Audio Engineering Society

Convention Paper

Presented at the 135th Convention
2013 October 17–20 New York, NY, USA

Sound identification from MPEG-encoded audio files

Joseph G. Studniarz and Robert C. Maher

Department of Electrical & Computer Engineering, Montana State University, Bozeman, MT 59715 USA
rob.maher@montana.edu

ABSTRACT

Numerous methods have been proposed for searching and analyzing long-term audio recordings for specific sound sources. It is increasingly common that audio recordings are archived using perceptual compression, such as MPEG-1 Layer 3 (MP3). Rather than performing sound identification upon the reconstructed time waveform after decoding, we operate on the undecoded MP3 audio data as a way to improve processing speed and efficiency. The compressed audio format is only partially processed using the initial bitstream unpacking of a standard decoder, but then the sound identification is performed directly using the frequency spectrum represented by each MP3 data frame. Practical uses are demonstrated for identifying anthropogenic sounds within a natural soundscape recording.

1. INTRODUCTION

Audio monitoring for noise levels at a specific location is not uncommon. However, analyzing these recordings is a much more difficult task. As researchers record audio data at a particular location, the analysis time of these sounds becomes exponentially greater as the recording length increases. While analysis software that provides basic statistics is common, little exists to actually identify and classify the sounds in the recording. In soundscape analysis and audio forensics investigation, this issue is continuing to grow as the ability to record audio data long-term becomes more feasible with inexpensive equipment. These long-term recordings can be created anywhere and contain all kinds of natural and culturally created sounds [1]. While the possible applications for analyzing long-term recordings are vast, the focus in this paper will be directed toward the National Park Service's interest in long-term recordings and analysis in the National Parks.

1.1. National Park Service and Natural Sound

The National Park Service Natural Sounds and Night Skies Division (NSNSD) is interested in scientifically measuring background sound levels in the parks and determining how the levels of cultural sounds affect the environment. The National Park Service describes these interests in their *Management Policies 2006* report, "The Service will restore to the natural condition wherever possible those park soundscapes that have become degraded by unnatural sounds (noise), and will protect natural soundscapes from unacceptable impacts" [3]. Later in the report, the National Park Service also states, "The Service will preserve soundscape resources and values of the parks to the greatest extent possible to protect opportunities for appropriate transmission of cultural and historic sounds that are fundamental components of the purposes and values for which the parks were established" [3].

The National Park Service has set high goals to preserve the National Parks as much as possible, including the soundscapes these parks contain. The Service has conducted many short term monitoring projects to investigate cultural sound impacts, but a research bottleneck always occurs in the corresponding quantity of analysis time required. Consequently, short term monitoring is not sufficient to draw meaningful conclusions from these projects. In almost all instances, one monitoring location is insufficient because of the size and soundscape diversity of the National Parks. Rapid analysis of long-term recordings is strongly desired to carry out single and multi-location monitoring investigations.

To minimize initial complexity, a smaller site to test out rapid analysis methods on one long-term recording was chosen. The Grant-Kohrs Ranch is a National Historic Site in Deer Lodge, Montana where no previous characterizing data has been collected to document the effects of natural and cultural sounds [3]. Researchers from Montana State University recently performed an audio study of the soundscape at the ranch from March of 2009 through March of 2010. During this time span, MP3 audio data was recorded using a flash memory-based digital recorder and a single (mono) microphone.

Therefore, the audio analysis can be based on the structure of the MP3 audio file. All sounds are composed of multiple frequencies and the composition of these frequencies is what gives an object a unique sound. To find these unique differences, the compressed audio can be transformed from the time domain into the frequency domain. This technique to analyze the compressed MP3 audio can be utilized to quickly search through the data collected at The Grant-Kohrs Ranch Site.

1.2. The GRKO Site

Grant-Kohrs Ranch National Historic Site (GRKO) is located in Deer Lodge, Montana. This site provides some unique opportunities for an audio study. The site is located 1.5 miles away from the Deer Lodge airport, 0.7 miles from an interstate highway, and less than 0.5 miles from a railroad [2, 3]. The site is also relatively small and homogenous, making it an ideal candidate for a single point monitoring investigation.

1.2.1. Data Collection

Equipment for performing the long-term recordings were provided by the NPS NSNSD and set up at the site. The system used a 12-volt battery that was recharged by a photovoltaic panel. The battery's capacity allowed the system to run even if there was no sunlight for several days from overcast skies. A sound level meter was also used to capture 1/3rd octave audio levels once every second. This data was recorded in a comma separated values format that can be analyzed with a spreadsheet program or other custom software. Continuous audio was recorded in the MP3 audio format during this time frame on a flash recorder using a single microphone. The flash recorder captured audio continuously in blocks of 9 hours. Data was organized into these blocks to prevent extensive data loss in case of a technical malfunction. All of the MP3 files together form 8,760 hours of continuous audio that can be sequentially analyzed. Rapid analysis of the audio data will provide statistics of the natural and cultural sound levels [1, 3].

1.2.2. Other example Collection Sites

In 2005, the National Park Service set up guidelines for acoustical studies in the parks. At that time, power and storage limitations made long-term continuous recordings impractical. Short sample recordings were made instead for 10 seconds every 2 minutes. For comparison, another acoustical study performed by the NSNSD at Sand Creek Massacre National Historic Site utilized data collection equipment similar to the equipment used at GRKO. At the Sand Creek site, a sound level meter to collect 1/3rd octave audio levels was the primary instrument used for collection. A continuous audio recorder was also used to collect MP3 data between 2009 and 2011. However, during post-analysis of this data, researchers listened to the audio for 10 seconds every 2 minutes [7]. While this is sufficient to draw meaningful collusions for frequent sounds, a more thorough analysis technique is still needed.

1.3. The MP3 Audio File

All of the continuous audio data collected at GRKO was recorded in the MP3 format. Therefore, the audio analysis is based on the structure of the MP3 audio file. MPEG-1 Layer 3 (MP3) is an audio format that utilizes lossy compression, with uncompressed audio being converted into the MP3 format. MP3 was developed and deployed by the Moving Pictures Experts Group (MPEG) in the mid 1990's. MPEG was formed by ISO, the International Organization for Standardization and IEC, the International Electrotechnical Commission in 1988 [10]. Other forms of audio compression exist, such as Windows Media Audio (WMA), but MP3 is a widely accepted and published standard. This allows for alternative analysis methods to be developed and tested on the MP3 files collected at GRKO.

To encode a standard MP3 file, audio is first sampled from the original audio file at a specified sample rate. Next, the sampled audio is passed through a filter bank that converts audio from the time domain into the frequency domain. The sampled audio is also passed through a psychoacoustic model. This model adjusts for sounds that are frequency masked by the human ear. When multiple sounds occur simultaneously, the human ear has limited detection of multiple simultaneous sounds. The psychoacoustic model allows for extra pieces of data to be eliminated and the output sound remains unchanged to the human ear. Next, the audio is organized and separated into frames that compose the compressed bitstream of the MP3 file [9].

To decode an MP3 file, the compressed bitstream is converted back to the frequency domain. The frequency domain is broken up into 576 different frequency "bins". In normal operation, the decoding process continues on to recreate the time domain audio from data in the frequency domain [9]. However, the frequency domain is the key focus of interest for the audio analysis methods that are used, so continuing the decoding process is not necessary. The step after the reordering of the frequency spectrum is the area of interest for our investigation.

2. AUDIO ANALYSIS FROM PARTIALLY-DECODED MP3 DATA

The MP3 audio at GRKO was sampled at 44.1kHz, encoded at 16-bit, and output at a bitrate of 64 kbps [3]. For analysis, the technique to partially decode the MP3 audio can be utilized. During the MP3 decoding process, the frequency domain is revealed. These frequencies can be analyzed directly and the decoding process will stop here and the identification method will begin at this point in the program before moving on to the next frame in the bitstream. This prevents extra data from being written back to the disk for identification later since the process is in succession with the short-circuited decoding process. There is also no need to convert the frequency data all the way back to an uncompressed audio waveform since the frequency domain is the main point of interest. Figure 1 below shows a high-level view of the analysis process. After the reordering of the frequency spectrum, the identification process begins. Identified sounds are recorded and the process is repeated until the end of the bitstream is reached. The decoding process could be continued after the reordering of the frequency spectrum to create an uncompressed audio file, but this requires additional computation time. For even faster analysis, the current software searches one out of every five frames.

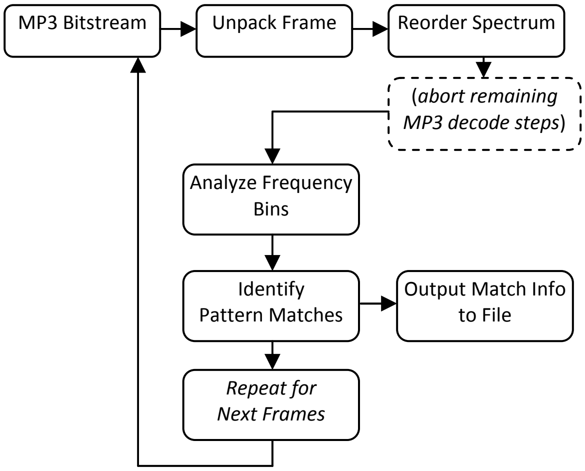

Figure 1: Overview of MP3 spectral data examination

Other researchers have utilized a similar method for analyzing compressed audio data by intercepting the decoding process and examining data in the frequency domain. Performing the analysis directly on the compressed audio allows for faster analysis by exploiting the previous calculations performed during the encoding [5]. Also, a United States patent application submitted in 1998 summarizes a similar searching technique using compressed audio data. The technique described in the patent was used specifically for fast speech recognition and speaker identification [6].

2.1. Analysis Software and Hardware

The MP3 file format is well known, so there are a variety of software programs that are open source for decoding MP3 files. Instead of writing our own MP3 decoder, an open source decoder was chosen instead. Several programs were considered, but ultimately, the “MPEG Audio Decoder” (MAD), was chosen for its simplicity and organized source code. MAD is managed by UnderBit Technologies, Inc. and is available under the terms of the GNU General Public License [11]. The source code was then modified, changed, and extended to achieve the results described in this paper.

The software is installed on a Unix based system, Mac OS X running on a 2009 MacBook Pro. The notebook has a 2.8Ghz Intel Core 2 Duo, 4GB of 1067 DDR3 memory, and a 7200-rpm hard drive. For extensive and prolonged use of the analysis software, faster computer hardware would decrease analysis time greatly. For initial testing and prototyping, the current hardware is sufficient.

2.2. Interpretation

After the frequency spectrum is reordered, the frequency domain is accessible. The frequencies are compared to values that depict a visual representation of the data [2, 8]. Natural and cultural sounds are composed of multiple frequencies, and a composition of multiple frequencies is unique to objects that have different sounds. To identify the different frequencies emitted by objects, a spectrogram can be used. The spectrogram provides a visual representation of all the frequencies recorded and displays all 576 frequency “bins” vertically vs. time.

Figure 2 shows a spectrogram of approximately 30 seconds of audio collected from GRKO on the morning of July 3rd, 2009. On the spectrogram, light shades represent spectral energy and dark areas represents the absence of energy at that time and frequency. Frequency is plotted over time and the frequencies increase from low to high on the vertical axis. Across the top, several short inflections show a bird chirping close to the microphone. The energy in the lower half is noise from the interstate highway.

Figure 3 shows a spectrogram from the same day, but these data were collected later that evening around dusk. The highway noise is considerably less than in the morning. The sounds that appear in this spectrogram are crickets and other birds chirping. The large spike near the end is a branch cracking under the weight of a larger animal at the ranch.

2.3. Detection

The ability to visualize frequency data with spectrograms assists in the process of setting parameters to identify specific sounds. The spectrograms of Figures 2 and 3 provide helpful views of the overall frequency spectrum, but specific levels in the frequency “bins” are difficult to interpret from a pattern-matching standpoint.

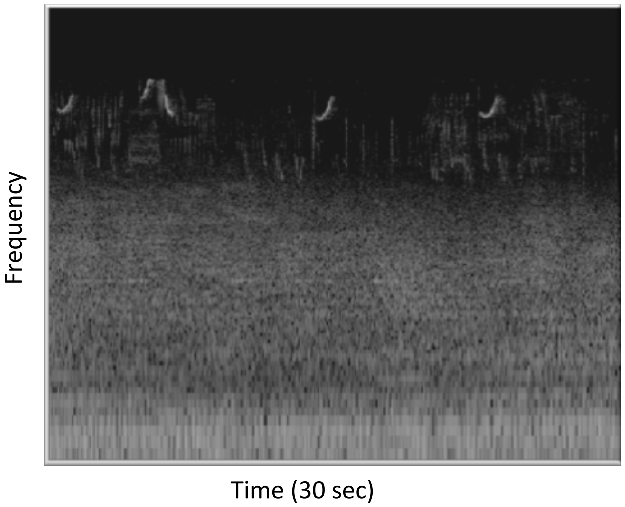

Figure 2: MP3 spectral data, GRKO site, July morning, 30 second excerpt

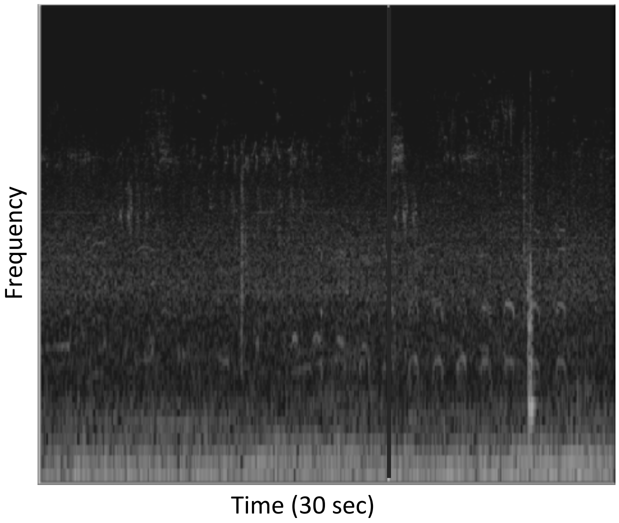

Figure 3: MP3 spectral data, GRKO site, July evening, 30 second excerpt

The same frequency data can be displayed differently to more closely view the levels in specific frequency ranges. Instead of plotting all 576 frequencies vertically vs. time, the amplitude of each individual frequency bin can be examined as a function of time.

As mentioned previously, there is a railroad that traverses a portion of the Grant-Kohrs site, so one specific and easily recognizable sound is the noise of the locomotive and train cars passing by. A visual representation of the frequency spectrum for a train passing by the site is shown in Figure 4. The gray shades represent very low frequencies from the rumble of the train. The four light gray peaks indicated by arrows near the beginning represent higher frequencies that are four horn blasts from the train at a nearby railroad crossing. All of the shorter peaks at the bottom of the plot are higher frequencies that contain relatively little energy for the train sound compared to the low frequency bins.

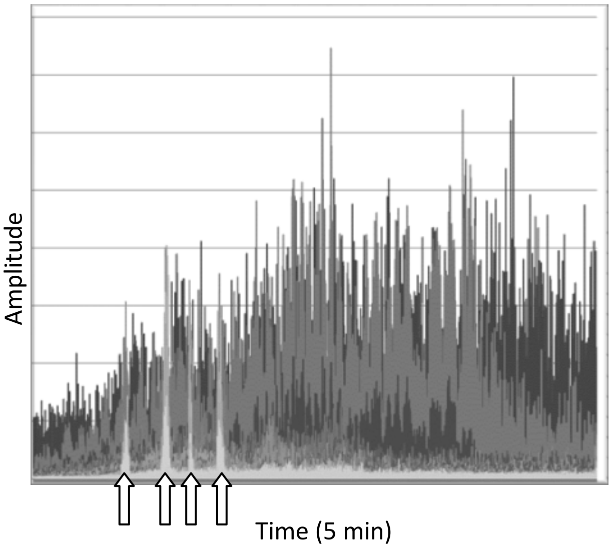

Figure 4: MP3 spectral data, energy level predominantly in the low frequency range as a function of time for a passing railroad train

The frequency bins are what the analysis program "sees" as it searches through the frequency domain. By setting spectral threshold trigger points for this data, the program is able to interpret the time-variant spectral content of the MP3 file without completely decoding the file. This allows the program to search for specific spectral sound template patterns at a rapid rate.

The plot in Figure 4 allows for some initial empirical amplitude thresholds to be set in software to detect a passing train. Portions of the code in the program are triggered when several frequency bins exceed a specific value.

For example, if the darker gray frequency "bins" in Figure 8, which correspond to low frequency energy, both exceed a particular amplitude threshold several times in a specified time interval, the detection algorithm will infer that a train is passing by. However, this sort of simple threshold rule is fragile, and resulted in many false triggers because simple thresholds alone do not exclusively trigger the likelihood of a passing train.

To make the identification more accurate, additional examinations of the time-variant frequency spectrum were made for a variety of sonic conditions. The frequencies of interest for a passing train are low frequencies, so the first 8 out of 576 frequency bins were examined. Two separate files from different months were used as well. This allows for a visual comparison between the different data sets. Figure 5 shows the resulting plots. The dark gray data represents the frequency amplitudes for a passing train in August and the light gray data represents the frequency amplitudes for a passing train in December. All of these plots are from the same range in time, but allow for comparison between different frequency bins. The frequency amplitudes greatly decrease as the frequency increases. These plots allowed more specific thresholds to be set that must be breached in order to record the presence of a train.

2.4. Sound level adaptation and averaging

The background sound level at GRKO changes greatly over the course of each day and it also changes throughout the year [3]. The amplitude thresholds that were set for the train may be true at the points initially tested, but these thresholds will not hold true all of the time. To account for this, a window averaging method was implemented. The average background sound across all 576 frequency "bins" is averaged and used over a defined span of time. As the software continues to process the bitstream, the sliding window is shifted through the file.

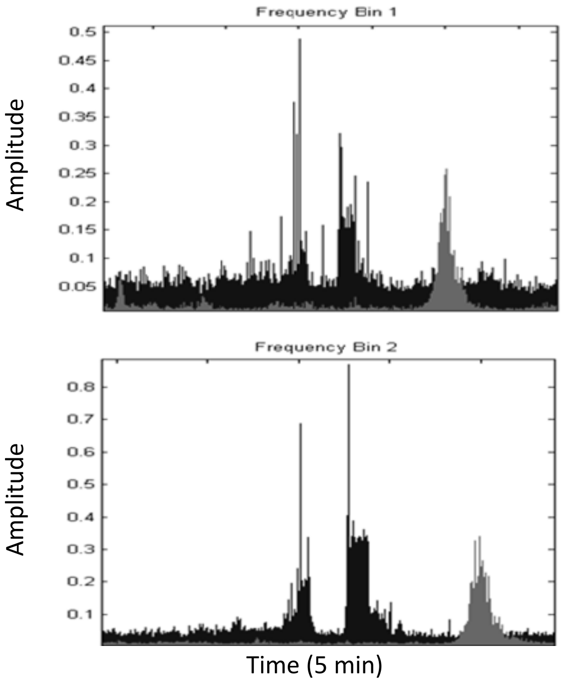

Figure 5: MP3 spectral "bin" amplitudes vs. time for two examples of a passing railroad train

Thus, background noise values outside of the time-averaging window are not factored into the average. The average is then applied to the thresholds that were previously set. When the background sound level is louder, the threshold is increased. When the background sound level is quieter, the threshold is decreased. This empirical technique makes the detection less prone to error when data from different times are examined and compared.

3. RESULTS AND DISCUSSION

Initially the software was tested with custom created MP3 files. These files were completely silent except for several seconds at a random point in the file. The detection thresholds were set very low to see if the code worked at all. This technique was also used to test the hour, minute, and second calculations in the software so a specific place in time can be recorded. After this proved successful, actual GRKO MP3 files were used. Some files contain very low background levels with limited sounds while others have high background sound levels and a plethora of sounds. Initially, a file with two passes of a train was chosen to help eliminate initial problems with false triggers. The time domain representation of the file is shown in Figure 6. This data was collected on August 1st, 2009 and the two large peaks are passing trains.

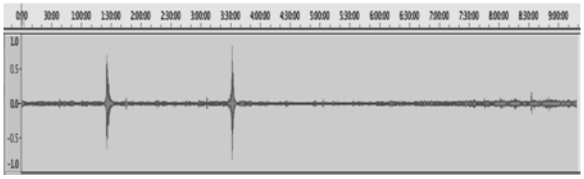

Figure 6: Audio waveform example, 9 hours elapsed time, with two passing trains

The software correctly detected the presence of a train at these two locations in the 9 hour file. However, the software triggered several times during the passing of each train. A timeout was placed in the software to prevent repeat triggers from happening. When a train is detected, the software waits for a defined amount of time before it begins looking for a train again.

For further testing, a file with a louder background sound level and sounds in addition to the train was used. A time domain representation of this file is shown in Figure 7. All of the peaks in the file were listened to by the human ear and documented for comparison. The MP3 file was then analyzed with the software. This data was collected on December 2nd, 2009. The results can be seen in Table 1.

In the table, it may appear that the software is "slower" to hear the train than the human listener. The human listener hears the first train two seconds before the software and the second train almost 2 minutes before the software. The thresholds that are set for the train in the software cause these differences. Also, the key purpose is to record the presence of a passing train and not necessarily the first instance of it being heard.

The truck that is heard in the file drives around inside the ranch and occasionally stops while the engine continues to idle. Initial testing with this file yielded many false triggers on the truck. The key was to prevent a trigger from the truck by ignoring anything with too many triggers in higher frequency bins. More energy in slightly higher frequencies helps distinguish the train from the truck.

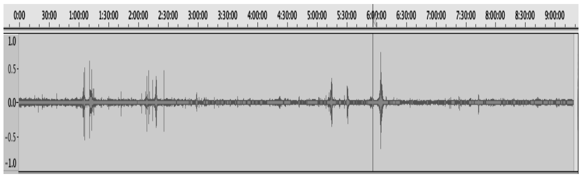

Figure 7: Audio waveform example, 9 hours elapsed time, for a noisy interval with various vehicles and trains passing by

	Human Listener			**Train-spotter Algorithm**		
Sound	*H*	*M*	*S*	*H*	*M*	*S*
Truck	1	3	37			
Truck	1	10	57			
Airplane (High overhead)	1	29	10			
Truck	2	7	19			
Train Whistle	2	16	35			
Rumble of Train	2	17	30	2	17	32
Train Passing	2	18	40			
Airplane (Low prop over)	2	59	5			
Airplane (High overhead)	4	22	30			
Loud Geese/Ducks	4	42	10			
Truck	5	11	25			
Airplane (Low prop over)	5	23	0			
Truck	5	30	5			
Airplane (High overhead)	5	54	30			
Train	6	2	30	6	4	11
Truck	6	9	50			
Airplane (High overhead)	7	34	30			
Airplane (Low prop over)	7	42	45			
Airplane (High overhead)	8	43	0			

Table 1: Comparison of identification of train sounds by human audition and by the proposed automatic algorithm

Additional testing has been done on other files with promising results. Three additional files from other points during the year work successfully with the software correctly pinpointing the time of the passing trains. However, there have been several tests on other files in which the software fails to recognize a train is passing. Some of these files contain high amounts of wind noise that cause the microphone to clip and the audio signal becomes very distorted. This has lead to false triggers of a passing train. These clipped signals need to be more carefully examined to see if the software can be modified to handle these cases.

3.1. Computational Performance

The current analysis software will analyze 9 hours and 19 minutes of audio, the length of one GRKO MP3 file, in 55 to 60 seconds. Timed trials have been performed on multiple files with consistent results. These trials were performed utilizing the hardware mentioned in the previous section. This is nearly 600 times faster than conventional human listening techniques.

3.2. Software Improvements

As mentioned, there are some issues with the software falsely triggering on signal clipping from wind noise. This will need to be addressed using similar frequency plotting techniques that were used to determine threshold trigger points for a passing train.

Locating passing trains is a useful metric to have, but being able to locate other sounds in addition to a passing train is strongly desired. With the current code framework set up, searching for additional sounds can be done simultaneously. The implementation of the additional identification only requires the correct thresholds for the new object to be determined.

The window averaging method appears to be working well, but more consideration needs to be given to the effect it has on thresholds. It may need to have a greater weighting so that the thresholds are based more on a ratio of the background noise level to the sound rather than depending primarily on the threshold.

Further improvements can be made to decrease analysis time. The original frame rate of the GRKO data is about 78 frames per second. Currently, one out of every five frames of the audio is being examined. When searching for sounds that last for several seconds or even minutes, it may be possible to search one out of every 50 frames and still accurately detect and identify everything that is being searched for.

4. CONCLUSIONS

This project increases the feasibility of long-term audio monitoring because of the significant decrease in time required for post-analysis. While many software improvements still need to be made, the method of performing analysis on compressed MP3 data is fully operational. The majority of the time and effort put into this project was focused on taking a MP3 file and viewing the frames in the frequency spectrum as quickly as possible. The identification method can be improved upon and expanded to use other techniques besides specific thresholds to indentify and classify the sounds.

Most long-term audio studies use the MP3 file format, but there could be benefits of recording uncompressed audio at the monitoring sites instead. As storage space becomes less of a concern, capturing uncompressed audio long-term becomes more practical. When an audio study is performed using lossy data compression, the potential to miss important data is possible. If a transition were made to capturing uncompressed audio long-term in the future, this current MP3 analysis method would see very little use. Currently however, the NSNSD can make use of this analysis method by using it on MP3 data that has already been collected and is still being collected. There are many additional challenges that lay ahead for cataloging the final results and drawing conclusions from this data.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Rocky Mountain Cooperative Ecosystem Study Unit, with the U.S. National Park Service.

We gratefully acknowledge the ongoing collaboration and encouragement of the Grant-Kohrs Ranch National Historic Site, Christine Ford, Integrated Resources Program Manager.

6. REFERENCES

[1] R.C. Maher and J. Studniarz, "Automatic search and classification of sound sources in long-term surveillance recordings," Proc. Audio Engineering Society 46th Conference, Audio Forensics—Recording, Recovery, Analysis, and Interpretation, Denver, CO, June, 2012.

[2] R.C. Maher, "Acoustics of national parks and historic sites: the 8,760 hour MP3 file," *Proc. 127th Audio Engineering Society Convention*, New York, NY, Preprint 7893, October, 2009.

[3] R.C. Maher, "Baseline Sound Monitoring at Grant Kohrs Ranch National Historic Site," *Final Project Report,* 181 pages, September, 2012.

[4] U.S. National Park Service, *Management Policies 2006*, U.S. Government Printing Office, 2006 (http://www.nps.gov/policy/MP2006.pdf)

[5] G. Tzanetakis and P. Cook, "Sound analysis using MPEG compressed audio," *Proc. ICASSP*, Istanbul, Turkey, pp.II761-II764, 2000.

[6] Gregory L. Zick and Lawrence Yapp, "Speech recognition on MPEG/Audio encoded files," U.S. Patent 6,370,504, issued April, 2002.

[7] E. Lynch, "Sand Creek Massacre National Historic Site," *Final Acoustical Monitoring Report,* 47 pages, May, 2011.

[8] J.P. Ogle and D.P.W. Ellis, "Fingerprinting to Identify Repeated Sound Events in Long- Duration Personal Audio Recordings," *Proc. ICASSP*, Honolulu, HI, pp. I-233—I-236, 2007.

[9] ISO/IEC JTC1/SC29, Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s-IS 11172 (Part 3, Audio), 1992.

[10] B. Grill, S. Quackenbush. (1995, Oct.). *MPEG-1 Audio* [Online]. Available: http://mpeg.chiariglione.org /standards/mpeg-1/audio

[11] Underbit Technologies, Inc. *MAD: MPEG Audio Decoder* [Online]. Available: http://www.underbit.com/products/mad