AUTOMATIC SEARCH AND CLASSIFICATION OF SOUND SOURCES IN LONG-TERM SURVEILLANCE RECORDINGS

ROBERT C. MAHER AND JOSEPH STUDNIARZ

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

The increasing availability of forensic audio surveillance recordings covering days or weeks of time makes human audition impractical. This paper describes the rationale and potential application of several techniques for high-speed automated search and classification of sound sources and sound events in long-term forensic audio recordings. Methods that can operate directly on perceptually compressed bitstreams without full decoding are of particular interest. Example applications include identification of aircraft overflights, the presence of human speech utterances, gunshot sounds, and other forensically relevant audio signals.

INTRODUCTION

An emerging problem in audio forensics involves the virtual explosion of data made available by long-term surveillance and soundscape audio recordings [1-5]. Long-term audio surveillance recordings may contain speech information and also non-speech sounds such as environmental noise, audible warning and alert signals, footsteps, mechanical sounds, gunshots, and other acoustic information of potential forensic interest. Manually searching and auditioning hours and hours of audio material to detect subtle, telltale sound events can be both tedious and error prone.

In this investigation we assess several techniques for high-speed automated search and classification of sound sources and sound events in long-term forensic audio recordings. The fundamental objective of these techniques is to transform the audio signal into a domain in which the desired signal attributes can be identified and classified.

The most promising current techniques use some form of short-time spectral analysis to localize in time, amplitude, and frequency the signal parameters of interest, and then to match these parameters to the classification dimensions. Although many techniques have been applied to this problem, the need for improved identification and classification techniques remains an open research question.

1 SEGMENTATION AND INTERPRETATION

Prior work that is relevant to forensic analysis of long-term surveillance recordings comes from several different research areas. These include techniques to process an audio recording to detect the presence of speech or music, distinguish between speech and non-speech sounds, recognize a particular speaker or song, and identify portions of the recording that have attributes of interest for subsequent manual investigation or follow-up [6-14]. Insights and techniques are also borrowed from the field of computational auditory scene analysis [15].

Three example systems and applications are representative of the prior research in this field. In 2005, Härmä, et al. [1], reported on an experiment with an automatic acoustic surveillance system that used a microphone to monitor the acoustical environment in an office for a two month period. The system identified “interesting” acoustical events and recorded them for additional processing. The basic detection and segmentation process used an adaptive spectral model to track the slowly-varying background noise profile, and an acoustic activity detection algorithm based on the departure of the currently measured spectral snapshot from the background noise profile. The acoustic events indicated by the detection algorithm were then analyzed to create a feature vector consisting of parameters such as RMS value, spectrum centroid, duration, and bandwidth. A k-means clustering algorithm was applied to the event parameters to classify acoustic events based on their parameter similarity.

A system described in 2010 by Wichern, et al. [10], used a variety of derived parameters, such as loudness, spectral centroid, and harmonicity, to enable classification and potentially identification of environmental sounds in continuous audio recordings. Like Härmä, et al., the Wichern experiment relied upon a parameterized representation to enable clustering and classification, but unlike Härmä’s background spectral profile, Wichern, et al., calculated the parameters continuously for the entire audio stream, and used the observed changes in those parameters to infer the onset and end of the acoustic events. Wichern, et al., also experimented with a framework to retrieve audio events based upon user query information.

Also in 2010, we developed a procedure to identify changes in *sonic texture* as an indication of audio events for an application involving extreme time-scale compression of long-term audio recordings [16]. Our definition of sonic texture was the time-variant fluctuation in the 1/3rd octave band levels. The method used a one second time interval and a 1/3rd octave spectral average to capture the time-variant spectral character of the signal. The event criterion was to monitor a threshold change in one or more of the 1/3rd octave bands. For each 1 second frame, the process examined the next several frames to determine any repetitive fluctuations. The result was a map of the textural transitions, with a goal to retain the time segments with a high number of sonic texture transitions, at the expense of the segments with lower activity.

2 SPECTRAL TEMPLATE CONCEPT

The insights derived from these existing techniques have led us to consider a time-frequency orientation in which two separate two-dimensional filters, or templates, are constructed. One template is designed so that it preferentially selects spectral components that are narrow in frequency but relatively broad in time, corresponding to tonal or quasi-harmonic content, while the other 2-D filter is designed to pass spectral components that are broad in frequency but relatively narrow in time, corresponding to impulsive, abrupt, and other similarly brief events. This approach is intended to uncover coherent acoustic information in the presence of incoherent broadband noise, and is in several ways an extension of our prior work on forensic audio quality enhancement [17]. The current spectral implementation uses similar data handling and formatting procedures to the audio enhancement system described before.

2.1 Application principles

The key features of the proposed audio identification and classification system are:

- Implementation of temporal and spectral filters that operate in the time-frequency domain.
- Description of the sonic events in terms of time-frequency templates, treating the spectrographic information as an image processing task.
- Use of multiple time-frequency resolutions in parallel to match the processing resolution to the time-variant signal characteristics.

These features are intended to mimic the approach used by human forensic examiners when considering long-term surveillance data. The examiner can benefit by examining a spectrographic (frequency content vs. time) display to look for features and patterns worthy of detailed examination.

2.2 Example framework

A spectrographic representation of one hour of audio recorded outdoors in a semi-rural area is shown in Figure 1. The recording contains a plethora of sound sources that overlap in time and in frequency. There are frequent bird vocalizations and the sounds of domestic animals, mechanical sounds from vehicles, and a variety of sounds attributable to wind.

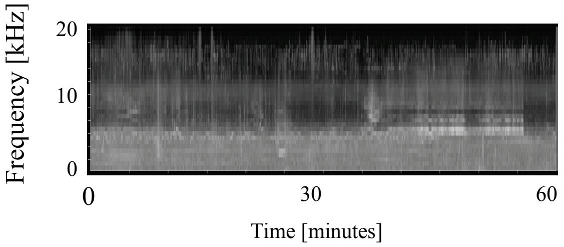

Figure 1: Spectrographic representation.

The spectrogram was created using the 1/3rd octave filterbank with 1 second average spectra. The resolution of each "pixel" in the spectrogram is therefore 1 second in width by 1/3rd octave in height. The white and light gray portions indicate time-frequency intervals in which the spectral energy is high, while the dark gray and black regions indicate low energy. The horizontal textural bands indicate steady-state or quasi-stationary sound sources, while the light colored vertical lines indicate impulsive sounds (relatively short duration with relatively broad frequency extent).

In Figure 1, consider the event visible as a bright interval at approximately 36 minutes in the 4-10kHz range. That sound source corresponding to that distinctive feature is probably not immediately recognizable from the spectrographic information alone, but additional context information and analysis can be employed to suit a particular investigation. In this case, the particular sound is an overflight by a single-engine piston-powered aircraft. The subsequent bright areas between 40 and 50 minutes are a power lawnmower's internal combustion engine starting up, followed by the sound of the mower moving with the spinning blade engaged.

Figure 2 shows an enlarged time scale spectrographic representation of the aircraft overflight and the beginning of the lawnmower event from ~36 minutes into the record shown in Figure 1.

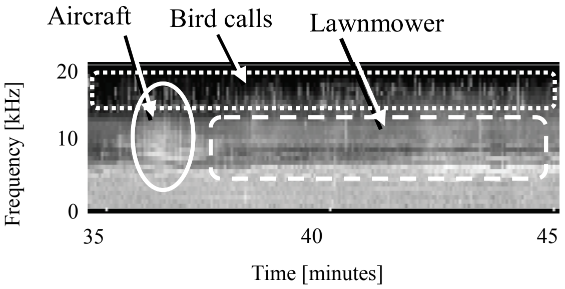

Figure 2: Enlarged portion of the spectral segment shown in Figure 1.

Decomposing the spectrographic representation of Figure 1 and Figure 2 can be thought of as an image processing exercise. For example, we can be interested in locating the vertically-oriented *edges* in the spectrographic image. The edges can be interpreted as the onsets and ends of sonic events as they evolve with time. Applying a high pass filter to the each of the 1/3rd octave filter output sequences produces the magnitude display shown in Figure 3, where the steepest detected edges are indicated as light gray and white.

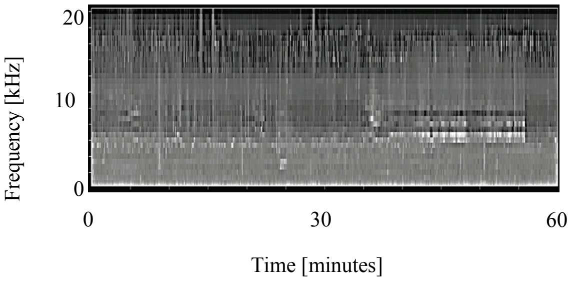

Figure 3: Spectral change "edge" representation.

The temporal edges can be further refined to enhance particular onset requirements. We currently determine the most appropriate detection settings for a particular recording by manual iteration. Fully automating the event/edge detection selections is a key part of our ongoing research.

2.3 Application considerations

Two fundamental issues arise when applying this spectral method. First, the time scale aspects must be adjusted to provide a meaningful level of detection. For example, the onset information depicted in Figure 3 includes speckle texture due to the quasi-stationary noise present in the recording, which can tend to obscure narrowband events of short duration in noisy recordings.

The second fundamental issue is the way in which sound events with a very gradual amplitude envelope appear from the spectrographic viewpoint. The proposed spectral method is best suited for sonic events with a sharp onset, and events that occur gradually, such as a high altitude aircraft approaching from a distance, do not provide a suitable "edge" in the 2-D framework. Developing appropriate techniques to reveal information regarding low-level sound events and sounds with slow amplitude envelope onset and/or release remains a research topic.

2.4 Implementation Optimization

The work on this identification and classification procedure has been conducted via software written in the C language and in Matlab. We have deliberately avoided any hardware-specific implementations up to this point because the research is driven by algorithmic flexibility and experimentation rather than time and efficiency constraints. A fully functional system would clearly require fast algorithms and implementation details that would allow processing at many times faster than real time. There is great potential for using graphics-oriented signal processing hardware in this regard.

A related development has been to use perceptually compressed audio bitstreams as the front-end for the analysis process without the need to decode the audio waveform itself [18]. Perceptual compression systems such as MP3 produce a bitstream containing a time-variant spectral analysis of the original signal and scale factors for the encoded audio, and this information can be extracted from the bitstream and interpreted without synthesizing the audio material itself. Since the 2-D filters used in the proposed identification and classification system operate in the frequency vs. time domain, allowing the initial processing to occur on the spectral parameters extracted from the perceptual coder bitstream could lead to greater potential efficiency compared to a process involving reconstructing the compressed bitstream into a time signal, only to have to perform a subsequent spectral analysis again on the decoded audio.

3 CONCLUSIONS

The availability of long-term audio surveillance recordings presents both opportunities and challenges for the field of audio forensics. Research to extract meaningful and forensically useful information from long-term recordings via automated processing remains an essential goal. The work described in this paper includes several features that will be a useful basis for future comprehensive solutions.

Many audio forensic investigations operate with an official query, such as "Are there any gunshots present in the recording between time *X* and time *Y*?" In these cases it would be desirable to have a suite of gunshot-related templates to use for the identification and classification task, and the proposed method will potentially be very useful. Similarly, in cases where the request is to determine if there are any distinctive sounds in a certain long segment of a recording, an automated method to identify candidate sound events can save time and potentially improve the examiner's efficiency and reliability. Nevertheless, in other cases the audio forensic examiner may be asked a more general question, such as "Can you help identify any of the audible background sounds between time *X* and time *Y*?" In response to such a query, the use of automated methods may be of less applicability due to the highly subjective nature of the query.

REFERENCES

[1] A. Härmä, M.F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *IEEE Int. Conference on Multimedia and Expo*, 4 pp. (2005).

[2] D.P.W. Ellis and K. Lee, "Accessing Minimal-Impact Personal Audio Archives," *IEEE Multimedia*, vol. 13, no. 4, pp. 30-38 (2006).

[3] 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).

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

[5] 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, (2009).

[6] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27-36 (1996).

[7] J. Foote, "Content-based retrieval of music and audio," Multimedia Storage and Archiving Systems II, *Proc. of SPIE*, vol. 3229, pp.138-147 (1997).

[8] L. Zhu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 20, no. 1/2, pp. 61-79 (1998).

[9] B.J. Gregoire and R.C. Maher, "Map seeking circuits: a novel method of detecting auditory events using iterative template mapping," *Proc. IEEE Signal Processing Society 12th DSP Workshop*, Jackson Lake, WY, pp. 511-515, (2006).

[10] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 688-707 (2010).

[11] S. Davies and D. Bland, "Interestingness Detection in Sports Audio Broadcasts," *Ninth Int. Conference on Machine Learning and Applications*, Washington, DC, pp. 643-648 (2010).

[12] G. Pietila, G. Cerrato, and R.E. Smith, "Detection and identification of acoustic signatures," *Proc. 2011 NDIA Ground Vehicle Systems Engineering and Technology Symposium*, Dearborn, MI, (2011).

[13] A. Muscariello, G. Gravier and F. Bimbot, "An efficient method for the unsupervised discovery of signalling motifs in large audio streams," *9th Int. Workshop on Content-Based Multimedia Indexing (CBMI)*, Madrid, Spain, pp. 145-150 (2011).

[14] Z. Chen and R.C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2974-2984 (2006).

[15] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[16] R.C. Maher, "Maintaining sonic texture with time scale compression by a factor of 100 or more," Preprint 8250, *Proc. 129th Audio Engineering Society Convention*, San Francisco, CA (2010).

[17] R.C. Maher, "Audio enhancement using nonlinear time-frequency filtering," *Proc. Audio Engineering Society 26th Conference, Audio Forensics in the Digital Age*, Denver, CO (2005).

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