Audio Forensics

By Staff Technical Writer

Last summer, Denver played host to the first AES conference on audio forensics, emphasizing the way that this specialized field has adapted to the digital age. In the September 2005 *Journal* (p. 838) we provided a detailed report of the proceedings of that conference, and now in this article we review in greater depth some of the work presented there in the fields of voice identification, spectrographic analysis, and signal-enhancement techniques.

Defined at the conference by Rich Saunders as "the study and examination of audio, recorded or otherwise, as it pertains to finding a truth," audio forensics is strongly linked to the legal profession. It has risen in importance over the years as methods of validating audio material have gained credibility and as such material has increasingly been regarded as admissible by courts. Challenges facing the audio forensics examiner include a wide range of identification tasks as well as the need to validate the authenticity of recorded materials and to determine whether evidence has been modified or tampered with. In the digital era, some of the tools used by the examiner will have changed and many of the formats used for recording will bear little resemblance to analog tape. Furthermore, there are problems to contend with such as the effect of mobile phone speech codecs and other forms of modern communication.

VOICE IDENTIFICATION

The ability to identify the person who is talking in an audio recording is a crucial part of many forensic examinations. This can be a relatively straightforward task if the question is one of determining which member of a known group is speaking at any instance, or much more complicated if it requires identifying an unknown speaker in a very large population. A more common task is to determine whether a voice sample taken from a forensic recording purporting to contain evidence is likely to be from the same speaker as exemplar recordings made of potential suspects. This becomes a question of matching or detection, with the conclusion reached by the examiner having a certain probability of being accurate.

Tito and Begault describe a technique used in forensic voice identification that uses aural-spectrographic protocols. It relies on a trained examiner's ability to determine whether known and unknown speech exemplars were produced by the same speaker or by two different speakers. The authors explain that this method uses both aural comparison and spectrographic analysis to inform the decision. The expert attempts to form an opinion about the similarity between spectrograms using a form of gestalt pattern matching rather than by specific quantification of individual features. Apparently, listening to the overall or gestalt speech has been shown to be more reliable than attending to individual physical or psychophysical measures of speech, probably because it relies on the many cognitive processes that humans have developed over the years to assist them with talker identification. Although there is no scientific evidence indicating that trained examiners do better than laymen at discriminating voices, it seems that, in general, aural voice identification can be quite accurate. In studies reported by the authors, false identification of a voice is less common than missing a correct identification, which is in line with the belief that it's better to lean toward not convicting a guilty person rather than taking a chance of convicting an innocent one.

Bias in identifying voices is covered by Tito and Begault at some length, as they feel this topic is not sufficiently appreciated. For example, they point out that while a line-up of suspects is often used for visual identification, it is quite common to compare a voice sample with only one suspect's exemplar recordings. This is partly because of the costs of undertaking more comparisons, and also seems to have something to do with increasing the chance that an examiner will make a false identification (regarded as a potentially career-threatening situation). The authors argue that the human costs of such corner-cutting are potentially high and that it is ludicrous to attribute infallibility to examiners. Consequently they propose three different types of line-ups that can be employed: closed set, closed sequential, and open sequential. Each of these has different degrees and types of potential bias, the first type probably being the least desirable because it includes the a priori probability that the guilty party is one of the known voices. The latter two require the examiner to make a judgment on each pair of ➥

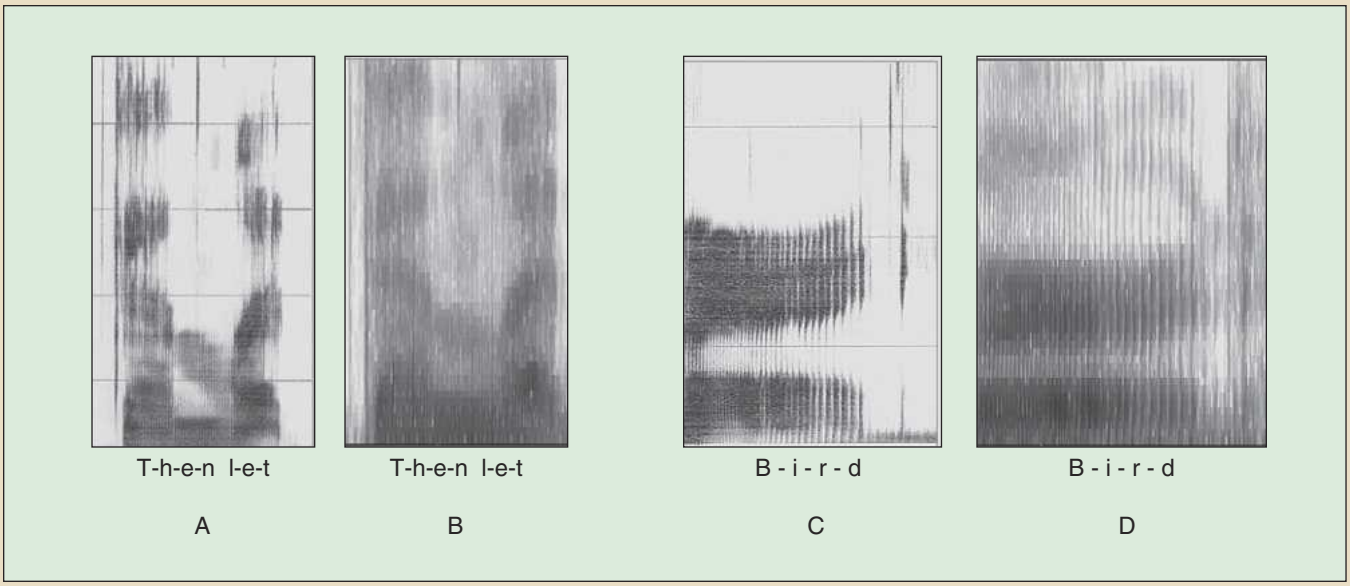

Fig. 1. (A) Analog spectrographic representation of subject Heather reciting the words "then let." (B) Digital spectrographic representation of same phrase. (C) Analog representation of the diphthong in subject Shawn's reciting of the word "bird." (D) Digital spectrographic representation of same word. (Courtesy Harper et al.)

known-unknowns in a sequence before moving on to the next, and long-term memory limitations make it more difficult to remember earlier exemplars when listening to later ones. This makes the chances of a false identification less likely.

COMPARING ANALOG AND DIGITAL SPECTROGRAPHS

Analog and digital spectrographs are compared by Harper et al., in an attempt to evaluate the potential for loss or modification of spectrographic data when using digital systems. They cite the situation with digitized fingerprint photographs in this connection, noting that in some cases serious damage had been caused to their accuracy. In order to make the comparisons between analog and digital systems as fair as possible the authors attempted to make the settings of filters, ranges, windows, and bandwidths as close as possible. When carrying out the comparison between one pair of extant systems, they evaluated four basic differences between the appearances of spectrographs: general change in definition, change of shape, blurriness, and definition between transitions. They concentrated on four areas of speech representation: vowel formants, diphthong formants, transitional areas, and consonants. In a second comparison they evaluated general change in definition, pixelation, change of curvature, and definition between transitions; these being the dominant differences in the second case. The comparisons in Fig. 1 show that the differences in representation can indeed be quite marked.

In order to gather quantitative data the authors coded differences likely to result in a mistake in identification of the voice using the value 1. Total differences were then counted over the different voice features and voices used. Overall the trend in the analysis showed slightly more differences between analog and digital spectrographs for female voices, leading the authors to conclude that although a more detailed study is necessary to arrive at reliable answers, there are enough differences between the displays of spectrographs to warrant careful further study.

CAN YOU TELL IF I HAVE A BLOCKED NOSE?

Smith et al. were interested to find out whether spectrographic analysis can assist in discovering the effect of subtle changes or alterations of a voice, and whether factors such as a blocked nose, or mimicking another person, can affect the speaker's "voiceprint." Ten males and ten females recorded four phrases with a range of vowel and diphthong sounds in three versions, including a mimicked sample. The first version involved the use of a natural voice. The second two consisted of a version with the speaker having a pinched nose and one spoken in a higher register. Spectrograms of the different versions were visually analyzed, and linear-predictive-coding analysis was undertaken to discover the changes in mean formant frequencies in each case.

Results of these experiments suggest that those involved with voice identification do not need to be extremely concerned when a subject has a blocked nose, as the main features of the spectrogram are maintained. Nonetheless, some subtle changes in formant structure were observed as a result of changes in the vocal resonators in the head. Of greater concern is the observation that a skilled mimic can closely imitate another person's voice in terms of formant frequency comparison. One subject, for example, was able to match the person she was imitating to within ±49 Hz in terms of the four vocal formant frequencies.

POSSIBILITIES FOR AUTOMATIC VOICE IDENTIFICATION

It is interesting to consider whether automatic systems might be capable of distinguishing voices more accurately than human examiners. In "Speaker Recognition Method Combining FFT, Wavelet Functions, and Neural Networks," Grubesa et al. describe a system that decomposes an averaged spectrum of a speaker's voice using wavelet functions, after a subdivision of the spectrum into 22 subbands consistent with human auditory filters (Bark bands). Approximations to the spectral function in each

TABLE 1
COMPARISON OF TEST RESULTS FOR DIFFERENT METHODS OF SPEAKER RECOGNITION

	System 1 / %	System 2 / %	System 3 / %
Normally read text	75.0	87.8	94.6
Text read in an attempt to deceive the system	26.6	28.3	48
Text read in an acoustically different environment	62.6	85.2	92
Normally read text with white noise added	42.5	60	58.1
Normally read text with distortion added	62.5	48.7	66.4
Total	47.5	52.5	65.8

subband are used as input data to a pretrained decision-making neural network, and the decisions made for each band are weighted and summed to make a final decision.

In a first system developed by the authors a neural network attempted to make a decision based on the averaged spectrum of the recorded voice of the speaker. Unfortunately, this was highly erroneous and could easily be deceived by a speaker wishing to fool the system. Furthermore, the system was intolerant of distorted or noisy recordings. Dividing the spectrum into Bark bands helped in improving immunity to noise and distortion, but otherwise the results were still relatively poor and the system showed some problems distinguishing similar speakers such as brothers, or fathers and sons. Improvements were therefore sought and it was found that immunity to noise and distortion could be greatly helped if only a few characteristic points were isolated from the spectrum, but this reduced accuracy and precision of identification so a compromise was required. Static or averaged spectral characteristics are of limited value in voice analysis so the authors looked at changes of those characteristics over time, which are distinctive.

Testing the resulting versions of the neural-network-based system, the network first had to be trained for every speaker included in the database. During recognition, values for the test sample were compared with coefficients from the database and likely matches were determined if the result was above some threshold calculated for each speaker. The probability that the decision is correct could also be shown. The system was tested using a number of different versions of a spoken phrase, including text read in an attempt to deceive the system, text read in an acoustically different environment, normally read text with white noise added, and normally read text with distortion applied. The results in Table 1 show that the final system performed well for the normally read text, and also well for text read in a different environment. It is a good compromise for samples with noise and distortion added, and it is much better than the other two systems at detecting people attempting to deceive the system. It is not clear how many voices were used in the training of the network or in the test exercise.

DOES LOW BIT-RATE SPEECH CODING CAUSE PROBLEMS?

Eddy Bøgh Brixen and Durand Begault investigated the validity of bit-compressed digital voice recordings for spectrographic analysis. This is because they were concerned about the ways in which voice codecs, such as those used in mobile phones and other portable devices, might modify the way in which the voice signal is reconstructed. In a preliminary investigation of this question they evaluated G.723-encoded recordings at a maximum bit rate of 6.3 kbit/s, as used in IC recorders and surveillance equipment, and the MSV LPEC-SP codec used in the Sony Memory Stick format. LPEC stands for long-term predicted excitation coding. G723, designed as a dual-rate speech codec for multimedia applications, uses linear prediction analysis-by-synthesis coding.

Voice exemplars—one in Danish-accented English and the other in American English—speaking the phrase "forensic audio analysis" were recorded by both Brixen and Begault. Background noise was introduced in some samples. In order to analyze the samples, a package called STx from the Acoustic Research Institute of the Austrian Academy of Sciences in Vienna, was employed. The authors concluded that there were minor differences with the two compression formats tested, but the formant shaping and positions were very similar. The fundamental pitch was unaffected, and it was indicated that these two schemes did not significantly change the key features of the spectrograph in comparison with the linearly coded reference, leading to the suggestion that one can reliably compare exemplars across the schemes. An example is shown in Fig. 2, which shows that the main difference with the version subjected to low bit-rate coding is a general increase in noise-like artifacts across the spectrum. Brixen and Begault, however, noted that an aural comparison of these schemes had not been conducted and would be strongly influenced by the ➥

THE WATERGATE TAPES

The most famous example of audio forensic analysis—and the legal and political ramifications of such an analysis—is the Watergate tapes. In 1973 the U.S. Distric Court for Washington D.C. assigned six technical experts—Richard Bolt, Franklin Coopeer, James Flanagan, Jay McKnight, Tom Stockham, and Mark Weiss—the task of verifying the integrity and originality of audio tapes reccorded in the Nixon White House. Among these tapes was one with an 18-minute gap, which the experts determined to be an erasure "so strong as to make the recovery of the original conversation virtually impossible." Jay McKnight, chair of the AES Historical Committee, has posted a downloadable scan of the report, under the heading "Forensic Audio Engineering," at http://www.aes.org/aeshc.

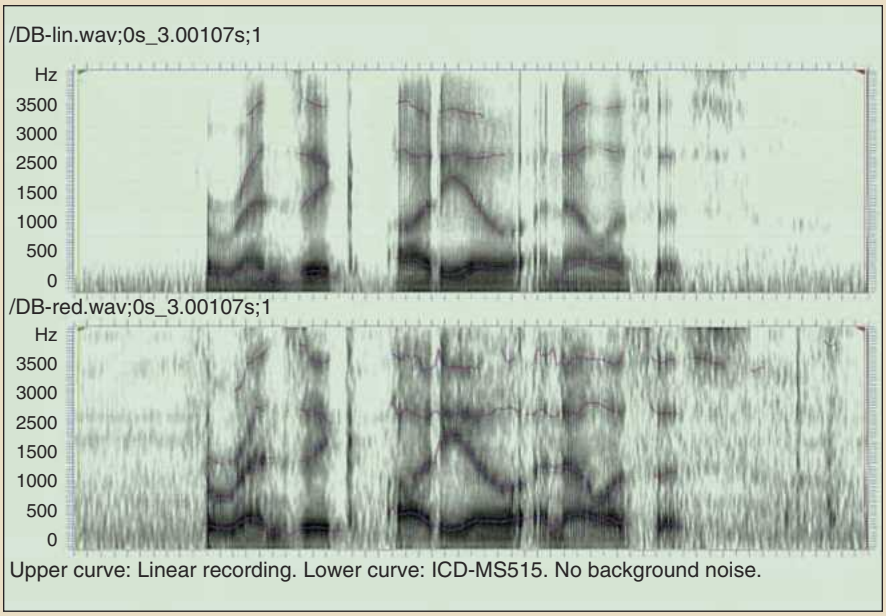

Fig. 2. Comparison of spectrographs of a linear PCM recording and a bit-rate-reduced version of the phrase "forensic audio analysis" (courtesy Brixen and Begault).

overall bandwidth differences concerned.

IMPROVING FORENSIC RECORDINGS USING ADVANCED FILTERING

Audio enhancement using nonlinear time-frequency filtering was described by Robert Maher. He addressed situations in which forensic audio recordings contain undesired noise that can impair source identification. In his paper he states the need for a single-ended noise-reduction approach that can operate with no information other than the noise-degraded audio signal itself. This is not a new problem, but one that is tackled in a novel way for a particular forensic application. Maher also explains that an approach is needed that can adapt to the short-term signal behavior and enhance the signal in such a way that an examiner will find it better and more useful than the unprocessed original. Starting off by introducing the concept of spectral subtraction, he shows how the noise level as a function of frequency must be estimated before subtracting it from the received spectrum. Sometimes the result of this process can be a residual bird-like or tinkling noise when the noise level varies or when the match between the subtracted and actual spectra is not correct. He makes the valuable point that the quality and effectiveness of the spectral subtraction technique is dependent on the forensic task in hand. Different tasks may give rise to a greater or lesser tolerance for noise and for different types of residual distortion and side effect.

The technique proposed by Maher attempts to distinguish between coherent elements in the desired signal and incoherent (noise-like) elements in the unwanted signal. To achieve this he looks for signal components in the short-time Fourier transform (STFT) that behave consistently over a short time window. A two-dimensional filter is used, having different time and frequency resolutions, which can be

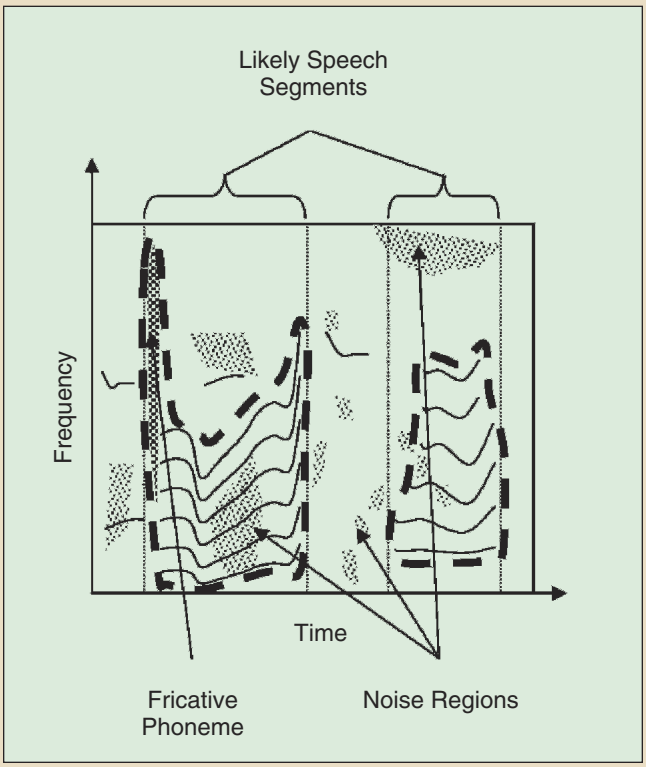

Fig. 3. Noisy speech with segments identified (Figs. 3 and 4 courtesy Maher)

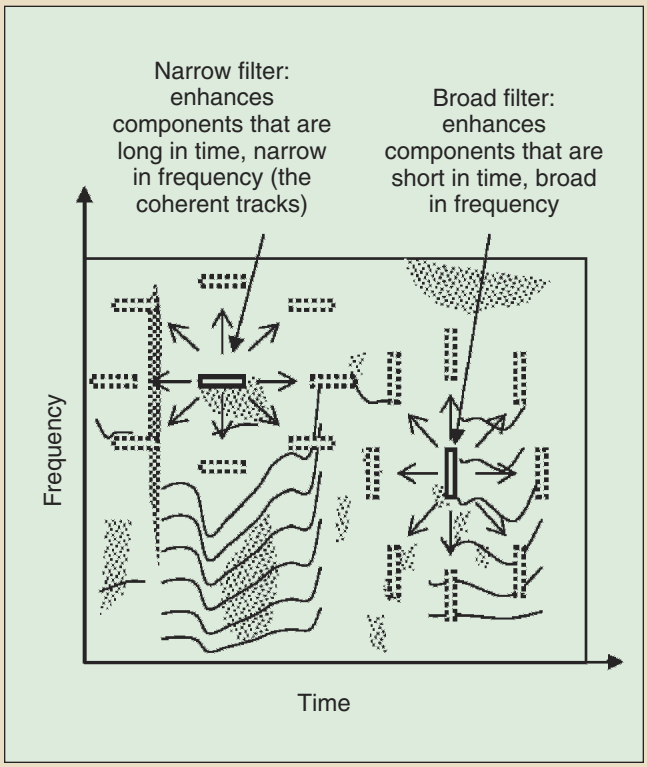

Fig. 4. Representation of Maher's 2-D filtering concept

adapted to the signal characteristics at any point in time. The system identifies and treats in different ways those sections of a speech recording that are transitions from voiced-to-silence, voiced-to-unvoiced, unvoiced-to-voiced, and silence-to-voiced. He shows examples of spectra in which likely areas of speech and noise are identified (see Fig. 3). Typical features of voiced speech are sets of parallel tracks in the spectrum, representing harmonics of the voice spectrum. Consequently, one filter is designed to be narrow in frequency and broad in time. Fricatives will generally appear as patches of noise-like information that are narrow in time but broad in frequency, whereas unwanted noise tends to appear as randomly located patches. Fricatives are very important for speech intelligibility and good voice identification, so they must be correctly preserved. In Maher's system they are preserved by using a filter that is broad in frequency but narrow in time, and by means of a pattern-detection process that allows such components to remain at boundaries between unvoiced and voiced speech elements. An example of Maher's 2-D filtering process is shown in Fig. 4.

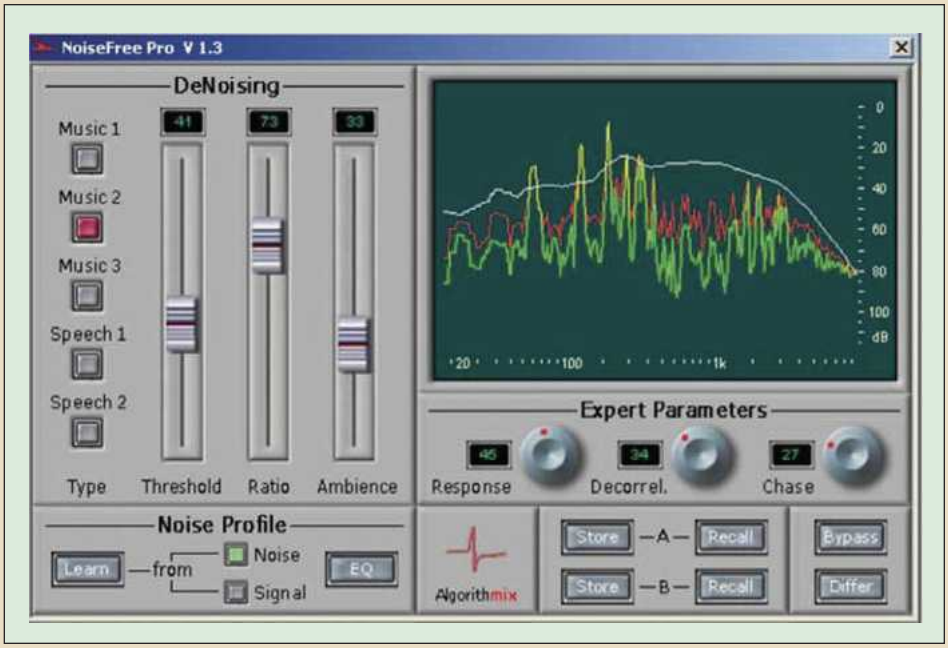

Fig. 5. Screen shot of NoiseFree noise remover tool (Figs. 5 and 6 courtesy Musialik and Hatje)

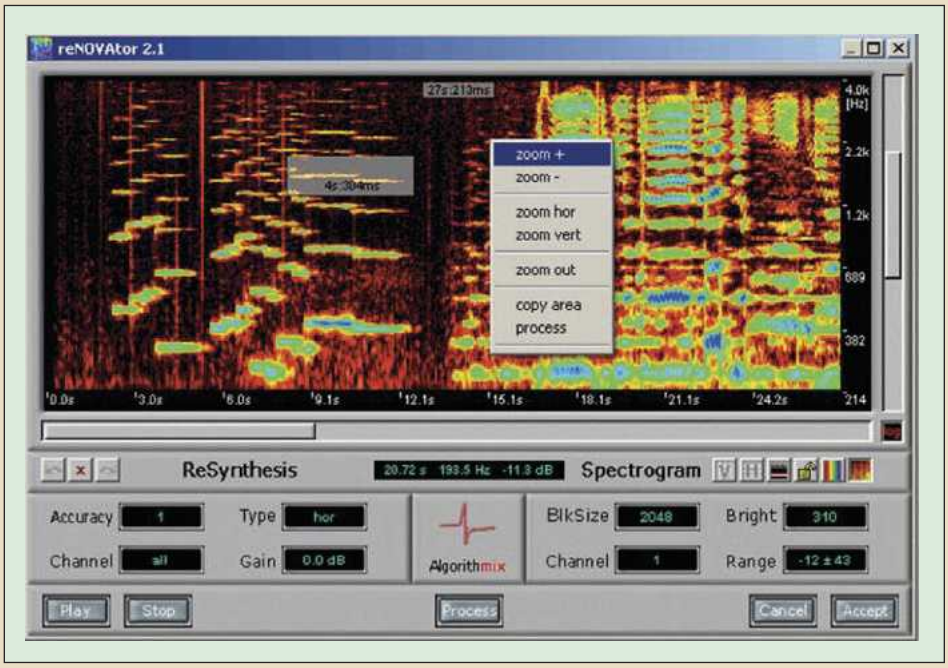

Fig. 6. Screen shot of the reNOVAtor spectrographic editing tool

Related approaches to noise removal were also covered by Musialik and Hatje. In their paper they describe the use of frequency-domain processors for this application. In particular they deal with two products known as NoiseFree and reNOVAtor, which are designed to offer a number of flexible options to the professional sound-restoration engineer or forensic examiner. The user interface of NoiseFree is shown in Fig. 5. The processor enables the single-ended removal of noise from a wanted signal. The noise profile can be derived in one of two ways: either by capturing it from the input signal or by "tailoring" it from flat white noise using a noise-profile EQ interface. As shown in Fig. 6, the user interface of reNOVAtor enables a spectrogram to be zoomed and edited graphically. Specific spectral components or broad band clicks can be highlighted and removed, with the system interpolating appropriate material to fill the hole left behind. There are various advanced features such as gain-selective interpolation that limit signal-repair operations to a certain gain range only, so that as much of the original signal as possible can be left intact. There is also a means of automatically selecting harmonics of the fundamental signal so that complex tones can be effectively removed, as well as tools for automatically selecting clicks.

POSTSCRIPT

With this conference on audio forensics in the digital age, the Audio Engineering Society has raised a number of interesting challenges to audio engineers as they adapt to the use of digital tools and analysis methods for tasks traditionally undertaken in the analog domain. For those wishing to study the topic in more detail, the full set of papers from this conference is available at http://www.aes.org/publications/conf.cfm.