 represents the Crossfade window shape selection stage, and also, the application of the selected Fade-in, Fade-out, and Crossfade. As an example, a logarithmic crossfade is used to ilustrate in Fig. 4. This is defined between two chunks. Three sections in the figure are shown: Fade in, Crossfade, and Fade out. Fade in is the time interval during which the amplitude increases. Fade out is the time interval during which the amplitude decreases. Crossfade is an interval during which the signal transition between two chunks is expected to be as gradual as a possible. In the T_4 stage, the Ec. (4) is used to compute the number of chunks that the STL will have and whose value depends on the input parameters and preliminary data of each i signal. T_5 is the segmentation stage, i.e., chunks of the signal are selected based on the input parameters (see Fig. 5). Then, T_6 stage represents the superposition of the selected chunks using Crossfades (see Fig. 5).

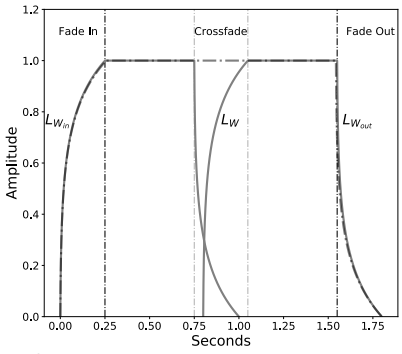

Figure 4. A logarithmic crossfade defined between two chunks.

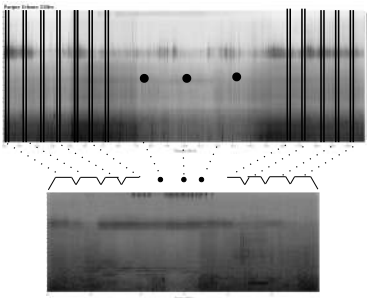

Figure 5. Signal chunks and crossfading.

3.3 User Interactivity Tools

We implemented an interface to visualize and interact with the sound information of the spectrogram through a Web browser. The interface and STL algorithm are distributed as an open source project written in `Python` and available at GitHub [1] . We use the `Python` libraries `Kivy` to implement the file selection process, `Librosa` to compute Fast Fourier Transform (FFT) and `Holoviews` and `Bokeh` to produce dynamic and interactive graphics visualization. We use `HoverTool`, a function of `Bokeh`, that allows for the mouse cursor to move freely inside a region of the spectrogram showing values of time (in seconds/minutes), frequency (in Hz), and sound intensity (in dB). Spectrograms are plotted using a logarithmic frequency scale resembling the way in which auditory perception works, representing also, in a better way, the lowest section of the spectrum.

4. STL APPLICATION

We test the algorithm using 6 consecutive field recordings audio files obtained from Parque Urbano wetland, located in the city of Valdivia. The recordings were carried out on March 19th, 2019. The total duration of the original continuous field recording is approximately 21 hours. As an example, we choose the parameters shown in Table 1

[1] `https://github.com/Nigglea/EspectrogramaUACh`

for illustrating a STL application. Additionally, the three types of crossfade window shape were tested.

P_0	L_{Ch}	ΔP	Lw
Min 10	22 Seg	60 Min	6 Seg

Table 1. Input parameters.

4.1 Results

Using the information contained in Table 1, and Eqs. (3) and (4), the values shown in Table 2 were obtained.

Total duration	20 hrs 47 min 12 seg
STL duration	5 min 42 seg
Chunk number	21
Crossfade window	6 seg
Constant	10 seg

Table 2. Application results.

We created three STL audio files each corresponding to a one different crossfade window shape. Then, we computed their respectives spectrograms and the corresponding spectrogram from the continuous wetland signal. Despite the fact that the time scale of the wetland signal is $\sim$ 21 hrs., and the time scale of the STLs is $\sim$ 6 min, we observed visually morphological similarities. To measure which crossfade window shape showed great similarity with respect to continuous wetland signal, we calculated the temporal alignment of the variation of the energy frame-by-frame between the wetland signal and each of the STL signals, using the dynamic time warping method [7]. The results were normalized by the minimun temporal alignment and showed that exponential shape was 1.0, while logarithmic and linear shapes were 1.002 and 1.005, respectively. Figures 6 and 7 show the spectrograms both of the original continuous field recordings, and the resulting exponential STL audio file, respectively.

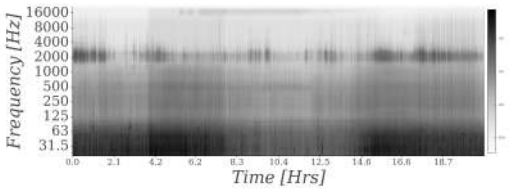

Figure 6. Wetland spectrogram, total duration of input file is $\sim$ 21 hrs.

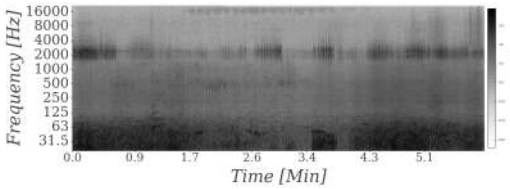

Figure 7. STL Wetland spectrogram, total duration of STL is $\sim$ 6 min.

5. CONCLUSIONS

The work presented here describes a Sound-Time Lapse algorithm that combines soundscape signal processing techniques and the use of high-performance tools to generate spectrograms and visualize acoustic information of wildlife sonic environments. Future developments of the project aim to optimize the algorithm by implementing non-periodic fragment collection and novelty detection techniques.When comparing the spectrogram morphologies between the original long field recording and the generated STL signal discussed above, considerable similarities were observed. In order to improve the application further, future versions of the algorithm will integrate machine learning-based methods to automatically learn representation of the rich acoustic diversity of wetland soundscapes.

Acknowledgments

The research that led to this paper was funded by the Chilean National Commission for Scientific and Technological Research under grant FONDECYT 1190722 and 1170305. The authors would like to thank Rodrigo Torres and Víctor Vargas for their help to carry out the research activities presented in this article.

6. REFERENCES

[1] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe, "Assessment of convolution neural networks for wetland mapping with landsat in the central Canadian boreal forest region," *Remote Sensing*, vol. 11, no. 7, Apr 1 2019.

[2] Fundación Cosmos, "World Wetlands Day 2017: what is celebrated and why," https://fundacioncosmos.cl/noticias/dia-mundial-los-humedales-se-conmemora/, 2017.

[3] The Convention on Wetlands, called the Ramsar Convention, "Wetland: Values and Functions - Water Purification," https://www.ramsar.org/, 2001.

[4] A. Vélez, N. M. Gordon, and M. A. Bee, "The signal in noise: acoustic information for soundscape orientation in two North American tree frogs," *Behavioral Ecology*, vol. 28, no. 3, pp. 844–853, 03 2017.

[5] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, "Soundscape ecology: The science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 03 2011.

[6] M. Southworth, "The sonic environment of cities," *Journal of the Environment and Behavior*, vol. 1, no. 1, pp. 49–70, 1969.

[7] V. Deecke and V. Janik, "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," *Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 645–653, Jan 2006.