Audio Engineering Society
Convention Paper

Presented at the 133rd Convention
2012 October 26–29 San Francisco, CA, USA

A method for enhancement of background sounds in forensic audio recordings

Robert C. Maher

Electrical & Computer Engineering, Montana State University, Bozeman, MT 59717-3780 USA
rob.maher@montana.edu

ABSTRACT

A method for suppressing speech while retaining background sound is presented in this paper. This procedure is useful for audio forensics investigations in which a strong foreground sound source or conversation obscures subtle background sounds or utterances that may be important to the investigation. The procedure uses a sinusoidal speech model to represent the strong foreground signal and then performs a synchronous subtraction to isolate the background sounds that are not well-modeled as part of the speech signal, thereby enhancing the audibility of the background material.

1. INTRODUCTION

The common situation in forensic audio enhancement is the need to remove background sounds and interference to improve the intelligibility and/or quality of a desired speech signal. Many techniques are available for single-ended noise reduction in these cases. Most techniques can improve the perceived quality of the foreground speech in a surveillance recording, although it is often found that the perception of quality does not indicate an improvement in intelligibility.

However, there are other circumstances in audio forensic analysis in which the investigation focuses on the background sounds and noise, not the foreground speech. In these cases it would be desirable to use an adaptive noise cancelling procedure to retain and enhance the background "error" (or residual) signal, but in general the only signal available is the mono noisy recording itself.

1.1. Adaptive interference cancelling

Traditional adaptive noise and interference cancelling systems [1] require two inputs: a *primary* input signal $d[n]$ (sometimes referred to as the *desired* signal) and a *reference* input signal, $x[n]$, as depicted in Figure 1. The reference signal is passed through a variable digital filter, $w_k[n]$, producing a processed signal, $y[n]$. The sample by sample difference sequence $d[n] - y[n]$ is referred to as the *error* signal or the *discrepancy* signal, $e[n]$. The idea is somehow to adjust the filter $w_k[n]$ so that $y[n]$ is as good a match as possible to $d[n]$, meaning that the power of the discrepancy signal $e[n]$ is

minimized. The mean-square value of $e[n]$ is often chosen to represent the discrepancy signal power [1].

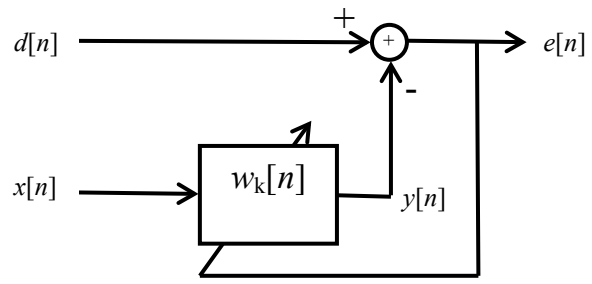

Figure 1: Basic adaptive filter structure.

In a widely used noise cancelling configuration the primary input $d[n]$ is assumed to be the sum of an unknown signal $s[n]$ and an interfering noise signal $n_1[n]$ (that is, $d[n] = s[n] + n_1[n]$), while the reference signal is a noise signal $n_2[n]$ that is *correlated* with the interfering noise signal $n_1[n]$ but *uncorrelated* with the unknown signal $s[n]$. In this configuration, we want the filter $w_k[n]$ adjusted to minimize the mean-square of the discrepancy $e[n]$, which means that $y[n]$ is made to be as close as possible to $n_1[n]$. When the discrepancy is minimized, the correlated noise $n_1[n]$ is minimized, and ideally $e[n] \approx s[n]$.

The application of interest in this paper is the situation in which the unknown signal $s[n]$ is very weak compared to the interfering noise signal $n_1[n]$. A classic example of this situation is detecting the subtle *in utero* fetal heartbeat in the presence of the strong heartbeat of the mother [1].

In the case of audio forensics, the situation of foreground sounds overpowering desired background sounds occurs when a foreground sound or conversation obscures the audibility of low-level sounds that are of potential interest to the investigation.

It is possible that an interference cancellation scheme may be effective for recovering background sounds if two or more simultaneous audio recording channels are available from different spatial vantage points [2, 3]. Unfortunately, many audio forensic cases have only a single monophonic audio channel, and since only a single input signal is available, it is not feasible to use one of the standard adaptive noise cancellation approaches that require both a primary input and a correlated reference input.

To accomplish interference cancellation in the single-channel case, we have investigated several ways to derive and synthesize a pair of signals from the original monophonic file. Our approach proposed in this paper uses a sinusoidal model for the strong foreground sounds, and either a synchronized subtraction process or an adaptive filter process to reduce the strong foreground signal while retaining the low-level background sounds.

Specifically, the method proposed in this paper uses the single noisy recording to generate a sinusoidal model capturing the strong foreground speech and other sounds, thereby creating a pair of signals that can be processed in a manner similar to an adaptive noise canceller: the original noisy recording becomes the primary input, while the sinusoidally modeled portion becomes the reference signal.

1.2. Sinusoidal modeling

The general sinusoidal modeling technique is based on the principles of short-time Fourier analysis and synthesis. The most widely used sinusoidal analysis/synthesis framework was described by McAulay and Quatieri and by Smith and Serra in the mid-1980s [4, 5]. Although originally intended for modeling speech signals, the McAulay and Quatieri (or **MQ**) representation has also been applied to music, seismic signals, bioacoustic sounds, and other audio signal processing applications [6].

The MQ sinusoidal representation captures time-variant spectra and possibly non-harmonic partials. The model assumes that short segments of a signal can be approximated by a sum of sinusoids, where each sinusoid has time-variant amplitude, frequency, and phase. If the sinusoidal model is a good representation of the signal, the Fourier transform magnitude of the signal will consist of "peaks" that are attributed to the presence of underlying sinusoidal components.

The MQ sinusoidal analysis procedure (see Figure 2) begins by divided the input signal into short-time overlapping sections of samples, referred to as **frames**. Each frame is multiplied ("windowed") by a tapered time window function to reduce spectral leakage, then processed with a zero-padded Fast Fourier Transform to obtain a spectrally oversampled discrete Fourier

transform (DFT) of the frame. Next, all "peaks" in the magnitude spectrum are carefully identified, often using interpolation to estimate the precise frequency and magnitude of the peak [5]. The amplitude, frequency, and phase corresponding to the peaks in the magnitude spectrum are then tabulated for subsequent use.

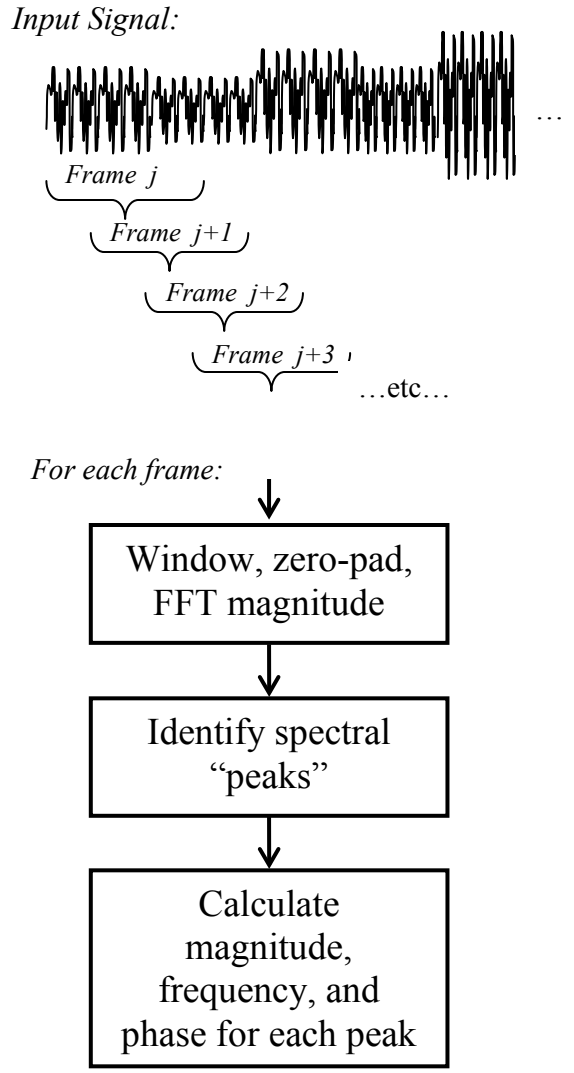

Figure 2: MQ sinusoidal analysis procedure.

The analysis and peak-picking process is repeated for each of the input frames and the spectral peak information (amplitude, frequency, and phase) is matched from frame to frame in order to follow changes in the input signal.

If the sum-of-sinusoids signal model is appropriate for the input signal, the frame-to-frame matching procedure will result in sets of peaks that track the amplitude, frequency, and phase of the underlying sinusoids. These *peak tracks* can appear and disappear as the signal spectrum varies with time.

A modeled version of the input signal is reconstructed by additive synthesis, with each sinusoidal oscillator controlled by the peak track's amplitude, frequency, and phase information interpolated smoothly from frame to frame.

In general, the signal analyzed and reconstructed by the MQ procedure does not represent a mathematically perfect analysis/synthesis system, but the resynthesis results have been found to be sufficiently good for a variety of audio signal processing tasks [6, 7]. Signals that are not well-represented as a sum-of-sinusoids, such as broadband noise, are usually not a good choice for MQ analysis and synthesis.

1.3. Decomposing additively combined signals

If the signal to be analyzed contains more than one source, such as a recording of two simultaneous talkers or a recording of a music ensemble, the sinusoidal model will be based upon the mixture of peaks in the spectral magnitude. In some cases it may be possible to sort out which peaks go with which source, but this is generally a difficult problem because the underlying signal components may interfere with one another, resulting in amplitude beating and other complexity [7].

Similarly, if the signal contains a component that is well-modeled by the sinusoidal procedure (e.g., voiced speech) and a concurrent component that is stochastic and not likely to appear as a peak track in the spectral magnitude (e.g., fricatives or background noise), it may be possible to separate the deterministic portion of the signal using the sinusoidal model from the stochastic portion by subtracting the resynthesized MQ signal from the original input signal [8, 9].

In cases where the additively combined signals comprise a relatively strong *foreground* component that is appropriate for MQ peak tracking and a relatively weak but still interfering *background* component that is well below the level of the foreground signal's spectral peaks, the MQ procedure can provide useful noise reduction if a spectral threshold is used. The

resynthesized signal will be based on the strong spectral components, while portions of the spectrum that are below the threshold will be discarded, similar to the action of a multiband spectral noise gate [6, 10].

1.4. Forensic enhancement for weak background sounds

The particular situation of interest in this paper is one in which a forensic examiner is called upon to enhance a weak background sound component in the presence of a strong and interfering foreground component. This situation arises, for example, when a poorly-placed surveillance microphone picks up a loud conversation or mechanical noise that obscures a relatively quiet background conversation of interest. The situation may also occur in an emergency call center recording in which the utterances or sounds of the dispatch center overlap the desired sounds from the caller's phone.

If it happens that the interfering foreground sound occupies a different frequency range or time interval than the desired background sound, separating out the background material may be accomplished by careful filtering or windowing. However, in general the strong foreground component will overlap and overpower the desired background sound both spectrally and temporally, making it unlikely that any method can recover the weak signal perfectly. Thus, the goal of the proposed sinusoidal analysis plus adaptive filtering procedure is to achieve forensically useful enhancement of the weak background signal.

2. TEST IMPLEMENTATION

The proposed test implementation to enhance a weak background signal in the presence of a strong foreground interfering component begins with the single channel (mono) input signal. This composite signal is modeled using the MQ sinusoidal procedure and a spectral threshold level chosen to be between the foreground and background signals. The foreground signal is resynthesized from the sinusoidal component model, creating the reference signal, $x[n]$, for the adaptive filter. The original mono composite input signal is used as the desired signal, $d[n]$. This basic structure is shown in Figure 3.

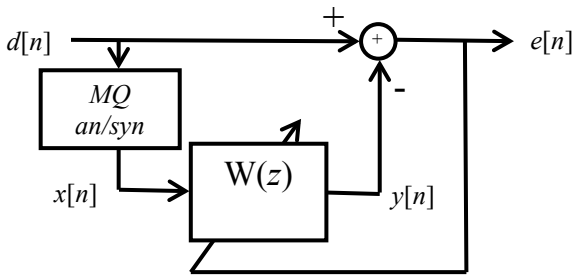

Figure 3: Proposed MQ+adaptive filter background signal enhancement structure.

In the test configuration, W(z) (z-transform of $w_k[n]$) is a 128-tap FIR adaptive filter, and a standard least mean-square (LMS) algorithm is used for the filter updates [11]. The FIR length is chosen to be sufficiently long to allow good adaptation to the gain and phase to match the sinusoidally synthesized $x[n]$ so that $y[n]$ is a solid match to the foreground signal portion of $d[n]$. The adaptive filter length and the adaptive rate-of-convergence parameter are adjusted to suit the particular signal.

2.1. Sine wave test signal

The first test to demonstrate the proposed system is a signal containing a single 1 kHz sinusoid, close to full scale, 16-bit amplitude quantization, with a 16 kHz sample rate. The MQ spectral threshold is set to model the 1 kHz signal while exceeding the window function's spectral sidelobes. If the system is working properly the MQ resynthesis should already be a good match to the original input signal, so the adaptive filter output discrepancy signal $e[n]$ should be essentially zero. The result for this synthetic test is shown in Figure 4, which confirms the baseline expectations.

2.2. High-level + low-level sine test signal

Next, a low-level 500 Hz sinusoid is added to the high-level 1 kHz sinusoid, creating a simple background + foreground combination. In this case we would expect the adaptive filter to cancel out the 1 kHz component, leaving $e[n]$ to contain only the low-level 500 Hz component, as verified in Figure 5.

2.3. Speech + low-level sine test signal

Finally, we verify the performance of the system with a segment of high-level speech to which a low-level 500 Hz background component is added. The MQ analysis threshold is set to pass as much as possible of the high-level speech while staying above the low-level component. The performance of this test is shown in Figure 6.

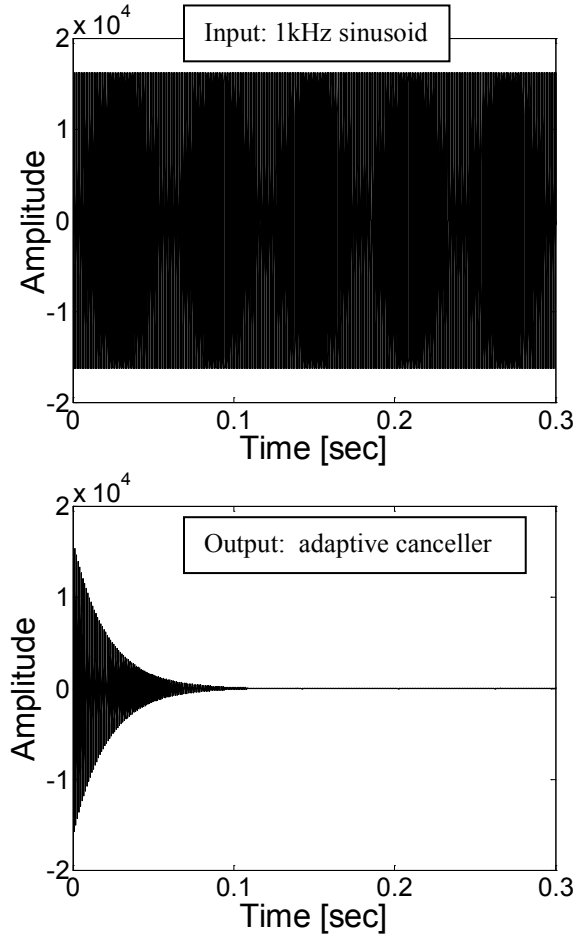

Figure 4: Synthetic example for a single 1 kHz sinusoid.

The low-level 500 Hz component is present in the output signal as desired, but the adaptation to eliminate the high-level speech only appears to be effective for certain portions of the utterance. The results of this demonstration show the interaction between the MQ sinusoidal analysis/synthesis parameters and the adaptation rate of the filter. Specifically, if the frame rate of the MQ analysis is not sufficient to track rapid variations with time in the underlying spectrum, the synthesized signal does not track the time-variant spectral envelope of the signal, and the adaptive filter tends to be influenced by the synthesis-dependent effects rather than the foreground vs. background distinction.

This indicates that a short hop between sinusoidal analysis frames may be desirable for the best quality, even though this increases the computation rate for the MQ process. Since the audio forensic applications are generally off-line anyway, the computation issue is not found to be a problem in practice.

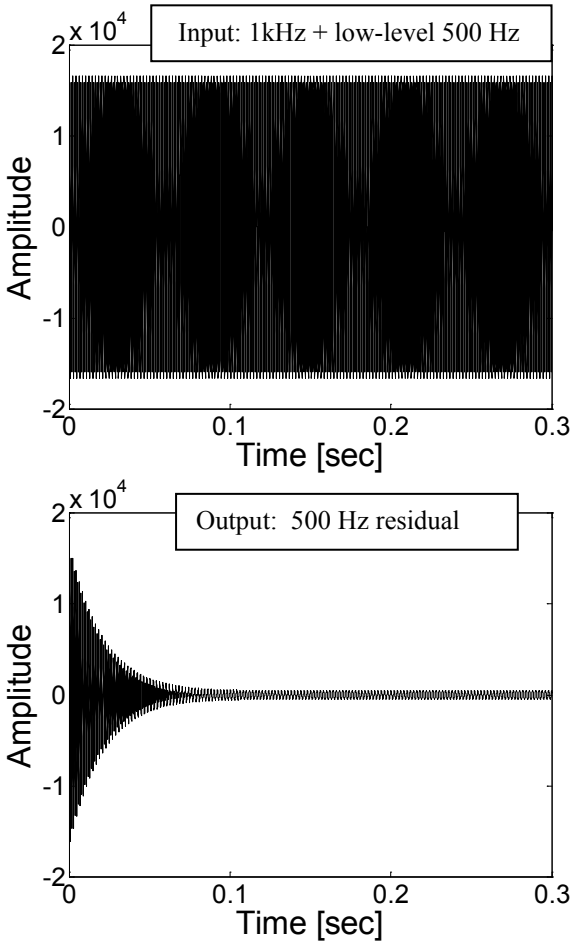

Figure 5: Synthetic example for a high-level 1 kHz sinusoid + low-level 500 Hz sinusoid.

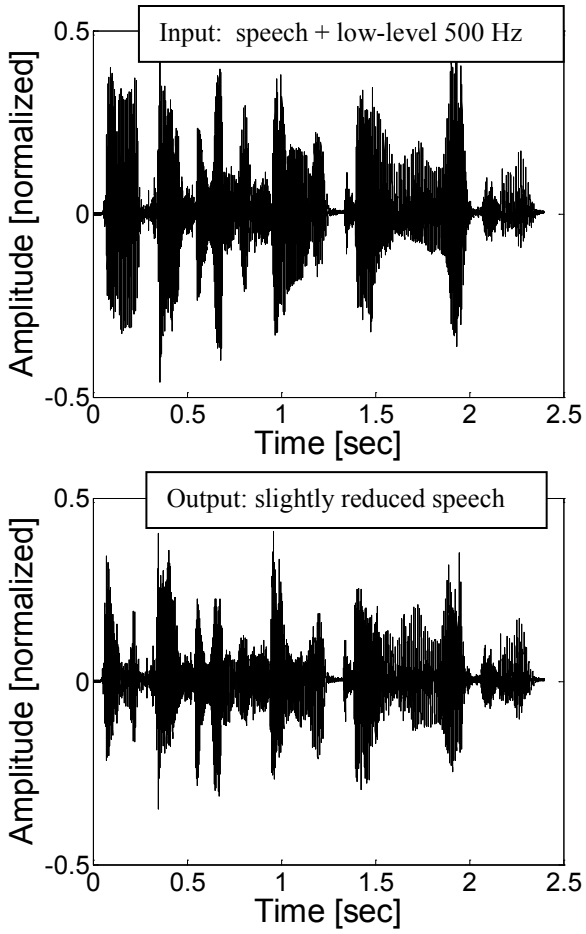

Figure 6: Synthetic example for a low-level 500 Hz sinusoid + high-level speech signal.

3. PERFORMANCE ASSESSMENT

With the basic enhancement system functioning as expected, the next step is to assess the performance with two forensically-relevant test signals.

3.1. High-level DTMF + low-level speech

The first test signal consists of a low-level speech signal that is obscured by a sequence of high-amplitude dual-tone multi-frequency (DTMF) telephone dialing tones. This situation can occur if one of the parties in the phone conversation accidentally or deliberately touches the phone's keypad during the recording. The input signal is shown in Figure 7.

The MQ sinusoidal analysis/synthesis of the input signal with the peak-picking threshold set to be above the low-level speech spectrum is shown in Figure 8. Note that the high-level tones are clearly presented while the low-level speech is missing both in the gaps between tones and during the notes themselves. Also note that the tone onsets and releases are smoothed compared to the original signal, due to the amplitude interpolation used in the MQ analysis procedure.

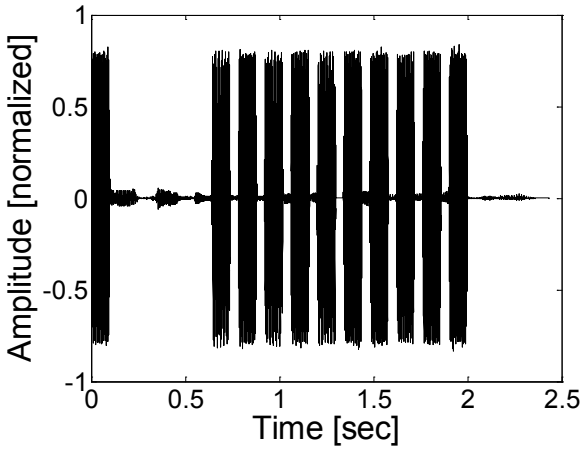

Figure 7: DTMF signals overlapping low-level background speech.

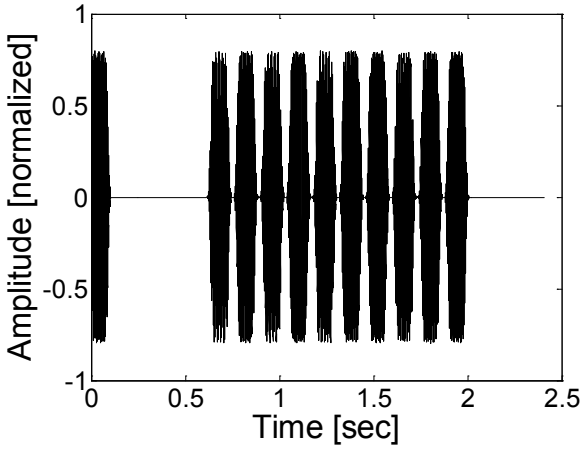

Figure 8: MQ synthesis of the input signal using a threshold higher than the speech signal spectrum.

Applying the adaptive cancellation process shown in Figure 3, with FIR filter length=32, to the signals depicted in Figures 7 and 8, results in the output signal shown in Figure 9.

The output signal shown in Figure 9 has brief clicks that occur at the start and end of each DTMF tone due to the

smoothing of the MQ synthesized signal and the convergence time of the adaptive filter. The onset and release behavior of the MQ analysis could be improved, for example, with a shorter hop size between analysis frames. This is also a topic for further algorithmic refinement. Nevertheless, the intelligibility of the underlying low-level speech signal is significantly improved, as the clicks are relatively easy for a human listener to accommodate.

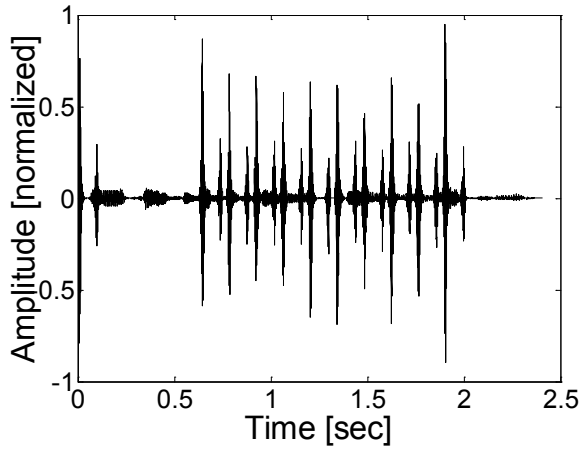

Figure 9: Adaptive cancellation output.

3.2. High-level speech + low-level background

The second forensic test signal is a 911 emergency call center recording consisting of a high-level utterance by the dispatcher (saying "hello?") that overlaps a weak background signal of interest to an investigator. The composite signal is shown in Figure 10.

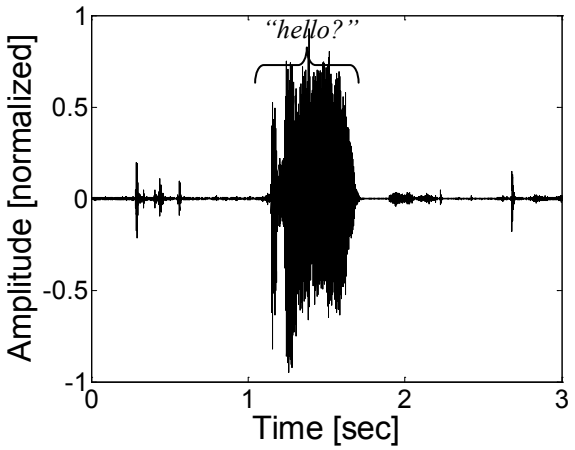

Figure 10: Segment of a 911 emergency call center recording with a high-level speech signal interfering with a low-level background signal of interest.

The MQ analysis/synthesis of the signal with a threshold just above the level of the background sounds is shown in Figure 11, and the foreground sound reduction result is shown in Figure 12.

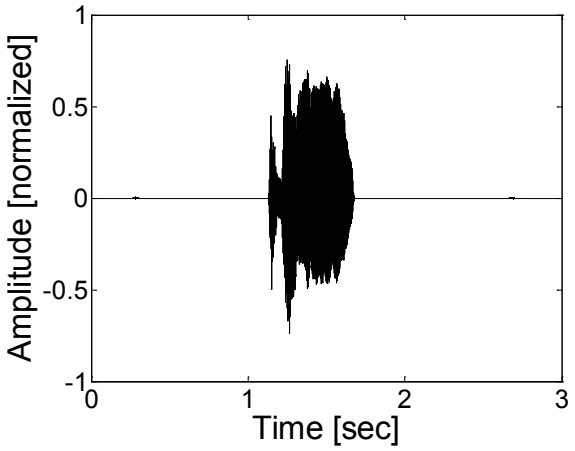

Figure 11: MQ synthesis of the high-level portion of the input signal.

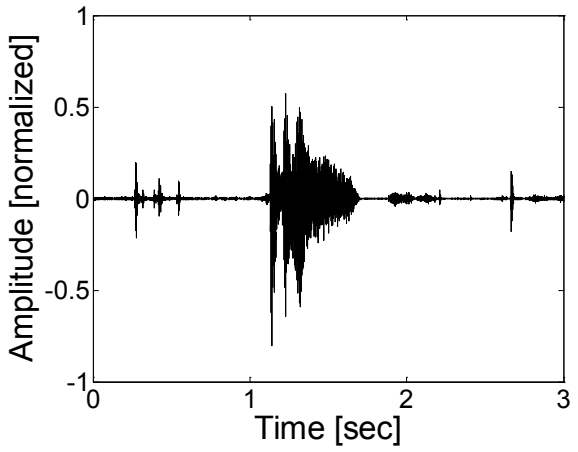

Figure 12: Adaptive cancellation output.

The output signal shown in Figure 12 still contains elements of the high-level "hello?" utterance from the dispatcher, but the aural clarity of the underlying low-level sounds and speech is improved sufficiently that the forensic examiner can better interpret the background sounds.

4. CONCLUSIONS AND FUTURE WORK

The proposed system is demonstrated to be useful in a variety of audio forensic scenarios. Although not a panacea, the MQ + adaptive formulation can be a useful addition to the audio forensic examiner's toolbox.

Typical audio forensic applications allow for iterative enhancement, meaning that the examiner can apply a variety of thresholds, adaptive convergence factors, and analysis parameters, seeking the most appropriate combination for a particular case. Nonetheless, it would be helpful to develop a broader adaptation strategy so that the basic parameters would be selected automatically. Future work will involve techniques for more automated processing and evaluation.

5. REFERENCES

[1] Widrow, B., Glover, J.R., McCool, J.M., Kaunitz, J., Williams, C.S., Hearn, R.H., Zeidler, J.R., Dong, E., and Goodlin, R.C., "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE* , vol.63, no.12, pp. 1692- 1716, Dec. 1975.

[2] Alexander, A, and Forth, O. "'No, thank you, for the music': An application of audio fingerprinting and automatic music signal cancellation for forensic audio enhancement," *International Association of Forensic Phonetics and Acoustics Conference*, Vienna, Austria, July 2011.

[3] Alexander, A., Forth, O., and Tunstall, D., "Music and noise fingerprinting and reference cancellation applied to forensic audio enhancement," *Audio Eng. Soc. 46th Int. Conf.: Audio Forensics*, Denver, CO, pp. 29-35, June 2012.

[4] McAulay, R., and Quatieri, T., "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.34, no.4, pp. 744- 754, Aug. 1986.

[5] Smith, J.O., and Serra, X., "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," *Proc. of the 1987 Int. Computer Music Conf.*, Computer Music Assoc., pp. 290-297, Aug. 1987.

[6] Maher, R.C., "Sinewave additive synthesis revisited," *Proc. 1991 Audio Eng. Soc. Conv.*, New York, NY, Preprint #3128, pp. 1-19, Oct. 1991.

[7] Maher, R.C., "Evaluation of a method for separating digitized duet signals," *J. Audio Eng. Soc.*, vol. 38, no. 12, pp. 956-979, Dec. 1990.

[8] Serra, X., "A system for sound analysis-transformation-synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989.

[9] Serra, X. and Smith, J.O., "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12-24, Winter, 1990.

[10] Moorer, J.A. and Berger, M., "Linear-phase bandsplitting: theory and applications," *J. Audio Eng. Soc.*, vol. 34, no. 3, pp. 143-152, March 1986.

[11] Treichler, J.R., Johnson, C.R., and Larimore, M.G., *Theory and Design of Adaptive Filters*, Prentice Hall, 2001.