A Model for Speech Reverberation and Intelligibility Restoring Filters

Owen.P. Kenny
Signals Analysis Discipline
Communications Division
Defence Science and Technology Organization
PO Box 1500, Salisbury, 5108
South Australia
Australia.

Douglas.J. Nelson
Department of Defence
9800 Savage Road
Fort George G. Meade, Maryland 20755
USA.

Abstract

The problem of removing channel effects from speech has generally been attacked by attempting to recover a time-varying filter which inverts the entire channel impulse response. We show that human listeners are insensitive to many channel conditions and that the human ear seems to respond primarily to discontinuities of the channel. As a result of these observations, a partial equalization is proposed in which the channel effects to which the ear is sensitive may be removed, without full inversion of the channel. In addition, it is shown that it is possible to build filters of arbitrary length which do not reduce speech intelligibility and do not produce annoying artifacts.

1.0 Introduction

Reverberation can have a dramatic effect on speech intelligibility. Even a small amount of additive noise combined with a moderately reverberant channel can completely destroy speech intelligibility. Speech reverberation is generally modeled according to the three-dimensional billiard table model. In this model, the received signal is the sum of signals which propagate along many different paths, each of which satisfy Snell’s law of reflection. In this model, each reflection is a point reflection in which the angles of incidence and reflection are the same. To further simplify the model, the absorption loss is considered to be constant across the entire spectrum. Finally, the source and receiver are modeled as points. The system is excited by the speech signal, and the channel is determined by the union of the reflections which are intercepted by the receiver. Since the signal loses power at each reflection, only the first few reflections need be modeled. Since the room may be considered to have finite dimensions, the channel response is well approximated as a finite, but long finite impulse response (FIR) filter. The impulse response can easily be a half a second or more.

There are other possible models for the acoustical reverberations. Instead of a point Snell’s law reflection, reflections may be scattered off of a surface which distributes the energy continuously in time. Such a model can arise if the reflecting material is rough, causing reflections in a continuum of directions. In addition, the situation could arise in which the signal propagates along the surface, re-radiating its energy over a distributed area of the surface. The effects of these models are very different, and the observed channels they produce may be mitigated by very different methods.

Regardless of the reverberation model used in channel equalization attempts, the universal approach has been to attempt to recover a time-varying adaptive filter which inverts the channel to reproduce the original speech signal. This approach is inspired by the blind equalization problem for communication signals.

In the modem problem, the channel impulse response is generally very short, having a typical effective duration on the order of 1/100 second. In this problem, the signal is assumed to be in one of several states, for a baud duration, which may be approximately 1/300 seconds. If the signal can be equalized to produce a rectangular pulse, the resulting equalized signal is insensitive to sampling phase, and the demodulated data has minimal dispersion about the signal constellation points.

For speech, there is no requirement that the reconstructed signal faithfully reproduce the original signal. It is only important that the reproduced signal sound like the original signal to a human listener. Since humans have a remarkable ability to ignore many differences in signal conditions, equalization for human listeners is a much weaker condition than full channel equalization in the modem sense. The primary thrust of this paper is to describe the effects of several types of channels on human perception and to re-pose the de-reverberation problem so that its solution is a less daunting task.

2.0 Simple channel simulation experiments

In modeling the reverberation problem, a series of simple channels were simulated and clean speech was convolved with the channel impulse responses to produce corrupted signals

$$\tilde{s}_{CHAN}(t) = s(t)*CHAN(t) \quad , \tag{1}$$

where * is the convolution operator

$$X(t)*Y(t) = \int X(\tau)Y(t-\tau)d\tau \quad . \tag{2}$$

The resulting audio signals were played for human listeners, who evaluated the results. Because of the simplicity of the channels, the results were dramatic, resulting in no ambiguity in interpreting the results.

2.1 Differentiation

For the first simulation, the speech signal was subjected to a simple differentiator

$$\tilde{X} = x(t) - x(t+\tau) \quad , \tag{3}$$

where x is a clean speech signal, and τ is the signal delay. For small values of τ, the frequency response of this filter is essentially

$$F(\omega) \approx i\omega \quad . \tag{4}$$

It was easily verified, by playing the resulting signal that the ear may be able to detect slight changes in the spectral shaping, but there is no effect on the intelligibility and no noticeable echo in the resulting signal for values of τ less than 1/50 sec. In addition, the ear could not distinguish the sign of the delayed signal, so the signal

$$\tilde{X} = x(t) + x(t+\tau) \tag{5}$$

is perceived by the listener to be indistinguishable from the signal in equation 3.

For small values of τ, the ear can not perceive the delayed signal as an echo, but for larger values of τ, the delayed signal is perceived to be an echo of the un-delayed signal. The the critical delay is approximately 1/10 sec.

2.2 Hilbert transform and signal projections

The Hilbert transform of a real signal $x(t)$ is the unique transform which produces a signal $y(t)$, which is $\pi/2$ radians out of phase with the original signal. The analytic signal

$$X(t) = x(t) + iy(t) \quad , \quad i = \sqrt{-1} \quad , \tag{6}$$

has properties which are very important in signal processing. The analytic signal $X(t)$ may be represented by the Euler identity as

$$X(t) = A(t)e^{i\omega t} \quad , \quad A(t) \geq 0 \quad . \tag{7}$$

The signal $X(t)$ may then be projected onto any axis by

$$X_{\varphi}(t) = \text{real}\left\{ X(t)e^{-i\varphi} \right\} \quad . \tag{8}$$

It was verified experimentally that the projected signal $X_{\varphi}(t)$ is indistinguishable by a human listener from the original signal $x(t)$ and that this result is independent of φ.

2.3 Rectangular pulse

The signal was subjected to a channel consisting of a single rectangular pulse of duration T seconds.

$$CHAN1_T(t) = \begin{cases} 1 & t \in [0, T) \\ 0 & t \notin [0, T) \end{cases} \quad , \tag{9}$$

where T was chosen initially to be 0.75 second.

The expectation of the authors was that the rectangular impulse response would produce a low-pass filtered (smoothed) version of the signal and that the channel would be reverberant due to the long response time of the filter. This intuition was completely false. The signal $\tilde{s}_{CHAN1}(t)$ was perceived by all listeners as clean un-delayed copy of the speech signal superimposed on a clean echo of the signal with the same power as the un-delayed signal, but having a delay of 0.75 second. There was no apparent reverberation, and there was no perceptible low-pass filter effect.

The signal was then convolved with two unit impulses separated in time by $T = 0.75$ seconds

$$CHAN2_T(t) = \begin{cases} 1 & t = 0, T \\ 0 & \text{otherwise} \end{cases} \quad . \tag{10}$$

The signal filtered with this channel response was not distinguishable from the signal convolved with $CHAN1$.

From the perceived results of convolving speech with $CHAN1$ and $CHAN2$, it is obvious that, in this simple case, the ear appeared to respond to the endpoints of this particular channel, which are discontinuities in the channel response. This observation is predictable if we note that, for sampled signals,

$$CHAN1_N(n) = \begin{cases} 1 & n = 0, 1 \ldots N-1 \\ 0 & n < 0, n > N-1 \end{cases} \quad . \tag{11}$$

The convolution of the channel with the signal can then be obtained by first computing the first N samples of the convolution

$$\tilde{X}(0) = \sum_{k=0}^{N-1} x(n) CHAN1_{Nn}(N-k) \quad . \tag{12}$$

The full convolution can then be computed as a recursion

$$\tilde{X}(n+1) = \tilde{X}(n) + x(N+n) - x(n) \quad . \tag{13}$$

This expression can be differentiated to produce

$$\tilde{X}(n+1) - \tilde{X}(n) = x(N+n) - x(n) \quad , \tag{14}$$

Since differentiation with small delay produces a signal which is perceptually identical with the undifferentiated signal, the differentiated signal represented by formula 14 is perceptually indistinguishable from the filtered signal $\tilde{X}(n)$, but the RHS of equation 14 is the original unfiltered signal differentiated with a delay of N, For large N, we know that the RHS of equation 14 represents a signal with a simple echo at delay N. so the predicted perceptual effect of convolution of speech with the rectangular pulse

3.0 Slowly varying channels with endpoint discontinuities and long time constants

The above argument is a bit lengthy, but important since it provides insight into the ear's perception of speech in reverberant channels. It should be expected that the ear does not respond to slowly time-varying changes in the channel. Discontinuities in the channel impulse response are perceptually the same as impulses in the channel response, and are perceived as echoes. The magnitude of the perceived echo is the magnitude of the discontinuity of the channel impulse response.

To test this assertions, speech signals were subjected to a variety of channels, each of which was selected to have a response time $T = 0.75 \sec onds$.

3.1 Cosinusoidal, single complete cycle

$$CHAN(t) = \begin{cases} \cos\left(2\pi \frac{t}{T}\right) & t \in [0, T) \\ 0 & t \notin [0, T) \end{cases} \tag{15}$$

Result:

Signal and clean echo at delay T, with no apparent distortion and no reverberation.

3.2 Sinusoidal, single complete cycle

$$CHAN(t) = \begin{cases} \sin\left(2\pi \frac{t}{T}\right) & t \in [0, T) \\ 0 & t \notin [0, T) \end{cases} \quad . \tag{16}$$

Result:

Signal nearly undetectable due to extremely low perceived power. No echoes and no apparent distortions and no reverberations.

3.3 Ramp

$$CHAN(t) = \begin{cases} \dfrac{t}{T} & t \in [0, T) \\ 0 & t \notin [0, T) \end{cases} \quad . \tag{17}$$

Result:

One observed signal, with no echo and no apparent distortion and no reverberation.

3.4 Sinusoidal, partial cycle

$$CHAN(t) = \begin{cases} \sin\left(2\pi\dfrac{t}{T-\varepsilon}\right) & t \in [0, T) \\ 0 & t \notin [0, T) \end{cases} \tag{18}$$

for $|\varepsilon| \ll T$.
Result:

One observed signal, with amplitude equal to magnitude of sinusoidal discontinuity at $t = T$. No echo, no apparent distortion and no reverberation.

In addition to these simple channels, the signal was subjected to several more complicated channels, which were slowly varying and continuous, except for discontinuities at the endpoints, $t = 0$ and $t = T$. In each of these cases, the ear responded only to the discontinuities at the endpoints, perceiving echoes at the discontinuities. The magnitude of the discontinuities were proportional to the magnitude of the perceived echoes, as was determined by comparing the signal convolved with the synthesized channel with the echo signal

$$\alpha_0 x(t) + \alpha_T x(t+T) \quad , \tag{19}$$

where α_0 and α_T are the magnitudes of the dicontinuities at the endpoints of the synthesized channel response. In each case, the signals convolved with the synthesized channels were perceptually indistinguishable from the corresponding echo signal. The conclusion is that the does not respond to slow continuous changes in the channel impulse response. The ear does perceive discontinuities in the channel as echoes.

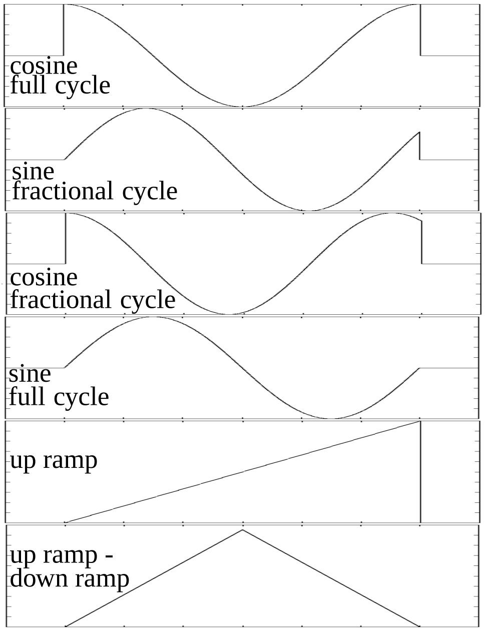
3.5 Up ramp, down ramp

$$CHAN(t) = \begin{cases} \dfrac{2t}{T} & t \in \left[0, \dfrac{T}{2}\right) \\ 1 - \dfrac{2t-T}{T} & t \in \left[\dfrac{T}{2}, T\right) \\ 0 & t \notin [0, T) \end{cases} \quad . \tag{20}$$

Result:

Signal nearly undetectable due to extremely low perceived power. No echoes and no apparent distortions and no reverberations.

There is essentially no signal observable to a human listener.



4.0 Partial equalization of an impulsive channel

It is clear from the above examples that the discontinuities on the channel result in echoes and that the ear tends to ignore the effects of the channel which are slowly varying in time. This assertion can be predicted since discontinuities in the channel impulse response result in a broad band spectral response, The slowly varying portions of the impulse response act as a low-pass filter. For FIR bandpass filters, the bandwidth is approximately related to the filter length as

$$BW \propto \frac{1}{L} \ . \tag{21}$$

For each of the channels the slowly varying, continuous portion of the channel response results in an FIR filter, whose pass band is near DC and only a few Hertz wide. This energy is generally out of the normal response of the ear. Even if the passband of the slowly varying portion of the channel is as wide as (***) Hz, it has been verified that the ear very effectively ignores the resulting spectral energy, resulting in little or no effect on the intelligibility of the received signal. The dicontinuities, however can not be removed by the ear, and are perceived as echoes.

The above discussion suggests that in some cases, the channel effects can be mitigated by removing the channel discontinuities, while ignoring any slowly varying portions of the channel. To test this hypothesis, a channel consisting of a sequence of exponentially decaying uniformly spaced impulses was synthesized

$$CHAN(t) = \begin{cases} e^{-\alpha t} & , \quad t \in 0, T, 2T, \ldots \\ 0 & \quad \text{otherwise} \end{cases} \quad . \tag{22}$$

The signal convolved with the channel sounded very reverberant, with so many echoes that the signal was completely unintelligible. The filter

$$FILT(t) = \begin{cases} e^{-\alpha t} , & t \in [0, T] \\ 0 & \text{otherwise} \end{cases} \qquad (23)$$

was convolved with the signal to which the channel filter (22) had been applied, with the result that the signal produced sounded like clean speech, with no noticeable channel artifacts. Clearly the filter (23) is not the inverse filter of the channel filter (22), so the processed speech signal had only been partially equalized to remove discontinuities. The implication is that complete channel equalization of speech is not necessary to restore intelligibility.

A corollary of this discussion is that filter length need not affect speech intelligibility. To test this, a prolate-spheroidal filter basis of narrow bandpass filters was constructed. each filter in the basis was constructed to be zero phase and have length approximately a quarter second. Clean speech signals were selected from the TIMIT database and these signals were filtered by a variety of narrowband notch filters, bandpass filters and high and lowpass filters, which were constructed as linear combinations of the prolate spheroidal basis filters. Care was taken to insure that none of the filters removed more than 10 percent of the spectral bandwidth. In each case, there was no noticeable loss of intelligibility. The loss of spectral energy was noticeable, but the signals did not sound muffled or distorted.

5.0 Conclusionsand Future Research

It has been demonstrated that the human ear appears to respond to the discontinuities of the channel impulse response to a much greater degree than it does to smooth.y time-varying effects. In addition, it is possible to mitigate channel effects, and in some cases restore intelligibility of speech subjected to channels with isolated discontinutites without complete equalization of the signal.

As a result of the experiments documented in this paper, the authors have conducted experiments to attempt to mitigate reverberation resulting from an unknown channel. These results appear promising, and will be the focus of future research.

References

[1] A. Akbari, J. LeBouquin,G. Faucon,”Optimizing speech enhancement by exploiting masking properties of the human ear”,ICASSP-95,Vol.1,pp 800-803,May 1995.

[2] D. J. Darlington and D. R. Campbell, “The Effect of Modified Filter Distribution on an Adaptive Sub-band Speech Enhancement Method”, Proc. of IEEE Dig. Proc. Workshop, pp 153-156, Sept. 1996.

[3] A. H. Hermansky and N. Morgan, “RASTA Processing of Speech”, IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp 578-589, October 1994.

[4] C.D. Yoo, ”Selective all-pole modeling of degraded speech using M-band decomposition”,ICASSP-96, Vol.2,pp 641-644,May 1996.

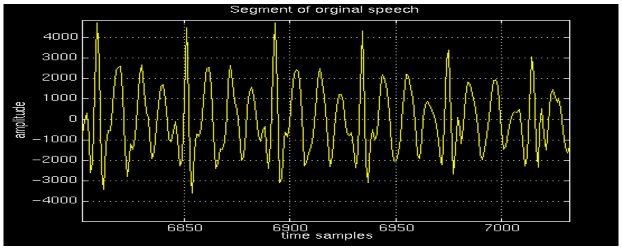


Figure 1 Clean TIMIT data

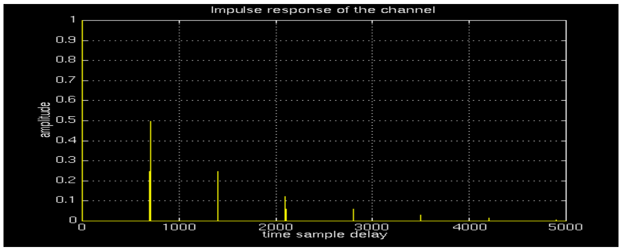
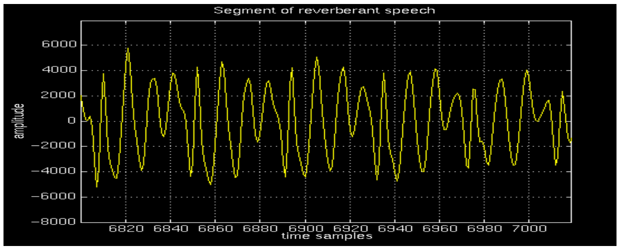


Figure 2 Cnannel impulse response



Figue 3 Clean signal convolved with channel impulse response

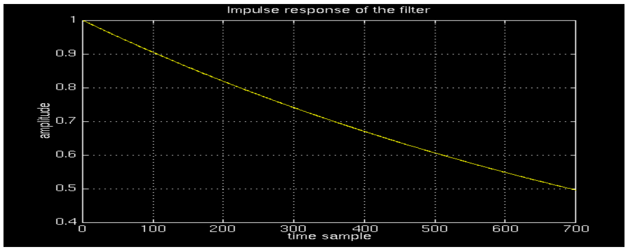


Figure 4 Partial equalization filter

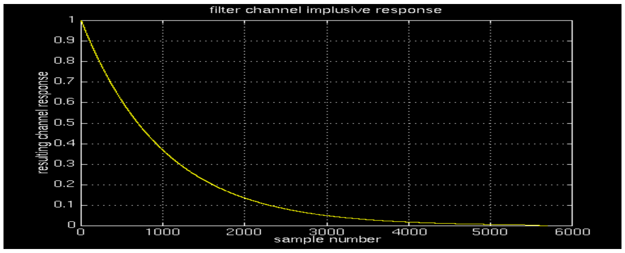


Figure 5 Partially equalized channel response

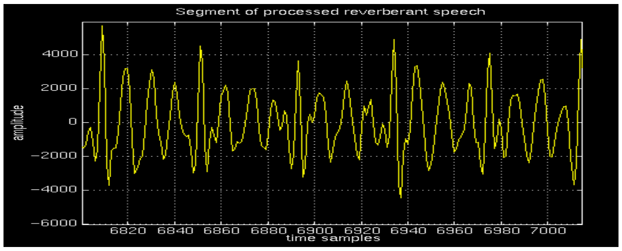


Figure 6 reconstructed partially “equalized” signal