Audio Engineering Society UK

Title: 'An introduction to forensic audio'
Location: Royal Academy of Engineering, London
Description: Lecture by Gordon Reid, Managing Director of CEDAR Audio Ltd

Speech enhancement has come a long way in the digital era, but it is not the 'magic wand' depicted on TV and in Hollywood movies. Adaptive filters have traditionally been the basis of forensic audio work, but a combination of techniques – including broadband noise reduction, buzz removal, equalisation and background noise suppression – can provide superior results when compared with any single approach. This introduction, illustrated using examples processed in real-time on a CEDAR Cambridge Forensic system, aims to shed light on this, demonstrating how signal processing can aid investigators in areas including criminal investigation, counter-terrorism and air accident investigation.

Lecture Report

Note: we are unable to provide a recording of this lecture because some of CEDAR's police and security customers place strict constraints on the public dissemination of the audio clips and details of cases used in demonstrations of CEDAR's forensic technology.

Gordon Reid is the Managing Director of CEDAR Audio, a leading manufacturer of audio restoration and speech enhancement products. He kicked off his lecture with a scenario of how video surveillance, without audio content, can give ambiguous or even completely misleading indications of intent.

Audio forensics is a relatively new field that first entered common use in the 1960s/1970s. Thanks to the technology of companies like CEDAR, audio forensics is now an established field, and the most recent trend is for audio and video surveillance data to be integrated. Before the arrival of digital technology in the 1990s, audio forensics was relatively crude, using often poorly-maintained analogue tape recorders, no single-ended noise reduction, and often just analogue EQ and dynamics processes for clean-up.

Nowadays, recordings are mostly digital, and can be made using low-cost consumer equipment. But this brings some new problems. Recordings are often made by untrained people using small, cheap recorders: he highlighted a divorce litigation case in which a woman concealed a recorder at the bottom of her handbag, covered by a scarf and jumper to make sure it wasn't found. Unsurprisingly, there was almost no discernable speech data on the resulting recording. So there are new problems to face, but fortunately, DSP algorithms and powerful computers can help get around many of these. But even these have limits: Gordon described a phenomenon known as the "CSI

Effect", whereby the public has unrealistic and fantasy-based expectations of surveillance restoration technology. He cited the apparently genuine example of a person who'd snapped a photo of the side of a speeding getaway vehicle on a mobile phone, and handed it to the police in the expectation that by rotating the side-on (and blurred, low-quality) image in a 3-D computer imaging system, they could read the license plate! But absurd cases aside, there is an increasing problem: the bad guys are increasingly aware of surveillance techniques, making (for example) body wires impractical because criminals know how to frisk for them effectively. They also know to hold sensitive conversations in locations where there is loud, effective masking noise such as running water or TV noise.

Gordon broke noise reduction technologies for audio forensics into two main applications: real- time surveillance and non-real-time laboratory investigation. Surveillance systems have live listeners (typically police or security officers), who may need to make fast, accurate and life- critical decisions based on what is heard. The principal requirements are low latency, high intelligibility and low listener fatigue. Non-real-time systems are typically used to produce evidence admissible for the courts, so the requirements are for high transcription accuracy, the retrieval of otherwise unintelligible speech, and to reduce transcriptor fatigue. Also, jurors are not trained listeners and courtrooms typically have very poor acoustics, so the presence of background noise may affect their judgement. He cited the case of a defence lawyer who used the presence of modest traffic and street noise on an intelligible recording of incriminating statements to cast doubt on the transcription of the recorded speech – and won.

Gordon listed the long-established principles of good non-covert audio evidence: a suitable recorder, competent operator, authentic recordings, recordings preserved such that they are demonstrable in court, speakers identified, evidence made voluntarily and in good faith – and no edits or changes made. The last point is potentially problematic as, in principle, it could exclude the enhancement processes that render noisy evidence intelligible. This is a grey area, with the degree of processing admissible dependent on the judge, court and jurisdiction. Clearly, there is a need to demonstrate that the processing has not modified the meaning of the evidence. For example, it's not possible for the microscopic editing of a real-time declicking algorithm to change phonemes, and so change the meaning, but the court may need to be convinced of this. Additionally, proposed UK government regulations on handling evidence may be applied to audio evidence, potentially causing substantial problems when regulations designed to protect physical items are applied to digital media.

Gordon moved on to talk about the specifics of the technology used: it's usually some combination of noise reduction, equalisation and level processing (e.g. dynamics processing). Dialogue noise suppression is a technology originally developed for the film industry, and CEDAR's first product in this field, a real-time and very easy to use

device, was aimed at post- production for film, video and TV: the typical application was to save a take that had been spoiled by ambient sound intrusion. This was contrasted with lab systems: large computer- based systems intended for off-line batch processing rather than real-time use.

The use of declickers was demonstrated. The earliest algorithms in this field were originally developed for 78rpm archives, but have been developed much further and are now extremely helpful in removing GSM noise, the familiar buzzing/pulsing interference caused by mobile phones. GSM noise can be shown to comprise buzz at around 217Hz and a series of pulses. The declicker can remove the impulsive noises, and the buzz can be removed with a dedicated Debuzz algorithm. The results of this were demonstrated with a 999 call recording, originally almost completely inaudible, but which when processed revealed much more information and the presence of a second, previously-unheard speaker in the background – of crucial important to the court case in which the recording was presented as evidence!

Gordon next discussed the use of adaptive filters. If the statistics of the noise are relatively constant, it's possible to design a filter to separate speech (which tends to change rapidly) from the noise. Additional improvements can sometimes be achieved by treating low and mid frequencies differently to high frequencies, based on perceptual models of hearing and intelligibility.

Some of the interesting applications of adaptive filters include cleaning-up reverberant spaces such as holding cells and transfer vans, and removal of the 400Hz buzz from aircraft power systems that can degrade air traffic control recordings. And, in a reversal of the normal filtering, it was described how CEDAR removed the shouting from a cockpit voice recording in a helicopter that had just suffered a catastrophic mechanical failure, so the investigators could listen to the mechanical sounds to trace the cause of the accident.

Cross-channel adaptive filters can overcome steps taken to defeat surveillance, such as using loud radio or TV to mask a conversation. This type of filter exploits the correlation between the direct broadcast signal (if available) and the tonally altered broadcast signal present in the surveillance, and can effectively remove it from the surveillance signal. If there isn't a convenient reference of the broadcast, use of multiple microphone locations causes some to have more speech and others to have more interfering signal, giving the cross-channel adaptive filter enough to work with. A reconstructed demonstration was played in which, when using a single mic recording of some speech in the presence of loud music from a radio, a transcription expert obtained approximately 30% accuracy. Adding a second mic positioned closer to the radio than the first and using this as the reference channel for the cross-channel adaptive filter, the intelligibility was hugely improved, and the transcription accuracy increased to 100%.

The form of broadband noise reduction known as spectral subtraction is an impressive tool in music production and restoration, but in forensics its use can be more limited: although it improves listenability and reduces fatigue, the best that can be hoped for regarding intelligibility is that it doesn't damage it. Nonetheless, it has significant other uses in audio forensics, such as removing the hiss that can be added by adaptive filters. EQ, despite its simplicity and ubiquity, has been a staple processor for forensics since long before the days of DSP and adaptive filters. Removal of low frequencies and the addition of a little boost in the upper mids can hugely increase intelligibility. Limiters are used to reduce the impact of sudden loud noises. By its nature, forensic audio can involve extreme dynamic ranges. When a surveillance officer or transcriptor is listening closely to very low-level signals at very high gain, loud sounds such as gunshots/vehicle crashes/etc. can, without limiting, damage the listener's hearing. In other cases, such as a recording of a telephone conversation made using a hand-held recorder, balancing the levels of the local and remote speaker can help render the evidence more intelligible and therefore more useful in court.

Gordon mentioned the increasingly widespread suspicion that audio data mining is being deployed by security agencies: that is, mass interception of all voice communications with automatic recognition of certain key words (e.g. bomb, jihad, etc.). Gordon's view is that this is not currently technically practical, but that its use may increase within a decade or two. What is currently feasible, and is being used to an ever greater degree, is automatic speaker recognition: commercial solutions are developing fast, but their robustness to voice signals that have been altered by enhancement processing is an ongoing research field. Another significant recent development is the prevalence of low bit-rate, highly-compressed perceptual codecs, which can make both enhancement and automatic speaker recognition more problematic.

Gordon concluded his lecture with a mention of spectrographic editing, which was invented by CEDAR. Time-domain editing can be recognised in a spectrograph, making this kind of evidence-tampering obvious. But spectrographic editing allows powerful manipulation of the signal, often invisible to future investigation. This tampering can be very dangerous in the wrong hands, but when used ethically can reduce or remove masking signals, making it a powerful enhancement tool.

Many thanks to Gordon for an eye-opening lecture, and his fascinating insights into the remarkable technology his company has created.

Lecture report by Michael Page