

APPLYING SPEECH ENHANCEMENT TO AUDIO SURVEILLANCE

Douglas O'Shaughnessy, Peter Kabal, David Bernardi, Louis Barbeau, Chung-Cheung Chu, Jean-Luc Moncet

INRS-Telecommunications, Université du Québec
3 Place du Commerce, Nuns Island, Québec, Canada H3E 1H6

Abstract: Audio surveillance tapes are prime candidates for speech enhancement, due to the many degradations and sources of interference that mask the speech signals on such tapes. In this paper we describe ways to cancel interference where an available reference signal is not synchronized with the surveillance recording, viz. the reference is obtained later from a phonograph record or air check recording from a broadcast source. As a specific example, we discuss our experiences processing a wiretap recording used in an actual court case. We transformed the reference signal to reflect room and transmission effects, and then subtracted the resulting secondary signal from the primary intercept signal, thus enhancing speech from the desired talkers by removing interfering sounds. Prior to subtraction, the signals had to be aligned properly in time. The intercept signal was subject to time-scale modifications due to variable phonograph and tape recorder speeds. While these speed differences are usually small enough not to affect perceived quality, they adversely affect the ability to cancel interference automatically. Concerning recording devices, we took into account four factors that affect signal quality: frequency response, nonlinear distortion, noise, and speed variations. The two methods that were most successful for enhancement were the LMS adaptive cancellation and spectral subtraction.

Introduction

While speech enhancement has many practical applications, one of particular interest concerns processing audio surveillance tapes, due to the wide variety of degradations that may affect speech signals on such tapes. In many other applications, speech is degraded by a specific, continuous type of noise, facilitating its removal. For the case of designing hearing aids, for example, the interference can come from many sources, but the user can orient the microphone for best results. However, the conditions of surveillance applications are often severe and rarely permit manipulation of microphones during speech recording. Unlike in other noisy communication systems, speakers under surveillance make no effort to enunciate clearly and directly into a microphone; indeed, they may hinder the surveillance by speaking softly and by adding loud sound sources to the environment, such as music. Thus, audio surveillance is one of the more difficult tasks for speech enhancement [1]. In this paper we describe difficulties that arise in audio surveillance, and note ways to increase the intelligibility of speech so obtained. As a specific example, we discuss our experiences with a wiretap recording used in an actual court case.

In electronic surveillance, a microphone is often placed inside or behind an object in a room. The audio signal is typically transmitted by telephone or radio to a tape recorder. The recording conditions are usually far from ideal: 1) people in the room may not talk loudly or clearly and may move about the room, often facing away and being far from the microphone. 2) there may be competing sounds, such as other talkers, music, television, and noise from nearby rooms or the street. 3) the room may add reverberation distortion to the signal. 4) the placement and quality of the microphone can distort the audio signal. 5) the transmission medium may

restrict frequency range, have nonlinear characteristics, and add noise. 6) the tape recorder can also add distortion, especially in terms of variable drive speed and adaptive gain. In the particular case that we examined, the signal was seriously affected by music from a phonograph, by the acoustics of a reverberant room, by the frequency range limitations of the microphone and telephone lines, and by timing distortions in the tape recorder.

Enhancement process

The choice of enhancement methods is based on the nature of the distortions which degrade the desired speech signal and on whether more than one signal is available for processing. In most cases, a single microphone provides one *intercept signal* to be processed with equalization and filtering. By examining a portion of the signal containing only background noise (no speech or music), a spectral model of the linear distortions to which the original sound signal has been subjected can be obtained. If the distortions vary only slowly in time, the intercept signal can then be equalized to reduce the variation of the response at different frequencies, thus reducing some of the resonant qualities of the original recording. In addition, filters can notch out interfering hum components. The distortion model may have to be updated periodically to reflect changes in room acoustics and transmission characteristics. One-signal enhancement techniques are generally limited to noise reduction, and have difficulty eliminating an interfering sound source such as music or an unwanted voice.

To mitigate the subjective effects of reverberation in the intercept signal, one could use a hands-free telephony technique which filters the output signal so as to cause a low-frequency rolloff. This simple technique, however, is not entirely satisfactory; while the subjective effect is to reduce reverberation, speech intelligibility does not seem to improve.

The number of enhancement techniques and likelihood of success increase if two relevant signals are available for processing. Interference subtraction methods are of particular utility in two cases: 1) if a second microphone is available in the room, or 2) if an interfering sound source can be identified and recorded separately, e.g., a radio or television in the room, or some record or tape on a stereo system. A tape of the interfering sound can later be obtained from the radio or television station (U.S. law requires all broadcasts to be recorded), or by purchasing a record or tape at a music store. This provides a version of the interfering sound with no speech from the desired talkers. A second microphone near an interfering sound source may record little of the desired speech, and thus provide a good interference *reference signal*. In both cases, after suitably transforming the reference signal to reflect room and transmission effects, the resulting *secondary signal* can be subtracted from the primary intercept signal, thus enhancing speech from the desired talkers by removing interfering sounds. The subtraction may either occur directly on the signal (in the time domain) or in terms of the signal's spectrum (frequency domain).

Speech enhancement almost invariably uses digital signal processing. Both the intercept and any available reference recording are digitized, in which the signals are lowpass

the environment 5 dB typically, and assuming a 1000 Hz reference frequency, the gain at higher frequencies is reduced by 10 dB. Saturation at higher frequencies is irrelevant to speech perception. Prior to subtracting a digitized secondary signal from a primary intercept recording, the secondary signal has to be modified to align properly with its component in the intercept signal. Such a component of the intercept signal is subject to the frequency response of the recording environment (including room reverberation) and also to time-scale modifications due to variable phonograph and/or tape recorder speeds. While these speed differences and variations are usually small enough not to affect perceived quality, they are large enough to adversely affect the ability to cancel interference using the secondary signal.

The effects of room reverberation and of the recording path can be mostly modelled as a complex linear filter. The aim is to subject the reference signal to the same filtering as the corresponding component of the intercept recording and then to cancel the modified reference signal from the intercept recording. In the case of enhancement via adaptive filtering, since only the reference signal is filtered, speech in the intercept signal is not further distorted. Reverberation often adds a large number of irregularly-spaced echos, which can be difficult to estimate. Because the ensemble of distortions cannot be completely modelled by the adaptive filter, not all of the secondary component can be removed by this process.

Signal Characterization

An intercept signal typically has three components: a conversation between two or more people, interfering sounds (e.g., music), and background noise. Estimates of the noise can be obtained from portions of the recording with no speech or other sounds present. While interfering sounds can be quite diverse, one case of interest is that of a single speaker talking (e.g., TV or radio) or of one or more singers accompanied by music.

When dealing with recording or playback devices, four factors affect signal quality: its frequency response, nonlinear distortion, noise, and time-scale or speed errors. Nonlinear distortion can be caused by amplifiers or loudspeakers, as well as by the magnetic medium of tape recorders. Time-scale errors consist of wow (0.5–2 Hz), flutter (2–20 Hz), and speed-offset errors. There also can be electronic and mechanical noise generated by the amplifiers and transducers. While motors are the source of most speed changes, other sources of time-scale errors are due to tape stretch and the sampling clock at the digitizer. The cumulative time-scale error between the digitized reference signal and the digitized intercept signal is a major problem that must be overcome before cancellation techniques based on adaptive filtering can be applied, since practical considerations require that these signals be as nearly synchronized as possible.

From the point of view of the surveillance microphone, the room acoustics are different for each sound source due to correspondingly varied geometry of the reflecting paths. In particular, the room acoustics change as the talkers move about. For this and other reasons, the amplitude of the surveillance signal is often highly variable in time, which can cause problems in the recording device of either overload saturation (for periods of high gain) or inaudibility (for periods of low gain). In many recording devices, Automatic Gain Control (AGC) is employed to avoid such major distortions. AGC tries to optimize the amplifier gain according to the input signal level. It slowly increases the gain when the input

signal is low, and rapidly decreases the gain when the input is high (fast attack, slow decay). Variations of 1–5 dB are common.

Time alignment

Adaptive filtering techniques used for cancellation require good temporal alignment to be successful. Because of variability in the speeds of various recording and playback devices, alignment at only one temporal location is not adequate. The gradual divergence from synchrony must be compensated for by stretching or 'warping' one of the signals into time alignment with the other. The reference signal $r(n)$ and intercept signal $s(n)$ can be aligned by first locating corresponding and clearly identifiable points present in both signals and by then shifting and stretching one signal so that these points coincide. These events should be selected so that they are very precise in time, particularly in $s(n)$ where events tend to be temporally smeared due to reverberation.

One method of assessing the degree of match is to cross-correlate $s(n)$ and $r(n)$, over a block of samples short enough so that the signals do not drift excessively with respect to one another. Any gradual drift of the correlation peaks is a measure of the speed differences. Given two points of coincidence of the reference and intercept signals, warping to approximately synchronize the signals over the intervening segment is used. To be effective, the drift must be predominantly linear in nature.

Stretching is achieved using interpolation/decimation: 1) the sampling rate of the signal is increased by inserting a fixed number of zero-valued samples after each sample, 2) the signal is then smoothed with a lowpass filter, and 3) the increased-rate signal is then decimated or subsampled. The interpolation/decimation allows for sample-rate conversion in which the ratio of the resultant sampling rate to the original sampling rate is a ratio of integers. An interpolating linear-phase, FIR (finite impulse response) filter is designed to minimize the mean square error in the interpolated signal, given a power spectral model for the input signal.

A simple change of sampling rate is appropriate for short segments of the intercept recording. However, for processing longer segments, anchor points defining points of coincidence must be found periodically throughout the recording, with spacing sufficiently small so that the drift is less than a fraction of the response of the adaptive filter. Between anchor points, the reference signal is linearly stretched or shrunk, allowing piecewise-linear changes in the sampling rate. This alignment method does not compensate for wow and flutter, however.

Noise Reduction and Whitening

To reduce background noise in an intercept recording, enhancement techniques often use a segment of the recording containing only background noise for training. A transformed noise spectrum is subtracted from the spectrum of the intercept signal [2], where the phase of the original spectrum is retained. Spectral subtraction can remove a large part of the noise, but this process often adds brief bursts of tones. With acclimatization, listeners can to some extent block out this type of degradation to concentrate on the remaining signal. However, this tonal noise can sometimes be more distracting than the original broad-spectrum noise.

The alternative process of noise whitening is concerned with the reduction of the perceived effect of the background noise and the equalization of the signal. Assuming that the background noise resulted from a process which generates

a flat (white) noise spectrum, the spectral coloring that is present in the intercept recording shows the effects of the room and recording system frequency response. By inverse-filtering the signal with the measured spectrum of the noise, the noise spectrum can be whitened, which renders the noise less disturbing. At the same time, the overall signal is equalized, with its original signal levels (as a function of frequency) approximately restored.

The first step in this technique is to build a crude estimate of the mean inverse spectrum of the background noise. Then the inverse noise spectrum is modified, e.g., the in-band response can be smoothed over a simple 5-band window. Finally, the passband frequency response can be normalized to bring the geometric mean of the in-band amplitudes to unity, to preserve the in-band gain.

The whitening process itself consists of filtering the intercept signal using the inverse noise spectrum. The overall energy distribution of the noise is much more uniform after whitening. When listening to the resulting signal, one can discern the difference in the quality of the signal. Some of the resonant qualities of the original are missing. There remain two problems, however, with these approaches: 1) the noise spectrum tends to change with time (and is therefore not white), and 2) the room acoustics for the noise differ from the acoustics for the speakers.

Time - Domain Cancellation

Adaptive filtering in the time domain can attenuate a secondary component of the intercept signal $s(n)$, using the filtered and time-aligned reference signal $r(n)$. Time alignment may be relatively coarse, but if speed variations are small and slow enough, the adaptation process will partly compensate for them. Any reverberant effects in the intercept signal are modelled as an FIR filter acting on a clean reference signal. The adaptive filter tries to track the coefficients of this filter and to filter the reference signal to produce a reverberated signal which matches the secondary component of the intercept signal. This can then be subtracted from the intercept signal to reduce the level of the secondary sound.

Effective reverberation may extend over a relatively long time period. The speed of sound corresponds to about 1 ms/foot. The path difference between a direct path and a reflected path can correspond to a large number of ms, and furthermore multiple reflections can extend the reverberant effects to a significant fraction of a second. This means that, at a sampling rate of 10,000 samples/s, effective filter lengths must correspond to hundreds of samples. For a filter with M coefficients, the time averages used to update the filter coefficients must extend over intervals significantly larger than M samples. If we violate this condition, the filter has enough degrees of freedom to synthesize other components of the intercept signal and cancel them also. (In the context of secondary sound cancellation, both the speech components and the background noise are interference.) Also working against the use of long time averages is the loss in synchrony between the reference signal and the secondary component of the intercept signal due to time alignment changes.

Consider two time-domain approaches to adaptive signal cancellation [3]. If we assume that the reference signal is uncorrelated with the speech and noise components of the intercept signal, then when the adaptive filter represents the reverberation effects of the room, cancellation of the secondary component is possible without affecting the speech or noise components. The first method uses a block based

approach to update the filter coefficients. The second approach uses the stochastic gradient (LMS) technique, where the adaptation process is performed on a sample-by-sample basis. Block Least-Squares Methods

In block least-squares methods, one examines successive short time frames over which input data is assumed to be stationary. The covariance method leads to an exact solution for the problem of minimizing the error between the degraded speech $s(n)$ and a transformed version of the reference signal $r(n)$, in the least-squares sense, over a block of finite length N . It gives the set of coefficients $\{h_0, h_1, \dots, h_{M-1}\}$ such that the quantity

$$\epsilon = \sum_{n=1}^N [s(n) - \sum_{i=0}^{M-1} h_i r(n-i)]^2$$

is minimized. Those coefficients are found as the solution of the following set of M linear equations

$$\sum_{i=0}^{M-1} h_i \sum_{n=1}^N r(n-i)r(n-j) = \sum_{n=1}^N s(n)r(n-j) \quad \text{for } j = 0, 1, \dots, M$$

The effect of the filter is to remove correlations between the reference and intercept signals. As long as the speech component of the intercept signal is uncorrelated with the reference signal, the speech will not be affected. In this block-based method, correlations are in effect estimated by using time averages; if the frame length is too short, residual correlation may be present between the speech component and the reference signal. This method, however, has some drawbacks: it is computationally intensive, and limits the maximum number of coefficients that can be used. Large errors may occur when solving the equations with finite-precision arithmetic for a large number of coefficients; to avoid numerical difficulties, a maximum of 30 coefficients is best.

Differences in time scales between the reference and intercept signals affect the process in two ways. First, because the intercept signal drifts slightly within a frame as compared to the reference, the correlation terms are smeared, which decreases the amount of possible cancellation. Thus, the signals tend to match better in the middle of a frame, which may be very annoying perceptually. With a typical linear drift on the order of 0.2%, those effects are still significant even for a short frame size of 400 samples. The second effect of asynchrony is the problem of time alignment exceeding the span of the adaptive filter. With the intercept signal drifting linearly compared to the reference, the offset, in samples, between the two signals increases from frame to frame until it becomes greater than the number of taps. Beyond this point the adaptive filter becomes much less effective.

Least Mean-Squares (LMS) Adaptive Cancellation

With the LMS algorithm, for every pair of input samples (one each from the intercept and reference signals), the gradient-descent technique updates the filter tap coefficients. In comparison, the block-algorithm approach updates the filter frame by frame. Although the filter realized using the gradient descent is not truly optimal in terms of the minimization of output signal energy, the learning process can be carried out smoothly and continuously. Also, the filter in this approach can have a large number of coefficients, whereas a practical implementation of a block-based algorithm limits the size of the filter that can be used.

The time-varying response of the M -coefficient linear filter is

$$y(n) = \sum_{i=0}^{M-1} a_i r(n-i).$$

The intent of the gradient update scheme is to minimize the energy in the difference between the output signal $y(n)$ and the intercept signal $s(n)$. Let the mean-square value of the output be

$$\varepsilon = E\{(y(n) - s(n))^2\}.$$

For the purposes of developing a practical LMS algorithm, the expectation operator in the gradient update scheme is omitted, and the instantaneous value of the squared error is used as an estimate of the mean-square error. The coefficients are updated to move in the negative gradient direction, decreasing the error at each step:

$$a_i' = a_i - \mu_i \frac{\partial \varepsilon}{\partial a_i},$$

where μ_i is the step size used for coefficient a_i . A system based on typical parameters such as $M = 201$ and $\mu = 10^{-10}$ provides noticeable speech enhancement, for our 16-bit, 10 000 sample/s system.

Frequency-Domain Techniques

Frequency-domain techniques operate on the Discrete Fourier Transform of the signals. These methods have one basic advantage over their time-domain counterparts: they can, to some extent, ignore phase. Thus, signal time alignment is less critical. Two frequency-domain techniques were tested: 1) comb filtering of intercept signal $s(n)$ to remove harmonics of an undesired component, and 2) spectral subtraction of the time-aligned and equalized reference signal from the intercept signal.

Spectral Subtraction

This technique attempts to remove the interference component of the intercept signal by spectral subtraction. At the input to this process, the intercept signal $s(n)$ has been pre-filtered, inverse-filtered to whiten the noise, and equalized in order to match the level of the reference signal $r(n)$. Spectral subtraction of the magnitudes frame by frame yields

$$|S'(k)| = \begin{cases} \beta |R(k)| & \text{if } |S(k)| \leq \alpha |R(k)| \\ |S(k)| - \alpha |R(k)| & \text{otherwise} \end{cases}$$

The resulting spectrum, $S'(k)$, takes on the phase of the original signal $S(k)$. Our best results were obtained with the parameters α and β set to 7 and 0.005, respectively. When listening to the resulting signal, one can hear the speech more clearly. However, it is surrounded by bubbling sounds. In the case of interfering music and song, the singer's voice virtually disappears, but tonal noise remains.

Different values for the factor α which determines the fraction of the reference music subtracted can be tried. Recall that the reference music signal has been equalized to match the music component of the intercept signal, which means that if a perfect match were present $\alpha = 1$ would be appropriate. However, experiments indicate that overcompensation is preferable to remove most of the music component. A value of 7 gives good music suppression but adds tonal noise; a lower value results in less music attenuation but also less tonal noise. A value of 0.5 gives reasonable amount of music suppression with little or no tonal noise.

Another experiment changed the value in β in the spectral subtraction algorithm, to attenuate the perceptibility of the tonal noise by increasing the background level. The value $\beta = 0.005$ (a very low value) gave the best results.

An increase of β tends to generate a distinct third signal in the background: a low level version of the reference music signal. This signal coexists with the speech and tonal noise components without significantly masking the tonal noise.

The net result after frequency-domain operations (i.e., noise inverse filtering, music equalization, and spectral subtraction) is definitely enhanced speech, in terms of intelligibility. For large values of α , tonal noises are introduced that the listener must try to ignore; this can be achieved after a few minutes of listening, because the tonal noises are somewhat unstructured. A better compromise is to use a smaller value of α which gives less music suppression, but also less tonal noise.

Conclusion

Adaptive cancellation with anchor points only at the ends of long segments produces an annoying time-varying cancellation, with the interference level changing in bursts. In the middle of the segment, if the time alignment is sufficiently in error, little cancellation is obtained for complex music passages. However, interfering sustained musical notes are cancelled well. With a larger number of anchor points, these problems do not manifest themselves as severely. Ultimately, it is the lack of good time synchrony which limits the suppression possible with the adaptive cancellation technique.

In long segments, changes in the gain produced by the AGC can reduce the effectiveness of spectral subtraction. The adaptive filtering strategy can cope with the gain changes for the most part, although perhaps a resetting of the step sizes is warranted if the gain changes radically. The inappropriateness of a single step size for the longer segments manifested itself as instabilities in the adaptive filtering, which can be avoided by fixing the step sizes at some loss in suppression capability.

The two methods that were successfully applied for enhancement were the LMS adaptive cancellation and the spectral subtraction technique. (The block least-squares and other methods yielded minimal enhancement in our case.) Although the overall intelligibility of the speech after processing is about the same for both successful methods, the results are somewhat complementary. The adaptive filtering approach has the least effect on the speech but does not achieve as high a level of interference suppression. The spectral subtraction method achieves higher levels of suppression with some local loss of speech content (whenever the speech spectrum significantly overlaps the interference spectrum). This means that some portions of the speech are more intelligible in one processed signal than the other. Listening to one and then to the other can enhance overall intelligibility.

References

1. H. Hollien and J. Fitzgerald, "Speech enhancement techniques for crime lab use," in *Proceedings of the IEEE Carnahan Conf. on Crime Countermeasures*, Lexington, Ky, pp. 21-29, 1977.
2. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, ASSP-27, no. 2, pp. 113-120, 1979.
3. B. Widrow & S. Stearns, *Adaptive Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1985.