Audio forensics
Not an episode from CSI

Francis Rumsey
Consultant Technical Writer

Audio forensic techniques rarely work like an episode from CSI. Professional members of the AES Technical Committee on Audio Forensics presented two tutorials on the topic at the 139th Convention, offering an excellent primer for those inside the field and out. The 2017 International Conference on Audio Forensics will be held in Arlington, VA, USA, June 15-17, 2017 (go to http://www.aes.org/conferences/2017/forensics/)

Audio forensics is a growing and increasingly mature field of work that has been represented by a number of international conferences over recent years. Leaders of the AES Technical Committee on the topic offered a pair of valuable tutorials during the 139th Convention, introducing people to the methods and techniques used in different aspects of such work. In the first, chaired by Jeff Smith of the National Center for Media Forensics at the University of Colorado Denver, with panelists Gordon Reid and Catalin Grigoras, the topics included speaker analysis and the application of Bayesian likelihood, best practices and future challenges in forensic authentication, noise reduction, and speech enhancement. In the second, chaired by Eddy Brixen, with Durand Begault, Rob Maher, and Keith McElveen, we learned among other things about recorded gunshots, microphone applications, musicological forensics, and acoustical issues.

AUDIO FORENSICS IS NOT LIKE AN EPISODE OF CSI

Kicking off the first tutorial Gordon Reid of CEDAR Audio emphasized that there are a lot of myths about dealing with surveillance audio. Unfortunately the world doesn't work like an episode of CSI where an expert can just turn a knob and pure noise miraculously turns into perfect, noise-free, wanted audio. There are in fact many reasons why you might want to adopt different approaches in improving the sound of surveillance audio, and none of them are magic.


Gordon Reid

Real-time latency-free surveillance equipment is vital when listening to live communications, for example. This is because you're often dealing with time-critical information in the field and can't wait for long processing times, but you might be willing to trade some lack of quality for being able to hear immediately what people are saying. After the event, perhaps when a crime has already been committed, then you may be able to wait longer to obtain the best quality results. If something is very difficult or unpleasant to listen to because of noise then an examiner will get tired very quickly, so there are advantages to delivering high sound quality in processed material, even if it is already intelligible. You can then listen "longer and better," said Reid. Presentation of material to the courts is also a critical area, and if non-trained listeners can't hear what is going on they could well come to the wrong conclusions. In such a case you might be wanting to maximize "listenability," whereas for examiners you might prefer "intelligibility."

Automated speaker recognition is a relatively new field, and recent research has suggested that some types of noise reduction can improve the performance of such systems. However the processing may change the tonal quality of the voice of the person talking, which may mitigate against the court agreeing that the person identified is actually the one talking. The courts have a general principle that material presented as evidence should not have been modified, but it is difficult to determine how far to take this when preparing material. Is it acceptable, for example, to edit out noise before and after the time of interest in a recording? There has, then, to be some interpretation of what changes are acceptable to recordings if they are to be presented

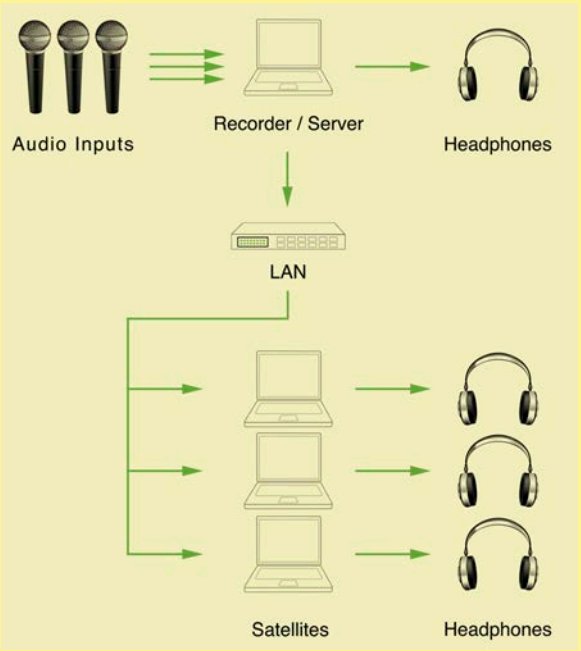

Fig. 1. Multi-input, multi-output surveillance system with networked satellite listening stations (courtesy Gordon Reid)

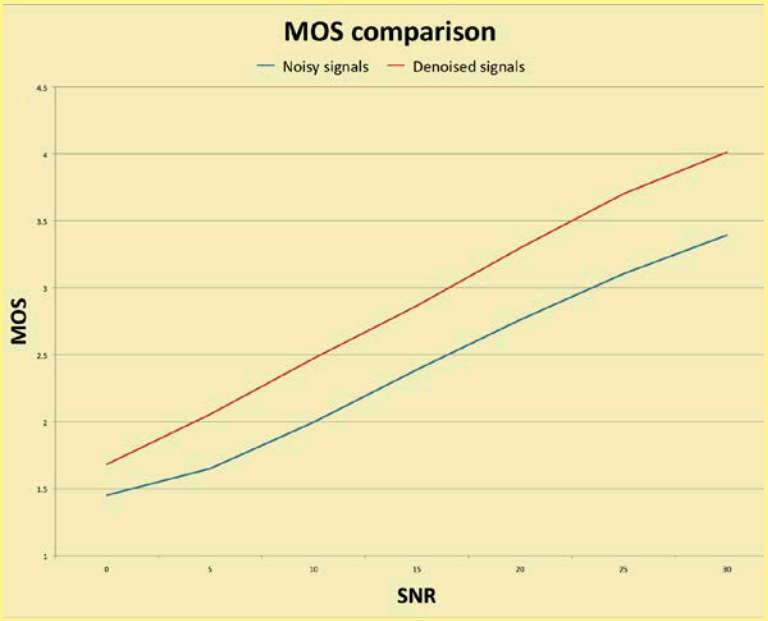

Fig. 2. MOS improvement with denoised signal (courtesy Gordon Reid)

in court. The criterion should be that people making any changes have to be experts—they have to be able to explain what they did and why in a clear way, and there has to be an exact evidential trail.

In the case of real-time surveillance noise reduction there are devices available with a latency of a few microseconds. Simplicity and speed of use are vital in such devices, because they are likely to be operated by field agents who may not be audio experts. Devices are coming onto the market that enable multiple feeds and multiple listeners, so that different people involved with a case can listen to different parts of a multi-source operation, with or without filters, and at different times, even while a recording is continuing (Fig. 1).

Gordon gave examples of different types of noise reduction, including removing GSM phone noise interference on a speech recording using a combination of de-buzzing and impulsive noise reduction algorithms. Considerable improvements are possible enabling one to hear what is being said quietly behind a strong buzzing tone. Single channel adaptive filters only have one input, and while the results can be impressive, they are less flexible than cross-channel adaptive filters that allow you to use two or more microphones, say with one of them having more of the interfering signal and the other with more of the wanted signal. With this type of filter, it's possible to use the former as a reference to help clean up the latter. Getting microphones in the right position to make this possible can be a real challenge, but if it can be done the results can be spectacularly good. Unfortunately, the number of occasions on which something like this can be used are very small, for practical reasons. Broadband noise reduction using processes based on spectral subtraction can give rise to unwanted artifacts so it has to be used with care, suggested Reid. It is possible to use speech quality prediction systems, such as those using PESQ (Perceptual Evaluation of Speech Quality) and MOS (Mean Opinion Score) measures (Fig. 2), to give an indication of whether processing has improved the result.

Compression and limiting can have advantages to manage the large dynamic range of some surveillance recordings, such as when a gunshot interrupts an otherwise quiet background. Balancing up the levels of the two sides of a telephone conversation can also benefit from it. There's also spectrographic audio editing to consider. Very small parts of a recording's time-frequency makeup can be accentuated or airbrushed out, and the process is very much harder to detect in authentication than is straightforward editing. It can be useful as a way of bringing out identifiable features to enhance them, but it's important to be aware that the "bad guys" also have access to such tools to hide things or change the evidence for nefarious purposes.

GOOD PRACTICE IN AUDIO FORENSICS

Catalin Grigoras

"We are right at the intersection of arts, sciences and justice," explained Catalin Grigoras of the National Center for Media Forensics at the University of Colorado Denver. "What happened, how, where, when, and who?" are the five big questions that people want the answer to in forensics. Keeping the audio evidence free from contamination is one of the critical mantras of the trade. For example, people use thumb drives a lot for storing data, including audio data, and when handling these in the forensic evidence chain it is vital not to modify the contents of such a drive. If you connect one directly to your computer, the chances are that something could be inadvertently changed, unless you employ hardware or software that prevents any writing of data to the drive—a so-called "write blocker" (Fig. 3).

During the analysis stage the highest accuracy and precision are needed, with the measurement errors and any possible bias being as low as possible. Repeatability is the third principal criterion to ensure during analysis—that is the possibility to get the same result when doing the same analysis again and again, and also possibly when

Fig. 3. A "write blocker" can be used to prevent anything being altered on a USB stick being examined. (Courtesy Eddy Brixen)

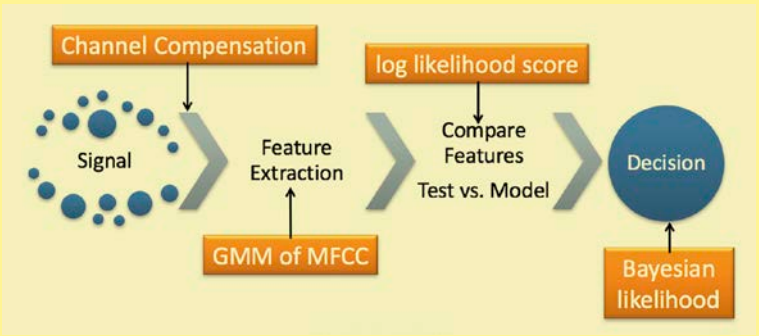

Fig. 4. The typical steps used in machine-based forensic speaker comparison (courtesy Jeff Smith)

conducted in another lab somewhere else in the world. Finally there is interpretation of the results where standing back from the data in an objective fashion is the key.

Do not change the evidence, be competent, document everything, and assume responsibility for what you do, it was suggested should be the bywords of the forensic audio expert. Taking photographs of the evidence at different stages is a very good way of providing evidence, such as whether a CD had scratches when it was first acquired rather than during the laboratory processes. Concluding his introduction, Grigoras emphasized that the way in which an expert presents his results is also a vitally important factor. If you don't know how to "sell" your results, the work can be compromised, he said. This means staying within a strict ethical framework and presenting the findings in a dispassionate, unbiased, and scientific way, in which case it is hard to challenge the motivations of the examiner.

FORENSIC AUTHENTICATION

Two signals that appear to be the same may need to be compared if one wants to say that they match, said Grigoras. While golden ears listening may be one option, mathematical comparison is very helpful. Likelihood ratios and correlation coefficients can be employed here, among other tools.

The big difference between analog and digital evidence is that it is impossible to clone analog evidence. We can copy analog evidence with very good quality but we can't clone it. It is, however, possible to clone a digital recording—a USB thumb drive could be cloned in such a way that its contents were bit for bit indistinguishable from the original. All one can say about a digital recording used in evidence, then, is that its characteristics are "consistent with" an original recording. When making copies of digital evidence in the lab it is almost always attempted to copy the bitstream directly, so that the copy is an authentic copy. Grigoras, though, was keen to point out the difference between intentional manipulation and other forms of change to digital audio data when discussing authentication. For example, one could re-encode a recording when copying it, in which case the data might be different (it would not be an authentic copy of the original), but this would not be considered a manipulation in forensic terms because the intention was not to deceive anyone using it down the line. In the same way, audio enhancement in the lab is not necessarily considered counterfeiting or manipulation, as the intention or effect is not to change the meaning of the information.

Grigoras also took a look at what can be done with the metadata of sound files. This is non-audio data that usually exists in the header of the file and can be used to gather information about parameters of the recording. Here you can find information about things like starting timecode, sampling rate, time of recording, and encoding method. Some DAW packages and file formats leave more "traces" than others of what has been done to the file and when. Additionally there are other methods of authentication such as compression analysis and electric network frequency analysis.

FORENSIC SPEAKER COMPARISON

Confirming the identity of the person talking on a recording is one of the key challenges of forensic audio work. Jeff Smith, of the National Center for Media Forensics at the University of Colorado Denver, introduced the principles, explaining that the human brain has evolved as an outstanding pattern-recognition system with an innate ability to recognize identities familiar to them based on voice alone. However, since this system can be influenced or biased by various cognitive factors it must be combined with more objective computer analyses in forensics to provide a potential for generating convincing evidence that can be presented in court.

Speech-related information is a form of biometric data, suggested Smith, but it's dynamic, and it is not as strong or reliable a biometric as something like a genetic fingerprint or a physical fingerprint. Speech provides information about the speaker's mood and there is a wide range of cultural differences in the ways people speak, for example, so the challenge is to discover what are its unique characteristics. Formants are the peaks in the frequency spectrum that correlate to particular vowel sounds, and they help to define the unique character of a person's speaking voice. It's actually the relationship between the formant peaks that are the most important factor, rather than their absolute frequency, as the absolute pitch of speaking varies with the person and their emotional state.

Features and metrics can be extracted from speech recordings that computer models can use to predict whether one speaker is the same as that in another reference recording or suspect database (Fig. 4). There are various problems, though, in using this information to identify a speaker, not least that the recording quality may be poor, many speakers are uncooperative (they don't want to help the exam-

Jeff Smith

iner), they may disguise their voice, or they may be in a highly emotional state. The term "identification" is consequently rarely used in forensics, because there is never 100% certainty that one has identified a speaker, there is only a degree of likelihood. The decision about similarity between one voice and another is usually expressed in the form of a Bayesian likelihood, in other words a statistical estimate of how likely it is that one voice is the same as another.

If you find metrics that show a good match between one recording and another it's important, said Smith, to determine whether or not that is due to the channel characteristics rather than the speech itself. Compensating for the differences in channel characteristics between the recordings being compared is therefore an important aspect of the work. Asked whether the bandwidth of the recording made a lot of difference to the process of comparison, Smith said that in practice it doesn't because the analysis sampling rate is usually limited to 8 kHz and the majority of useful formant information to be analyzed is consequently limited to 4 kHz.

GUNSHOT ANALYSIS


Rob Maher

A question that had intrigued Rob Maher of Montana State University when he first got into the world of audio forensics was how much it might be possible to identify about a gun from a recording of it. An attorney had asked him whether it might be possible to identify the exact model of weapon and its serial number from a recording, and his initial reaction had been "of course not, one gunshot is pretty much the same as another." His work since then has uncovered a lot more about what is and isn't possible in this domain. Because increasing numbers of recordings are available from law enforcement agencies these days, the audio forensics world needs to know how to handle them.

The muzzle blast of a gunshot is extremely short, lasting only a few milliseconds, said Rob. Recording such a shot in an anechoic environment makes it possible to discover the inherent characteristics of the impulse and its aftermath, free of any reflections. Once you put it in a reflective environment most of the information in a recording tells you things about the acoustics of the space rather than the shot itself, although gunshots are quite directional so there are some features that can be determined about the relationship between the gun and the recording microphone. The microphone may or may not be placed in a desirable location, often being on the dashboard of a police vehicle or picked up from a device carried on someone's clothing. It's quite common for recordings to be clipped because of the dynamic range of the impulse, and there may be automatic gain control (AGC) as well as various forms of perceptual or speech coding. Typically AGC takes time to react so a gunshot recording clips initially, then the gain is progressively reduced, which can affect the level of the succeeding reverberation. Maher gave an example of a gunshot recording made in reverberant surroundings with a digital voice recorder, showing that it was actually quite hard to determine the exact start time of the shot. In many cases recordings from such devices make it almost impossible to determine when a shot occurred, how many shots there were, and where. Sometimes what seem like multiple shots are in fact a single shot with subsequent "echoes" that result from device overload, AGC reaction, and recovery.

A questioner raised the important point about why the recording quality of the devices used by law enforcement agencies in these situations is so low. Why use a simple speech recorder when much more advanced systems with microphone arrays and wide range recording are available? Rob didn't have an easy answer to this, but we must assume that a lot of the time these are low-cost, general purpose systems that have to be issued to large numbers of people and have to be able to record for many hours at a time. It is to be hoped that as costs of more sophisticated devices come down and they become more widely available the quality of available recordings will go up. It's possibly also the case that the video recording capabilities of so-called "dash-cams" and "body cams" have been prized above the quality of their audio recording. Another questioner felt that there really isn't much guidance about what is acceptable technology for making such recordings and that audio experts could lead in this area. Maher agreed that some sort of standard or guidelines would help and perhaps the AES could move in this regard. There's also the question, he said, of how agencies should archive such material and with what degree of quality, another area where developments are slowly happening.

MICROPHONE TECHNOLOGY

The improvement of microphones for forensics purposes was the topic of Keith McElveen's (Wave Sciences) contribution. The differences between restoration and enhancement are critical here, and Keith likened the processes to those used by plastic surgeons. Restoration involves putting something back as close as possible to how it should be, after having been damaged in some way, whereas enhancement involves making some sort of cosmetic improvement, changing the material from how it was originally. Improving the signal-to-noise ratio generally helps the situation in almost all circumstances, and if you can do that with the microphone in the first instance this is better than trying to do it using magic processes afterwards. Subsequent processing can often introduce artifacts so it's desirable to try to improve the acoustic S/N ratio during recording by improving the microphone selectivity.

One of the biggest challenges in audio forensics is to separate one person's voice from another, particularly when there is a lot of speech babble in the recording environment. Poor operator technique is one of the most common reasons why recording quality is bad, with microphones badly located and having clothing noise, for example. Getting the microphone closer to the person talking is clearly the most valuable thing to do, but not always possible. McElveen gave an example of a woman who had confessed to killing her child in a statement to an officer, but where the recording device was close to the officer interviewing her and he had started talking over the top of her. She was speaking quietly because she was emotionally distraught and the result was a recording in which what she said was indistinct. A lawyer had subsequently asked for the recorded statement to be suppressed. Having the microphone closer to her would have made a lot of difference.

Spatial filtering can be used, said Keith, as a way of processing a signal arriving at a microphone based on the angle of inci-

dence. The further away a talker is from a microphone the poorer its S/N ratio will be compared with static ambient noise, so the aim is to increase the gain only in the direction of the talker. With two or more microphones the pickup can be made more selective as a result of the phase difference between them. The more distance there is between wanted talker and interferer the more microphones will be needed to improve the selectivity. Cardioid or shotgun microphones are useful up to a point, but are not particularly useful beyond a relatively short distance, which points to the need for microphone arrays.

ACOUSTIC CONSIDERATIONS


Eddy Brixen

Among the issues discussed in relation to acoustic analysis, Eddy Brixen of EBB Consult mentioned the need to identify the room in which a recording was made. For example it might be that a person had made a phone call from one of a number of possible rooms in a house but couldn't remember which one, leading to a need for forensic analysis to help determine it. Such a recording could involve the need to understand the mobile device's audio coding system and its possible effects on the audio signal. Trying to evaluate the room's reverberation time or reflection patterns from recordings made of phone calls in situations such as these requires the taking into account of multiple factors. You have to be able to demonstrate that the signal chain and coding/recording process makes such measurements reliable. Similar things are true of gunshot recordings, where features of the recording and transmission channel can smear or distort features of the wanted signal, making it difficult to determine what are real acoustical reflections and what are by-products of the channel. It's particularly important in such cases to evaluate the effects of the channel before undertaking formal analysis.

FORENSIC MUSICOLOGY

Apart from the more standard areas of investigation discussed above, there are less widely known or discussed areas of forensic audio analysis, such as forensic musicology. Music copyright infringement cases involving famous artists is often newsworthy, but the techniques and testimony of the forensic musicology experts themselves is often not discussed. Durand Begault of Audio Forensic Center at Charles M. Salter Associates in San Francisco pointed to a particularly useful web resource provided by the USC Gould School of Law known as the Music Copyright Infringement Resource, to be found at http://mcir.usc.edu. This gives a background on how cases relating to pieces of music have been contested over the years.


Durand Begault

Compositional analysis, said Durand, is sometimes needed when someone is trying to claim that a piece of music has been copied by another composer. It typically looks at the sheet music independent of the recording. Recording analysis looks at the recording itself, independent of the content, whereas production analysis partly overlaps with computer forensics and has to do with the path from creation to distribution.

Compositional infringement is a bit more difficult to prove than recording infringement, as it requires one to prove that the infringer had previous access to the claimed music, that there is a striking similarity in terms of melody, harmonic progression, rhythm or structure, and that the elements are indeed copyrightable. In one case mentioned a particular musicological expert on one side had pointed out a number of similarities between two songs, and concluded that they were essentially the same song, whereas an expert for the other side had concluded that the two were not "meaningfully similar." Arguments against infringement included that the structure was a commonly used format, that no consecutive words were the same, and that none of the infringements claimed were unique to this song. Two people who seem eminently qualified in the field can apparently analyze the same thing and come up with strongly held and contrasting opinions. Begault contrasted this with the scientific field where one would expect this to be much less likely. Although there have been many attempts to make musical analysis "scientific," forensic musicologists often disagree in their conclusions regarding infringement.

Very few forensic science methods, though, have come up with adequate measures of the accuracy of their inferences, Begault claimed (as did the NRC report with reference to forensic science in general: see http://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward report). He recommended that all analyses should include an indication of the degree of uncertainty involved, and a determination, at the outset of an analysis, of the criteria to be used for the expert's decision. In the musicology field as in many areas of forensics, there is often no established "ground truth," usually because each case is unique, whereas the scientific method allows verification through the use of repeated analyses in controlled experiments and the application of statistics. Forensic musicology experts can improve their reportage by setting out their methods in advance, for example, in order to support transparency in their analyses. Ultimately, in the U.S. courts, the role of the forensic musicology expert is to provide an extrinsic analysis to aid the trier of fact, who is typically a lay person. But it is an intrinsic analysis made by the lay person/trier-of-fact as to whether or not a plaintiff's claim of infringement is valid.

CONCLUSION

From the tutorial material presented at the convention by the professional members of the AES's Audio Forensics Technical Committee, it is clear that the scientific credibility of the work has come a long way in a relatively short time. That said, it is still a relatively young field that appears to be capable of improvements in its analytical methods and evaluation of evidence, particularly with regard to ways of stating uncertainty.

Editor's note: to purchase recordings of the tutorials go to http://www.mobiltape.com/conference/Audio-Engineering-Society-139th-Convention
2017 Audio Forensics Conference, go to http://www.aes.org/conferences/2017/forensics/