

Improving Performance of Voice Echo Cancellers.

There is nothing more practical than a good theory.

James C. Maxwell

Introduction.

The Voice Over Packet (IP/ATM/FR/Ethernet/DSL/Cable/etc) technology is undergoing fast changes in recent years. There is less need in voice compression due to the higher bandwidth on local loops / last mile and the telecom infrastructure backbone. But any VoX system introduces some delay, and therefore such a system shall follow the PSTN rule: if extra delay exceeds 5 ms, an echo canceller shall be employed. The delay is usually larger than that, and an EC is a must. It may sound paradoxically, but the quality of EC starts to determine the overall Voice Over Packet system quality and becomes the bottleneck. Many EC in field provide poor quality of voice, and customers complain on echo, unnatural sound, voice clippings, etc, and report that the quality is very far from a local TDM call quality, despite the fact that the same EC has successfully passed lab tests.

Let's have a look on the ITU-T G.168 standard, defining EC test procedures and criteria. On one hand, it requires (test 2B) that an echo canceller should achieve at least 30 dB of echo cancellation in a second on some artificial signal (CSS). On another hand, test 3A assumes that a far-end signal, which is only 15dB lower than near-end signal, should be considered as a low-level double talk and it might be clipped off. As we know that a median echo return loss is about 11 dB, we should raise another question: "How different are a G.168 complying echo canceller and a simple traditional G.164 echo suppressor?" The difference in double talk performance is only 4 dB, while the echo itself can be cancelled by 26 dB more. This is indeed very strange and something here is indeed quite wrong. What is the reason of such striking discrepancy?

The theoretical base of underlying adaptive filtration, usually presented in papers on voice echo canceller, consists of only 2 quite simple equations in matrix form and therefore raises legitimate questions "What's the problem with voice echo cancellers? Why can't they be perfect?"

First, the simplicity of the equations for adaptive filtering is deceitful, and this is not only due to their computational complexity and numerical problems with fixed-point implementations. Second, that set of 2 equations is not the entire set of equations. Third and the most important, the entire paradigm of adaptive control profoundly contradicts to the "normal" model of human behavior.

Humans prefer to learn first (and rather to be taught by someone else than to dive into research themselves), and act only afterwards, slowly improving their skills in safe conditions. Such normal human model would be fully applicable to the voice-band modem echo canceller. It is first trained in known conditions without disturbances, under good guidance of training signal that covers entire spectrum. Then it is explicitly switched to cancellation mode (or slow adaptation), and works in the same conditions with the signal that is very close to the training sequence. A basic theory would suffice in this case.

Voice echo canceller should work in an absolutely different set of conditions. The guidance given by human voice signal is poor due to its non-stationary nature and the spectrum rich in singularities and blind spots. The disturbance signal is also human voice, also non-stationary, also unpredictable, also without any known statistics that may serve as a good base for tuning of adaptive filter. The echo path may suddenly change, and voice echo canceller should detect this change and re-adapt quite quickly.

If we draw an analogy with car driving further, the modem echo canceller would correspond to the normal human way of learning: you should first take enough lessons with a good instructor, acquire required skills from him (rather than invent them by yourself), and practice them in the typical situations that you would face later alone under instructor's clear guidance and using his help (as duplicate pedals) to lower the cost of your mistakes.

The voice echo canceller would correspond to the situation when you are thrown in an extraterrestrial car with lots of unfamiliar controls and buttons of unknown purpose, and you have to drive fast and safely from the very beginning, the guidance is fragmentary and unclear, obstacles are various and numerous, the visibility is poor, the conditions may suddenly change from a high-speed freeway to a busy city street, and the car reaction on your actions may alter from time to time, as if the you suddenly turn from a good paved

street into an icy skating ring and back. Nevertheless, you have to drive on time, avoiding both hitting pedestrians (e.g. clipping double talk out) as well as damaging the car (e.g. passing echo through).

Most of people would justly avoid getting into such adventures. Those who take them should be very clear about what challenges they are about to meet. You should abandon your normal human learning models and start thinking differently. Of course it is not easy to break out of sacred time honored behavioral patterns. If you do not change the mindset, your extraterrestrial car driving would be dangerous and self-devastating and the echo canceller you would develop will perform very poorly on human voice despite its formal compliance with G.168 standard. If you do, soon you will find out that the key to success in driving extraterrestrial cars is not as much the speed of learning (adaptation) how to manipulate the available controls, as gaining and maintaining very clear detailed understanding how well you are driving and how well the car is handling the road, both in terms of what you know and can do, and what you don't. It is not so hard to learn how to drive a car, but it is known that young or drunk drivers' overconfidence is the major reason of accidents. Thus the overestimation or underestimation of your abilities can be troublesome.

This kind of understanding in adaptive filtering is represented by a co-variation matrix of the errors of adaptation, which knowledge supports the ability to assess the variance of disturbing far-end signal, compare it with the estimation of residual echo, and process it accordingly, thus maximizing the echo canceller sensitivity and widening the double talk range as much as possible. Unfortunately, the task of estimation the adaptation errors' co-variation matrix is much harder [$O(N^2)$ - without simplification] than the adaptive filtering [$O(N)$] itself, but the benefits of using the high quality echo canceller usually greatly outweigh the increase in demands for DSP resource as MIPS and memory, which, fortunately, become cheaper every year.

The rest of the paper is constructed as series of examples showing how often inappropriate are our assumptions and how useful it is to know the depths of adaptive theory. The exact math theory is avoided wherever possible because otherwise only few specialists would understand the discussion. Instead, the graphical illustrations are given throughout the paper because one good picture can replace a thousand words and may give better understanding than any formula.

Basic Equations of Adaptive Filtration

$$(1) \quad e_k = y_k - \mathbf{x}_k^T \underline{\mathbf{h}}_k;$$

$$(2) \quad \underline{\mathbf{h}}_{k+1} = \underline{\mathbf{h}}_k + \mathbf{a}_k \mathbf{x}_k e_k / \mathbf{x}_k^T \mathbf{x}_k;$$

where:

- e_k - error of echo cancellation at time k ;
- y_k - observation at time k ;
- \mathbf{x}_k^T - input vector of near-end voice signal, size N , $\{x_k, x_{k-1}, \dots, x_{k-N+1}\}$;
- $\underline{\mathbf{h}}_k$ - vector of estimation of true echo path response \mathbf{h} , size N ;
- \mathbf{a}_k - scalar step size;

The 2 equations above had been (as far as my knowledge stretches) first proposed by a Polish scientist Kaczmarz¹, and since were re-invented multiple times (and probably will be re-discovered more times in future). They are known under different names in different scientific fields. In the telecommunication theory they are known under the name of Normalized Least Mean Square (NLMS), given by Widrow in 60s.

Let's now obtain the co-variation matrix as math expectation $\mathbf{D}_k = E\{\mathbf{d}_k * \mathbf{d}_k^T\}$, where $\mathbf{d}_k = \underline{\mathbf{h}}_k - \mathbf{h}$ is the vector of errors of echo path response \mathbf{h} estimation. Why this matrix is so important? Because the math expectation of residual echo is simply a quadratic form: $\mathbf{x}_k * \mathbf{D}_k * \mathbf{x}_k^T$. So if we have good estimation of \mathbf{D}_k ,

¹ Kaczmarz, S. "Angenäherte Auflösung von Systemen linearer Gleichungen", Bulletin International de l'Academie Polonaise des Sciences, Lett A: 355-357, Cracouie, 1937.

we know almost everything. The corresponding equation for co-variation matrix of estimation errors can be easily obtained if we assume that (understanding that those assumptions are very far from reality):

- the echo path is absolutely linear;
- the echo path is fully described by moving-average model of dimension N ;
- the only source of distortions on output is stationary additive noise h_k with standard deviation of s_k , for sake of simplicity assumed zero-mean, truly white and uncorrelated with input sequence $\{x_k\}$;
- there are no distortions on input (or along the echo path) of whatever kind, i.e. we know exactly what signal enters the echo path.

Then:

$$(3) \quad y_k = \mathbf{x}_k^T \mathbf{h} + h_k;$$

$$(4) \quad \mathbf{d}_k = \underline{\mathbf{h}}_k - \mathbf{h};$$

$$(5) \quad \mathbf{d}_{k+1} = \mathbf{d}_k - \mathbf{a}_k \mathbf{x}_k \mathbf{x}_k^T \mathbf{d}_k / \mathbf{x}_k^T \mathbf{x}_k + \mathbf{a}_k \mathbf{x}_k h_k / \mathbf{x}_k^T \mathbf{x}_k;$$

or

$$(6) \quad \mathbf{d}_{k+1} = (\mathbf{I} - \mathbf{a}_k \mathbf{x}_k \mathbf{x}_k^T / \mathbf{x}_k^T \mathbf{x}_k) \mathbf{d}_k + \mathbf{a}_k \mathbf{x}_k h_k / \mathbf{x}_k^T \mathbf{x}_k;$$

from here we can obtain co-variation matrix as math expectation $\mathbf{D}_k = E\{ \mathbf{d}_k * \mathbf{d}_k^T \}$;

$$(7) \quad \mathbf{D}_{k+1} = (\mathbf{I} - \mathbf{a}_k \mathbf{x}_k \mathbf{x}_k^T / \mathbf{x}_k^T \mathbf{x}_k) \mathbf{D}_k (\mathbf{I} - \mathbf{a}_k \mathbf{x}_k \mathbf{x}_k^T / \mathbf{x}_k^T \mathbf{x}_k) + s_k^2 \mathbf{a}_k^2 \mathbf{x}_k \mathbf{x}_k^T / (\mathbf{x}_k^T \mathbf{x}_k)^2;$$

Let's note first that however paradoxically it sounds, in the first approximation the level of residual echo does not depend on the actual shape of echo path response or on the Echo Return Loss (ERL) (there are some limitations to be discussed later). Whether the echo itself is -6dB or -30dB , the residual echo would be approximately the same as far as we do not hit the limits dictated by echo path non-linearity and the background noise.

A Test Signal

So, let's take a life example (real recording) of a female talker, and assume that the echo signal is corrupted by the noise of known and constant level -46dBm0 (what is quite high), and see what happens to the estimation errors' co-variation matrix.

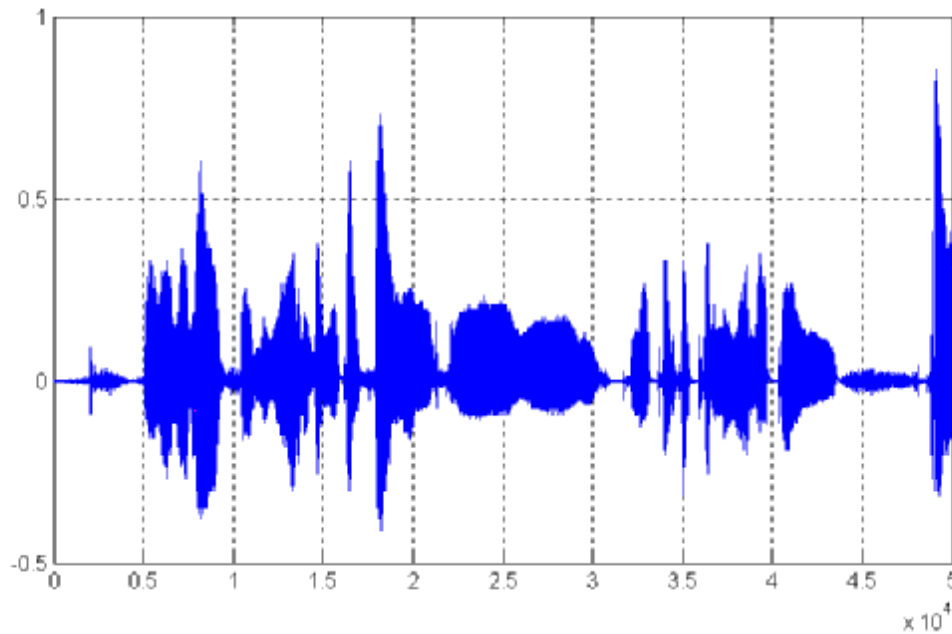


Figure 1. Recording of a female talker speech, to be used as Rcv signal for the further discussion. The amplitude is normalized; the time is in samples with sampling rate of 8000 Hz.

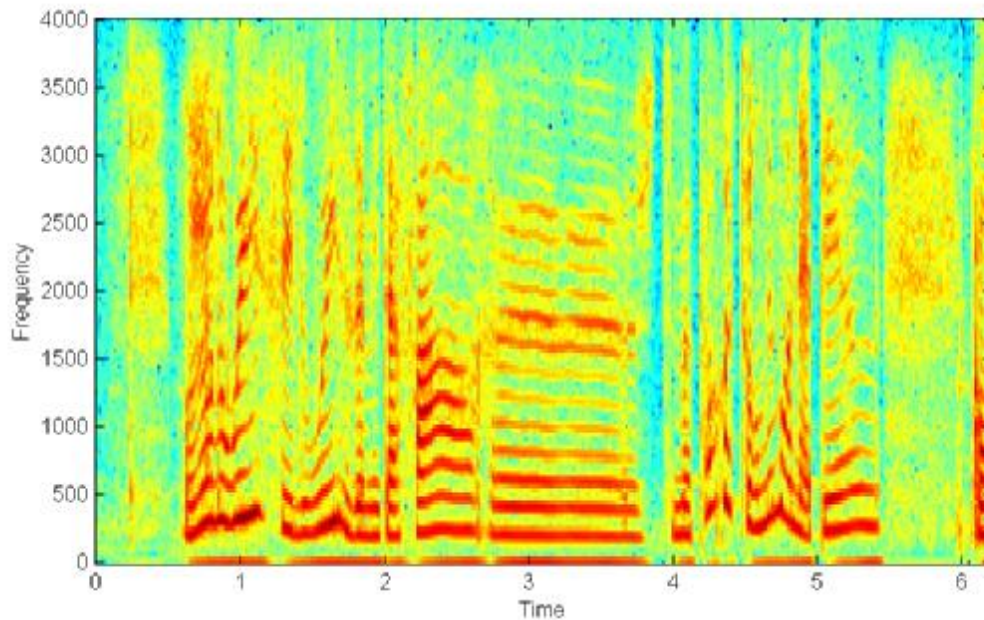


Figure 2. The corresponding spectrogram of the signal on Figure 1. Time in seconds. You can see how strong are the pitch variations. The blurred regions around $t=0.8$ sec or $t=4.9$ sec and $f=2500$ Hz are not noise but fast scanning harmonics become similar to a chirp signal. The signal around $t=0.4$ sec and $t=5.7$ sec are true noise-like fricatives or inhalation sounds. You can notice how short are the gluing consonants between voiced vowels, and how low they are in energy.

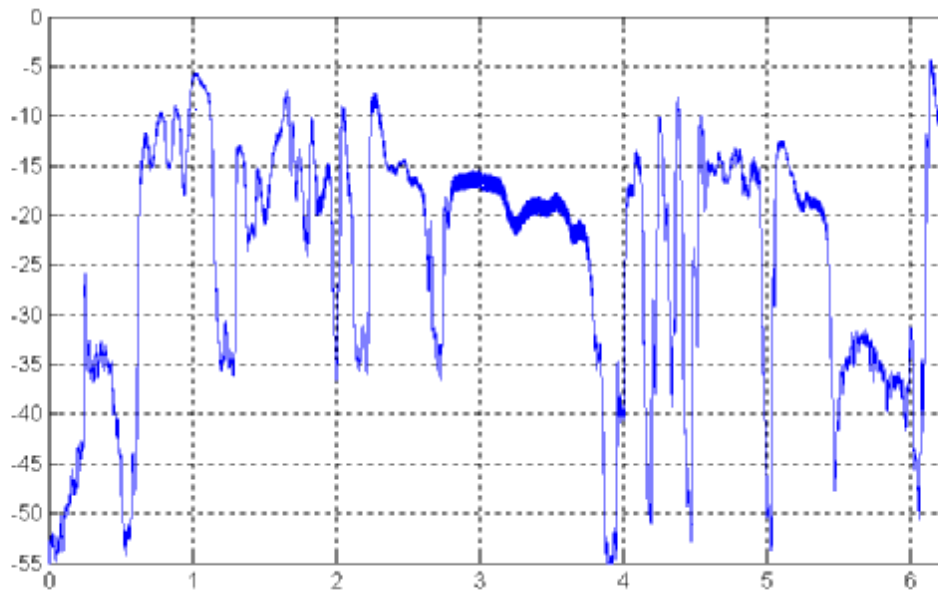


Figure 3. The energy of the Rcv Signal in dBm0, averaged over [adaptive filter length] $N=100$. The energy of noise-like voice signals is 20...25 dB lower (as at $t=1.2$; 2.2; 3.95 sec, etc) than the energy of voiced segments, and they are typically much shorter, except for the sounds of breathing.

High SNR Assumption

EC developers often assume that the higher SNR is at the moment, the better this signal is suited for training EC, the weak signals are not particularly useful, and in questionable situations it is better to disable adaptation entirely. Let's see if there is anything wrong with this assumption.

A Simple EC

Let's now assume that the adaptation is inhibited if the energy is lower than -30dBm0 (the threshold set by tests of G.168), and if the adaptation is enabled, the step size is equal to 0.25. These settings will allow successfully passing most of not all of the G.168 tests. Let's also assume that $\mathbf{D}_0 = \mathbf{I}$; what means that we know that the shape of echo path response is limited by 1.0 (absolute value). The first observation we can make over the co-variation matrix is that it is often (but not always) constructed from very similar and shifted rows of a fast-alternating waveform.

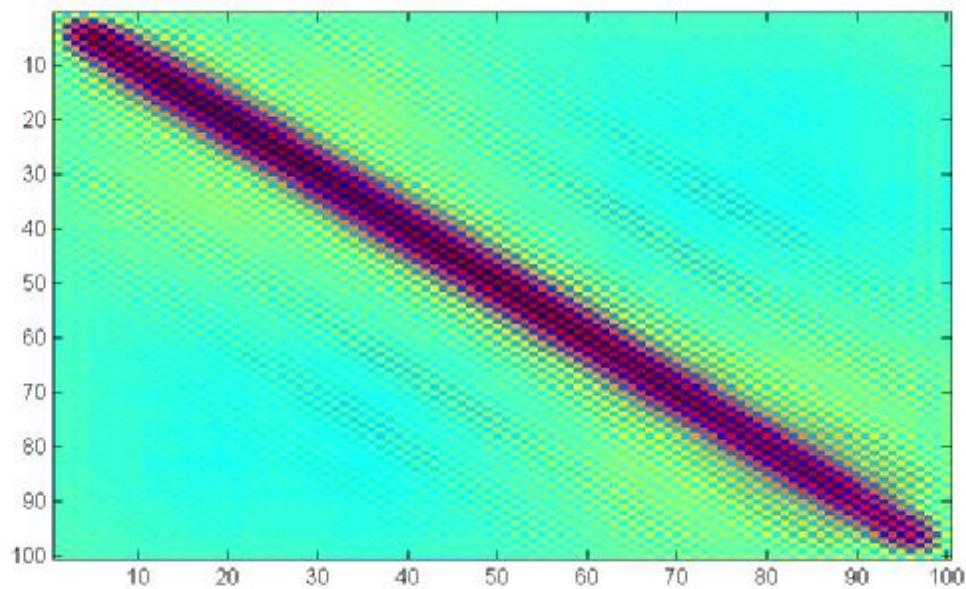


Figure 4. The co-variation matrix has almost diagonal shape.

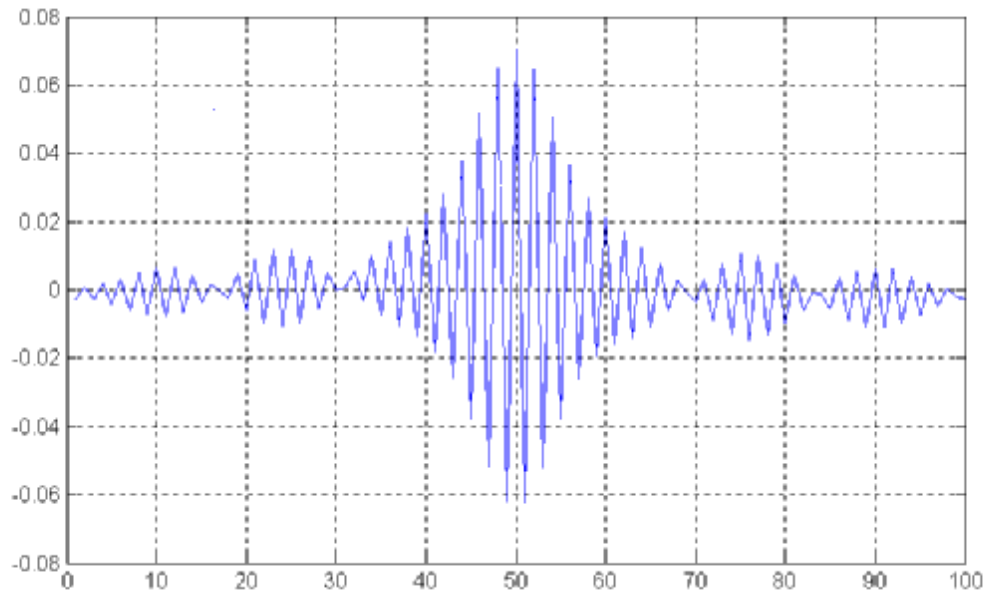


Figure 5. The middle line of the co-variation matrix has a shape of a high-pass filter.

The shape of the surface of the co-variation matrix indeed only tells us that the input signal does not have any components above 3500 Hz. The most interesting region of 300...3400 Hz is hidden inside the waveform of figures 4,5 as a second-order effect prevailed by factors completely out of our interest. We do not care if an echo canceller has poor convergence in areas where voice never occurs, whereas pure generalized math models cannot weight this circumstance appropriately. By the way, that is one of the reasons why either true of fast RLS is not given for implementation for voice echo cancellers.

To combat this trouble, let's consider instead the spectrum of co-variation matrix, e.g.

$j(\mathbf{D}_k, f) = |\mathbf{w}^H \mathbf{D}_k \mathbf{w}| / \mathbf{w}^H \mathbf{w}$; where $\mathbf{w} = \exp(-jk2\pi f/f_s)$, $k = \{1, 2, \dots, N\}$, f_s is the sampling frequency 8kHz.

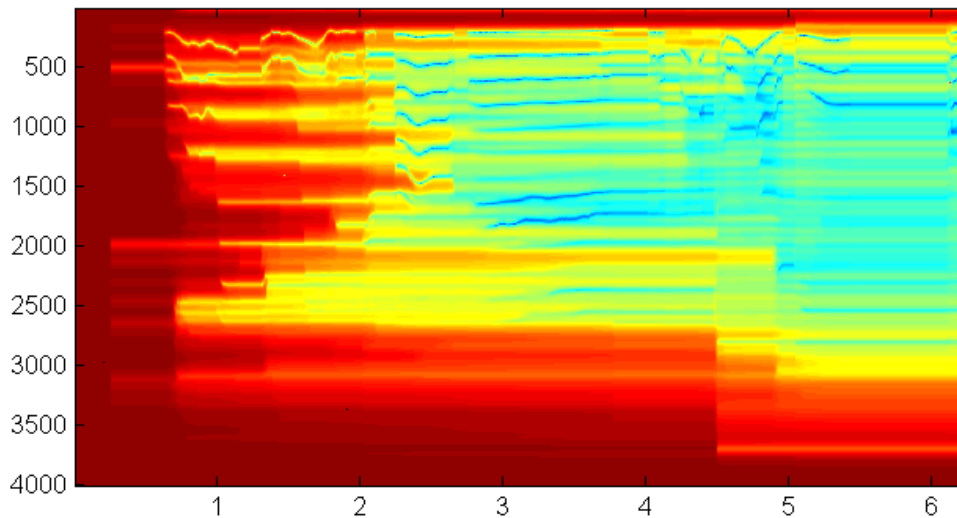


Figure 6. The spectrum of \mathbf{D}_k as the function of time in sec. Y axis is in Hz. Step size 0.25. Signal Energy > -30dBm.

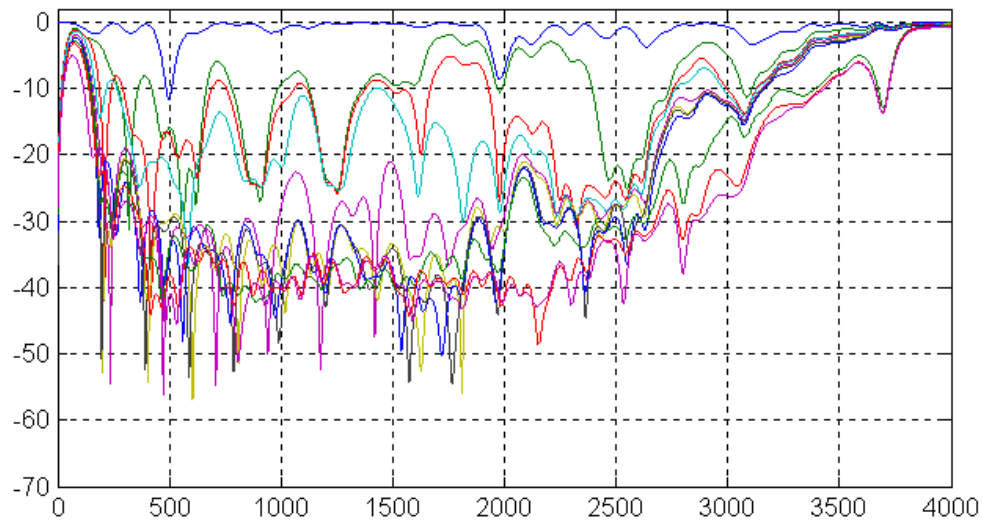


Figure 7. The slices of the co-variation matrix spectrum are taken every 0.5 sec. Step size 0.25. Energy > -30dBm.

It is obvious that the NLMS, as any other scalar step size algorithms, tracks the current signal and therefore loses the convergence gained previously (it is easy to see that the pattern of convergence of Figure 6 inversely repeats the spectral lines on Figure 2): if a EC was good once on a signal with a certain pitch, it may perform significantly worse later after it re-converges to another pitch. The difference is big: we can see that the convergence spectrum (Figure 7) have deep valleys, corresponding to the current pitch, and the hills in-between which are about 20 dB higher. So, obviously, the current ERLE cannot serve as a reliable predictor of future performance, unless enough time has passed, during which the signal covered entire pitch range, and the EC has converged in entire voice spectrum.

The resulting ERLE (+ERL) is not impressive. The ERLE is unstable, even if it achieves a certain value, it readily drops as soon the pitch undergoes changes, especially abrupt ones in the beginning of utterances, due to the speaker's emotional context.

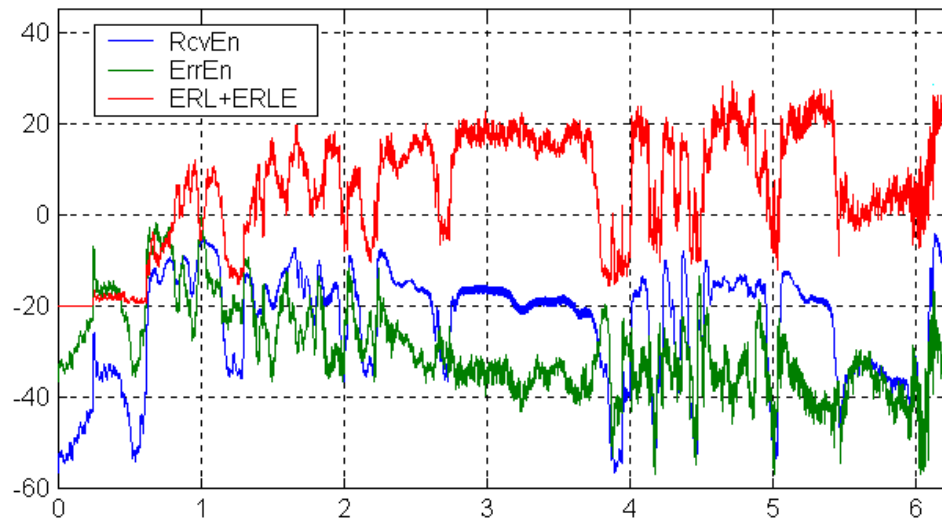


Figure 8. The signal energy, the expected error energy and the resulting ERL+ERLE curves, for the case of step size = 0.25 if energy exceeds -30dBm.

A simple EC with Optimized Step Size

Let's now consider a little bit the more sophisticated approach to determining of the step size. Let's chose step size a_k so that it minimizes the trace of co-variation matrix D_{k+1} , which is the sum of all its eigen values (omitting the intermediate math):

$$(8) \quad a_k^{opt} = \mathbf{x}_k^T \mathbf{D}_k \mathbf{x}_k / (\mathbf{x}_k^T \mathbf{D}_k \mathbf{x}_k + s_k^2);$$

That is the classical Wiener filtering equation: the signal and noise weighted in the quadratic sense. Lets' now remove the -30dBm limitation (the equation (8) must do all the work) and use this step size, assuming the same initial conditions and noise level. Note that ideally we had the weight the spectrum of co-variation matrix D_{k+1} to ensure its better minimization in the frequency range of interest: 300Hz to 3400Hz, but let's postpone considerations of such advanced topics.

It is easy to see that the performance in this case (figures 10,11) is different from a previous case as are the heaven and the earth, and this is what a good theory can do for a practical implementation if applied appropriately.

Of course we should not ever compare directly the energy of the signal with the energy of the noise to draw the decision on the step size value, because the signals shall be weighted according to equation (8). It means that initially, when D_k is close to unity matrix, we need to sum up the energy of the signal over the echo tail length (a hundred in our case) and only then the weight this sum and the noise level. Using that property, we can very successfully adapt on the background noise and the sounds of breathing (first 0.5 sec of the signal) even if their instantaneous energy is quite lower that the energy of the noise. As we can use the low energy but wide spectrum signals, we converge much faster and are affected much less by the influence of the pitch phenomena.

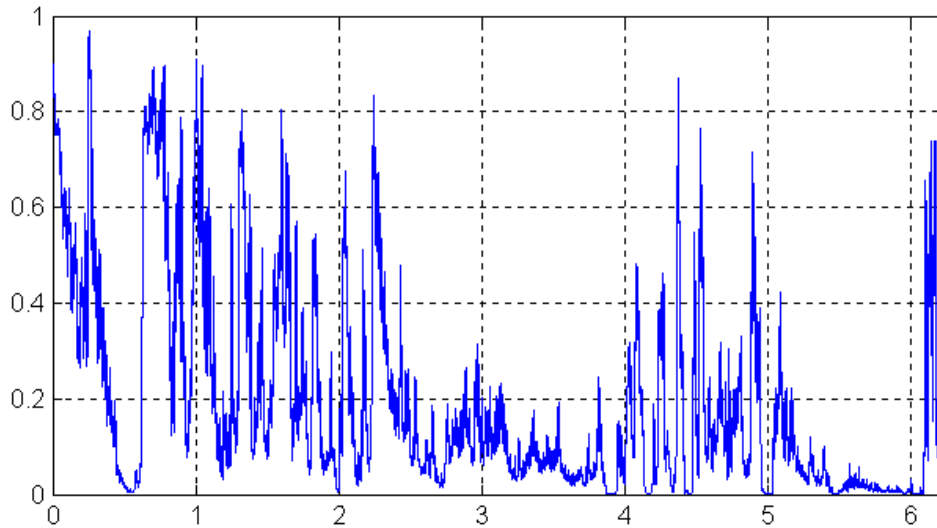


Figure 9. The optimal step size curve is very different from a simplistic 0 or 0.25 approach.

We can see that the step size curve is not simple. The step size stays high (about 0.8) for noise-like signals and for the duration of the very first voiced utterance. Then, as far as some properties of the signal change, step size jumps to high values about 0.6...1, and then it decreases as $O(1/t)$ if the signal stays approximately the same.

This is easy to understand: the strong signals are vowels or other voiced phonemes, which are constructed from nearly periodic waveforms, because human voice tract opens and resonates with the pitch frequency.

For unvoiced phonemes, the voice tract is nearly closed, so the signal does not resonate on any particular frequency, and also comes out much weaker.

As NLMS tries to combine them periodic signals to construct an echo path response (which does not have the same kind of periodic pattern), it cannot do it and inserts step size attenuation, to discard long signals, which bear very little information. But the random noise-like signals cover almost entire space, and thus an echo canceller can combine them to recover the echo path response much easier.

The $O(1/t)$ law is often met in the analysis of adaptive systems due to their accumulate-and-average nature.

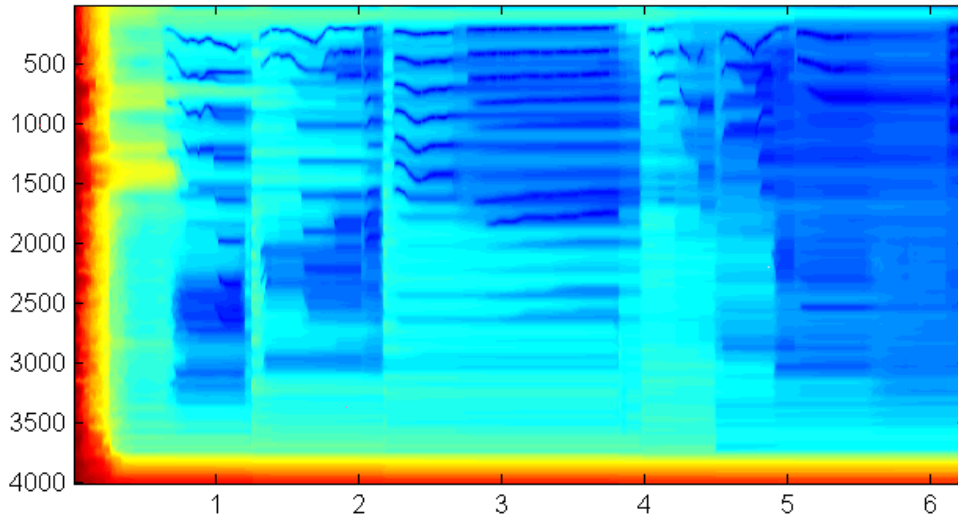


Figure 10. The spectrum of D_k as the function of time in sec. Y axis is in Hz. Step size optimal. Energy - any.

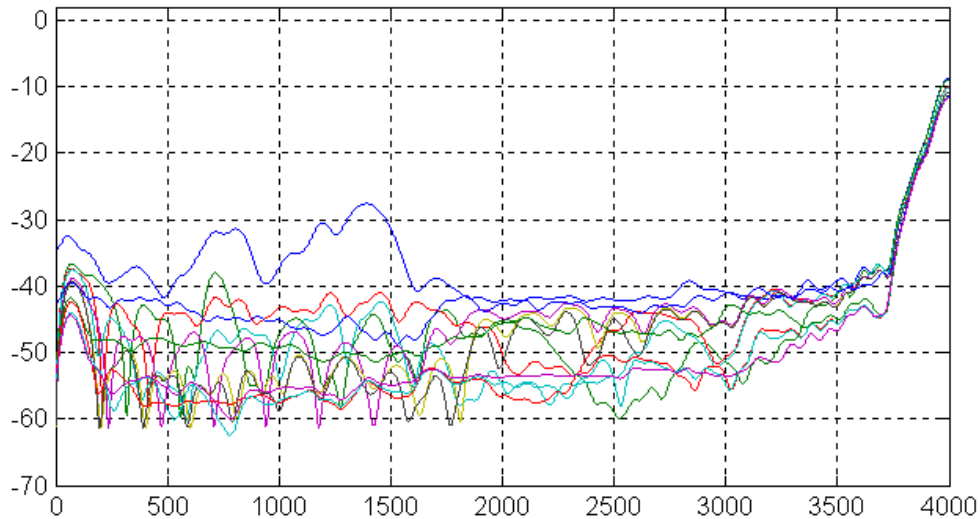


Figure 11: The slices of the co-variation matrix spectrum are taken every 0.5 sec. Step size optimal. Energy - any.

If we observe the energies of the signal and the expected error (Figure 12), we can note that ERLE also falls whenever the signal changes pitch, especially abruptly, but its stability is affected less than in the previous case. The error energy less than -46 dBm would not be observed in reality because the residual

error would be masked by noise. Note that this performance (about 40...45 dB of ERL+ERLE) is achieved in the conditions of high disturbing noise, which is usually about 20 dB lower.

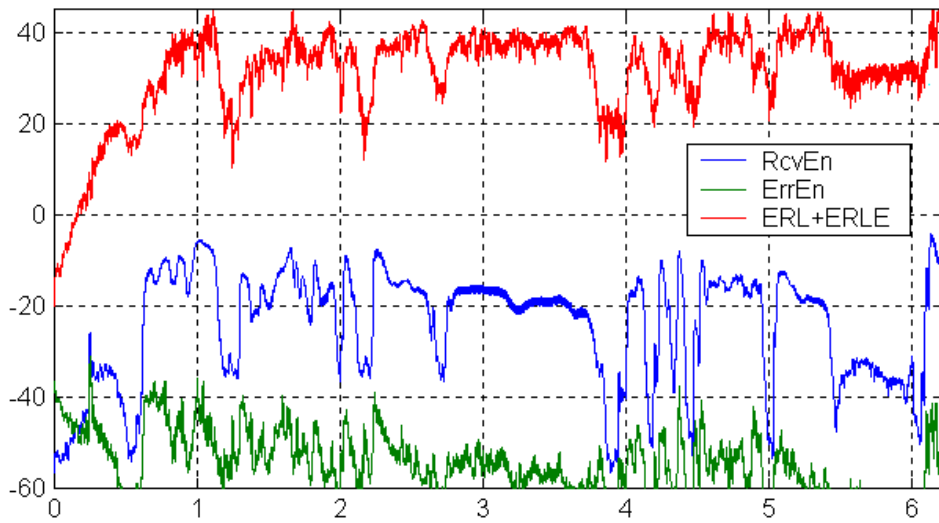


Figure 12. The signal energy, the expected error energy and the resulting ERL+ERLE curves.

Note that the co-variation matrix would lose its near-Toeplitz shape and become more complicated.

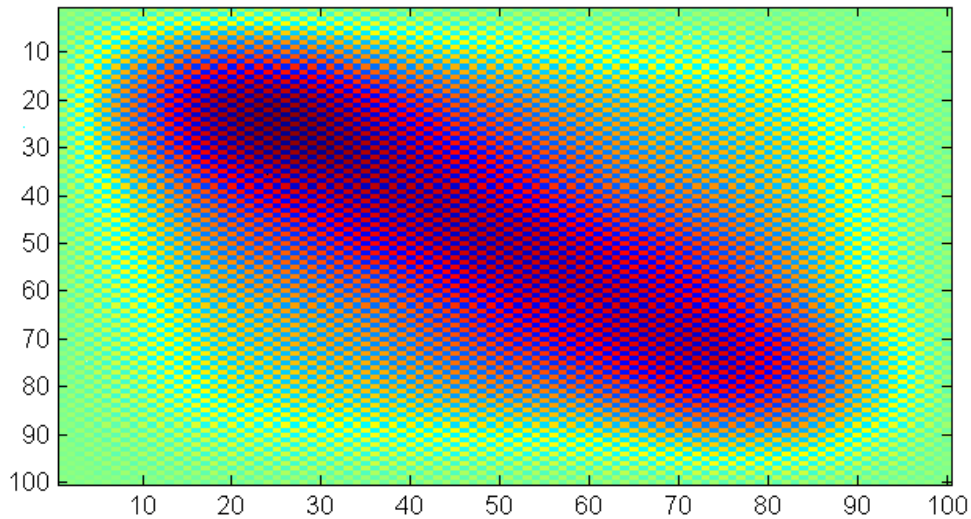


Figure 13. The co-variation matrix for optimal step size, any energy.

So we see how important it is to extend the dynamic range of voice echo cancellers and use close approximations of theoretically optimal step size.

Note that the test signal in these cases has relative high energy. In the real world, there are many soft speaking people (usually with keen hearing and soft personality as well) whose voice signal energy rarely exceeds -30 dBm, and only on voiced signals like vowels. The performance of overly simplified ECs (yet passing G.168 requirements) would be much more disastrous in those cases.

Let's assume that the signal was 14 dB lower (5 times smaller), the cut-off energy is -30 dBm, but let's still use the optimal step size.

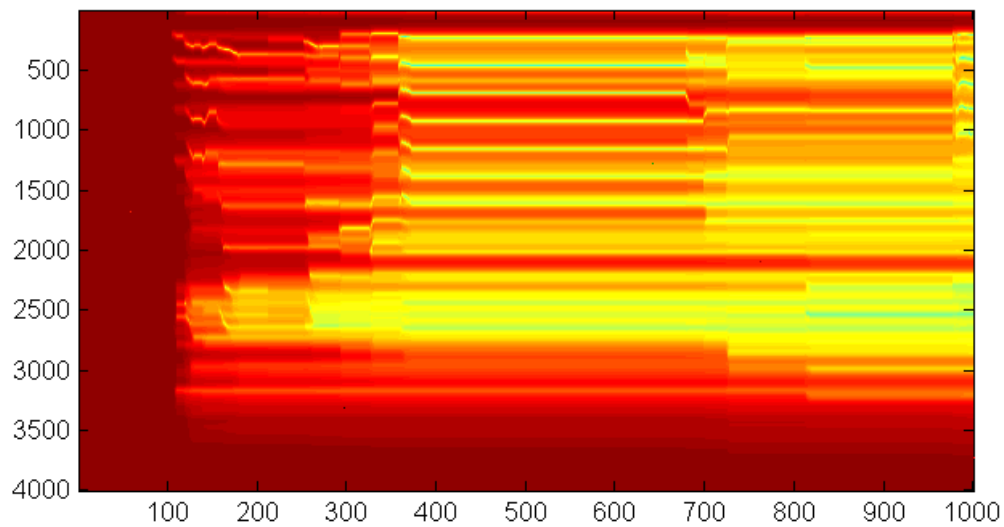


Figure 14. Co-variation matrix spectrum. The signal energy is 14 dB lower than in the regular case. Optimal step size, cut-off energy of -30 dBm.

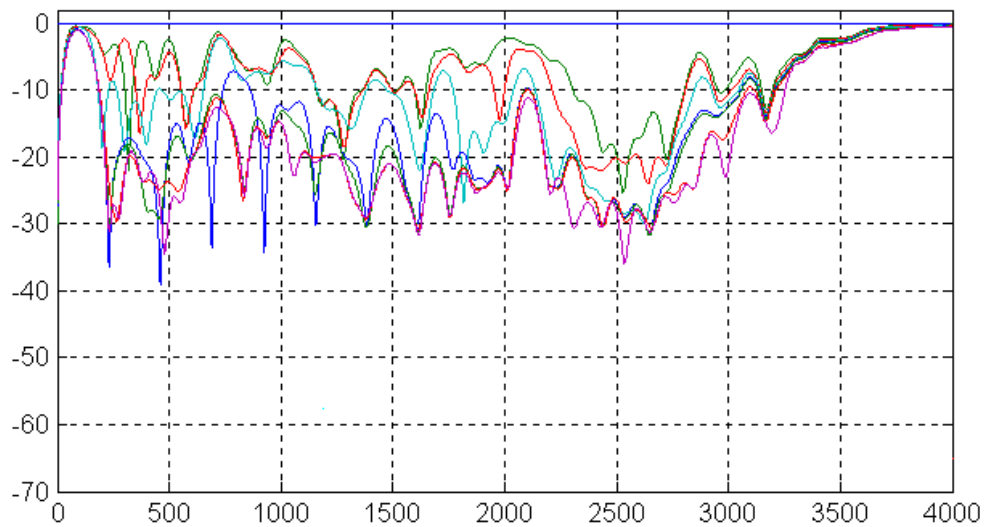


Figure 15. The slices of the co-variation matrix spectrum (from previous figure) are taken every 0.5 sec. As we can see, the ability of EC to provide adequate performance in such circumstances is debatable.

The Influence of Variations in Cut-Off Energy

Let's illustrate the degradation in EC performance (due to increased inability to exploit low-energy noise-like signals) with examples of co-variation matrix spectrum in the cases of optimal step size and cut-off energy, varying from -50 dBm up to -30 dBm.

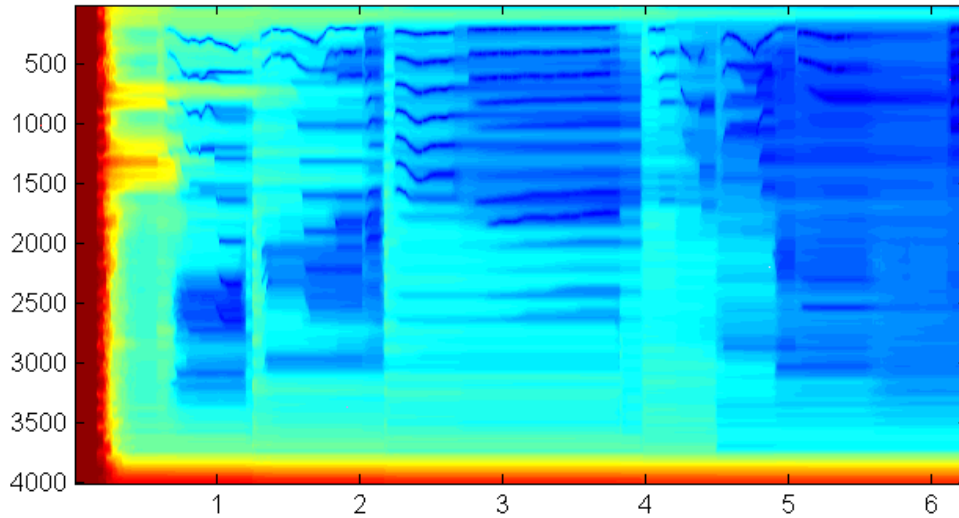


Figure 16. Co-variation matrix spectrum. Optimal step size, cut-off energy of -50 dBm

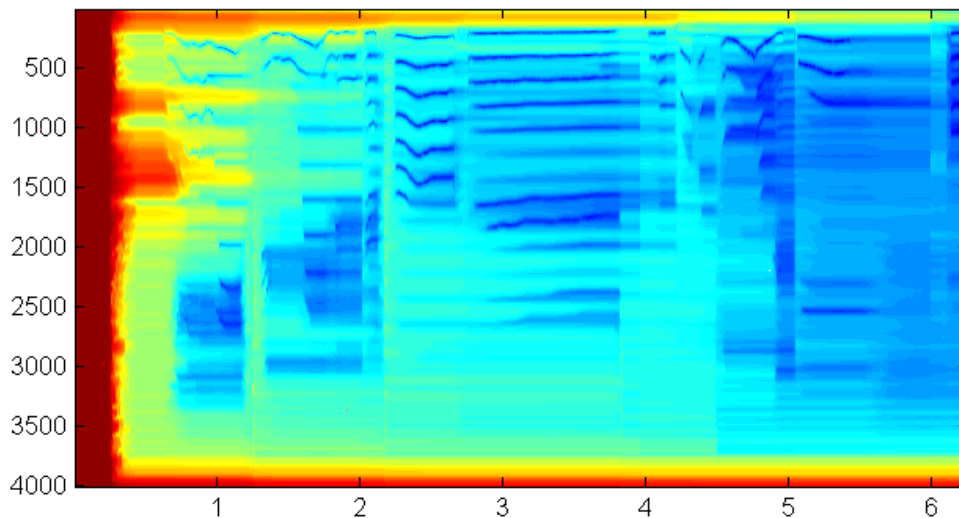


Figure 17. Co-variation matrix spectrum. Optimal step size, cut-off energy of -40 dBm.

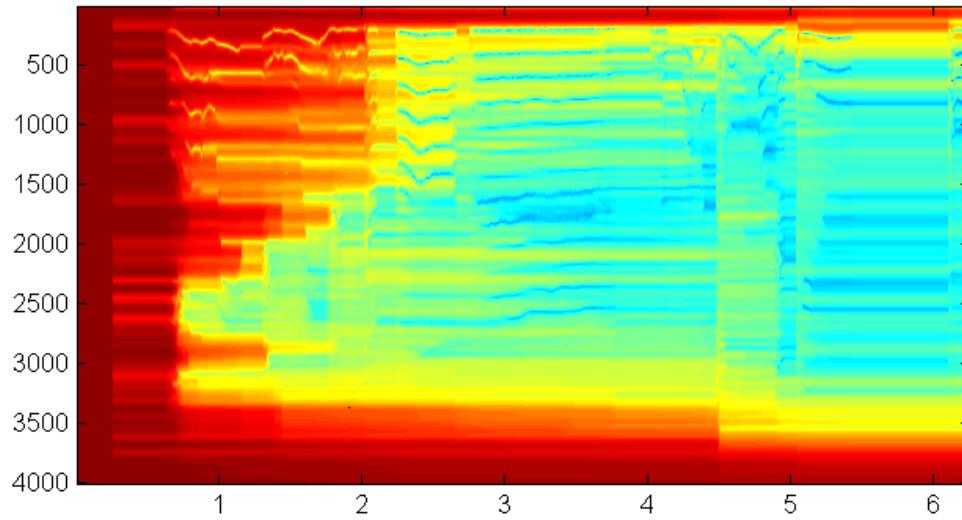


Figure 18. Co-variation matrix spectrum. Optimal step size, cut-off energy of -30 dBm.

It is obvious that the maximum of echo cancellation does not necessarily fall into the spectral regions with high energy but rather into the regions with proper signals, and that a minor increase in dynamic range would not solve the problem. Note that there are certain difficulties in fixed-point implementations of EC with extended dynamic range in the DSPs, which do not allow on-fly scaling of the multiplication product.

Neglecting A Priory Knowledge of Underlying Physics

Does the extension of the dynamic range represent the entire solution to the echo cancellation problem? No. There is another point that is often overlooked by EC developers, and it is the a-priory knowledge of the certain properties of the echo path response. If we know where the echo starts (this kind of knowledge can be achieved with several methods), then the envelope of the echo tails shape is mostly determined by the properties of codecs' receiving (Rx) and transmitting (Tx) filters.

The frequency responses of these filters have the narrowest transient bands (below 300 Hz and above 3300 Hz), compared to other devices along the circuit. They determine the echo tail disperse region, and otherwise are flat in their pass-band. Analog circuitry and local loops (which are shorter than 10 km \approx 40 μ s) may introduce gradual roll-off or emphasis. Two or more reflections with delay between them result in a frequency beat with spacing inversely proportional to the delay and the amplitude, depending on the difference between reflection magnitudes.

How long the real echo tail is? It is indeed infinite because those physical filters, mentioned above, are of auto-regressive nature. Then, let's see what happens if we use finite size filters do approximate the infinite echo tail.

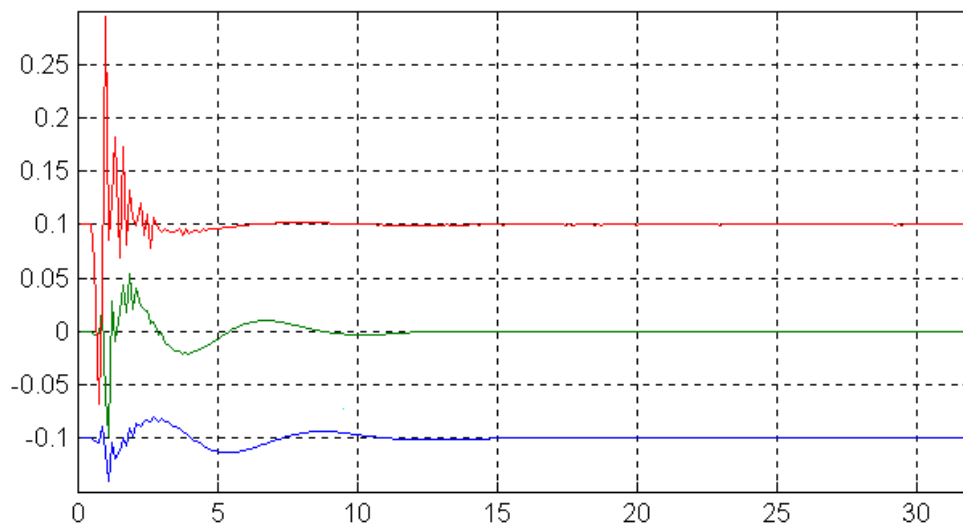


Figure 19. The echo tails, identified by least square method with white noise excitation, taken from real experiments with 3 phones from different vendors (time axis in ms).

The major differences between the echo tails are due to different circuitry matching, especially for low frequencies. It is obvious that 8 ms echo tail, recommended by G.168, would not likely suffice for all those cases. What will happen if an EC utilizes short echo tail to adapt on the echo response shown in green on the previous picture? The degradation in adaptation quality due to the EC's short echo tail will mostly affect low frequencies (see the following picture), and the resulting ERLE will heavily depend on the spectrum the signal has (ERL is also heavily frequency dependant in this case, what is not an exception).

There are so many combinations of codecs, physical lines and terminating phones. The high-pass filters (about 250 Hz) within this circuit is very often responsible for longer than 8 ms echo tail, and if the EC does not have enough "echo tail capacity per reflector", it can be very hard or impossible to build a decent double talk detector for such echo canceller, and that will result in low double talk range.

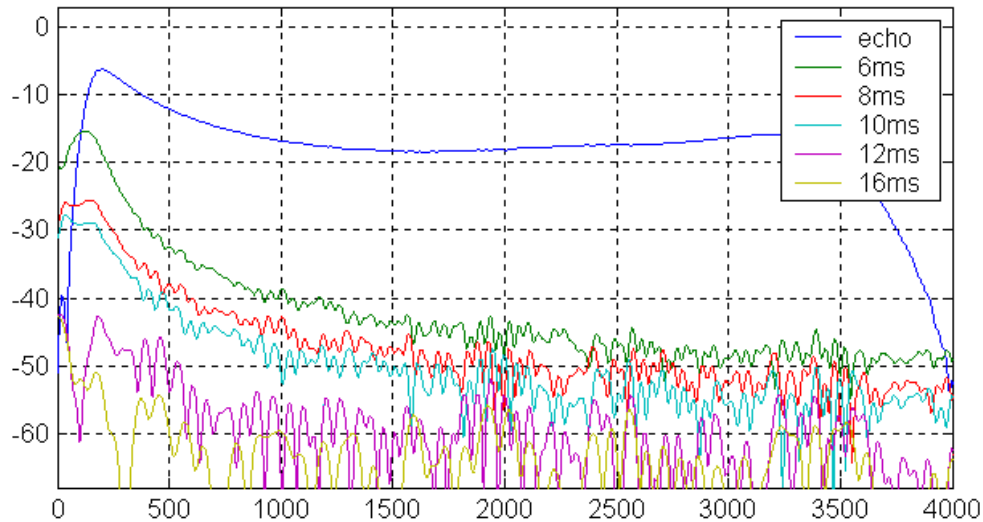


Figure 20. The echo tail response in the frequency domain is shown, along with the margins of adaptation in the cases when echo tail is too short (from 6 to 16 ms).

As we understand more about the nature of echo response, we may think about using this understanding, but to incorporate this knowledge into EC, we will need to refer the Recursive Least Square (RLS) algorithm.

A theoretically optimal RLS

The most optimal (in a certain sense) algorithm is known as Recursive Least Square (RLS). If we replace normalized scalar step size $a_k / \mathbf{x}_k^T \mathbf{x}_k$ with a (symmetric) matrix \mathbf{P}_k ,

$$(9) \quad \underline{\mathbf{h}}_{k+1} = \underline{\mathbf{h}}_k + \mathbf{P}_k \mathbf{x}_k e_k;$$

and seek the solution by minimizing \mathbf{D}_{k+1} , we will get (omitting tedious math):

$$(10) \quad \mathbf{D}_{k+1} = (\mathbf{I} - \mathbf{P}_k \mathbf{x}_k \mathbf{x}_k^T) \mathbf{D}_k (\mathbf{I} - \mathbf{x}_k \mathbf{x}_k^T \mathbf{P}_k) + s_k^2 \mathbf{P}_k \mathbf{x}_k \mathbf{x}_k^T \mathbf{P}_k;$$

$$(11) \quad \mathbf{P}_k = \mathbf{D}_k / (\mathbf{x}_k^T \mathbf{D}_k \mathbf{x}_k + s_k^2);$$

which (10) can be further simplified to:

$$(12) \quad \mathbf{D}_{k+1} = (\mathbf{D}_0^{-1} + \mathbf{S}_I^k (\mathbf{x}_k \mathbf{x}_k^T / s_k^2))^{-1};$$

The equation (9) can be re-written as (13), using familiar step size factor and rotation $\mathbf{D}_k \mathbf{x}_k / \mathbf{x}_k^T \mathbf{D}_k \mathbf{x}_k$, which does not depend on the \mathbf{D}_k scaling:

$$(13) \quad \underline{\mathbf{h}}_{k+1} = \underline{\mathbf{h}}_k + a_k^{opt} \mathbf{D}_k \mathbf{x}_k e_k / \mathbf{x}_k^T \mathbf{D}_k \mathbf{x}_k;$$

or, introducing $\mathbf{z}_k = \mathbf{D}_k \mathbf{x}_k$, as

$$(14) \quad \underline{\mathbf{h}}_{k+1} = \underline{\mathbf{h}}_k + a_k^{opt} \mathbf{z}_k e_k / \mathbf{x}_k^T \mathbf{z}_k;$$

which closely resembles the equation (2).

An RLS-inspired vector shaping

Let's assume that the envelope of the echo tail response is an exponentially failing curve $\exp(-f_T t)$ (it is indeed a bit more complicated, what is not so important now), where f_T is 300 Hz – the width of the transient band for the codecs' high-pass filters. Then we can let \mathbf{D}_0 to be not a unity matrix, but with diagonal elements as squares of the expected envelope function:

$$(15) \quad D_{0,ii} = \exp(-2ri);$$

Where r is equal to approximately $4/N$ in our case. Yet we do not want to run full-blown RLS, so we may form \mathbf{z}_k using only a normalized diagonal of \mathbf{D}_k :

$$(16) \quad \mathbf{z}_{k,i} = \mathbf{D}_{k,ii} \mathbf{x}_{k,i} / \max(\mathbf{D}_{k,ii}), \text{ or more empirical approach}$$

$$(17) \quad \mathbf{z}_{k,i} = \exp(-8i/(N+k/50)) \mathbf{x}_{k,i}$$

$$(18) \quad \mathbf{D}_{k+1} = (\mathbf{I} - \mathbf{a}_k^{opt} \mathbf{z}_k \mathbf{x}_k^T / \mathbf{x}_k^T \mathbf{z}_k) \mathbf{D}_k (\mathbf{I} - \mathbf{a}_k^{opt} \mathbf{z}_k \mathbf{x}_k^T / \mathbf{x}_k^T \mathbf{z}_k) + S_k^2 \mathbf{a}_k^{opt} \mathbf{a}_k^{opt} \mathbf{z}_k \mathbf{z}_k^T / (\mathbf{x}_k^T \mathbf{z}_k)^2;$$

As we know, out-of-voice-band regions determine the shape of co-variation matrix, so using of empirical (17) instead of (16) often proves meaningful because it follows what happens in the 300-3400Hz band of interest. We can use either old formula (8) for \mathbf{a}_k^{opt} or the new one:

$$(19) \quad \mathbf{a}_k^{opt} = \mathbf{x}_k^T \mathbf{D}_k \mathbf{z}_k / (\mathbf{x}_k^T \mathbf{D}_k \mathbf{z}_k + (\mathbf{z}_k^T \mathbf{z}_k / \mathbf{x}_k^T \mathbf{z}_k) S_k^2);$$

The spectrum of co-variation matrix is very similar to the spectrum on Figure 10, except for the very start of the adaptation period, and there are good reasons for it. From mathematical point of view, the exponentially decaying signal shaping is roughly equivalent to the shortening of the signal dimension. As all the problems with adaptive filtering grow (many, in a square proportion) with the increase of dimension, it is reasonably to expect some improvements if the effective dimension falls.

Let's look at the first, initial 0.5 seconds and closely compare the cases with and without signal shaping.

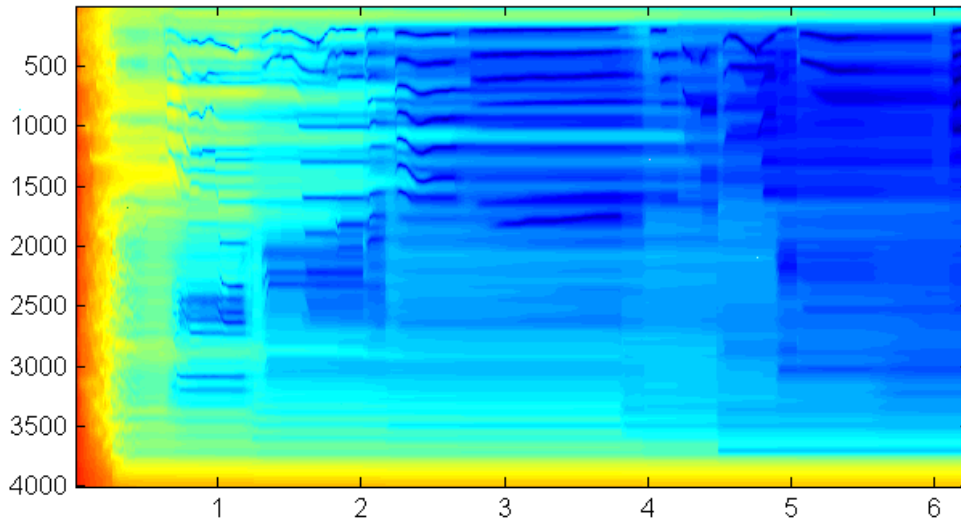


Figure 21. Co-variation matrix spectrum. Exponentially shaped signal according to formula (17). Optimal step size.

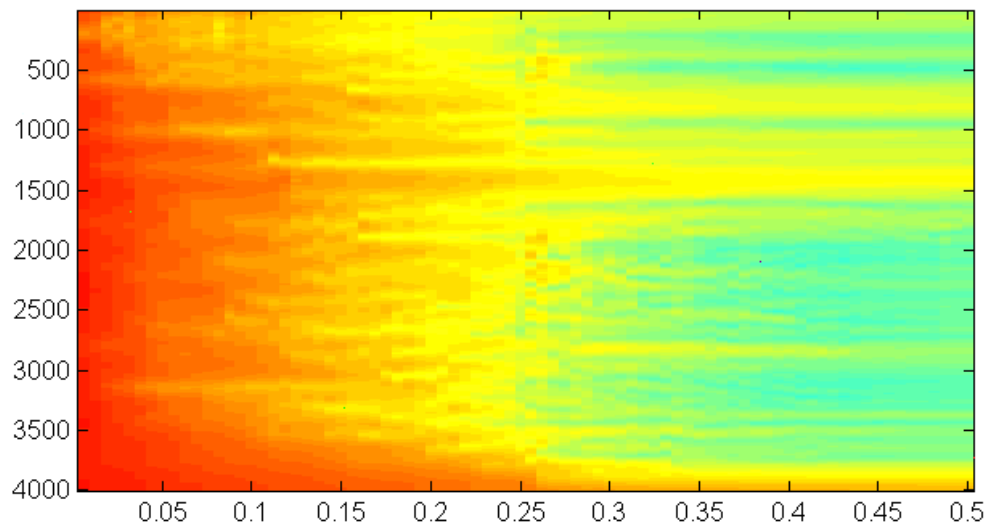


Figure 22. The first 0.5 seconds. Co-variation matrix spectrum. Exponentially shaped signal according to formula (17). Optimal step size.

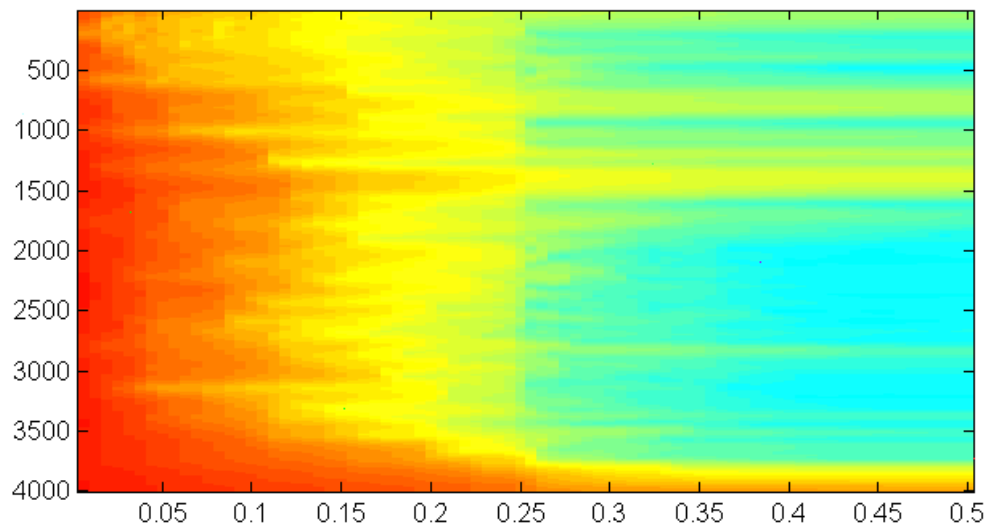


Figure 23. The first 0.5 seconds. Co-variation matrix spectrum. The shaped signal according to formula (16), following diagonal of the co-variation matrix. Optimal step size.

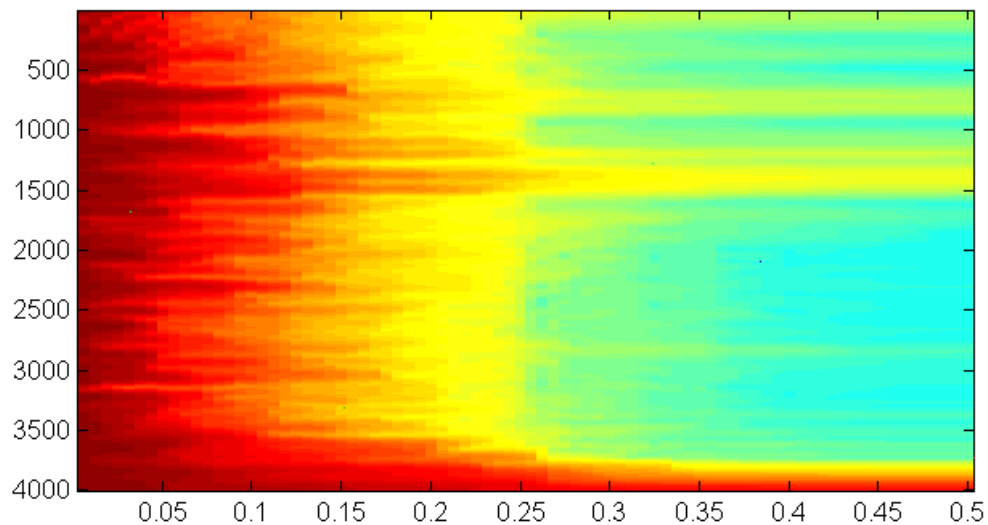


Figure 24. The first 0.5 seconds. Co-variation matrix spectrum. Optimal step size

It is obvious that the speed of convergence has improved (although final quality stays approximately the same). If we look at the shape of the signal on the Figure 1, we see that the phrase itself is preceded by some breathing sounds of relatively high amplitude and duration of about 200ms. It is a luxury rather than a regular case. As we have learned, the fricatives are much more important for voice echo cancellers than the vowels. But fricatives are not only 15...25 dB weaker than vowels, they are also much shorter. It isn't unusual if a fricative or plosive takes only 25...50 ms, and in those cases the increase in speed due to the signal shaping (and therefore effective dimension lowering) plays its major role (of course it shall be carefully applied). The result is simple: the users may not notice any echo phenomena at all, even when echo path changes, and the echo canceller may sound transparent from the very beginning of the call.

This is not the only way to incorporate the knowledge of echo path physics into an EC.

NLMS & RLS

The modern DSPs are reasonably fast at running LMS type adaptive algorithms. Many DSP have embedded instructions for delayed LMS and spend only 2 cycles per tap for adaptation and cancellation (what are expressed by the formula (1) and (2)). But can an algorithm like LMS serve as a base for high-quality voice echo canceller, or something closer to RLS-like algorithms shall be used instead?

NLMS and many other scalar step-size algorithms have been under intensive studies for many years. The algorithm shall in theory converge ($\text{trc}(E\{\text{var}(\mathbf{h}_k)\}) \rightarrow 0$ for $t \rightarrow \infty$) for any nonsingular Fisher matrix $S(\mathbf{x}_k^T \mathbf{x}_k)$, regardless of definiteness, symmetry, or localization of the eigenvalues of the coefficient matrix. In spite of this theoretically stated robustness and the simplicity of the algorithm, the area of its practical applicability shall not be overstretched. All algorithms with scalar step size have one very unpleasant property: they work as good as RLS if the signal is noise-like, but degrade very quickly if the signal has abnormalities.

Let's assume that \mathbf{x}_k is combined of only 2 orthogonal components, \mathbf{x}_{1k} and \mathbf{x}_{2k} , (for example, sampled from sine waves, $f_1 = n * f_2$). We can also project \mathbf{h}_k on the same vectors. Then the components will be converging as:

$$\mathbf{h}_{1k+1} = \mathbf{h}_{1k} + a_k * \mathbf{x}_{1k} * (y_k - \mathbf{x}_{1k}^T * \mathbf{h}_{1k} - \mathbf{x}_{2k}^T * \mathbf{h}_{2k}) / (\mathbf{x}_{1k}^T * \mathbf{x}_{1k} + \mathbf{x}_{2k}^T * \mathbf{x}_{2k});$$

$$\mathbf{h}_{2k+1} = \mathbf{h}_{2k} + a_k * \mathbf{x}_{2k} * (y_k - \mathbf{x}_{1k}^T * \mathbf{h}_{1k} - \mathbf{x}_{2k}^T * \mathbf{h}_{2k}) / (\mathbf{x}_{1k}^T * \mathbf{x}_{1k} + \mathbf{x}_{2k}^T * \mathbf{x}_{2k});$$

If the amplitudes of \mathbf{x}_{1k} and \mathbf{x}_{2k} are different ($\|\mathbf{x}_{1k}\| \gg \|\mathbf{x}_{2k}\|$), then \mathbf{h}_{1k} convergence will be fast, but \mathbf{h}_{2k} convergence will be slowed down by factor of $\mathbf{x}_{2k}^T * \mathbf{x}_{2k} / (\mathbf{x}_{1k}^T * \mathbf{x}_{1k} + \mathbf{x}_{2k}^T * \mathbf{x}_{2k})$, what is roughly squared proportion. Lets assume that $\text{std}(\mathbf{x}_1) = 5 * \text{std}(\mathbf{x}_2)$. A fixed point NLMS implementation will go through 3 distinct stages:

- EC achieves first 14 dB of echo cancellation (e.g. Echo Return Loss Enhancement, ERLE) on full speed.
- Then the convergence slows down 25 times.
- Then EC stops adapting entirely because the energy-normalized step size for the second mode is so low that the energy-normalized error drops below 1 bit resolution.

As soon as convergence on second mode stops, the convergence on first mode will stop as well because the error term $(y_k - \mathbf{x}_{1k}^T * \mathbf{h}_{1k} - \mathbf{x}_{2k}^T * \mathbf{h}_{2k})$ is combined of both.

The voice signal structure (even for periodic vowels) is more complicated than just two sine waves, but the source of the problem is traceable to this very simple case. As we saw it, an EC, based on NLMS, will be converging to a 'solution', mostly determined by the current pitch and the position of major formant.

ERLE may be quite good, but the spectrum of estimation error co-variation matrix will be poor. Low-energy high frequency components may be not cancelled, but they are more perceptually audible. As the result, double talk performance of a NLMS based EC is generally poor.

G.168 often specifies EC performance in terms of ERLE, but ERLE is an internal (to EC) measure, while double talk range is external, customer-perceivable measure. Of course, an EC with deep double-talk range usually has high ERLE value. The opposite statement is not true: an EC with high ERLE may have a mediocre or low double-talk range.

G.168 uses the test signal (CSS) with long and powerful noise part, and an algorithm like NLMS may look as an adequate solution. Unfortunately, this is not a realistic scenario, and voice EC test procedures shall either avoid using noise-like signals, or avoid drawing far-stretching conclusions from such experiments.

RLS performance

We can characterize the RLS performance for the same test case in the same terms of co-variation matrix spectrum, and it is obviously very different:

- RLS is not tracking the current pitch. From a formal point of view, the formulas of NLMS and RLS are similar, but the behavior is absolutely not.
- For any frequency, the error estimation is a monotonically falling curve. RLS is an algorithm with 'infinite' memory (what is both a blessing and a curse).
- The convergence is much faster.
- The valleys of the spectrum are wider. RLS can successfully converge even on vowels if the pitch changes significantly.

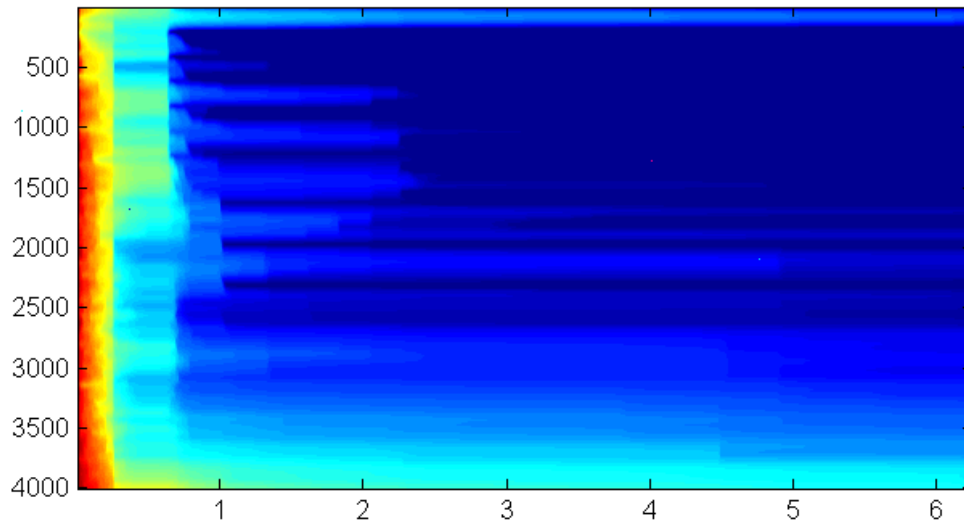


Figure 25. The theoretically optimal RLS' co-variation matrix spectrum.

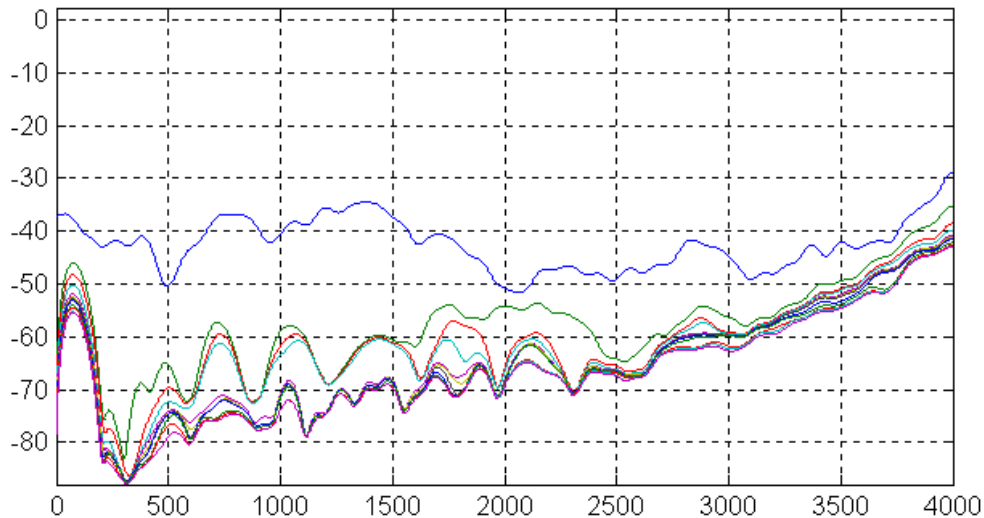


Figure 26. The 0.5 sec slices of the theoretically optimal RLS' co-variation matrix spectrum

Of course such degree of performance cannot be achieved in practice due to many reasons. Some of the RLS properties are not feasible for practical EC implementations with a 16 bit DSP and a realistic echo path circuitry, whatever MIPS and memory are spent. For example, the RLS algorithm is internally based on the so-called Gram-Schmidt orthogonalization process, which assumes that the identified system is truly linear. It is not true for any realistic echo path circuitry, and it limits the ability of an EC to use the

previously accumulated information (as RLS does) even if infinite precision is used for internal RLS calculations.

But some of RLS properties can be approached in many ways, from different sides, and to a degree by simplified and more robust RLS-derived algorithms. Although there cannot be an algorithm claiming to be the best RLS derivative, the closer a practical EC implementation is to the RLS (or Kalman filter), the higher may become the perceptual quality.

While deriving a practical algorithm, we shall not forget one very important point: the theory often assumes that the S_k^2 variance of noise is known, what can not be true because this is the statistic of the unknown far-end voice signal we need to pass through EC unmodified, so there is a need in the additional circuit of adaptation above the adaptive filter itself.

If an EC algorithm is limited to simple scalar step size method like LMS, it may bear close similarity to an anecdote about a drunken man who is searching for something lost under a light stand not because the thing was lost there, but because it is lighter there.

Conclusions

- Voice echo cancellation has a hidden shift in paradigm, which is often neglected.
- The key to the solution is the understanding of the voice nature, the physics of the signal reflection, and the fundamental properties of estimation error co-variation matrix.
- Learning of the co-variation matrix spectrum helps to enhance understanding of adaptive filtering and dissolve some popular illusions.
- Any empirical approach shall be verified against theory.
- Many good working simplified methods can be derived from theoretically optimized algorithms.
- Near-optimal step size control, precise double talk detectors and accurate non-linear processors can be derived from the studies of co-variation matrix properties, resulting in wide double talk range of voice EC and a smooth sound, undistinguishable from a local TDM call.
- If the theoretical base of a particular EC adaptive filtration method is not right, then no amount of algorithm tweaking, tuning and patching will help.
- If practice differs from theory, it is practice that suffers.