Audio Engineering Society

Convention Paper 10515

Presented at the 151st Convention
2021 October, Online

Forensic Handling of User Generated Audio Recordings

Benjamin F. Miller, Fraser A. Robertson, and Robert C. Maher

Electrical & Computer Engineering, Montana State University, Bozeman, MT USA 59717-3780

Correspondence should be addressed to R.C. Maher (rob.maher@montana.edu)

ABSTRACT

User generated recordings (UGRs) are common in audio forensic examination. The prevalence of handheld private recording devices, stationary doorbell cameras, law enforcement body cameras, and other systems capable of creating UGRs at public incidents is only expected to increase with the development of new and less expensive recording technology. It is increasingly likely that an audio forensic examiner will have to deal with an ad hoc collection of unsynchronized UGRs from mobile and stationary audio recording devices. The examiner's tasks will include proper time synchronization, deducing microphone positions, and reducing the presence of competing sound sources and noise. We propose a standard forensic methodology for handling UGRs, including best practices for assessing authenticity and timeline synchronization.

1 Introduction

The widespread use of handheld smartphones and other device capable of recording audio and video means that user generated recordings (UGRs) are increasingly presented as evidence in criminal investigations. Combined with other recordings from law enforcement body cameras, business surveillance systems, etc., the availability of user-generated recordings may offer important audio forensic insights.

UGRs of a public event will involve multiple recording devices at different and often imprecisely known spatial locations. The recordings may start and stop at different times, have differing technical format specifications, and will seldom have sufficiently reliable time stamp information for exact synchronization. Nevertheless, the multiple audio recordings could potentially be combined to provide details about the crime scene compared to analyzing each recording individually.

Figure 1: Common recording devices in audio forensic analysis

Several researchers have studied the use of multiple unsynchronized UGRs of public events, such as bootleg recordings [1, 2, 3, 4]. These prior non-forensic examples have generally involved recordings of the same event simultaneously from different vantage points, so one goal has been to create a post-produced video with simulated "cuts" [5].

Recent work in the forensic realm includes recommendations for handling UGRs in circumstances that might call into question the authenticity of the recorded material, as well as its quality and integrity [6, 7, 8]. Forensic work has also included comparison of gunshot sounds from multiple simultaneous recordings [9, 10], and examination of synchronization issues for user generated recordings involving multiple sound source positions [11].

The remainder of this paper is organized as follows. First, we consider the recommendations for handling audio evidence obtained from user devices that may not have started with a standard chain of evidence. Next, we address several authenticity concerns regarding material presented by users in various circumstances and formats. Finally, we present results and recommendations regarding timing synchronization among UGRs that cover all of, or just portions of, a time interval of interest to the audio forensic investigation.

2 Handling User Generated Audio

User generated audio recordings often come from mobile devices (e.g., smartphones, tablets, handheld recorders) that generally have the ability to edit or otherwise alter the recorded information. Such devices often have wireless data transfer capability, and operating systems that may do background synchronization to/from other wireless devices or "cloud" storage. The concern is, of course, that the integrity of the recording could be compromised, either deliberately or inadvertently, during the investigation. Therefore, a forensic examiner will need to take steps to reduce the likelihood of post-recording data manipulation by blocking communication signals to/from the device. For example, a mobile phone would be placed in "airplane mode," with WiFi and Bluetooth disabled. [6].

For the purposes of the following discussion, we assume that a cooperative device owner unlocked the system, or the owner was able to share an unaltered digital copy of the recording. The examiner would need to document the type, model, operating system version, serial number, recording app used, and related information about the recording device.

2.1 Hash Tag Original Files

Upon receipt of user generated recordings, most agencies have a policy to compute a hash value (checksum) of each data file. The hash value provides what the Scientific Working Group on Digital Evidence refers to as *fixity checking* [7], meaning the ability to verify that the file contents have not changed since the original receipt. The computed hash value, and the hash algorithm used, are stored with the other file-related information. Any time a copy of the UGR is produced, transmitted, and received, the hash value needs to be re-verified (fixity check).

2.2 Conversion of UGRs to a Standard Audio Format

When working with UGRs, the original files may contain audio data in a wide variety of possible formats, different sampling rates, data precision, lossless and lossy encoding, mono and stereo, etc. In our experience, we have found that it is a best practice to create a set of working files: a single high-resolution lossless .wav file version of each UGR. The conversion from the original file to the working file happens once, and no additional encoding and decoding (no tandem signaling) is subsequently involved.

The best available decoding and sampling rate conversion algorithms are used when creating the working files in order to minimize the risk of audio quality degradation. The original file and its hash tag are left unmodified, of course, in the secure data repository. Each working file has a hash code computed and saved with the project information for later verification, such as when the working files are moved or transmitted from one system to another.

The choice of .wav file parameters for the working files can vary from case to case, but we find that adopting a standard 48 kHz mono file with 32-bit floating point sample precision has worked well. This assumes that none of the original UGRs had a higher sampling rate, as the preference is to avoid downsampling and the inherent loss of bandwidth. If the original UGR was a stereo recording (not duplicate mono) with two or more separate channels, it is recommended to create separate mono files from each channel, labeling the files clearly as to the original channel configuration.

3 Authenticity Considerations for UGRs

As noted in the prior section, audio recordings and multimedia files that the forensic examiner receives from civilian bystanders or other amateur sources may raise concerns about authenticity. The examiner needs to be able to identify any reasons that the UGR might be altered from its original state, or that the UGR was made under circumstances different than what the user asserts it to be.

It is vital to note that one of the difficulties associated with digital files is the likely inability of an examiner to be able to distinguish between an authentic file and a forgery prepared by a skillful adversary. Edits can be made undetectable, metadata can be altered in such a manner as to appear consistent with an authentic recording, and false recordings can be presented as genuine. Thus, an examiner cannot absolutely guarantee authenticity even if no inconsistencies are found.

3.1 Chain of Custody

First, whenever a user generated recording is collected as evidence, it is vital to document the chain of custody. The chain of custody identifies how the recording was made and exactly who has had access to it. This record of custody may give investigators a better sense of the likelihood of deliberate or inadvertent tampering—or the opportunity for someone to have altered the recording. Moreover, identifying clearly which UGR files are original, and which files are decoded working copies, helps avoid having work files with enhancement or embedded markers being mistaken for original material.

3.2 Metadata Consistency

A routine way to check for file authenticity is through verification of the file's *metadata*. This is done by observing the file contents with a non-destructive editor or other software tool that can reveal the embedded meta-information. The examination includes looking for discrepancies in the date and time information, missing file parameters, geographic tags, and so forth. It is often helpful to obtain a comparison recording made on the same device model that was used to make the recording in question, and then compare the raw metadata looking for any significant differences in the contents and structure [8, 12].

Any discrepancies or significant differences in the metadata will be a strong indication that the UGR has been edited or is otherwise inauthentic.

3.3 Identifying Possible Insertions/Deletions

An audio forgery could consist of one or more edits made to an original recording by deleting certain time segments, by inserting audio material, or by additively mixing in the forged material. Detecting such modifications in UGRs can be easy or impossible depending upon the features of the editing software and the sophistication of the forger.

Critical listening, combined with waveform and spectrographic observation to note any changes in the audio signal and/or background noise, can sometimes reveal likely edit points. An abrupt insertion or deletion, often referred to as a *butt splice*, may leave behind telltale audible effects and discontinuities at the boundary [13]. A skillful cross-fade, on the other hand, may escape simple detection [8].

4 Synchronization Scenarios for UGRs

When two or more audio devices are operating concurrently from different spatial locations while recording the same sound source, we would not expect the recordings to be identical, but we would expect a good correspondence, or correlation, among the recordings. However, because the sound received at each microphone will differ due to the directionality of the source and microphones, the different distance between the source and each microphone, and the presence of sound reflections, noise and reverberation, there is a need to determine how best to combine the available information.

For example, the absolute time when a sound of forensic interest occurred will depend upon the relative position of the receiving microphones and the relative starting and stopping times of the various recordings. Without time synchronization among the recordings, the relative time-of-arrival of the sound at each microphone will be ambiguous.

In forensic examinations, UGRs may come from multiple recording devices at different spatial locations. The recordings will have unsynchronized start and stop times, different durations, and typically imprecise or unknown spatial location information. Therefore, the examiner must consider how to combine the multiple audio recordings in a manner that leads to a useful and meaningful investigative conclusion [10].

Unlike a simple audio mixing scenario, the needs of an audio forensic investigation may require an objective assessment of the timing between different sound events from the point of view of each recording device, as indicated in Figure 2.

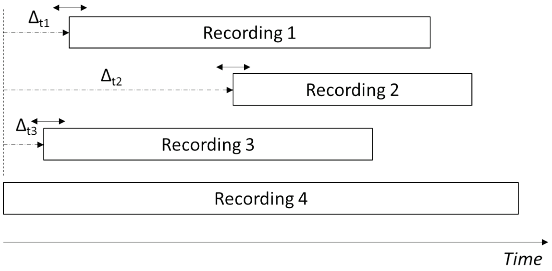

Figure 2: Time alignment problem

Our prior research has emphasized the fact that multiple sound sources can appear with different timing in various simultaneous recordings depending upon the relative position of each recording device, so simply mixing the recordings is not appropriate for a timing and sequencing investigation, as summarized in Figure 3.

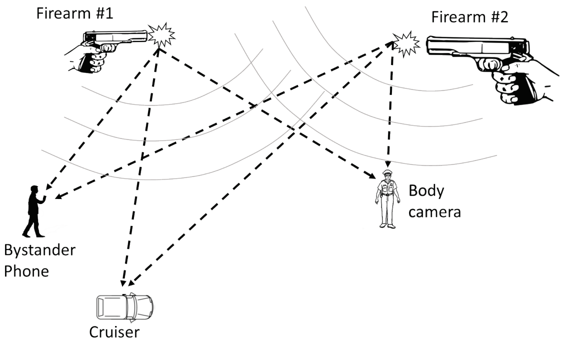

Figure 3: Multiple recordings of multiple sound sources [11].

In general, the audio signal processing requirements for forensic purposes are applicable to both UGRs and to audio evidence obtained from law enforcement sources (e.g., dashboard and body camera systems).

4.1 Sequence and Synchronization

There are four basic alignment techniques that we examined in this research: waveform correlation, audio fingerprinting, spectrographic structural similarity index measure (SSIM), and spectrographic correlation. Waveform correlation is a time domain technique, while the others involve short-time spectral analysis.

4.1.1 Correlation

Waveform correlation is done with traditional cross-correlation, or average magnitude difference function (AMDF). These calculations are purely waveform-based, so they have the highest time resolution. The correlation techniques are especially useful for recordings containing loud, impulsive, and distinct sound events that are short in time and broad in frequency. In practical cases, we find that the correlation technique is most useful for relatively low sampling rates, both in terms of minimizing computation time and locating unambiguous synchronization estimates. Signals with low signal-to-noise ratio may result in poor synchronization estimates.

4.1.2 Audio fingerprinting

An audio fingerprint is a compact representation of a particular audio recording that can be used to seek out similar audio recordings from a database [14]. Fingerprinting methods commonly involve calculating a spectrogram, identifying prominent spectral peaks, sorting by time, and using a hashing function to simplify the representation [15]. Depending on different parameters, the number of fingerprints constructed from each point and the maximum distance between each point can be adjusted.

4.1.3 Structural similarity index measure (SSIM)

In several experiments, we used a structural similarity index measure to compare the spectrographic structure using time windows taken from different recordings. SSIM is commonly used to compare images [16]. The rationale for its use in synchronization was to treat the spectrogram as an image: presumably the same sound event recorded simultaneously from different spatial locations would likely have similar time-frequency content represented in the spectrographic image. However, this technique was not found to be reliable due to its sensitivity to differences in background noise and sound level.

4.1.4 Spectrographic correlation

While the SSIM technique was not productive in this application, the concept of spectrographic similarity lead to experiments with image correlation. Spectrographic correlation is implemented by correlating the time magnitude in each frequency band for a short-time window of one audio recording with the matching frequency band in a short-time window from another file. This frequency-based matching captures the characteristics across the time window, as opposed to identifying individual peaks and features as is done in fingerprinting. We find that spectrographic correlation can be productive if the recording has relatively short sound events that correspond to the window length, but the spectrographic method is sensitive to the presence of noise, like waveform correlation.

4.2 Noise and clipping

In many cases, the UGRs (and recordings from law enforcement, too) will contain noise. Wind gusts, traffic, crowd noise, footsteps, and other unintended sounds are very common in most recordings, and these interfering sounds may be louder than the sound sources of interest for synchronization purposes. Recordings with a high level of noise will produce many more fingerprint features and will result in much larger correlation scaling factor magnitude. These effects of noise can confound the sequencing and synchronization attempts.

Similarly, UGRs sometimes exhibit waveform *clipping* due to the sound level exceeding the capabilities of the microphone and/or the recording system. Signal and spectrographic comparisons among clipped and unclipped recordings of the same sound event can be ambiguous.

Furthermore, the synchronization techniques typically perform best when the recorded signals are of comparable sound level. If the various UGRs are quite different in signal amplitude, we have found that some degree of gain compression may be useful for reliable synchronization.

There are, of course, numerous algorithms available for noise reduction, de-clipping (bandlimited reconstruction), filtering, and level normalization. We have found that applying some de-noising and de-clipping strategies can be helpful prior to attempting synchronization. However, it is important to understand the potential ways in which the "enhancement" can alter the temporal characteristics of the signals, which in turn will alter the reliability of the synchronization. One area of our future research will involve a two-step procedure to use signals after de-noising and de-clipping for initial synchronization, then to refine the sync using the preliminary estimate as the starting point for comparing the original recordings.

5 Conclusions

Audio forensic investigations increasingly include multiple concurrent user generated recordings (UGRs). The prevalence of unsynchronized body cameras, private surveillance systems, and handheld private recording devices at the scene of public incidents can only be expected to increase with the commercial availability of new and less expensive mobile recording technology.

We recommend that audio forensic examiners adopt a standard methodology for gathering and preserving UGRs [6, 7], and practice an internal workflow that helps verify and maintain the integrity of the audio evidence. Issues of authenticity are likely to arise when dealing with recordings provided by bystanders or other possibly unverified sources, and these concerns need to be considered as part of the standard operating procedures.

We are continuing our research using a variety of techniques to attempt synchronization of audible events in multiple UGRs. There are important considerations of what it means to synchronize recordings in a forensic context, because multiple recording positions with multiple sound source positions result in different relative perspectives of the sequence of events [10, 11].

Finally, ongoing research is also needed in the area of synchronization and interpretation of noisy and clipped audio from UGRs. Temporal and spectral alterations of time waveforms can enhance the perceived quality of a recording, but these alterations may or may not be significant in terms of the scientific reliability and integrity of the audio forensic investigation.

6 Acknowledgements

This research work has been supported in part by U.S. National Institute of Justice grants 2014-DN-BXK034, 2017-DN-BX-0179, and 2019-DU-BX-0019. The authors thank Ethan Hoerr and Charles Nicholson for their research and programming assistance. The authors also gratefully acknowledge the assistance of the anonymous reviewers in improving this paper.

References

[1] P. Shrestha, M. Barbieri, and H. Weda, "Synchronization of multi-camera video recordings based on audio," *Proc. 15th ACM Int. Conf. Multimedia*, pp. 545–548 (2007).

[2] L. Kennedy and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," *Proc. 18th Int. Conf. on World Wide Web*, pp. 311–320 (2009).

[3] S. Bano and A. Cavallaro, "Discovery and organization of multi-camera user-generated videos of the same event," *J. Inf. Sciences*, vol. 302, pp. 108–121 (2015).

[4] N. Stefanakis, Y. Mastorakis, A. Alexandridis, and A. Mouchtaris, , "Automating Mixing of User-Generated Audio Recordings from the Same Event," *J. Audio Eng. Soc.*, vol. 67, no. 4, pp. 201–212 (2019).

[5] P. Shrestha, P. de With, H. Weda, M. Barbieri, and E. Aarts, "Automatic mashup generation from multiple-camera concert recordings," *Proc. ACM Int. Conf. Multimedia*, pp. 541–550 (2010).

[6] Scientific Working Group on Digital Evidence (SWGDE), "Best Practices for Mobile Device Evidence Collection & Preservation Handling and Acquisition," v1.2 (2020a).

[7] Scientific Working Group on Digital Evidence (SWGDE), "Best Practices for Archiving Digital and Multimedia Evidence," v1.0 (2020b).

[8] R.C. Maher, *Principles of Forensic Audio Analysis*, Springer Nature Publishing (2018).

[9] S. Beck, "Who fired when: associating multiple audio events from uncalibrated receivers," *Proc. AES Int. Conf. Audio Forensics*, Porto, Portugal (2019).

[10] R.C. Maher and E. Hoerr, "Forensic comparison of simultaneous recordings of gunshots at a crime scene," Preprint 10281, *Proc. 147th AES Convention*, New York, NY (2019).

[11] R.C. Maher, "Forensic Interpretation and Processing of User Generated Audio Recordings," Preprint 10419, *Proc. 149th AES Convention*, Online (2020).

[12] B.E. Koenig and D.S. Lacey, “Forensic Authenticity Analyses of the Metadata in Re-Encoded WAV Files,” *Proc. AES 54th Int. Conf.: Audio Forensics*, Paper 4-2 (2014).

[13] A.J. Cooper, “Detecting Butt-Spliced Edits in Forensic Digital Audio Recordings,” *Proc. AES 39th Int. Conf.: Audio Forensics*, Paper 1-1 (2010).

[14] J. Six and M. Leman, “PANAKO – A scalable acoustic fingerprinting system handling time-scale and pitch modification,” *Proc. 15th Int. Soc. for Music Inf. Retrieval Conf. (ISMIR)*, Taipei, Taiwan (2014).

[15] W. Drevo, “Audio fingerprinting with Python and Numpy,” (2013) URL: https://willdrevo.com/fingerprinting-and-audio-recognition-with-python/. Accessed 2021 August.

[16] Z. Wang, A. Bovik, H. Sheikh, E Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600-612 (2004).