Audio Engineering Society

Convention Paper 10419

Presented at the 149th Convention
Online, 2020 October 27-30

Forensic Interpretation and Processing of User Generated Audio Recordings

Robert C. Maher

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

Correspondence should be addressed to rob.maher@montana.edu

ABSTRACT

For audio forensic analysis, it is increasingly likely that multiple user-generated recordings (UGRs) may be presented as evidence in a criminal investigation. Audio evidence may come from handheld smartphones, private surveillance systems, police body cameras, and other unsynchronized recording devices. When multiple UGRs are available, the recordings could provide spatial and temporal information about the location and orientation of sound sources, and potentially a means to increase intelligibility of spoken utterances. However, UGRs generally start and stop at different times, differ in technical format specifications, and seldom have sufficiently reliable time stamp information for exact time and position synchronization. We study these analytical and practical constraints, and develop forensic recommendations for combining and synchronizing multiple UGRs.

1 Introduction

The widespread use of handheld smartphones and other devices capable of recording audio and video has resulted in the likelihood that user generated audio recordings (UGRs) may be presented as evidence in a criminal investigation. Combined with other recordings from body cameras worn by law enforcement, dashboard camera systems, residential and commercial surveillance systems, etc., the availability of user-generated audio recordings may offer important audio forensic insights [1].

User generated audio-visual recordings of a public event will involve multiple recording devices at different 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. The precise spatial location of each recording device is also typically unknown. Nevertheless, it is important to consider how to combine the multiple audio recordings to yield important forensic insights. While there may be an initial assumption that simple additive mixing can combine various audio recordings, in practical situations there are multiple sound sources at different locations contributing to the sound received at a particular microphone, and the relative time-of-arrival of the signals therefore varies sufficiently that simple alignment by correlation may not be useful for forensic analysis.

1.1 Concurrent audio recordings

When two or more audio devices are operating concurrently from different spatial locations while recording the same sound source, the audio recordings will not be identical, but we would expect strong similarity among the recordings. The sound received at each microphone will differ due to (a) the directionality of the source and microphones, (b) the differing distance between the source(s) and each microphone, (c) the presence of acoustic noise and reverberation, and (d) the likelihood that one or more of the recording devices may be moving during the recording. Thus, there is a need to determine how best to combine the available audio information within the asynchronous framework.

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, and without knowledge of the microphone positions, the relative time-of-arrival of the sound at each microphone will be ambiguous.

For the research described in this paper, we consider the increasingly common situation in which several mobile recording devices capture the sound of gunshots in the vicinity that the shootings took place. We assume that the position of the recording devices is not known precisely and that the recorders are not synchronized, but, for our first analysis, we assume that the recording devices are not moving over the time interval of interest. Our research in this scenario is to find the best time alignment of the available recordings with respect to the specific sound of interest.

1.2 Relevant prior research

Since at least the early 2000s, published research describes the use of multiple unsynchronized user generated recordings. Several researchers have studied the use of UGRs of public events, such as live concert bootleg recordings, to create a composite mixture [2-7]. In addition to audio recordings, the prior non-forensic work has generally involved video sequences of an event that come from multiple vantage points, so the goal has been to create a post-produced video sequence with simulated "cuts" from one camera to another [8, 9].

Among the key prior research questions is the need to deduce the temporal relationship among different UGRs. Sometimes the recordings may include metadata, such as the file creation date and time. However, the timing precision and reliability of metadata may not be of sufficient quality for forensic audio analysis [10].

An emerging research challenge is to investigate different means by which correlations among the various audio streams can be revealed by exploiting available knowledge about the microphone positions. Also, there can be interest in using several low-quality UGRs from different spatial locations to create a single mixture of the captured sound scene, but with enhanced quality and/or voice intelligibility [11-14].

In the audio forensic realm specifically, the use of multiple simultaneous recordings has been described for an unsynchronized collection of law enforcement body camera and dashboard camera systems [1, 14], and simulation of position uncertainty [15, 16].

1.3 Audio sources and formats

A contemporary audio recorder, such as a personal memo recorder, smartphone, or home surveillance system, creates a digital file consisting of audio samples obtained from a microphone by sampling at a specified rate, such as 16,000 samples per second. Common audio recordings may be in the form of uncompressed pulse-code modulation (PCM), with either one channel (monophonic) or two channels (stereo). It is more common today to have the audio recording device perform lossy perceptual compression, meaning that the recording is processed to reduce the storage and transmission size while retaining as much as possible of the *perceived* audio quality of the original recording, within the constraint on reducing the file size. Common compressed audio formats include MP3, AAC, and WMA.

User generated audio/visual (motion picture) recordings start with a sequence of video frames (sequence of still pictures) typically encoded exploiting image frame-to-frame correlation via MPEG video, and a corresponding digital audio recording. As with contemporary audio-only recordings, the audio content accompanying the video sequence is generally encoded into a lossy compressed digital format. Because the video actually consists of a sequence of video frames (still pictures) displayed in rapid succession, the video file playback must be performed at a precisely controlled rate so that the duration of each displayed video field exactly coincides with the playback speed of the digital audio material in order to maintain audio-video synchronization. Moreover, lossy audio encoding generally involves block-based processing, so pre-echo effects may blur the precise timing of waveform events [17].

If two or more audio devices are operating concurrently from different spatial locations while recording the same sound source, we would expect a good correspondence among the recordings, as depicted in Figure 1.

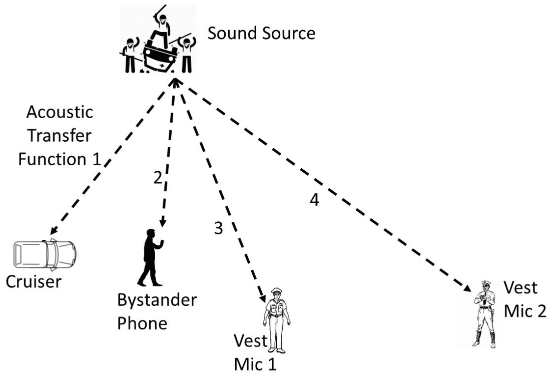

Figure 1: Concurrent but unsynchronized recordings from different spatial positions.

The sound received at each microphone will differ due to the directionality of the source, the different distance between the source and each microphone, and the presence of sound reflections, noise and reverberation. Thus, there is a need to determine how best to combine the available information. The basic concept is summarized in Figure 2, where the various transfer functions and noise contributions are generally not known. For example, the concept of when a sound occurred will depend upon the position of the receiving microphone with respect to the sound source. At 20° C, the speed of sound in air is 343 m/s, which adds a time delay of 2.9 milliseconds per meter. Nevertheless, without time synchronization among the recordings, the relative time-of-arrival of the sound at each microphone will be equivocal.

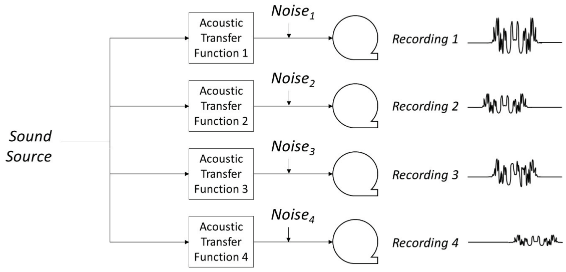
Sound Source
Acoustic Transfer Function 1
Acoustic Transfer Function 2
Acoustic Transfer Function 3
Acoustic Transfer Function 4
Noise$_1$
Noise$_2$
Noise$_3$
Noise$_4$
Recording 1
Recording 2
Recording 3
Recording 4

Figure 2: Block diagram depiction of multiple concurrent recordings.

If all of the UGRs were recorded simultaneously as concurrent tracks on a single recording device, an audio forensic examiner could simply determine the relative time of each detected acoustical event in each recording. If the ambient air temperature was also known or well estimated, the speed of sound could be used to convert the relative delay into the distance between the source and each corresponding microphone.

However, in the case of forensic reconstruction of interest here, we assume that the audio forensic examiner will need to use an *ad hoc* collection of UGRs from mobile audio recording devices. Along with the need to estimate proper time synchronization, the examiner will also have to estimate the microphone positions, potential reflections, and the presence of competing noise sources.

This is a time consuming process with obvious uncertainty, because even in what might appear to be relatively simple cases the ambiguous timing can become complicated. For example, the available recording segments may not actually overlap in time, or the sound sources and/or the recording devices may move with respect to each other during the recording. As described in the literature, some or all of the recordings may include a sequence of repetitive sounds that can lead to misalignment if a simple algorithm is attempted.

In most of the existing techniques applied to ascertain the appropriate alignment, the process involves determining a set of audio *features* from each of the recordings, and then comparing/correlating the features among the recordings [4, 6, 10, 18].

2 Application example 1: Sequence of gunshots: Who shot first?

The forensic situation is often different from the mixing processes described in the literature. Unlike a “concert bootleg” recording scenario in which the goal is to create a pleasing sound mixture, user generated audio recordings for forensic analysis require careful consideration of the meaning and interpretation of time alignment in a particular context. For example, consider a simulated “who shot first?” scenario in which two individuals discharge firearms within a sufficiently short period of time that eye and ear witnesses at the scene cannot agree upon the order of the shots, and no definitive video exists (e.g., camera not pointed in the direction of the firearms). In this simulated scenario we further assume that no precise geometric information is available, but that witnesses and other physical evidence gives *the general orientation and relative position* of the two shooters and the three audio recording microphones. For this scenario, we assume that the shooters were approximately 30 meters apart, and the observers spread out about 50 meters away, as sketched in Figure 3.

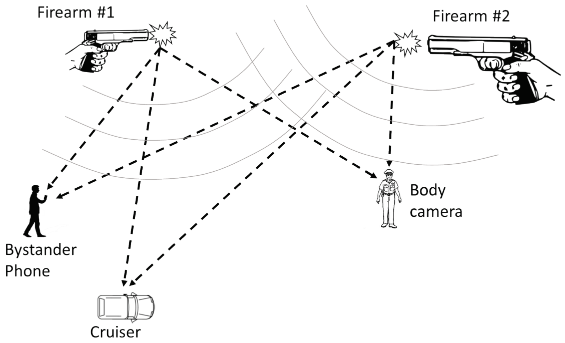

Figure 3: "Who shot first?" scenario (not to scale). Firearms are about 30 m apart; microphones are about 50 meters from the closest firearm.

2.1 Information obtained from the “`Body camera`” recording

Figure 4 shows approximately 1.5 seconds of the `Body camera` audio recording in this situation, and the corresponding spectrogram. The two identified gunshot sounds, labelled BC-1 and BC-2, are from Firearm #1 and Firearm #2, but we do not know which gun caused which sound based upon the `Body camera` audio alone.

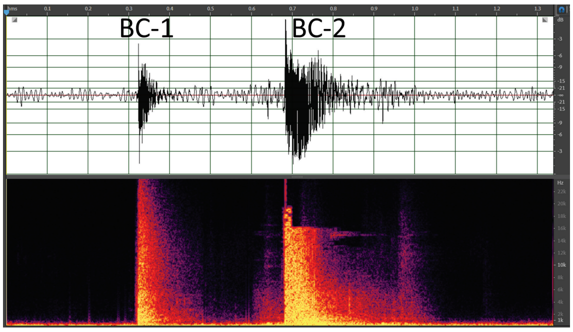

Figure 4: "`Body camera`" recording of the two shots (upper panel: time waveform, 1.5 seconds; lower panel: spectrogram).

The time difference ΔT between the onset of BC-1 and BC-2 can be determined from the `Body camera` recording by observing the recorded waveforms, as shown in the zoomed waveform images of Figure 5.

Note that despite the gunshot muzzle blast being a very abrupt sound, the recorded signal of each shot has noise and distortion preceding the high-energy portion of the sound, requiring some interpretation of the true time-of-arrival of the gunshot sounds.

Using the waveform observations for the `Body camera` recording, the ΔT between the two gunshot sounds, BC-1 and BC-2, *at the microphone position*, is found to be 374.3 ms. The `Body camera` in this simulation is known to have been located approximately 50 meters from Firearm #2, and that the two firearms were about 30 meters apart. Therefore, the distance from the `Body camera` to Firearm #1 is perhaps 58 meters.

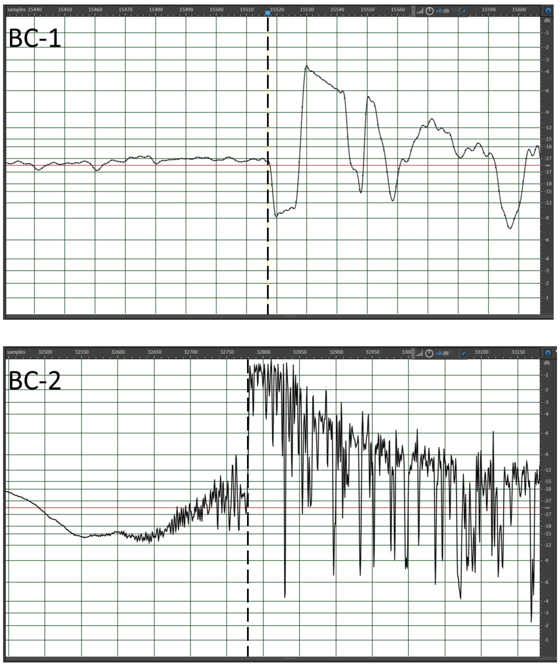

Figure 5: Zoom showing approximately 15 ms of the two gunshot sounds recorded by the `Body camera`.

Assuming a local speed of sound of 343 m/s, the time required for the muzzle blast of Firearm #1 to reach the microphone of the `Body camera` is 169.9 ms, while the time for the muzzle blast of Firearm #2 to reach the `Body camera` microphone is 145.8 ms. In our example scenario, we do not know which of the two firearms caused which of the two sounds in the recording.

The timing synchronization between the absolute trigger-pull times (which are not known, in general) and the observed time-of-arrival of the muzzle blast sounds is sketched schematically in Figure 6. The figure shows time on the horizontal axis, with the two gunshots indicated at times t_a and t_b, and distance on the vertical axis, with the relative distance of the firearms to the `Body camera` audio recording device.

If we suppose the sound BC-1 is attributable to Firearm #1 and the sound BC-2 to Firearm #2, we would determine that Firearm #1 was fired t_2-t_A = 169.9 ms before the sound propagated the ~58 meters to the microphone, and Firearm #2 was fired t_3-t_B = 145.8 ms before its sound arrived at the microphone.

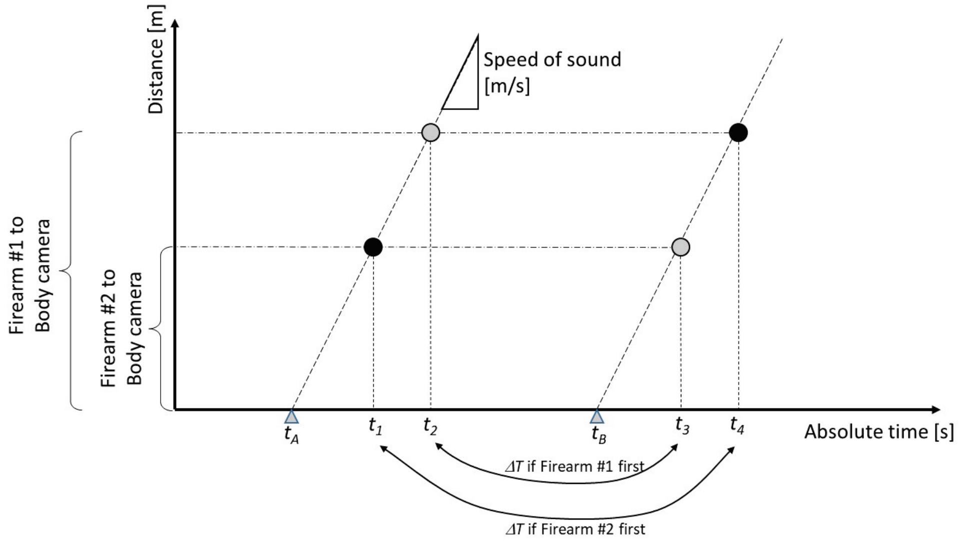

Figure 6: Schematic representation of time-of-arrival for gunshot sounds at the `Body camera` position, with unknown order of Firearm #1 and Firearm #2.

If Firearm #1 shot first, the actual time interval *between* the trigger-pulls of the shots (i.e., t_B-t_A) would be about 374.3 + 169.9 – 145.8 = 382.6 ms. On the other hand, if Firearm #2 shot first, then Firearm #2 was fired 145.8 ms before the sound arrived at the microphone, and Firearm #1 was fired 169.9 ms before BC-2. This would give the actual trigger interval t_B-t_A of 374.3 + 145.8 – 169.9 = 350.2 ms. With only a single audio recording and no other information, there does not appear to be a timing-related clue to determine which firearm was discharged first.

2.2 Information combined among multiple recordings

Now we consider the additional information provided by the other audio recordings. Performing the same timing analysis with the "`Cruiser`" recording and the "`Bystander`" recording, we observe the inter-shot timings as shown in Figure 7 and Figure 8, respectively.

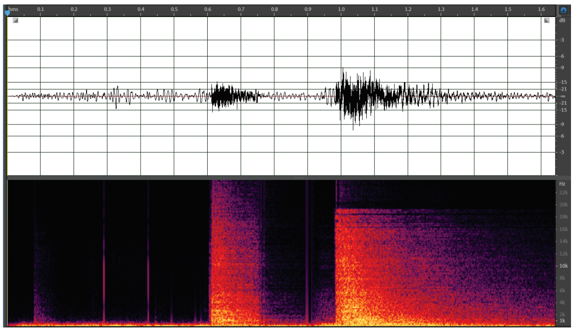

Figure 7: "`Cruiser`" recording of two shots (time waveform, 1.5 seconds, and spectrogram).

The inter-shot timing information for the three available recordings is shown in Table 1. Initially it may seem baffling that the three recordings of *the same incident* reveal significantly different timing, but the important realization is that because the relative distance of the sound sources differs for each recording, the time-of-arrival of the muzzle blast sound will also differ. We are able to utilize the information for forensic purposes if we at least know the relative position of the shooters and the recording devices to some stated degree of certainty.

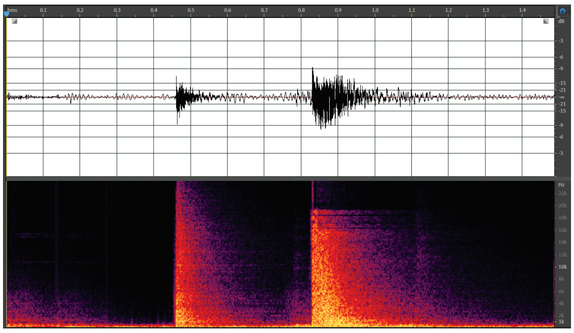

Figure 8: "`Bystander`" recording of two shots (time waveform, 1.5 seconds, and spectrogram).

	Shot-to-shot timing (ΔT) from recording [ms]
`Body camera`	374.2
`Cruiser`	320.9
`Bystander`	292.9

Table 1: Inter-shot timing for the three recordings.

Reiterating the forensic scenario, the three recordings of the scene were unsynchronized and made from different but imprecisely known spatial locations, so to address the forensic question of which of the two sounds in each recording corresponds to which firearm, we consider only the *relative* position of the microphones and the firearms.

For example, the `Body camera` is located closer to Firearm #2 than to Firearm #1, while the `Bystander` is located closer to Firearm #1 than to Firearm #2. If observing from the `Body camera` position and Firearm #1 was discharged first, the observed arrival of the second shot (Firearm #2) would be *advanced* slightly because Firearm #2 is closer to the `Body camera` than Firearm #1. At the same time, if observing from the `Bystander` position and Firearm #1 was discharged first, the observed arrival of the second shot (Firearm #2) would be *delayed* due to the greater distance the second sound must travel. Thus, if Firearm #1 was fired first, we would expect the inter-shot duration observed at the `Bystander` location to be *longer* than the inter-shot duration observed at the

`Body camera` location. However, we find that *ΔT* from the `Bystander` recording, 292.9 ms, is *less* than *ΔT* `Body camera`, 374.2 ms. Thus, the inter-shot timing would favor the forensic conclusion that *Firearm #2 was discharged first*.

If we perform a similar comparison of the shot-to-shot timing observed by the `Cruiser` and the `Body camera`, we see that *ΔT* `Cruiser`, 320.9 ms, is less than *ΔT* `Body camera`, 374.226. Since the `Cruiser` position is closer to Firearm #1 than to Firearm #2, the observation confirms that Firearm #2 was discharged first. Furthermore, the approximate position of the recording devices and the two firearms, and the observed timing differences in the recordings (Table 1), are consistent with an actual trigger interval of 350 ms between the two shots.

It is important to note that if the relative position of the firearms and the recording devices is not adequately dispersed, or if the relative positions are ambiguous, then the timing evaluation from UGRs may also be ambiguous.

3 Application example 2: UGRs and signal corroboration

When more than one user generated audio recording is available, an important advantage may be corroboration among the recordings. As noted in our prior experiments [16], recorded waveforms differ significantly for various microphone positions, automatic gain control settings, signal encoding formats, and other differences. The different devices generally have different spatial orientation, distance, wind turbulence, and acoustic shadowing, resulting in differing waveforms and clipping. In some cases, the availability of multiple independent recordings can help elucidate and corroborate the forensic findings.

For example, consider an actual example from an audio forensic examination case involving two recordings made during a gunshot incident. The first recording (Figure 9) came from a microphone located within a few meters of several loud gunshots and is highly clipped and distorted.

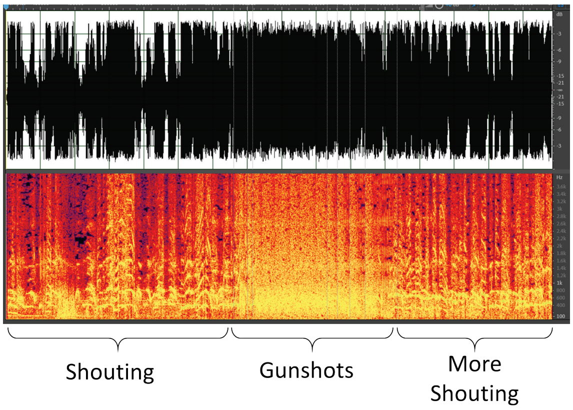

Figure 9: A distorted recording, ~5 seconds in duration, obtained near the shooting position (waveform and spectrogram).

Simultaneously, a second recording (Figure 10) came from a microphone located inside a closed vehicle about 40 meters away.

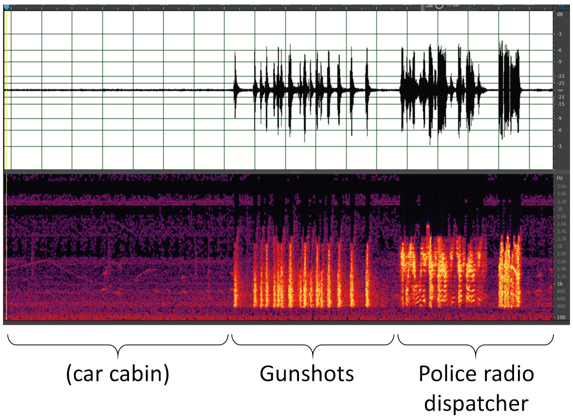

Figure 10: A recording from inside a closed vehicle 40 meters away from shooting position (5 sec time waveform and spectrogram).

The severely clipped recording in Figure 9 is still helpful because it captures several intelligible utterances in the immediate vicinity of the incident, but it is not possible to distinguish details about the number of gunshots or to determine the type of firearms involved.

The simultaneous recording of Figure 10, obtained from a closed vehicle away from the scene, does not capture the utterances at the scene, but it does allow unclipped resolution of the individual gunshots. The recording also includes police radio traffic contemporary with the shooting incident, thereby helping establish a comprehensive timeline.

4 Conclusion

The availability of multiple audio recordings, including user generated audio recordings, is increasingly likely in audio forensic investigations. As described in this paper, the aspects of audio forensic investigations may have different requirements and goals compared to other applications with unsynchronized audio data, such as combined renderings of live concerts or other entertainment events.

As demonstrated in Section 2, the definition of what it means to have "time synchronization" will depend upon the frame of reference of the recording. Specifically, the absolute time of an event is recorded with the propagation delay from the source to the microphone, and therefore different sources and microphones will show different timing relationships. The timing differences may be very helpful in discerning details about the incident from a forensic standpoint, even though these differences may not allow a meaningful "mix-down" of the multiple UGRs.

For future work, two additional areas are of concern for audio forensic examination. The first area is the need to provide speech enhancement in the presence of noise if multiple UGRs contain the same sequence of recorded utterances. The process would presumably start with a time alignment for each utterance (compensating for the time of arrival differences) and then involve a linear combination of the available recordings to maximize intelligibility. There will be the same caveats described above regarding the ambiguity of time synchronization because of the different source-to-microphone distances and their relative motion.

The second area of forensic concern is authenticity of the recordings. With user generated audio material recorded by bystanders who may or may not have extensive recording experience and potential conflicts of interest, the customary law enforcement standard operating procedures and chain of custody expectations will not be available. The investigators may contend with a dispute about whether a particular recording was obtained with an unreliable device, or if the recording was deliberately—or inadvertently—edited or otherwise manipulated after the fact. The integrity and authenticity of the UGRs is especially important in sensational criminal cases, where there may be ulterior motives for various parties to submit forged or altered UGRs

5 Acknowledgements

This research work was supported in part by U.S. National Institute of Justice grants 2017-DN-BX-0179 and 2019-DU-BX-0019. The author gratefully acknowledges the assistance of the anonymous reviewers who commented on the abstract and précis of this paper.

References

[1] R.C. Maher, *Principles of Forensic Audio Analysis*, Springer Nature Switzerland, 2018.

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

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

[4] C. Cotton and D. Ellis, "Audio fingerprinting to identify multiple videos of an event," *Proc. ICASSP*, pp. 2386–2389, 2010.

[5] J. Bryan, P. Smaragdis, and J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," *Proc. ICASSP*, pp. 2389 – 2392, 2012.

[6] J. Kammerl, N. Birkbeck, S. Inguva, D. Kelly, A. Crawford, H. Denman, A. Kokaram, and C.

Pantofaru, "Temporal synchronization of multiple audio signals," *Proc. ICASSP*, pp. 4603–4607, 2014.

[7] S. Bano and A. Cavallaro, "Discovery and organization of multi-camera user-generated videos of the same event," *Journal of Information Sciences*, vol. 302, pp. 108–121, 2015.

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

[9] C.A. Dimoulas and A.L. Symeonidis, "Syncing shared multimedia through audiovisual bimodal segmentation," *IEEE MultiMedia*, vol. 22, no. 3, pp. 26-42, 2015.

[10] N. Stefanakis, Y. Mastorakis, A. Alexandridis, and A. Mouchtaris, "Automated mixing of user-generated audio recordings from the same event," *J. Audio Eng. Soc.*, vol. 67, no 4, pp. 201-212, April 2019.

[11] N. Gaubitch, J. Martinez, B. Kleijn, and R. Heusdens, "On near-field beamforming with smartphone-based ad-hoc microphone arrays," *Proc. 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 94–98, 2014.

[12] M. Kim and P. Smaragdis, "Efficient neighborhood-based topic modelling for collaborative audio enhancement on massive crowdsourced recordings," *Proc. ICASSP*, pp. 41-45, 2016.

[13] N. Stefanakis, M. Viskadouros, and A. Mouchtaris, "A subjective evaluation on mixtures of crowdsourced audio recordings," *Proc. European Signal Processing Conference (EUSIPCO)*, 2017.

[14] S.D. Beck, "Who fired when: associating multiple audio events from uncalibrated receivers," Paper 9, *Proc. AES Int. Conf. Audio Forensics*, 2019.

[15] R.C. Maher and E.R. Hoerr, "Audio forensic gunshot analysis and multilateration," Preprint 10100, *Proc. 145th Audio Engineering Society Convention*, New York, NY, October, 2018.

[16] R.C. Maher and E.R. Hoerr, "Forensic comparison of simultaneous recordings of gunshots at a crime scene," Preprint 10281, *Proc. 147th Audio Engineering Society Convention*, New York, NY, October, 2019.

[17] R.C. Maher and S.R. Shaw, "Gunshot Recordings from Digital Voice Recorders," Paper 6-1, *Proc. 54th AES Int. Conf. Audio Forensics*, 2014.

[18] D. Basaran, A.T. Cemgil, and E. Anarim, "Multiresolution alignment for multiple unsynchronized audio sequences using sequential Monte Carlo samplers," *Digital Signal Processing*, vol. 77, pp. 77-85, 2018.