PART II - Adaptation

© Michael Tsiroulnikov a.k.a. MIKET DSP SOLUTIONS 2001-2020. Proprietary. All rights reserved.

GNU General Public License v.3+ https://www.gnu.org/licenses/ Only research, academic, and free-for-all open-source open-test-vectors usage and applications are allowed. Any commercial and/or for-profit usage of disclosed technology, directly or indirectly, in part or in whole, in whatever form it may take, is expressly prohibited unless a prior written permission has been obtained from the rightsholder. Cite this work as *"Michael Zrull (2020). Fast Subband Adaptive Filtering. Matlab Central File Exchange."*

1 PREFACE

1.1 BASICS

Read Part I first.

You can not proceed until you have read and understood the prerequisite Gustafsson's "Adaptive Filtering and Change Detection". Even if you know adaptive filtering for under-modeled stochastic non-stationary systems with weak nonlinearity pretty well, you are still advised to glance through it. I will not repeat anything discussed there.

Quite a bit of problematics can be discussed without taking sub-band limitations, subsampling, and aliasing into consideration, even without going into complex domain.

For as much as possible, the discussion will be leaning towards simplicity, delta-function responses, and white Gaussian noise as excitation.

The names of chapters with pictures include a reference [2xy] to a doc_p2xy.m script which was used to generate these pictures.

1.2 SUMMARY

- The system to be identified must be properly band-limited and sample-able, aliasing and singularity adequately accounted for. Adaptive filtering configuration plays a critical role. Lack of such understanding makes the rest meaningless.
- Scalar step size (not using Gram-Schmidt orthogonalization in any form) algorithms have spectral deficiency whenever excitation is colored at source or colored by band-limiting filters
- The acoustic model of RIR perturbations with exponential decay can and should be properly incorporated into adaptive algorithms
- A new vector step size class of Diagonal Least Square adaptive algorithms to account for meaningful RIR perturbation models is introduced
- The robustness of traditional and new adaptive algorithms to implicit and explicit assumptions is discussed, and it's found that all of the conventional single model algorithms are not robust.
- A new class of meta-adaptive modifications to the traditional and new algorithms is introduced by closing a feedback loop on the RIR perturbation estimations to make the meta-adaptive algorithms to be robust to errors in such estimations, resulting in a lower number of models for xxMM algorithms.

2 BAND EDGE EFFECT

2.1 BASICS

On one hand, the "band edge effect" is a direct consequence of parallel configuration and has nothing to do with subband processing.

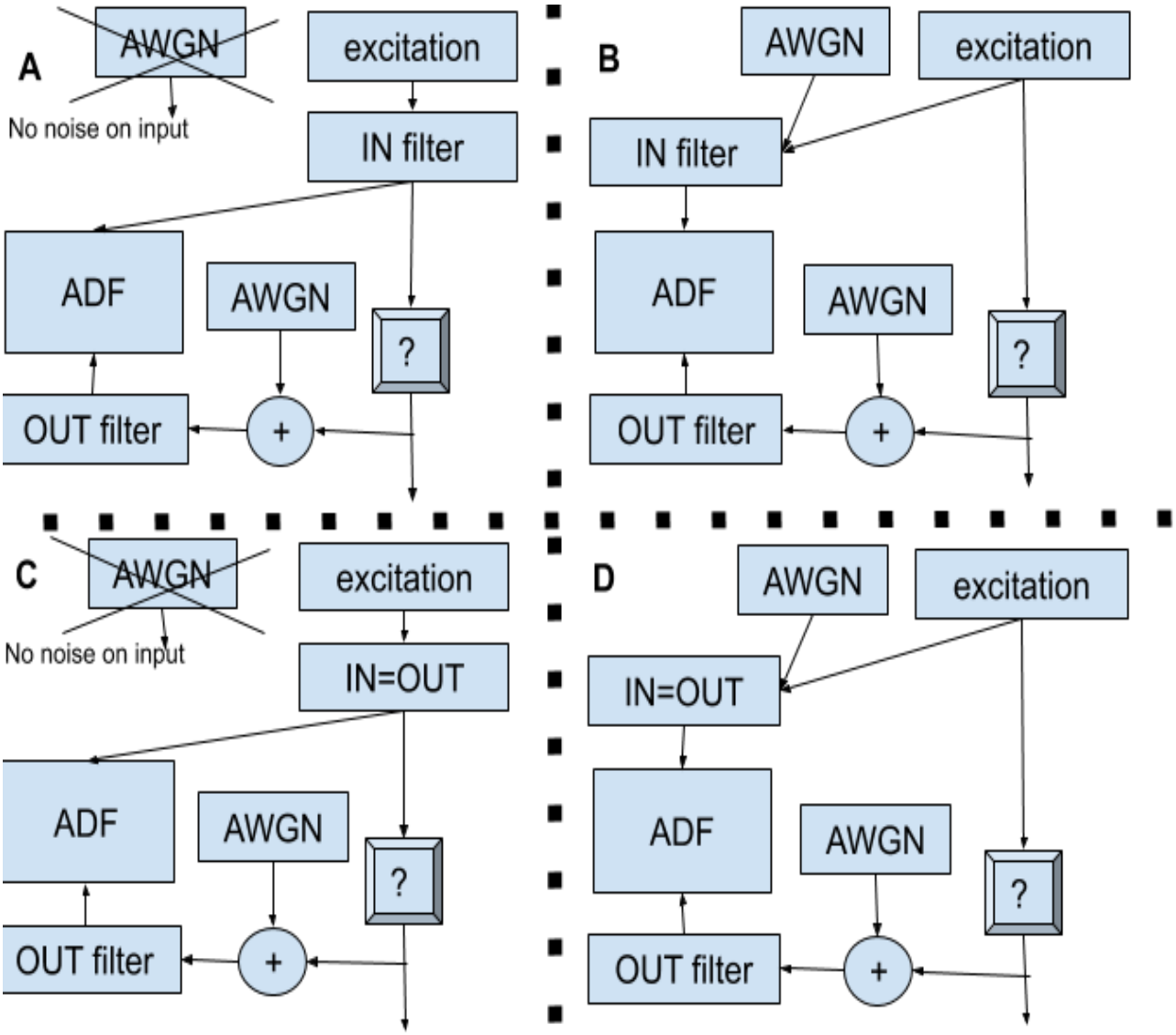
On another hand, the "band edge" effect is a direct consequence of LMS spectral deficiency and also has nothing to do with subband processing per se.

2.2 PARALLEL CONFIGURATION [201]

Let's consider the following 4 configurations:

- Left: Consecutive
- Right: Parallel
- Top: IN and OUT filters are different
- Bottom: IN filter is the same as OUT filter

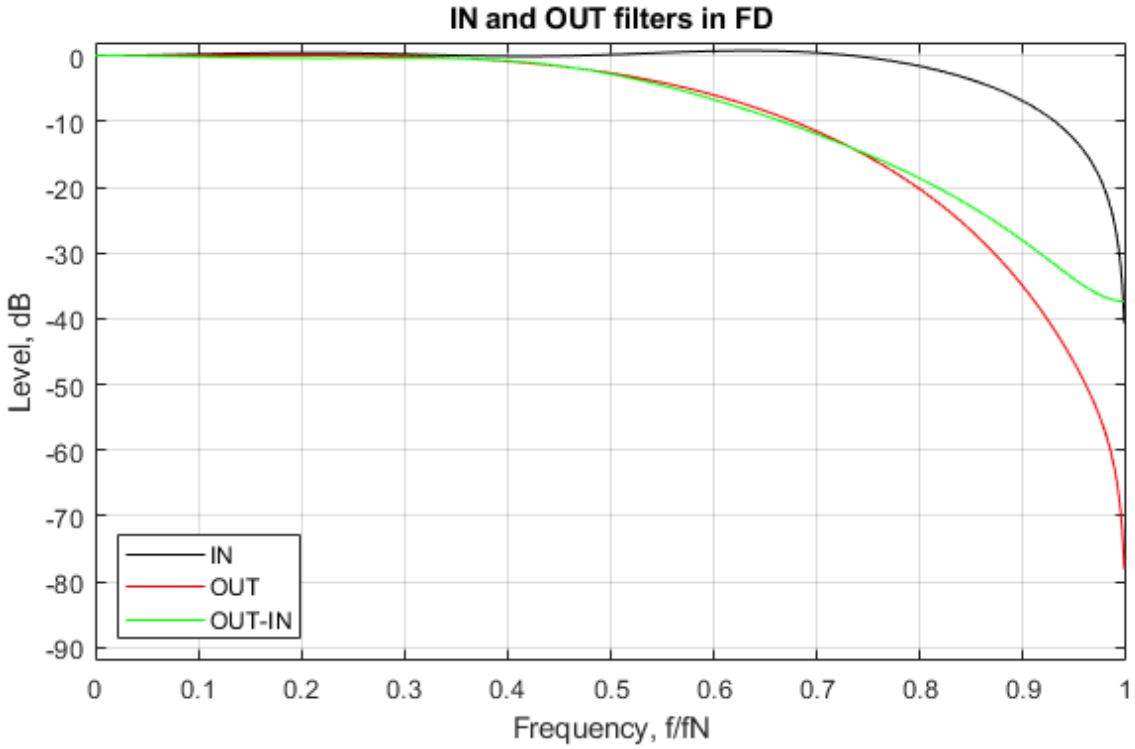
In top/bottom cases, ADF gets the same reference signal regardless of the consecutive/parallel configurations. If the old theory holds, both the dispersion matrix of estimation errors and residual error spectra shall be identical because the expectation of residual error variance is $var(res_t) = x_t{}^H D_t x_t$ where x_t is the excitation vector (after IN filter), and D_t is the dispersion matrix of estimation errors, $D_t = E\{(\widehat{h_t} - h)(\widehat{h_t} - h)^H\}$ regardless of the actual system response.



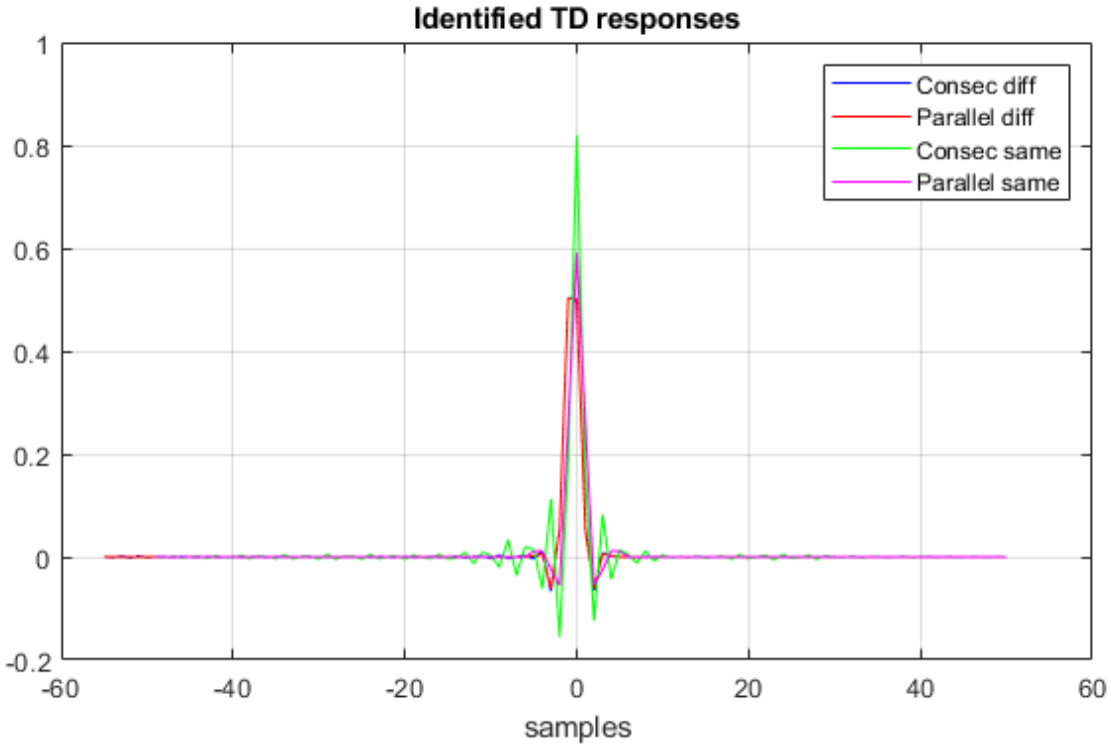
The OUT (QMF-ish) and a wider IN filters are chosen as

```
fout= fir1(FIRSZ-1,0.60,hann(FIRSZ).^1.3);
```

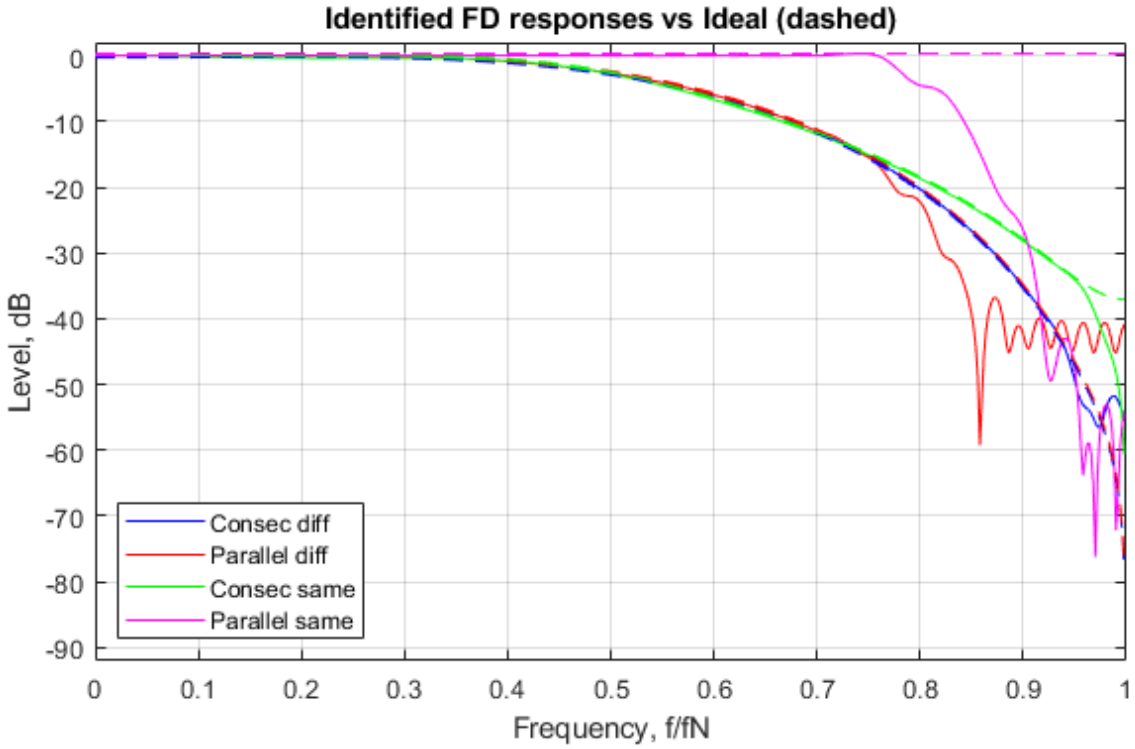
```
fin=fir1(FIRSZ-1,0.87,tukeywin(FIRSZ,.25));
```



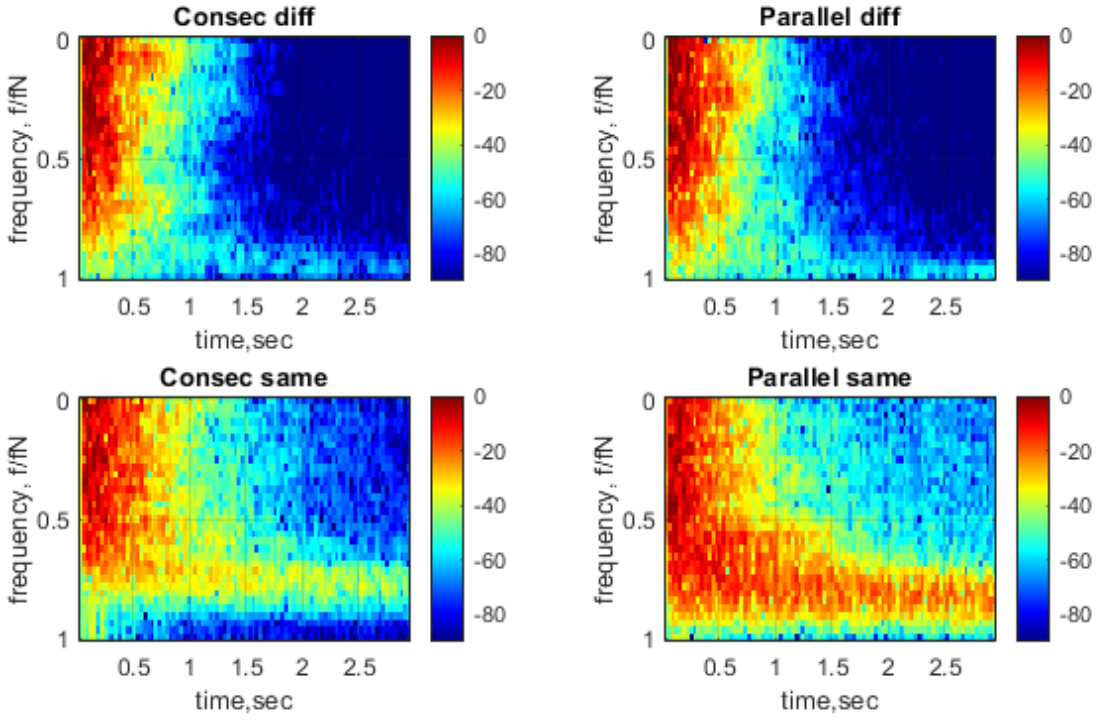
Then, we run LMS for several seconds and the results are, in the time domain:



... and frequency domains:



As we can see, the "same" IN and OUT of QMF-ish type filter simply do not work.



Well, actually, in any configuration. Yes, it's hopelessly worse in parallel with LMS.

2.3 LMS SPECTRAL DEFICIENCY [202]

Whatever IN filter we chose we can not make it flat edge-to-edge, even if we use per-subband internal equalization. Some of the sharp drop will remain. For RLS, the dispersion matrix of estimation errors $D_t$is the inverse of the Fisher matrix, where $\sigma_t$is (possibly, non-stationary) noise standard deviation;

$$D_t = (D_0{}^{-1} + \sum_{i=1}^{t} x_i x_i{}^H / \sigma_i{}^2)^{-1}$$

And therefore, if the frequency spectrum of$x_t$formed by IN filter, then the spectrum of residual error, with power expectation $x_t{}^H D_t x_t$, shall be flat (if no aliasing happens, of course). It's not so simple for LMS.

$h_{t+1} = h_t + \mu_t x_t (y_t - x_t{}^H h_t)/(x_t{}^H x_t)$;

where μ_t is the step size;

$\delta_t = h_t - h$;

where h is the true value of system response.

$\delta_{t+1} = \delta_t - \mu_t x_t x_t{}^H \delta_t/(x_t{}^H x_t)) + \mu_t x_t n_t/(x_t{}^H x_t)$;

where n_t is the noise sample.

$G_t = (I - \mu_t x_t x_t{}^H/(x_t{}^H x_t))$;

is the "dispersion squeezing" matrix, symmetric.

$D_{t+1} = G_t D_t G_t + \mu_t{}^2 \sigma_t{}^2 x_t x_t{}^H/(x_t{}^H x_t)^2$;

where for negligible noise we can see that

$D_{t+1} = G_t G_{t-1} G_{t-2} \dots G_2 G_1 D_0 G_1 G_{2\dots} G_{t-2} G_{t-1} G_t$;

and

$G_1 G_{2\dots} G_{t-2} G_{t-1} G_t \approx I - \mu_1 x_1 x_1{}^H/(x_1{}^H x_1) - \mu_2 x_2 x_2{}^H/(x_2{}^H x_2) - \dots \mu_t x_t x_t{}^H/(x_t{}^H x_t)$

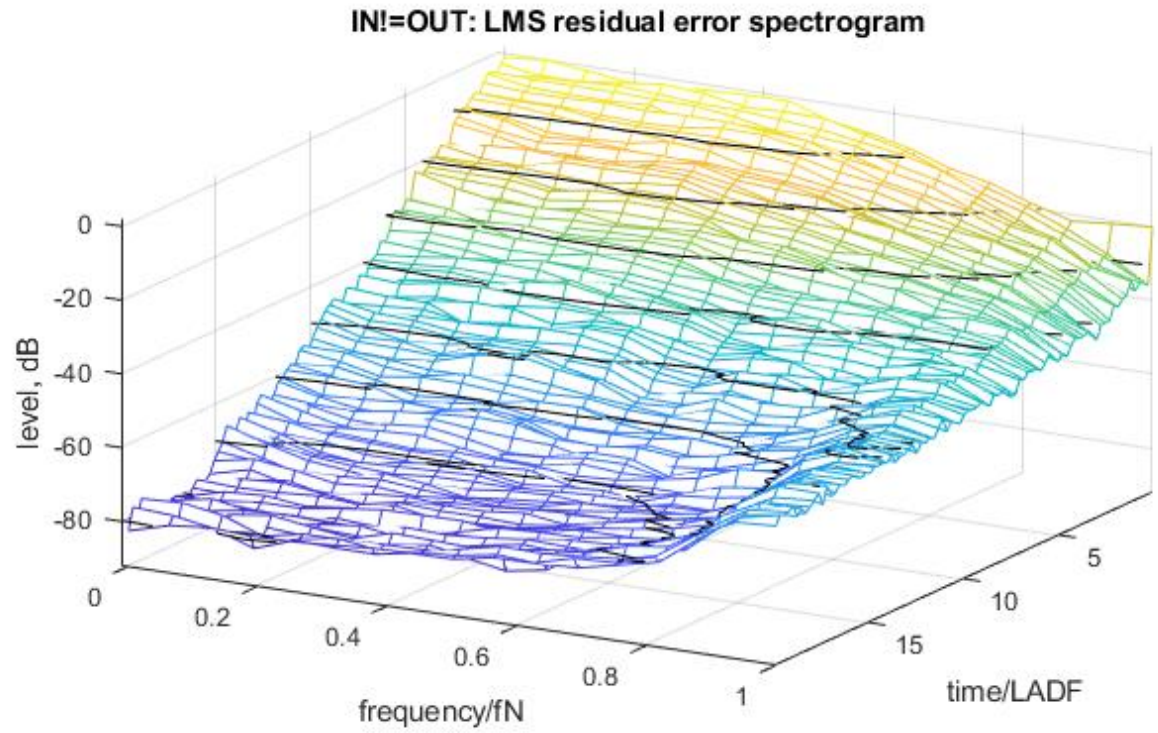
or, approximately for small step sizes:

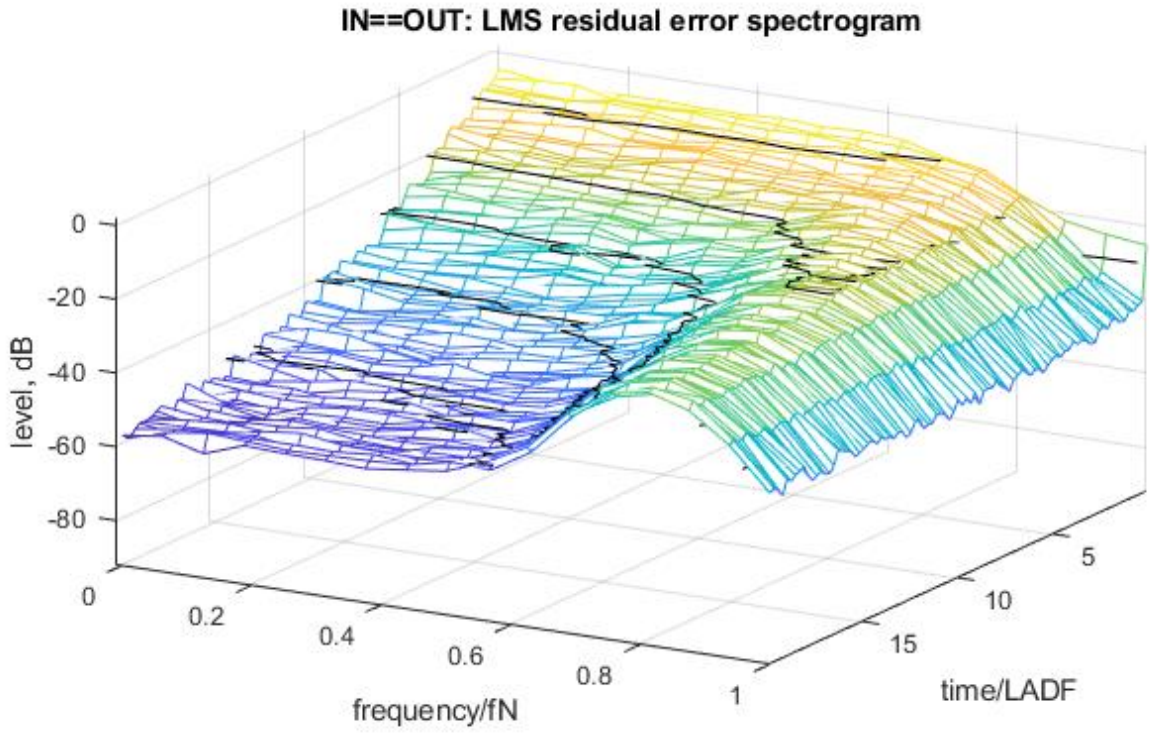
$$\approx I - \mu \ (\sum_{i=1}^{t} x_i x_i{}^H / \sigma_i{}^2)/\, mean(x^H x)$$

where we can see the familiar Fisher matrix again.

The power spectrum of LMS residual error is inversely proportional to the IN(f)^2 filter. This is the major difference between LMS and RLS, and therefore, the filterbanks (and/or per-band equalizer) for an FSAF application shall depend on the sub-band algorithm used for adaptation in subbands. RLS, despite being optimal, is more forgiving to the non-idealities in the filterbank design, while LMS and other non-matrix step-size algorithms require much more consideration, design efforts, and of course longer testing.

'Diff' case is very different from 'Same' [doc_p202.m]:





In LMS, there is practically no feedback on spectral error close to the band edge. Thus, the errors can grow unconstrained unless specific regularizations are applied. One of them is to add some noise to IN signal, close to the band edge, to enforce spectral zeros. We may allow IN filter to intentionally alias some signal into a subband, details to be discussed in Part III.

2.4 SUMMARY

The discussed material, due to some reasons beyond my understanding, is not covered in [most of] textbooks:

- The system to be identified must be properly band-limited and sample-able.
- Aliasing and singularities are not negligeable and must be adequately accounted for.
- Adaptive filtering configuration plays a critical role.
- If you don't understand it, read a book on statistics instead this one.

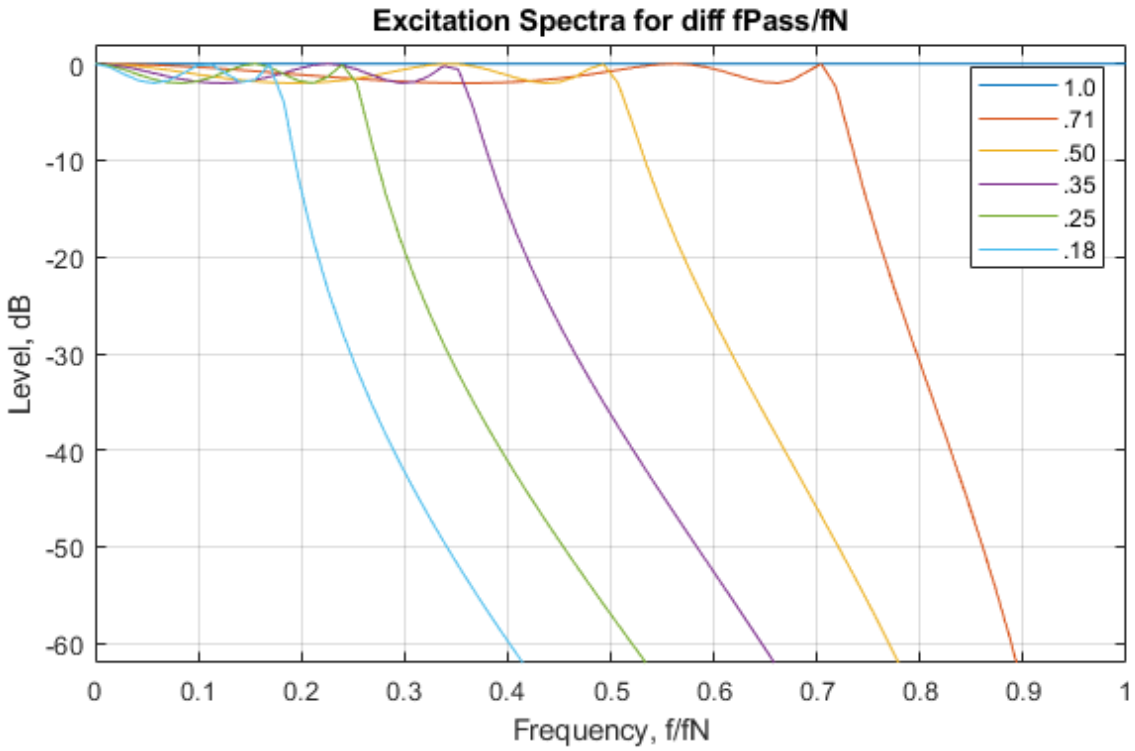
3 COLORED OR TONAL EXCITATION AND SCALAR STEP SIZE ALGORITHMS

3.1 BASICS

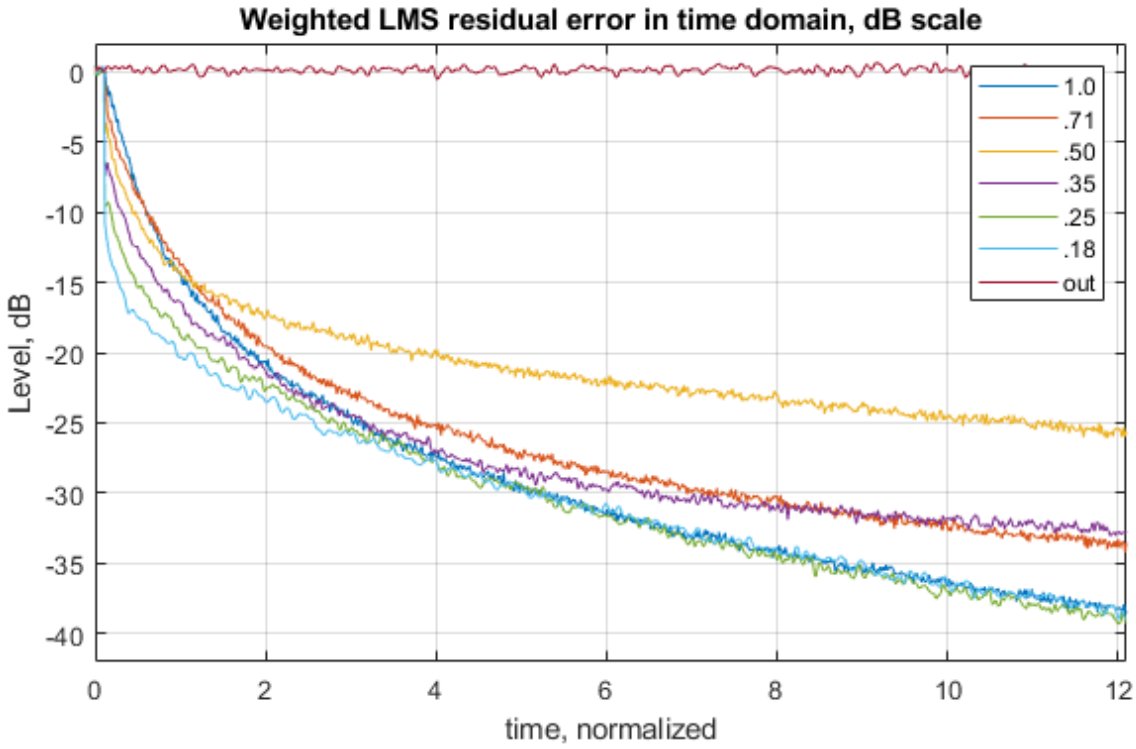
Generally, LMS convergence speed can be normalized to the L_{ADF} multiplied by the relative spectra width. For a white noise input, LMS with step size of 1.0 converges exponentially, approximately 5.2dB per L_{ADF}. I can not explain why it's ~5.2, not 4 or 6.

3.2 SUBSPACE CONVERGENCE [206]

When using speech as an excitation signal, we can reasonably assume that it is transformed to the white-ish sub-band excitation for higher frequency sub-bands. However, It's not the case for a few very first subbands. Thus, we shall see what we can expect there. Let's see a case with -48dB WLMS, AWGN=-80dB, L_{ADF}=RT_{60}, averaging over 900 runs.



The results of simulations show that the WLMS and LMS spectral deficiency effects are generally additive, except for some combinations that we need to be aware of (such as yellow curve below).

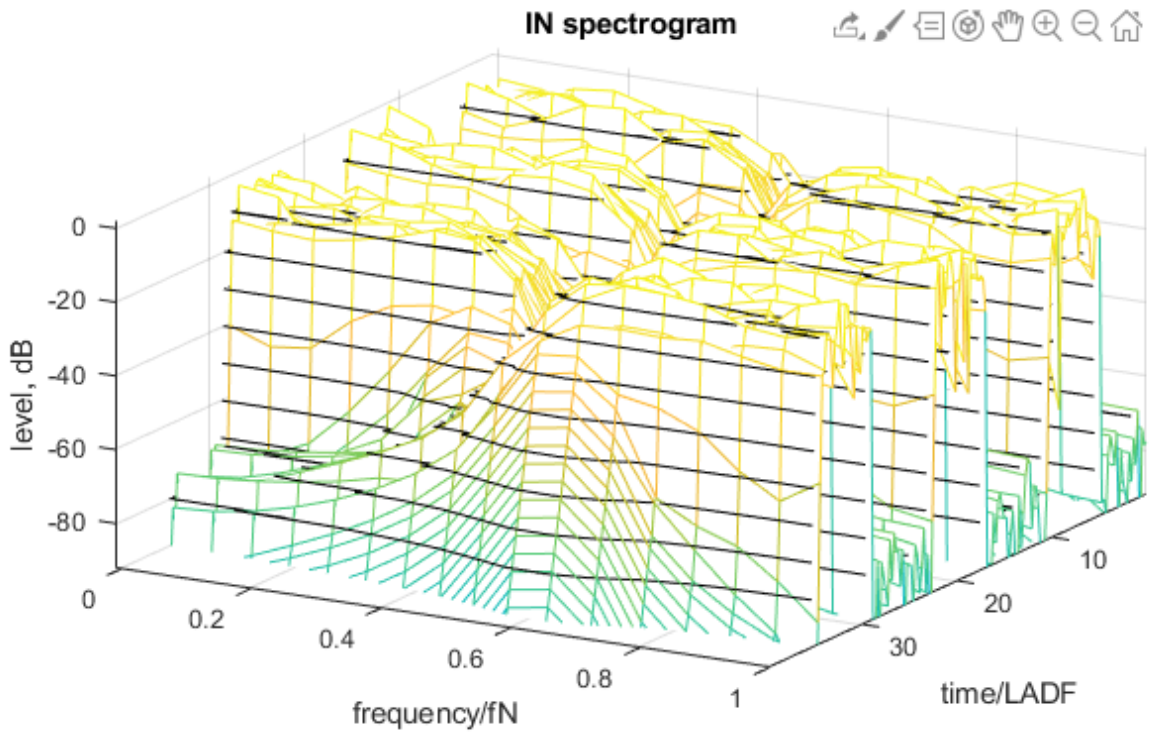


Generally, the spectral limits on excitation are interchangeable with model's length reduction i.e., if the excitation is always limited to $f_N/2$, adaptive algorithms converge twice faster, as if the model's length was halved, etc.

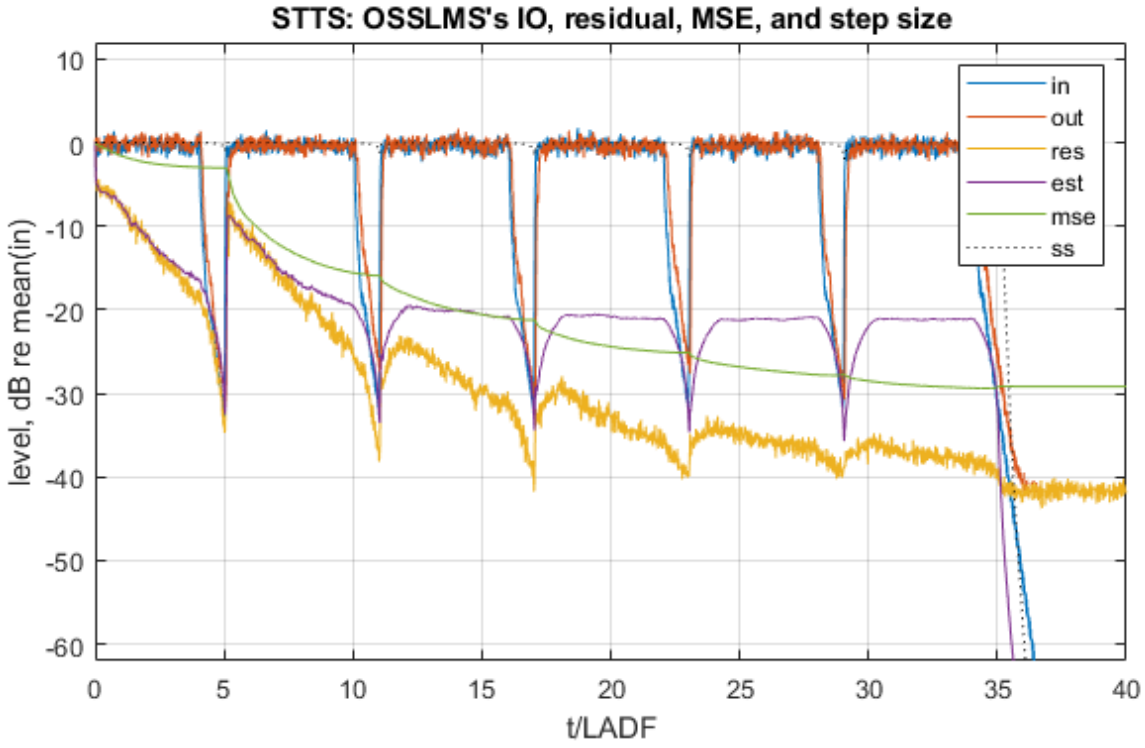
Used together, these effects pave the way to extremely high but still measurable performance of FSAF, and, alas, some of my licensees indeed put the 100dB/s convergence speed in their product specs, which is of debatable merit. To tell the truth, we can attain such high "true" acoustic (not mechanical coupling) convergence speed for usual rooms with RT_{60} of ~0.4sec on real speech only with RLS proper.

3.3 MEMORY-LESS-NESS [210]

We also need to remember about LMS/WLMS memory-less-ness, that after converging fast to a subspace defined by a certain frequency content, it will have to reconverge, if a new narrow band signal comes, and back, etc. Let's illustrate it on an example of excitation with interleaving frequency content:

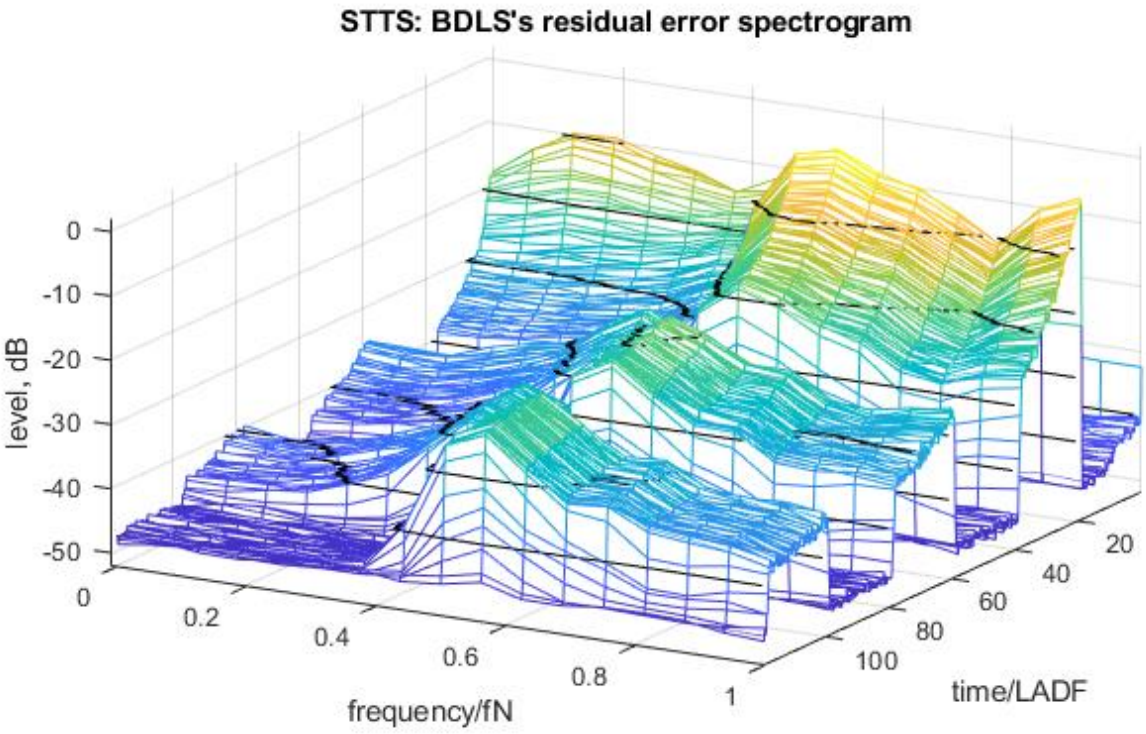


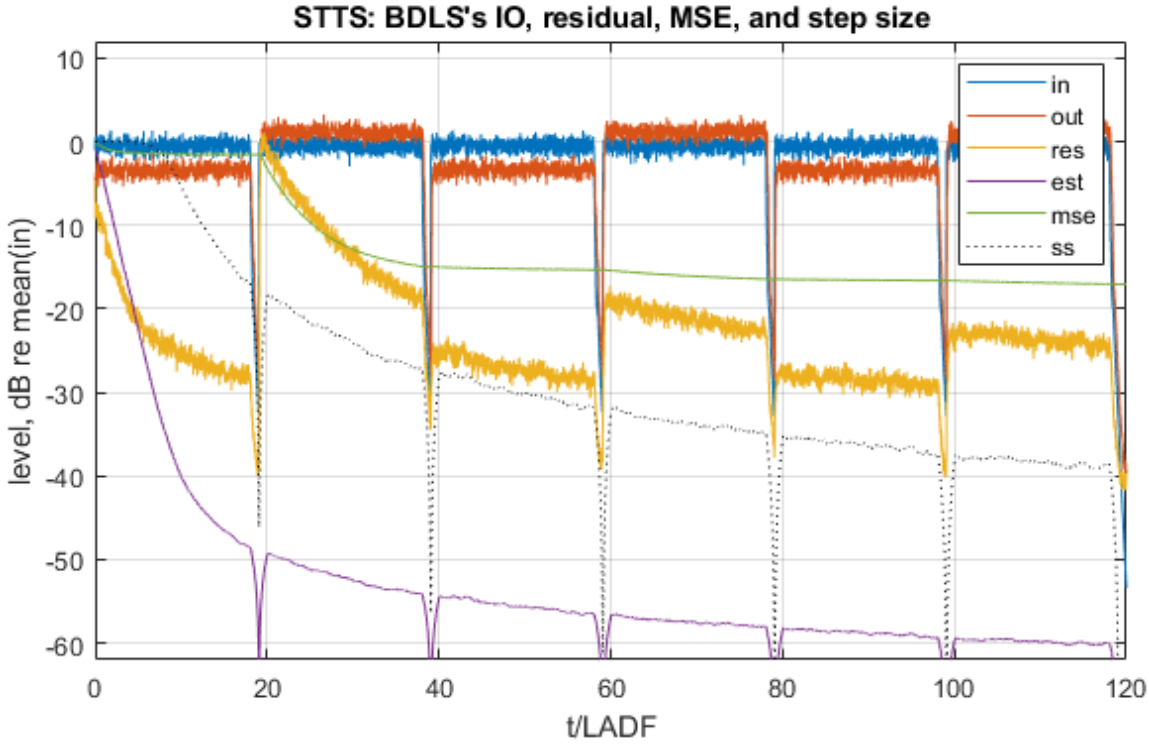
Then the convergence curves with optimal step-size LMS, which has memory, would be:



With higher noise levels, LMS would forget the previous convergence and the results would be even worse.

For simpler algorithms, like BDLS to be discussed later, which do not keep entire the dispersion matrix, the results are significantly worse:





RLS does not suffer from this effect: it keeps all necessary information in the $D_t$matrix. Various Fast RLS versions succeed to keep this information in a vector format but their proper initialization is not clear yet (I tried but failed, so far).

3.4 SUMMARY

Scalar step size algorithms are a can of worms.

4 INCORPORATING ACOUSTICS INTO STATISTICS

4.1 BASICS

An acoustic RIR consist of

- a spike of mechanical coupling between a loudspeaker and a microphone because the speed of sound in solids like plastic and wood is 10...20 times higher than in air. It could be huge in smartphones, tablets, speakerphones, and laptops and very small in the "installed sound" conference rooms.
- a spike of acoustic coupling via a direct acoustic path between a loudspeaker and a microphone
- a few spikes of first reflections of varying amplitude, depending on the geometry and cushioning of the room, lasting 20…50ms.
- an exponentially decaying reverberation "echo" tail. This one is characterized by RT_{60} - the time it takes to decrease "average" acoustic (excluding mechanical coupling) IR by 60dB.

When acoustics are altered by people moving in/out/siting/standing, moving chairs, tables, loudspeaker(s) and microphone(s), etc - all of which can and does happen during conferences, the impact on RIR is decaying with the same RT_{60}. It's easy to verify experimentally:

- turn loudspeaker (and whatever else needed) on,
- start microphone recording and start playing MLS/chirp/white noise,
- Exit the room and close the door
- Wait for a minute
- Enter the room:
  - move a chair 20cm away (or something like that)
  - Exit the room and close the door
- Wait for a minute,
- Enter the room again,
- stop microphone recording and stop MLS/chirp/white noise playing.
- Analise RIR before and after using any off-line methods.

So, the simplest model of RIR variations is an $RT_{60}*2$ exponent (because it's energy, for least SQUARES approach). A bit more realistic model would be an exponent with a flat shelf for the first 20...50ms, followed by the same $RT_{60}*2$ exponent[1]. Also, it's important to be acutely aware of proportionally higher temporal RIR variability on higher frequencies.

In real-world applications involving calculation of a RIR, the applicable mathematical apparatus of acoustics appears to be an unnecessary over-complication, a bad husband material for much simpler mathematical apparatus of statistics. At this point, usually, the entire domain of physics is thrown away, a well-lit room is seen as a black box, and RIR is considered as an arbitrary FIR to be identified by statistical means, with standard Least Squares or something else.

However, we'll see that the "bad husband material" appearances of physics are deceitful.

[1] I am not sure that there is a need to include the spikes of direct mechanical and/or acoustical coupling in the model of RIR variations because the adaptation to them happens only once.

4.2 REGULARISED RECURSIVE LEAST SQUARES (RERLS)

4.2.1 LS Basics

$$x_t{}^H h = y_t, \text{ for } t=1:T$$

$$Xh = y$$

$$X^H X h = X^H y$$

- ✓ Everything would be fine if the $X^H X$ matrix were full rank and well conditioned but, as the rule, it is not [at all], due to various reasons. Thus, a regularization comes into play:

$$h = (X^H X + \gamma I)^{-1} X^H y$$

…which is a standard LS. Let's define the physical sense of the I regularization term and γ.

We can think of y_t as microphone current, x as voltage on loudspeaker, and h as an array of conductivities separated by z^{-1} delays. So, γ should be measured in V^2. What is the physical reality that γ corresponds to? Do you have a clue? Me neither… and I'd rather avoid writing formulas whose physical meaning I don't understand.

4.2.2 Tikhonov no more

- ✓ We know quite a bit about room acoustic and audio propagation physics, and can formalize the applicable physics in many ways.
- ✓ Ignoring this physics and using γI instead would be … somewhat unproductive.
- ✓ … which isn't uncommon. Quite often, an author publishes a formulation of an idea, with less important parts simplified, the idea turns out great and becomes carved in stone as it was described initially – together with those simplifications which were not supposed to last.

Let's reconstruct an optimal estimate h_3 out of h_1 and h_2, which have dispersion matrices Φ^{-1} and D_0 correspondingly (and/or information matrices Φ and D_0^{-1}), where h_2 and D_0 corresponds to the RIR's physics. Let's use a Wiener filter[2]:

$h_3=(\Phi+D_0^{-1})^{-1}\,\Phi h_1+(\Phi+D_0^{-1})^{-1}D_0^{-1}h_2;$

h_3 information matrix $D_3^{-1} = \Phi+D_0^{-1}$.

$h_3=(\Phi+D_0^{-1})^{-1}\,\Phi h_1;$ % $h_2=0$ is a fully valid unbiased choice.

$\Phi=X^H X/\sigma^2;$ % for variable and correlated noise see R. Bellman 1961 book.

$h_1=(X^H X)^{-1}X^H y;$

$h_3 = (X^H X + \sigma^2 D_0^{-1})^{-1}X^H y;$

BTW, there is no need to see this bond through the lens of biased estimation. It's more like bounding a statistical approximation by the laws of physics (because for any v: $v'D_3v < v'\Phi^{-1}v$), or like narrowing down a general physical model by the collected site-specific statistics ($v'D_3v < v'D_0v$).

- ✓ Can we construct D_0 so that $(X^H X + \sigma^2 D_0^{-1})$ has overwhelmingly better chances to be a well conditioned, robustly invertible, productive solution? Easy.

[2] Omitting the boring formal math on $\arg(\min|_A\{\mathrm{var}(Ah_1+(I-A)h_2)\})$

The matrices X^HX and D_0 are not necessary of the same dimension. The meaningful length of RIR, which is the size of D_0, is limited by

- room type ($RT_{60} < 700ms$: usable for conferencing)
- by air's Brownian thermal noise:
  - The noise at $+20^0$C is estimated as -118 dB re 1Pa in the frequency range [2500 3500] Hz.
  - The noise spectral density grows by 6dB/octave.
  - A typical loudness of voice at 1m is about 60dB (A-weighted, re 20μPa)
  - The typical voice has about 10…15 dB dynamics and 10…15dB crest factor.
- … and other second order effects of the system itself.
- The $length(h_2) = size(D_0) = 2*RT_{60}$ is an adequate estimation for voice-centered applications.

The meaningful size of the $h_1 = size(X^HX)$ is limited by

  - Reasonably limited observation times
  - Noise: microphone, in-room, external noise leakage,
  - RIR variability due to the people's breathing and limbic-brain-controlled movements,
  - Nope, we are not interested in rooms without people inside,
  - Robustness to and acceptability of under-modelling related errors (any statements like $\Phi(t)^{-1} \to 0$ for $t \to \infty$ do not have a meaningful physical interpretation, here is Rhodos)
- The $length(h_1)$ is usually at or below RT_{60}. X^HX should be thought of as augmented to the D_0's size.

The $\sigma^2 D_0^{-1}$ may stay finite even when $\sigma^2 \to 0$, especially in the case of under-defined or singular systems of linear equations. Exploring and exploiting this regularization may bring a new classis of solutions which will be demonstrated in Part III.

4.2.3 RLS Basics

As usually, LS can be written in the recursive form.

$\sigma_{n,t}^2$ % estimation of variation of additive noise on the output, which is usually combined from microphone FET preamp (see B&K white papers) and HVAC noises.

$\sigma_{u,t}^2$ % estimation of variation of under-modeling error, which depends on RIR amplitude and L_{ADF} / RT_{60} ratio.

$\sigma_{a,t}^2$ % estimation of variation of aliasing error, which will be discussed in Part III.

$\Sigma_t^2 = \sigma_{n,t}^2 + \sigma_{u,t}^2 + \sigma_{a,t}^2$; % the total noise on the output

$D_t = E\{(h_t - h)(h_t - h)^H\}$; % dispersion matrix

$\nu_t^2 = x_t^H D_t x_t$; % a priory estimation of residual error, before the noises. This residual error variance does not depend on the RIR amplitude - directly.

$\mu_t = \nu_t^2/(\nu_t^2 + \Sigma_t^2)$; % Wiener optimal step size

$z_t = D_t x_t$; % projection vector

$h_{t+1} = h_t + \mu_t z_t (y_t - x_t^H h_t)/x_t^H z_t$; % adaptation step

$D_{t+1} = D_t - \mu_t z_t z_t^H / x_t^H z_t$; % Dispersion matrix correction due to adaptation,

which can be rewritten as

$D_{t+1} = D_t(I - \mu_t x_t z_t{}^H / x_t{}^H z_t);$

A posteriori residual error estimate is related to the a priori residual error estimate via usual information summation principle:

$$x_t{}^H D_{t+1} x_t = x_t{}^H D_t x_t - \frac{x_t{}^H D_t x_t x_t{}^H D_t x_t}{\Sigma_t{}^2 + x_t{}^H D_t x_t} = x_t{}^H D_t x_t \left(1 - \frac{x_t{}^H D_t x_t}{\Sigma_t{}^2 + x_t{}^H D_t x_t}\right) = x_t{}^H D_t x_t \frac{\Sigma_t{}^2}{\Sigma_t{}^2 + x_t{}^H D_t x_t}$$

$$\ldots = \left(\frac{1}{\Sigma_t{}^2} + \quad \frac{1}{x_t{}^H D_t x_t}\right)^{-1}$$

The expression $\frac{x_t{}^H D_t x_t}{\Sigma_t{}^2 + x_t{}^H D_t x_t}$ $\left(= \frac{x_t{}^H z_t}{\Sigma_t{}^2 + x_t{}^H z_t} \; for\ RLS\right)$ will be referred as the optimal / Wiener scalar step-size for all and any adaptive algorithm.

The step-size control in RLS is not adaptive but of program control type. After the assumptions on the noise variance $\sigma_n{}^2$ and initial dispersion matrix D_0 are done, D_t calculations ignore observations of the system's output y_t completely. D_t calculations depend ONLY on the excitation.

4.2.4 RLS Initialization

Consider the case where D_0 is a scaled unity matrix which has the radius larger or equal to λ_{max} - the largest eigenvalue of true D , (which we may know from physicists sitting next door) so that for any vector *v, v'Dv <= v'D₀v.*

- If the excitation is a step function, zeros for *t<=0*, and constant unity amplitude for *t>0*, then x_t, *t=1:L_{ADF}* form an orthogonal basis. Then the process of convergence consists of consecutive replacing λ_{max} with Σ^2, so that $tr(D_t) < \lambda_{max}(L_{ADF}-t)+t\Sigma^2$; and the a-priory residual error estimate follows the same rule.
- If the excitation x_t, *t=1:L_{ADF}* is a realization of Gaussian white noise, which is statistically delta-correlated and statistically of the same uniform amplitude, then, on average, the convergence follows the same rule if $\Sigma_t{}^2 \ll \sigma_{x^2}$ (more precisely, if non-orthogonal (to previous $x_{1:t-1}$) part of x_t has sufficiently good SNR)

Consider cases where D_0 is not a scaled unity matrix, and the things become much more interesting.

- ✓ RLS, in a form of oversimplified Kalman filter, already has everything "in" to incorporate the theory of acoustics in the algorithm, assuming RT_{60} is known with sufficient precision already
- ✓ RLS [re]initialization time[s] is[are] the perfect moment[s] for the acoustics incursion.
- ✓ Let's define $d0 = 10^{-6\tau/RT60}$; where $\tau = [0\ 1\ 2 \ldots LADF - 1]/FS$;
- ✓ $D_o = diag(d0)/sum(d0)$; % plus scaling when needed because the microphone signal after ADC shall be about or less than the speaker signal before DAC.
- ✓ Whenever RIR change is suspected, $diag(D_{t+1}) = diag\ (D_{t+1}) + \rho d0$; where ρ somehow reflects the RIR change magnitude.

Of course, multiple model adaptation is required because we can not distinguish between RIR change (of unknown amplitudes) and double talk easily, etc.

4.2.5 ReRLS time-domain kernel: exponential vs standard flat RLS [203]

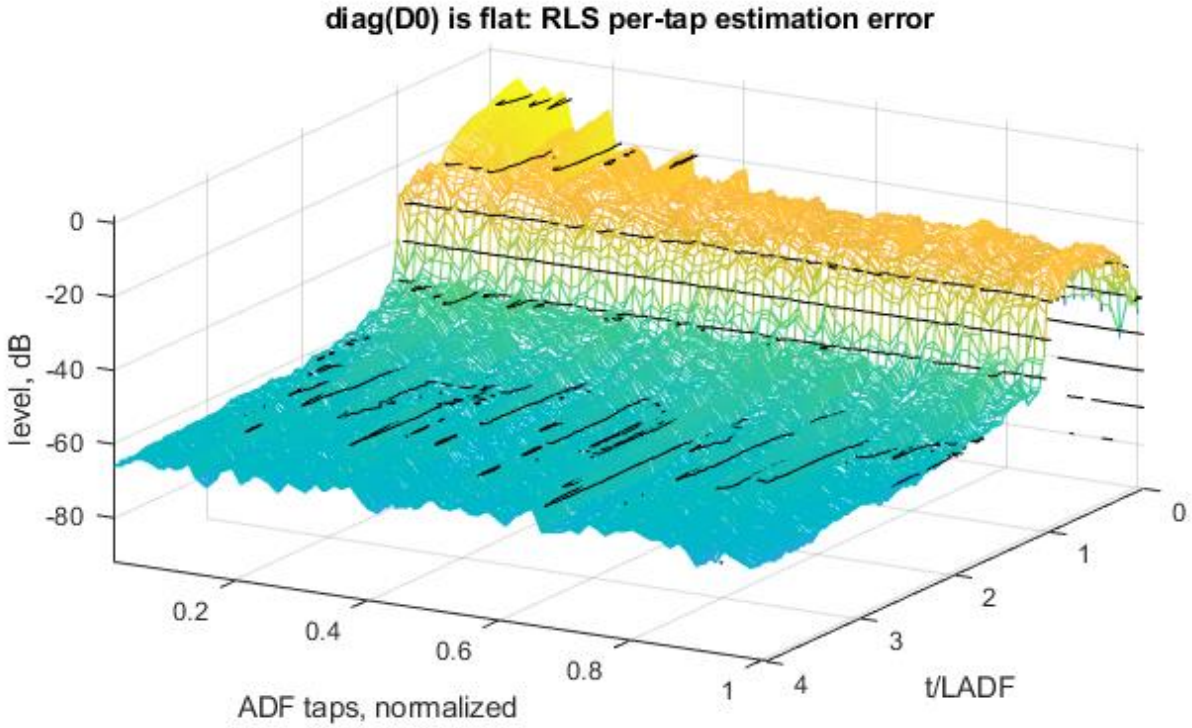
Let's use

- the simplest traditional, consecutive approach,
- a long auto-generated exponentially weighted RIR with RT_{60} slightly lower than L_{ADF},
- WGN excitation,

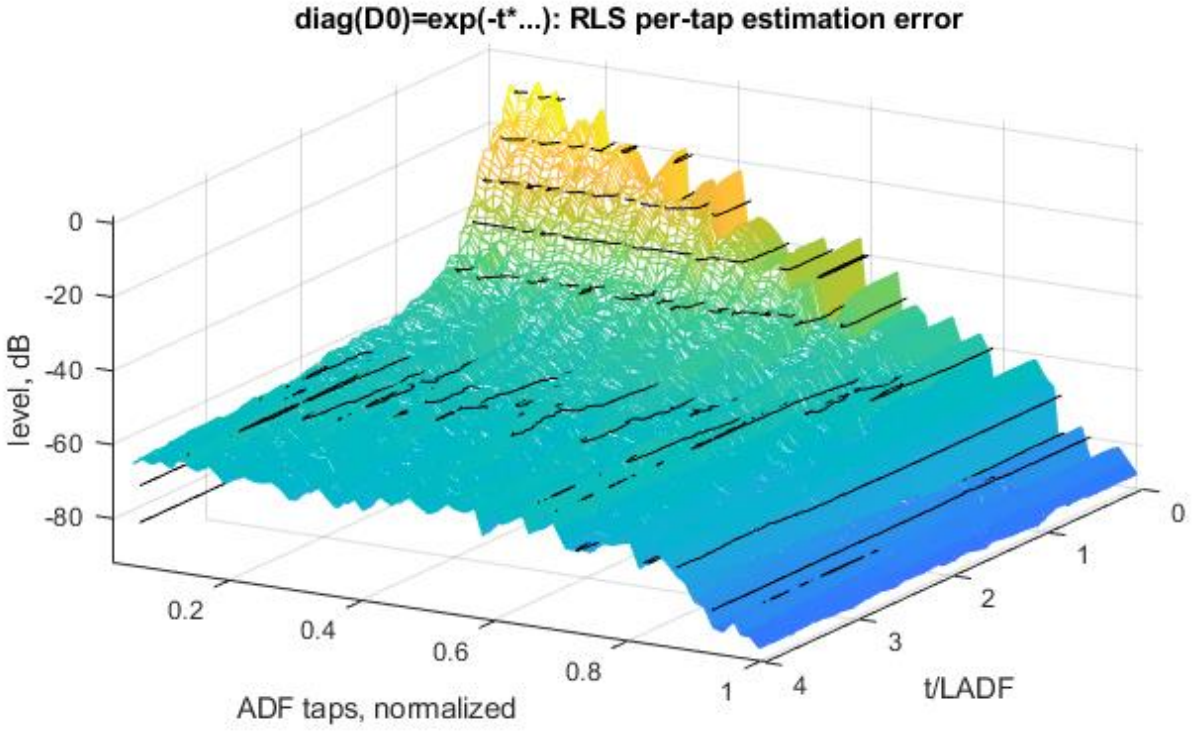
- -40dB AWGN,

…and compare two RLS filters which differ only by initial dispersion matrices:

1. $D_0=eye(L_{ADF})/L_{ADF}$; (flat), and

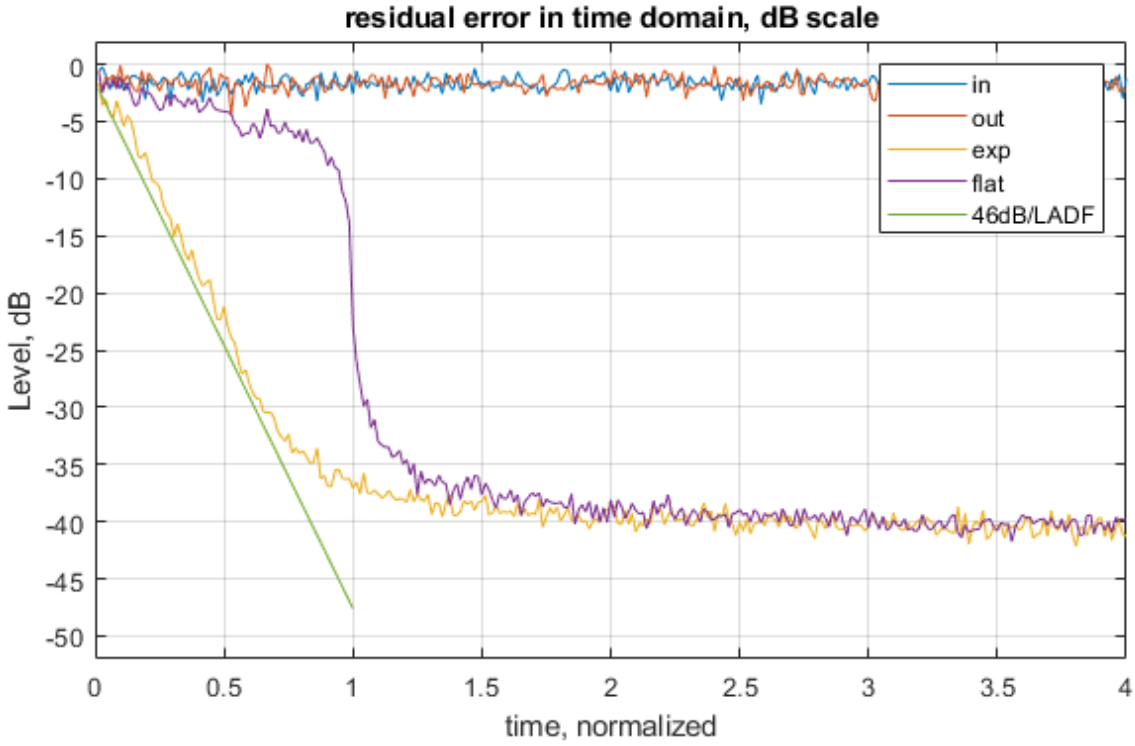


2. $D_0=diag(d_0)$; $d_0=10.^{(-t*60/(RT_{60}*10))}$; $t=(1:L_{ADF})$; and RT_{60} is assumed known well enough.

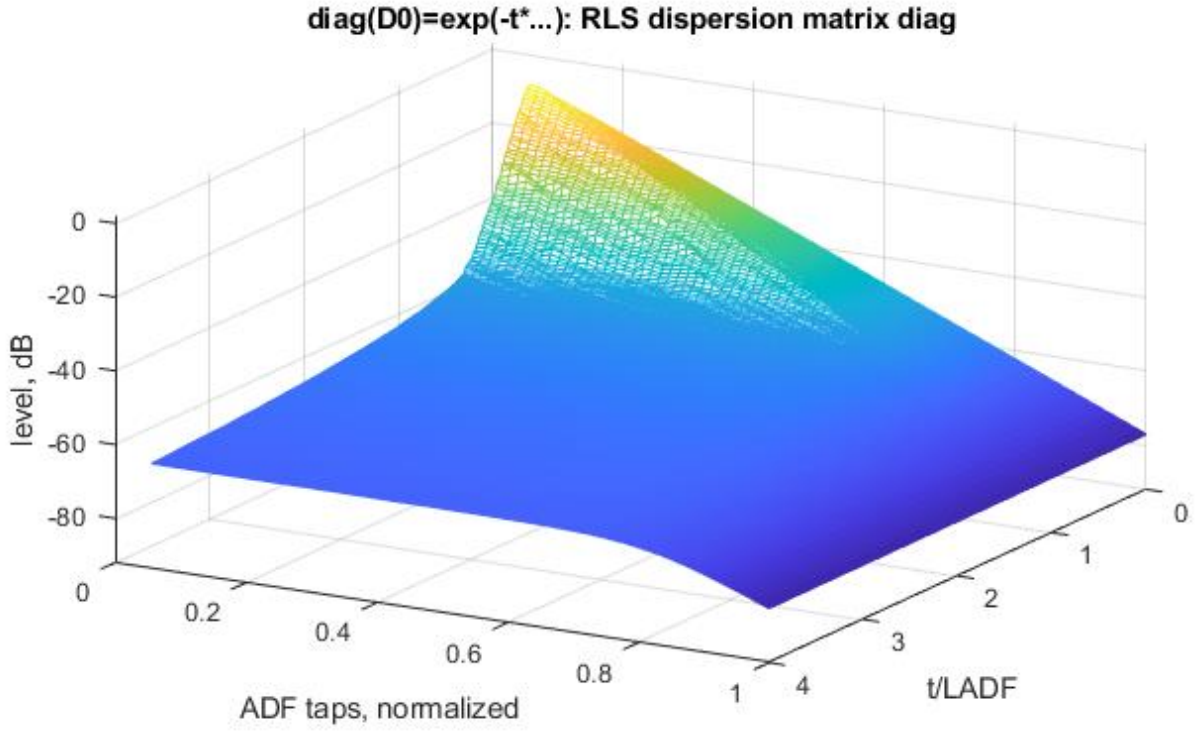


We can see the pyramidal shape of estimation errors for this weighted initialization. Note also that the estimation errors of the BLUE = flat, standard RLS at the end of RIR first go up, quite significantly (in full agreement with theory), and do not improve for quite a time.

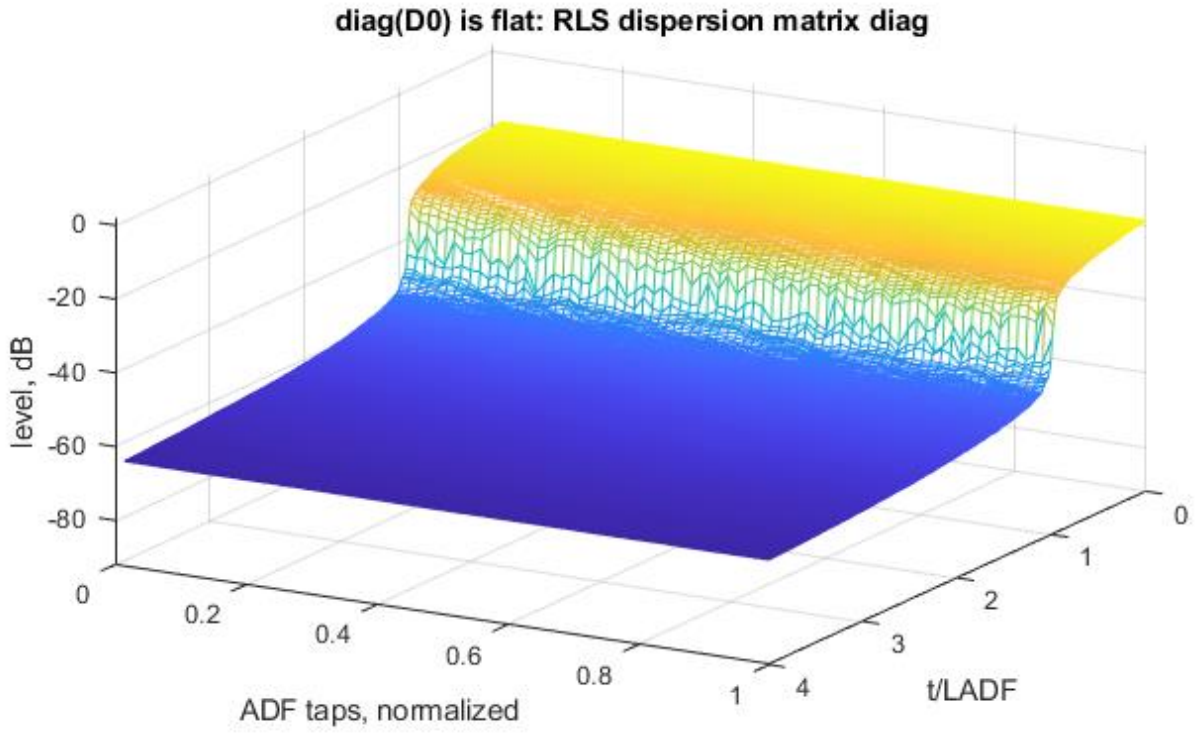
Moreover, the speed of convergence does not depend on L_{ADF} (degree of over-modelling) but on RT_{60}.



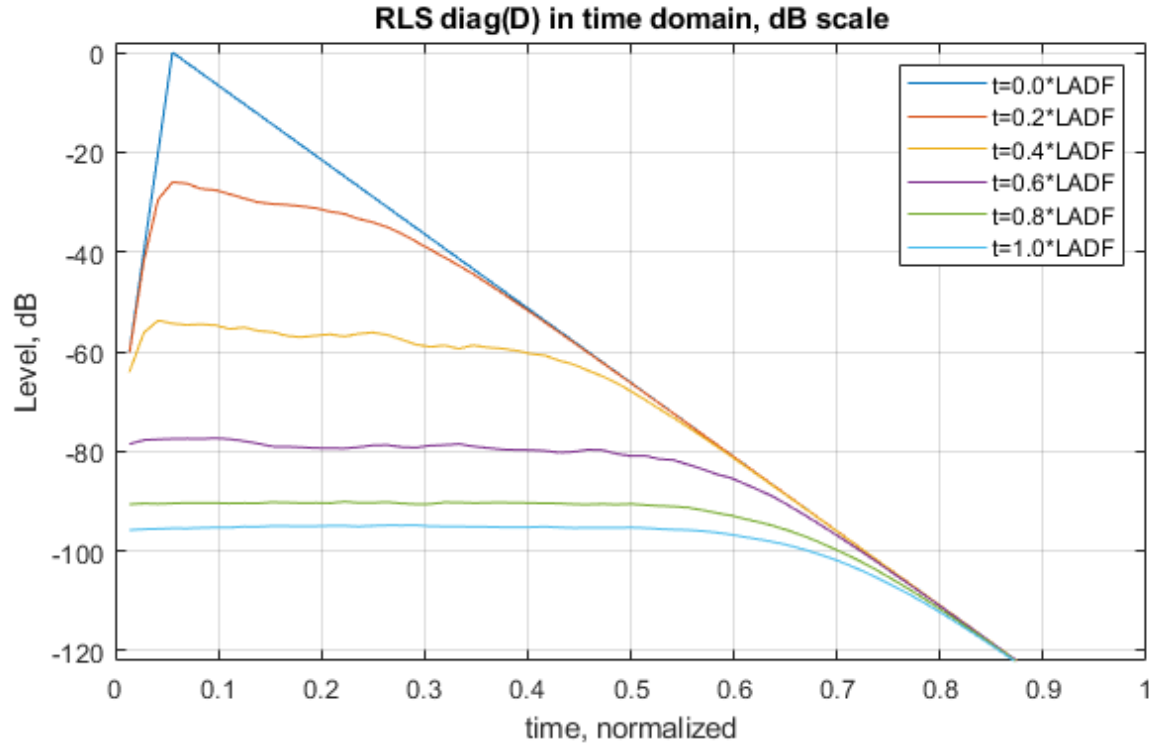
Here, on the example of [sqrt of] dispersion matrices, we can see a close fit to the observed data, clean pyramidal shape of error's dispersion,



and a monotonic shelf for flat D_0 initialization.



The close-up observation of the *diag(D_t)* allows us to see that, in full accordance with the Least Square approach, the efforts are directed to the most affected areas, in a square proportion of the relative 'damage'. And, it looks like RLS is moving a shelf down, right?



The slope of D_0 is squared relative to the RT_{60} curve because the dispersion matrices are energy. The speed of convergence is not affected by the degree of over-modeling, only by RT_{60} of the room. For typical living and conference rooms with RT_{60} of 0.4s, properly initialised RLS' convergence speed is above 100 dB/sec.

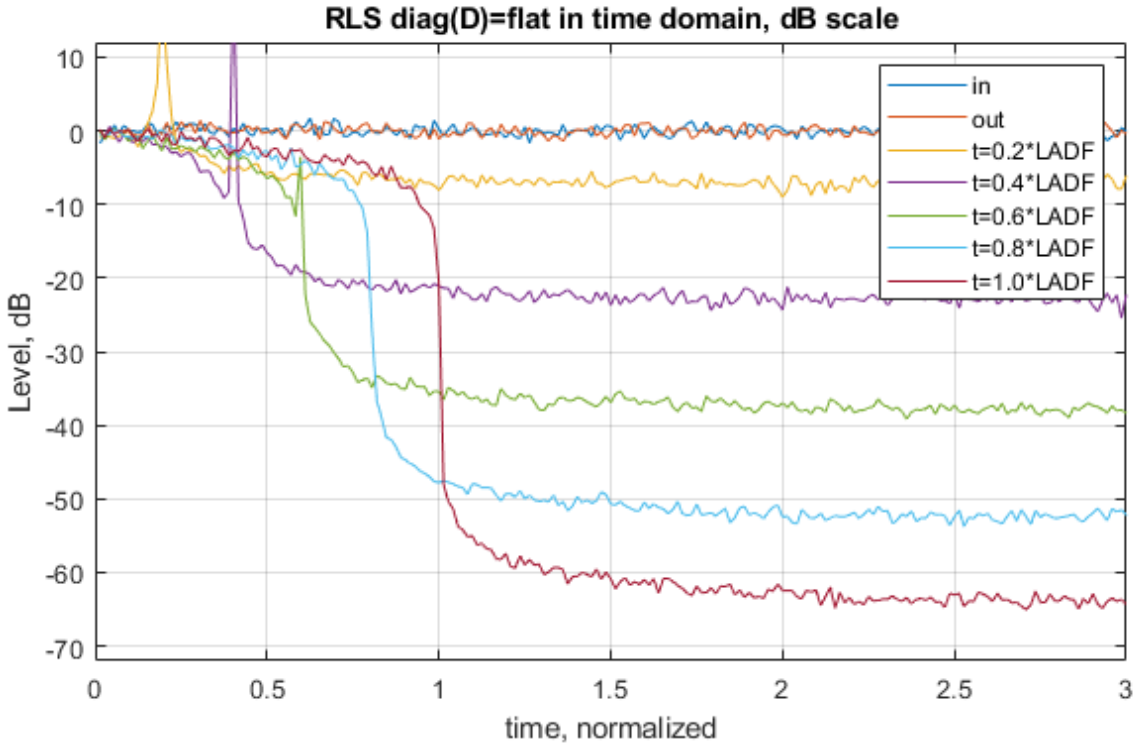
- ✓ We'll refer to such properly initialized RLS as regularised RLS, or ReRLS, to follow Prof. Lennart Ljung established terminology[3].
- ✓ RLS is a perfect bride for physics, with ReRLS as their child.

4.2.6 ReRLS vs. RLS for Under-Modeling [204]

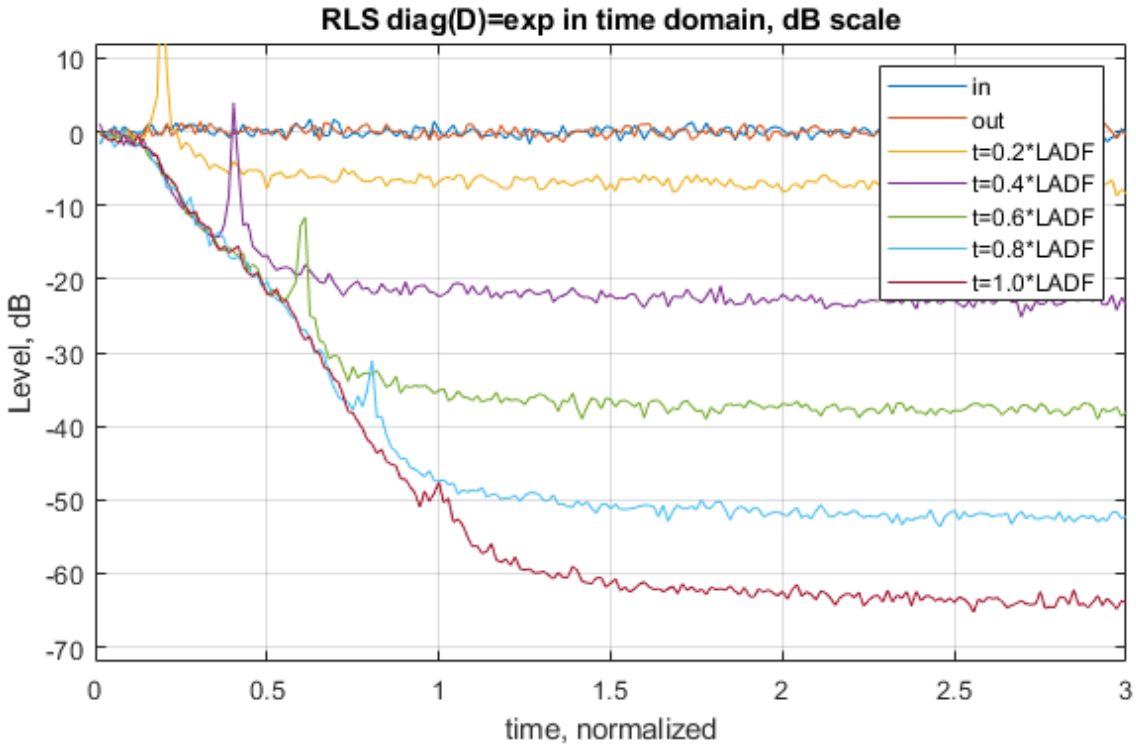
The effect of under-modeling shows the principal difference between exponential (ReRLS) and flat (RLS) initialization.

ReRLS is converging to the asymptote defined by under-modeling degree, relative to RT_{60} in a smooth uniform way, with exactly the same initial convergence speed. The flat has a shelf-like convergence property. The asymptote is the same and the step down happens as soon as the Φ_t matrix becomes full rank. AWGN = -80dB, RT_{60} = 0.8*L_{ADF}max, L_{ADF}=(0.2:0.2:1)*$L_{ADF}max$, time normalized to $L_{ADF}max$.

The spikes at the time when Φ_t becomes full-rank are due to the unintentional lack of under-modeling control in a standard RLS.



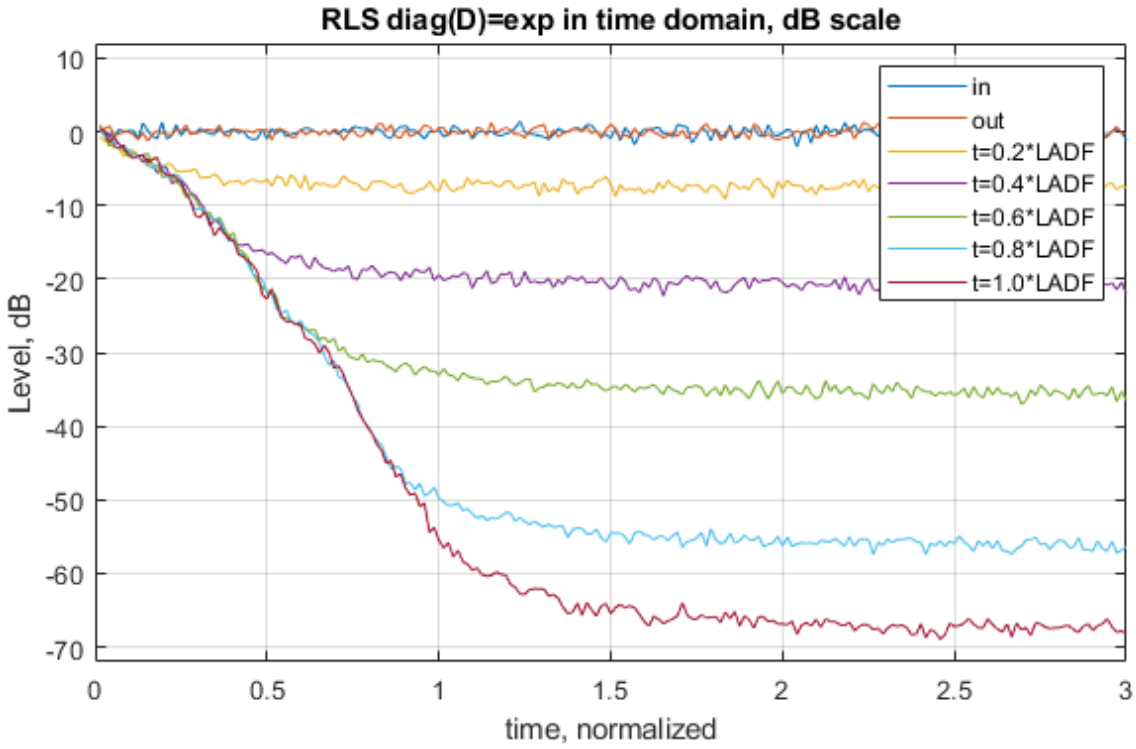
[3] In this version, graphs may still erroneously display "WRLS" or "RLS" in both cases



It is important to notice the merging of convergence curves into one curve, which conveys that for a physically meaningful RIRs, $sum(\lambda)/\lambda_{max} = tr(D_0)/\lambda_{max} < L_{IR}$ for any L_{ADF}, where L_{IR} is a property of RIR, a finite number, a kind of true, eigen size of RIR.

4.2.7 ReRLS Under-Modeling control [204]

We can account for under-modelling only in the ReRLS exponential case (not the flat one). The convergence curves start to look less troublesome:



The control of under-modeling noise is done by calculated as the RT_{60}-exponentially weighted sum of inputs beyond $(1{:}L_{ADF})$, given that the weighting curve $d0$ is meaningful.

$\sigma_{um}{}^2(t+1) = a\sigma_{um}{}^2(t) + b|x_{end}(t)|^2$; where

$\sigma_{um}{}^2$ is additive noise, not accounted in the standard RLS

$$a = \frac{d0(LADF)}{d0(LADF-1)};$$

$b = d0(LADF)$;

$TF = bz^{-1}/(1-az^{-1})$

The {a,b} are the coefficients of IIR filter to calculate exponential FIR response.

$x_{end}(t)$is the last value of IN to be shifted out at this iteration.

In practice, the under-modelling error shall also cover the effects of other deviations from ideal FIR linear models such as weak nonlinearities, etc.

TBD: ReRLS robustness to the (b,a) estimation errors.

The explicit under-modelling control is unnecessary when adaptive filter is long i.e., $L_{ADF} > RT_{60}$, for typical audio SNR.

4.2.8 ReRLS frequency domain kernel [216]

The same approach can be used for the incorporating of physically meaningful frequency response (which shall be ensured to be band-limited i.e., decaying to zero when $f \to f_N$) into D_0.

We can start with a diagonal, unity matrix, and see it as a frequency response of the RIR. Then we set the diagonal elements to the values of power spectrum of *DFT(RIR)* at the frequency bins *(0:2*f_N/L_{ADF}: end),* not forgetting to reflect the values for the folded part of the spectrum *[0…f_N] ->flipud(.)->[f_N…f_s];*

To transfer this frequency domain initialization into time domain, we form an IDFT matrix $F_{idft} = F_{dft}'/L_{ADF}$, and $D_0 = F_{idft}'*D_{0f}* F_{idft}$;

Again, the ratio of the square footage under the curve of power spectrum to the total envelope's footage is the 1/gain. The rest is the same.

4.2.9 JTF[4]-ReRLS: Joint Time Frequency kernel [217]

The key word here is not "Time" or "Frequency" but "Joint" because we can "rotate" our understanding of the RIR physics from any basis onto time basis.

- ✓ To apply both time and frequency domain regularization, we need to convert frequency-domain description of the system into time-domain, [physically meaningful] minimal phase.
- ✓ We are supposed to understand that the off-diagonal elements of D_t describe the frequency response of the transformation of the input x_t into Kalman gain z_t .

If we do, the rest is a series of simple steps:

1. Approximate the a-priory knowledge of the frequency response of RIR with a transfer function $b(z^{-1})/a(z^{-1})$, order not essential.
2. Find the impulse response of *IR=impz(b,a,L_{ADF});*
3. Form the filtering matrix F_{rir} as a simple equivalent of *filter(b,a,x)* operation when the IR is moving along stationary *x,* instead of usual visualization of input *x* moving along stationary filter.

IR(1:L_{ADF})						
0	*IR(1:L_{ADF}-1)*					
0	*0*	*IR(1:L_{ADF}-2)*				
…						
0	*0*	*0*	*0*	…	*IR(1)*	*IR(2)*
0	*0*	*0*	*0*	…	*0*	*IR(1)*

4. Form T_{rir} as a diagonal matrix, with exponentially decaying γ^t *(0<γ<1)* envelope of RIR (not squared yet!).
5. Finally: $D_0 = F_{rir}'*T_{rir}'*\sigma_v^2*I*T_{rir}*F_{rir}$;

Think of it as: to form a RIR=*h*,

- generate a random vector *v=randn(L_{rir},1);*
- condition it: T_{rir} **v*
- *filter(b,a)*: $h = F_{rir}*T_{rir}*v$;

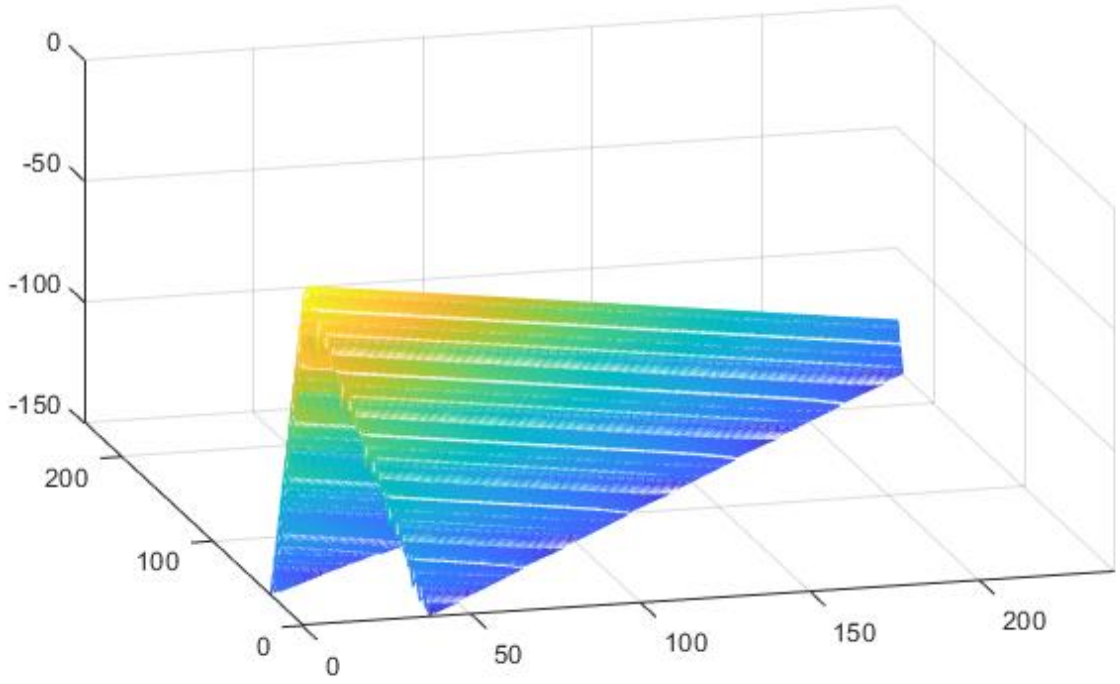
To form the corresponding $D_0 = E\{h*h'\}$, we would have to square the initial random sequence *v*. We don't know it and thus replace these squares, *v.^2* with their mathematical expectations σ_v^2 as our best guess.

- ✓ The off-diagonal elements [usually] decay faster than on-diagonal, and the very top become smoothed out due to application of filtering (otherwise, it's a step function with infinitely wide spectrum).

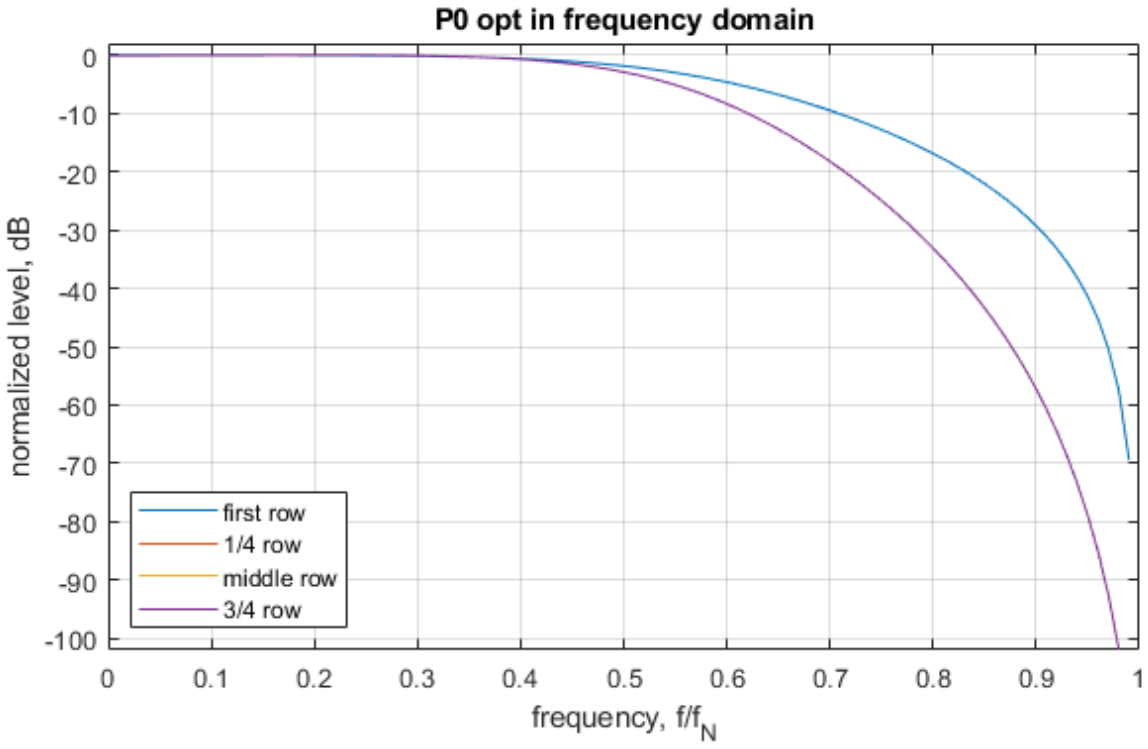
[4] The naming is temporary and will likely change in future. Prof. Lennart Ljung is de facto authority in giving names in this domain, and it seems he prefers 2-letter abbreviations.

- ✓ Slower decay of off-diagonal elements (time / column index wise) corresponds to faster decay in frequency domain, and vice versa.

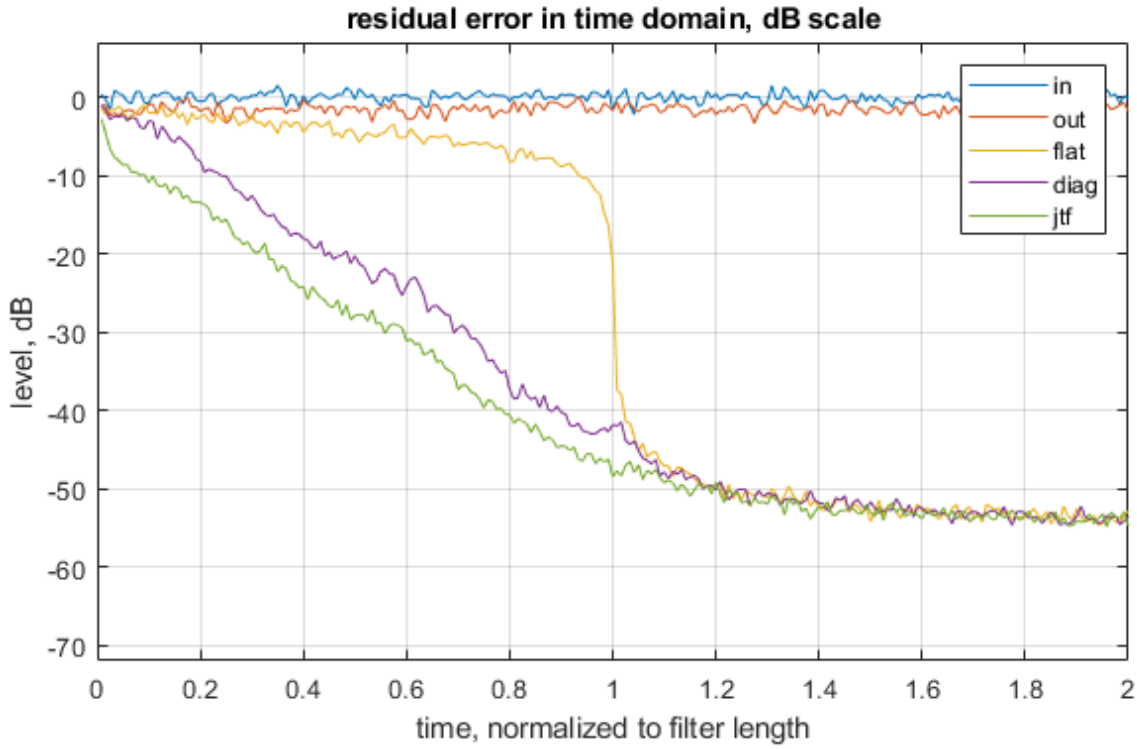
For the purpose of FSAF, we need to consider subband specific filters like QMF. An acceptable simple approximation would be [b,a]=butter(2,0.55). The result:



Now, each row of D_0 forms an FIR filter, and for the most rows the corresponding filters have identical magnitude frequency response and row-proportional group delay:

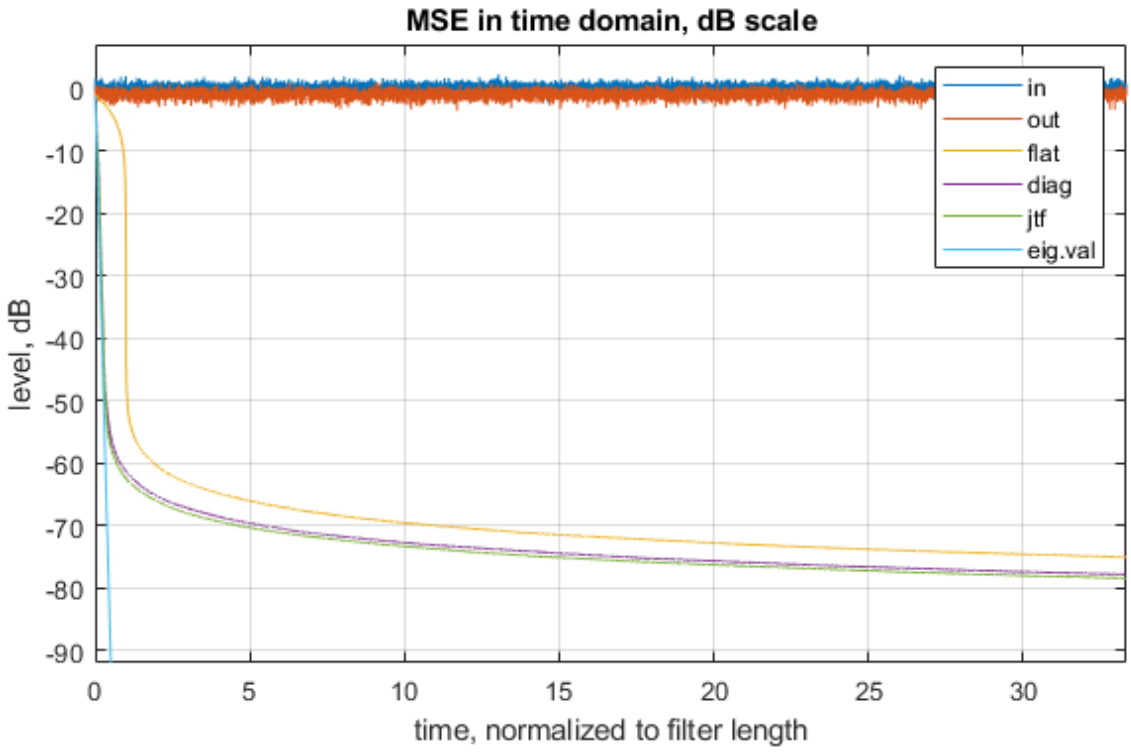


Convergence wise, the outcome, for $L_{ADF} = RT_{60}$, σ_{nse} *=-60dB re* σ_x*:*

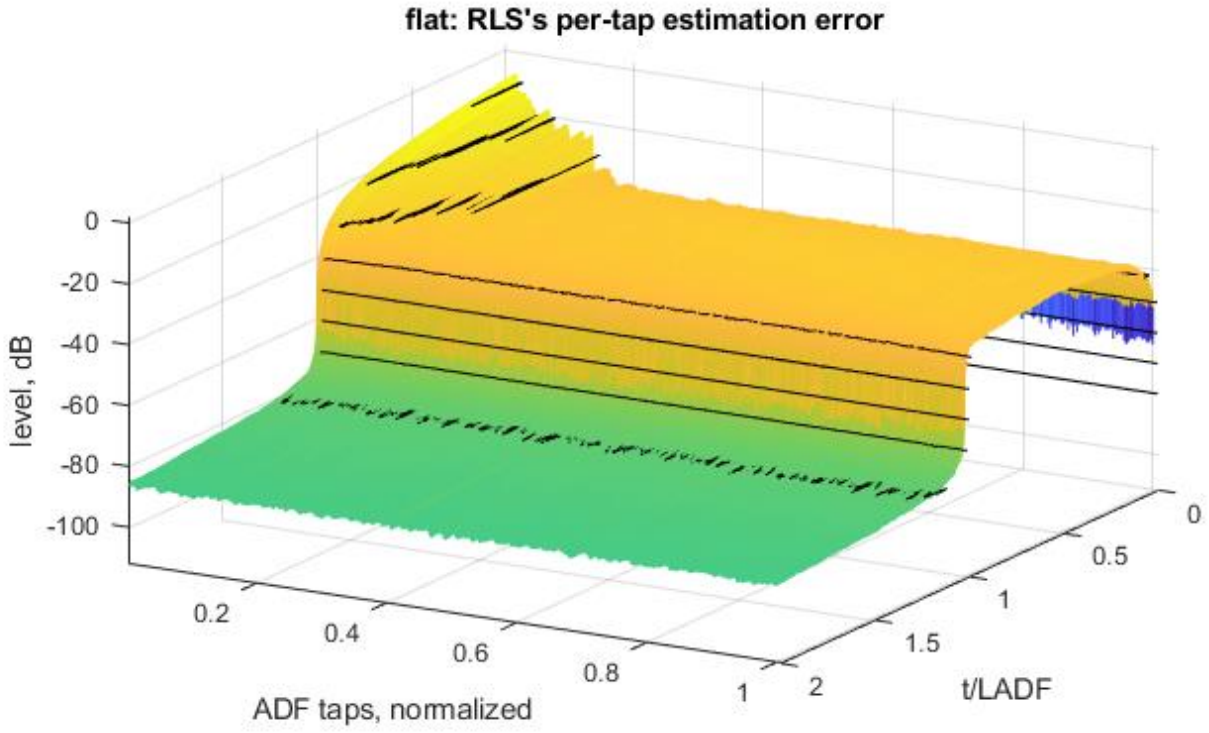


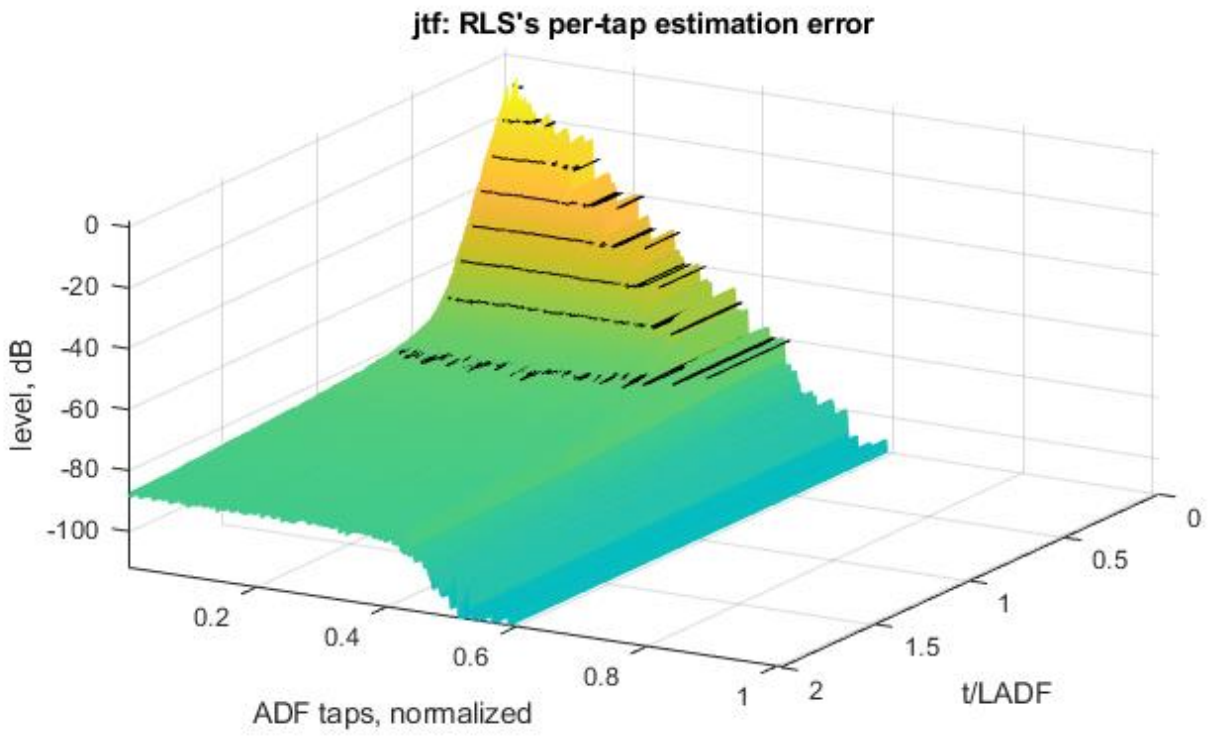
For adaptive control systems with long -20 (or -40--, after 0dB) dB/decade slopes the effect shall be much more pronounced. Moreover, for long adaptive filters ($L_{ADF} > 1.5*RT_{60}$), the effect of JTF initialization is the largest and it remains for a very long time (for standard flat RLS, each 3dB of MSE improvement beyond

SNR require doubling of observation time), being the meaningful gain of a biased estimator over unbiased:



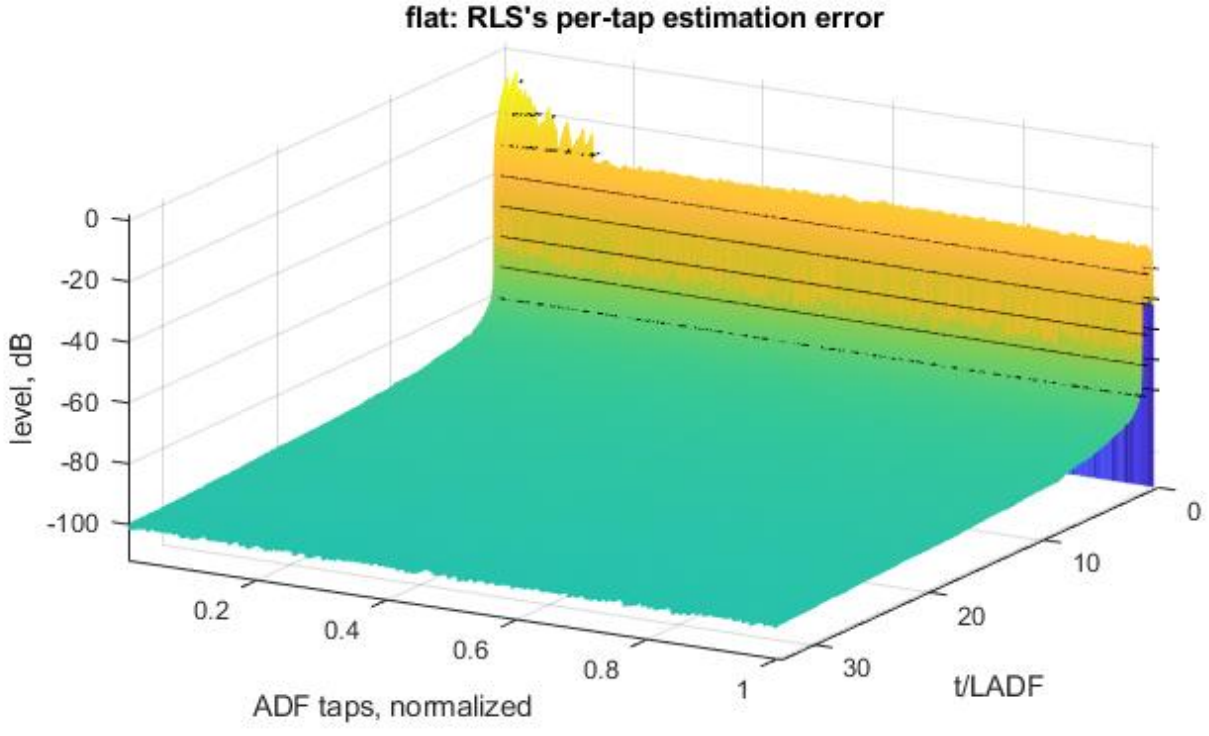
The internal mechanics become obvious after looking into the details of convergence. In the beginning:

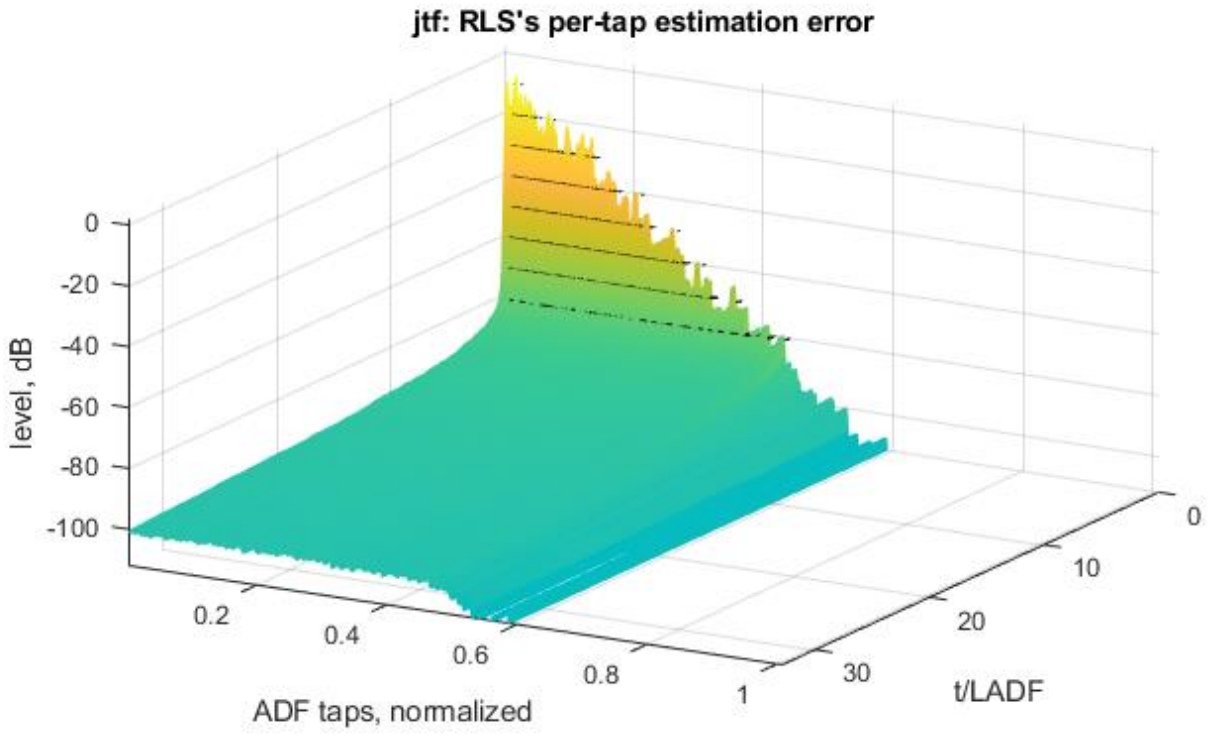




The frequency-depending effect is essentially the same as described in 3.2 “Subspace convergence [206]” because D_0 and Φ^{-1} are interchengeable.

…and at the long run:





It would take literally forever (if ever, accounting for the unaccounted tail) till flat RLS catches up to the proper JTF-ReRLS.

4.2.10 ReRLS and eigen spectrum

Now it's a good time to guess that this entire exercise was about applying our understanding of the concept of eigen-spectrum and eigen space to a real-world problem. Obviously, the eigen spectrum of the resulting D_0 is a multiplication of power spectrums of T_{rir} and F_{rir}. It's a little less obvious that the

- ✓ ReRLS (and RLS) initial convergence closely follows the distribution of the eigenvalues of the D_0.

1. For stationary white noise excitation, and known level AWGN, the average process of initial convergence is fully defined by the eigen-spectrum of D_0. For any $t>0$, the kth eigen value: $\lambda_t(k) <= \lambda_0(k)$; and for any vector v, $v'D_tv <= v'D_0v$. That's how statistics become bound by physics.
2. For stationary white noise excitation, the average set of eigen vectors of D_t is the same as for D_0 because the "averaged" Fisher matrix of white noise excitation is a unity matrix.
3. For stationary white noise excitation, ReRLS squeezes the D_t ellipsoid ($D_{t+1}= Dt-\mu*z_t*z_t'$), applying the force to the currently maximal radius (almost, $z_t=D_t*x_t$), squeezing $\lambda_{t+1}(1:t)$ down to the $\lambda_0(t+1)$ (actually, slightly below it).
    a. On the first iteration, consider D_0 in the basis of eigen vectors as a diagonal matrix with sorted down eigen values $\lambda_0(1) >= \lambda_0(2) >= .... >= \lambda_0(N)$.
    b. assuming x_t as gaussian white noise, and $\sigma_{nse} << \sigma_x$, we can approximate, in "average": $\lambda_1(1) <= \lambda_0(1) - \lambda_0(1)^2*x_1(1)^2 / (\sigma_{nse}^2 + \sigma_x^2*(\lambda_0(1) + \lambda_0(2) + ... + \lambda_0(N)))$; if $\lambda_0(1) << sum(\lambda_0(1:N))$.
    c. assuming $\sigma_x^2=1$; $\lambda_1(1) <= \lambda_0(1) - \lambda_0(1)^2 / (\lambda_0(1) + \lambda_0(2) + ... + \sigma_{nse}^2)$;
    d. $\lambda_1(1) <= \lambda_0(1) - \lambda_0(1)^2/(\lambda_0(1) + \lambda_0(2))$;

e. $\lambda_1(1) <= (\lambda_0(1)^{-1} + \lambda_0(2)^{-1})^{-1}$;

f. $\lambda_1(1) <= \lambda_0(2)$;

g. As far as the norm of the Gramm-Schmidt prediction error is >> σ_{nse}, next iterations follow the same mechanics.

h. finally: $sup\{\lambda_t(1:t)\} = \lambda_0(t)$; - which is the upper limit of ReRLS performance. Actual ReRLS is faster, actually. The assumptions become broken for $t >= L_{ADF}$. The applicability is also limited by the term "average" which is, alas, non-observable and non-falsifiable.

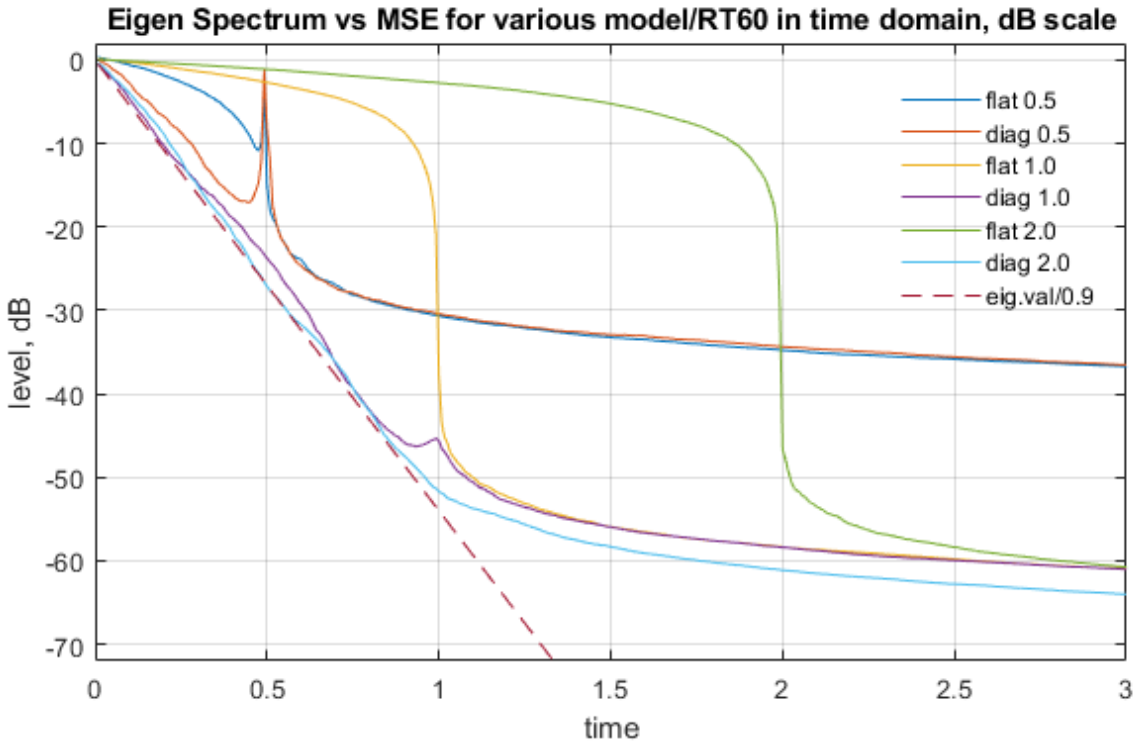
4. When the basis is rotated to the eigen-space, it does not matter if diagonal elements are sorted or unsorted. The effect depends of the ratio of square footage under power spectrum *(or* $sum(\lambda_0(1:L_{ADF}))$*)* to the enclosing $S_{xx,max}*f_N$ *(or* $\lambda_{0,max}*L_{ADF}$*).* The effect does not depend on the choice of basis vectors. However, the up-down sorting is easier to comprehend.

5. The wider is the spread of D_0 eigen-values, the faster is convergence, which is limited by the physical nature of audio waves propagation itself.

Here is the illustration of flat RLS and exponentially initialized ReRLS convergence, for model size / RT_{60} ratios of ½ , 1, and 2, and SNR of 60dB.

✓ The convergence curve decays ~10% slower than the distribution of the eigenvalues of the D_0.

I have no idea why the 10% (nor 5% nor 20%) of slowing down approximation "works" so well.



4.2.11 Summary

By now, we have seen how physics and statistics complement each other, that their marriage is made in heaven, united we stand and live happily ever after.

The Tikhonov regularization is equivalent to implicitly assuming that the variance of solution is uniformly distributed. That is rarely the case; thus, a more appropriate regularization shall be the scaled inverse of the variance εD_0^{-1}, in either full of diagonal form. It shall also be noted that ReLS is the solution to the optimal biased estimation problem (instead of [scalar] scaling of the solution vector).

It turns out that since at least 2010, there has been an independent effort in the same direction, and it is catching up fast. It was a great pleasure to find out that I am not the only one, and that "All roads lead to Rome". The overview of this alternative approach could be found in the Ljung et al (2019) "A shift in paradigm for system identification". The alternative approach is much more general but also so much more complicated that the achievements of the researchers, in my personal opinion, are a true wonder, a monument of incredible intellectual curiosity, hard work and persistence in face of devastating destructiveness of publish or perish, good enough, etc.

In my personal opinion, the key to this new paradigm is in the acknowledging of the physics' importance, in detailed understanding the meaning of dispersion matrices and eigenvalues, in understanding how to formalise the underlying physics into solvable mathematical expressions, overall, in understanding over formal knowing[5]. Audio RIR identification within adaptive filtering is just one of many examples of physically stipulated exponentially decaying IRs.

Since the version 1.3, I have started to change my naming conventions (like "properly initialised RLS", "exponential", "WRLS") to conform to the Prof. Lennart Ljung's ("ReLS", "kernel"). It may take few versions to complete the transition.

4.3 RELAXED KACZMARZ A.K.A. [N]LMS

4.3.1 Basics

The infamous algorithm was proposed in 1937 by Stefan Kaczmarz, a co-worker of Stefan Banach at KUL, for solving systems of linear equations. It is also known as "Raw Projection Algorithm". LMS, contrary to the misleading name "Least Mean Squares" has little or nothing to do with the Least Squares approach. LMS has an implicit assumption that the envelope of system response variance is uniform along the time axis (or any other axis), and thus her efforts are distributed uniformly, freely and promiscuously. Sure, there are compatible problem domains but acoustic is not one of them; it would not be a happy marriage.

Whenever the Fisher matrix is relatively well-conditioned, and the fidelity, sorry, speed of convergence, at any cost, is not required, like in CT, Kaczmarz's algorithm is well suited to the task. We shall understand the physical meaning of both explicit and implicit assumptions, clearly distinguish use cases, and have appropriate expectations.

[5] Around 2010, Signal Processing Magazine (SPM) published an article that you could greatly improve convergence of LMS by windowing the projection vector with a 7*gausswin(), if you knew that the IR was concentrated somewhere in the middle. Normalization remained unchanged. I wrote to the SPM. They replied that they didn't understand what I was talking about. I wrote to the author that all eigenvalues of *(I-μ*x*win.*x'/(x'*x))* must be <=1 for any *x*. He answered that I was talking nonsense because he tested his algorithm on randn(10000000,1) input, it worked, and that's it. That was (and still is, alas) the pathetic state of adaptive filtering in the signal processing community. I hope that unless understanding of what you are doing becomes a mandatory prerequisite to using ReLS, ReLS will remain an esoteric discipline. We saw the disastrous consequences of wide availability of advanced statistical packages, a dirty stream of "p < 0.05" pseudo-scientific publications full of complete rubbish, sick pandemic mortality predictions, etc.

4.3.2 Optimal Step Size (OSS LMS)

The optimal step size Kaczmarz was proposed in 1981 by Oleg Kulchitsky (or Kulczycky, if more appropriate Polish orthography is used) in the "A Kaczmarz - type algorithm for identification of linear objects in the presence of noise", 24 pages, LPI, VINITI No. 4170-81 Dep. 20.08.81.

$\nu_t{}^2 = x_t{}^H D_t x_t$; % estimation of residual error variance before the noises

$\mu_t = \nu_t{}^2/(\nu_t{}^2 + \Sigma_t{}^2)$; % the same Wiener step size to minimize $trace(D_{t+1})$;

$h_{t+1} = h_t + \mu_t x_t (y_t - x_t{}^H h_t)/x_t{}^H x_t$; % adaptation step

$G = I - \mu_t x_t x_t{}^H / x_t{}^H x_t$; % intermediate subspace compression matrix

$D_{t+1} = G D_t G + \mu_t{}^2 \Sigma_t{}^2 x_t x_t{}^H / x_t{}^H x_t$; % Dispersion matrix correction due to adaptation

While carrying out matrix operations in hard real-time is of somewhat excessive complexity, the value of OSS-LMS lies in being the reference for developing scalar step-size suboptimal algorithms.

4.3.3 Weighted LMS [205]

The LMS' disagreement with acoustics was noticed by many. The weighted variants of LMS blossomed, as WLMS. Let's see how and when it helps. Of course, WLMS completely misses the fact that the distribution of RIR estimation errors changes with time.

$d0 = 10^{-T/\tau}$; where $T = [0\ 1\ 2 \ldots LADF - 1]/FS$;and $\tau$is the time when the weight drops by 20dB.

$W = diag(d0)$; % vector to diagonal, time-invariant matrix

$z_t = W x_t$; % projection vector

$\nu_t{}^2 = x_t{}^H z_t$; % estimation of residual error before the noises

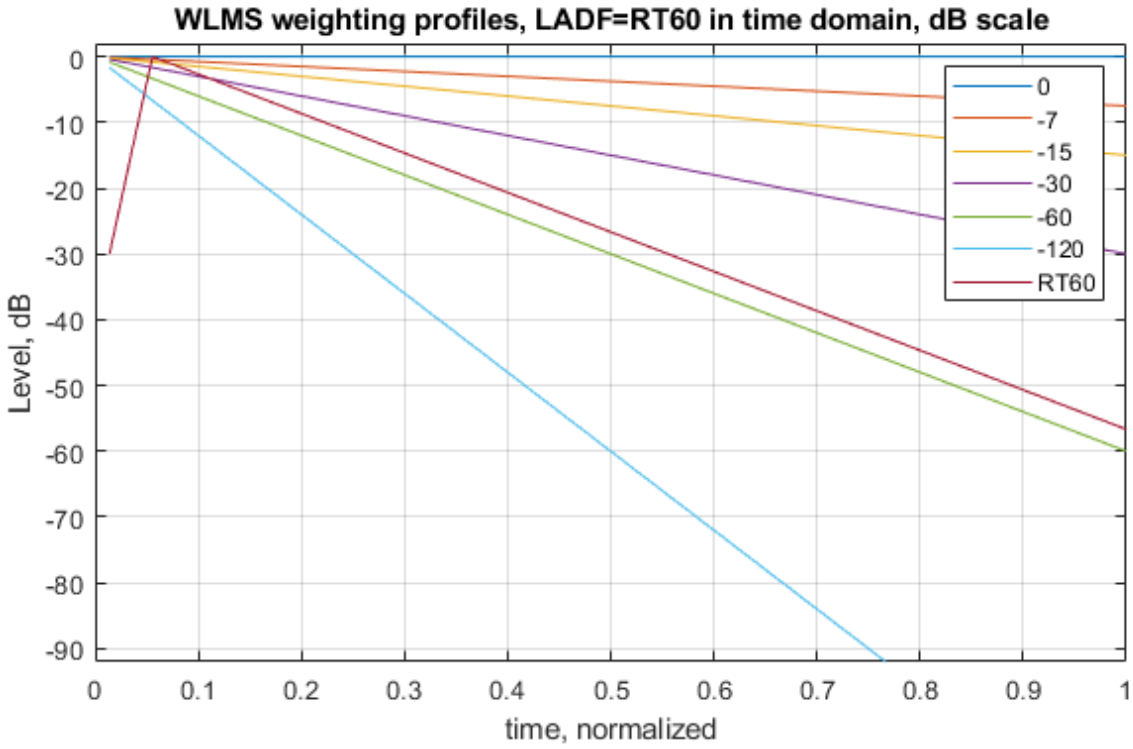
$\mu < 1$; % relaxed step size

$h_{t+1} = h_t + \mu \ \ z_t (y_t - x_t{}^H h_t)/x_t{}^H z_t$; % adaptation step

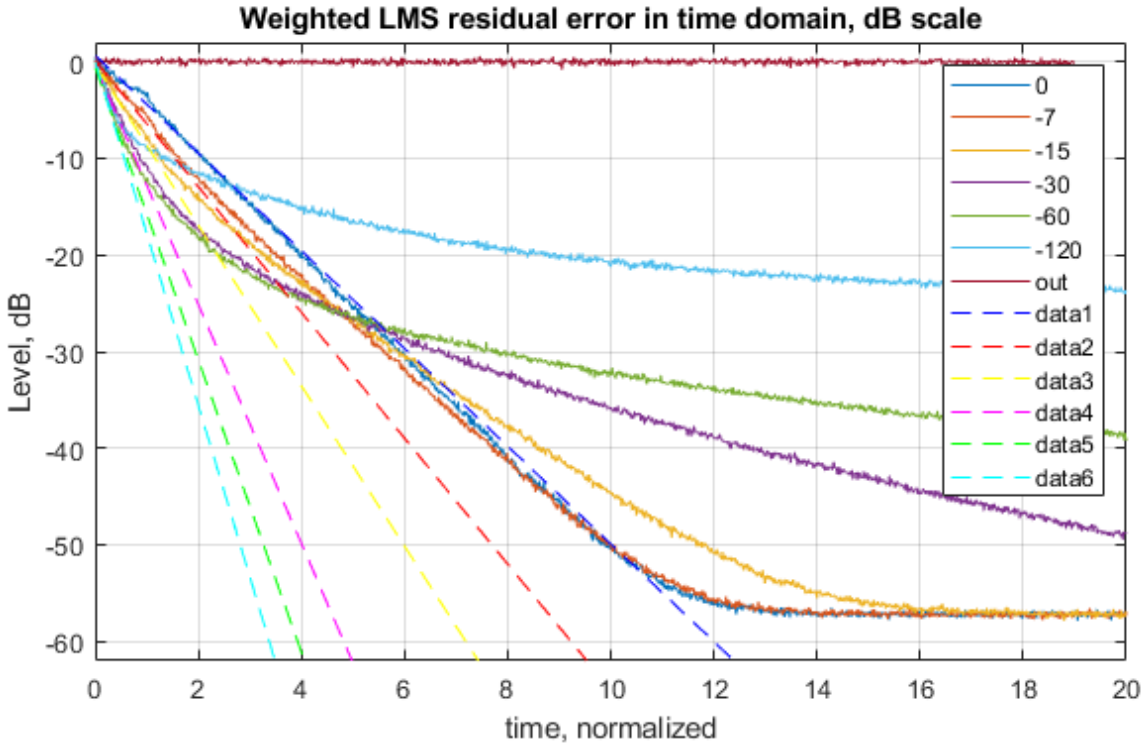
It should have been obvious that a W-modified OSS-[W]LMS had to be used for working out a suboptimal step size tuning.

Let's consider the case of a long adaptive filter, with $L_{ADF} = RT_{60}$. Such long filter can, potentially, achieve residual error level of ~ -60dB (due to unavoidable under-modelling of infinite RIR). AWGN is -80dB.

The weighing profiles are purely exponential, with varying slope, from -120 dB at the end (as for ReRLS at the t=0) to flat.



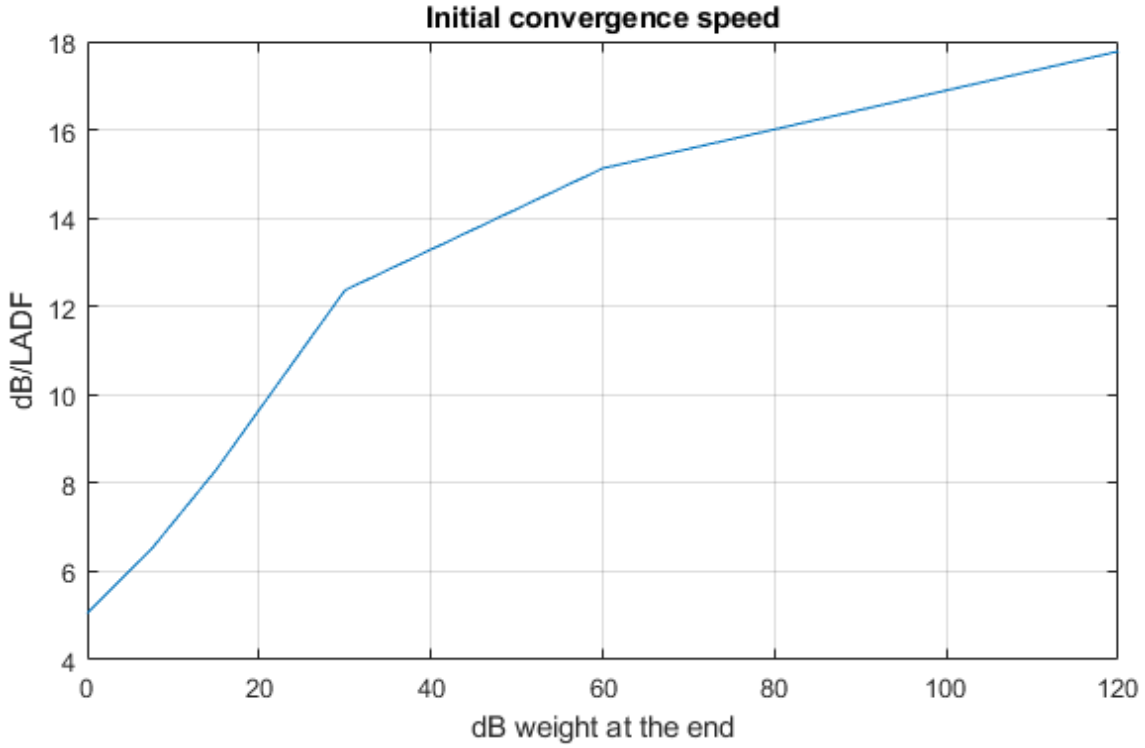
The results are expectedly somewhat similar to the illustrations of LMS spectral deficiency, for obvious reasons. Fast initial convergence is traded off with [very] slow, if any, late fidelity, sorry, convergence.



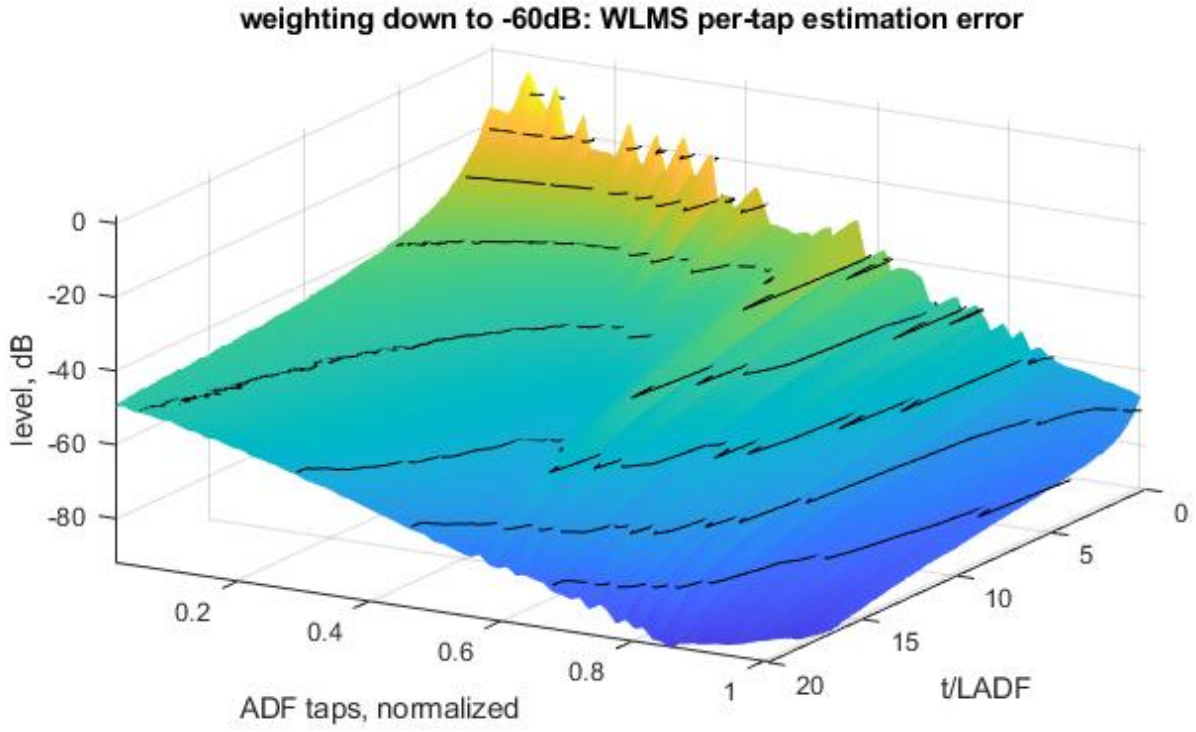
For WLMS with ReRLS-like weighting, the speed of convergence:

- For the very first 2-3 dB, is about the same as for reference ReRLS.
- For the following 3-10dB, it drops to about 15...20 dB/RT_{60}

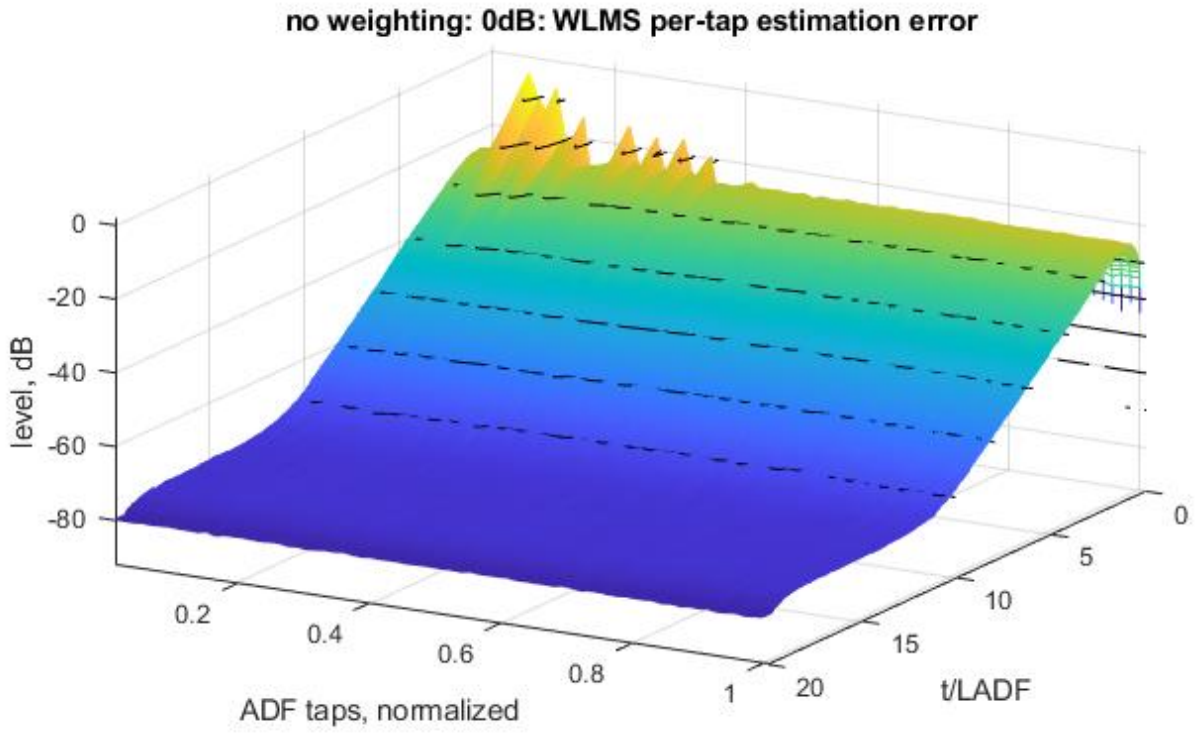
- Then it drops well below flat LMS's convergence speed
- Then it stops before achieving even 30dB, alas
- Note that flat LMS convergence speed is ~5.0...5.2 dB/L_{ADF}.
- Note that ReRLS convergence speed is about 45...55dB/RT_{60} (for same 60dB SNR).



The -20...-40dB WLMS appears quite reasonable, the robustness to RIR RT_{60} variations is ok but the limits of convergence in the beginning of RIR are poor.:



relative to the flat, unweighted LMS is opposite, with poor initial speed of convergence, etc:



4.3.4 Multiple Model WLMS

Obviously, ideas of gear-switching, or running multiple models in parallel, jump to mind immediately when we see convergence curves of WLMS with different weighting. The discussion of MM approaches can be found in Gustafsson, Part IV, Chapters 10 and 11. There is nothing of essence that I can add to his text.

4.3.5 Variable Length LMS

Variants of variable length LMS have been used for system identification with uneven distribution of response envelope. Initially, adaptation happens only in regions where the disturbances are the worst (the same least square (p=2) approach, in essence, where norm p -> inf). Then, little by little, the regions of adaptations spread and, at the end, the entire adaptive filter is used. The control of such algorithms is a bit non-trivial.

4.3.6 Summary

Hmm… sorry, [N]LMS has been the favorite of publications on the adaptive filtering. Too many bragged of solving hard problems with so simple means. Too many knew a way too little about what can of worms LMS really is. Too many simply did not test enough of corner cases before making claims of supposedly scientific value. I don't think I'd ever want to read any article containing [N]LMS in its name.

4.4 Diagonal Least Squares (DLS)

4.4.1 Basics

If LMS is so bad, are we constrained to MIPS and memory hangry ReRLS for any acoustics-related problem? Actually, no.

We can utilize the knowledge of time-domain acoustical properties of the RIR by simplifying the ReRLS to a "vector" step size adaptive algorithm, referred here as Diagonal Least Squares, or DLS.

DLS retains only the main diagonal between iterations, it stays “straight”, remembers very little of her past encounters with previous excitations, and does not collect nor keeps her baggage indefinitely. The average DLS fidelity, sorry, convergence is described by the upper limit for time-domain kernel-based ReRLS performance.

DLS is best suited for DSP implementations, which are memory constrained, or when external RAM exchange bandwidth limits applicability of proper FSAF ReRLS. The more perspective husbands are out there, sorry, the more of the narrower subbands we have, the better DLS performs.

DLS is the first choice for high-frequency husbands, sorry, subbands where excitation’s whiteness is expected. DLS should not be expected to behave ReRLS-like for either tonal, loud or highly colored excitation.

4.4.2 Naive DLS [207]

$d0 = 10^{-\tau\, RT60/3}$; where $\tau = [0\ 1\ 2 \ldots \text{LADF} - 1]/\text{fs}$; % It does not have to be a simple exponent, it may contain a diagonal from JTF-ReRLS, etc.

$d_o = diag(d0/sum(d0))$; % if needed (it may not) to normalize to 0dB IN->OUT gain, on average

$z_t = d_t . x_t$; % projection vector, d.*x

${\Sigma_t}^2 = {\sigma_{n,t}}^2 + {\sigma_{u,t}}^2 \quad + {\sigma_{a,t}}^2$; % the total noise on the output

${\nu_t}^2 = {x_t}^H z_t$; % estimation of residual error before the noise added

$\mu_t = {\nu_t}^2/({\nu_t}^2 + {\Sigma_t}^2)$; % Wiener optimal step size

$h_{t+1} = h_t + \mu_t z_t (y_t - {x_t}^H h_t)/{x_t}^H z_t$; % adaptation step

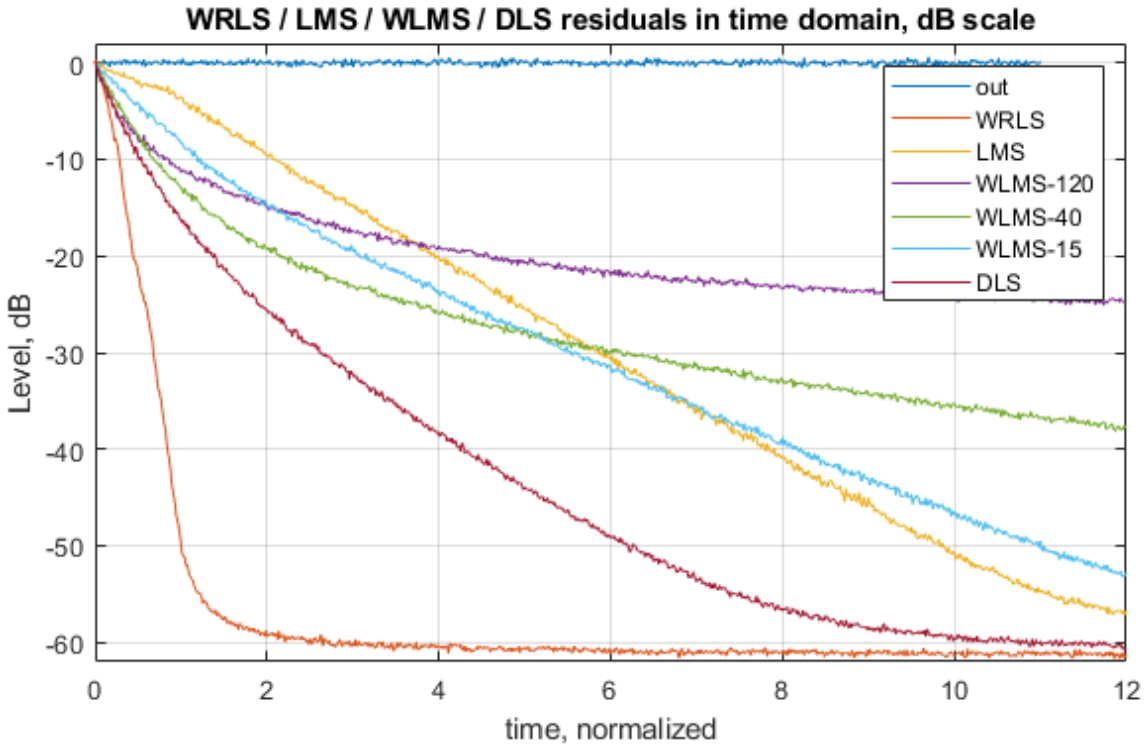
$d_{t+1} = d_t - \mu_t z_t . conj(z_t) \quad /{x_t}^H z_t$; % d=d-mu*(z.*z)/v2; or, in a clearer form

$d_{t+1} = d_t . (1_v - \mu_t x_t . d_t . conj(x_t) \quad /{x_t}^H (d_t . x_t))$; % d=d-mu*(z.*z)/v2;

$d_{t+1} = \max(d_{t+1}, d_{min})$;

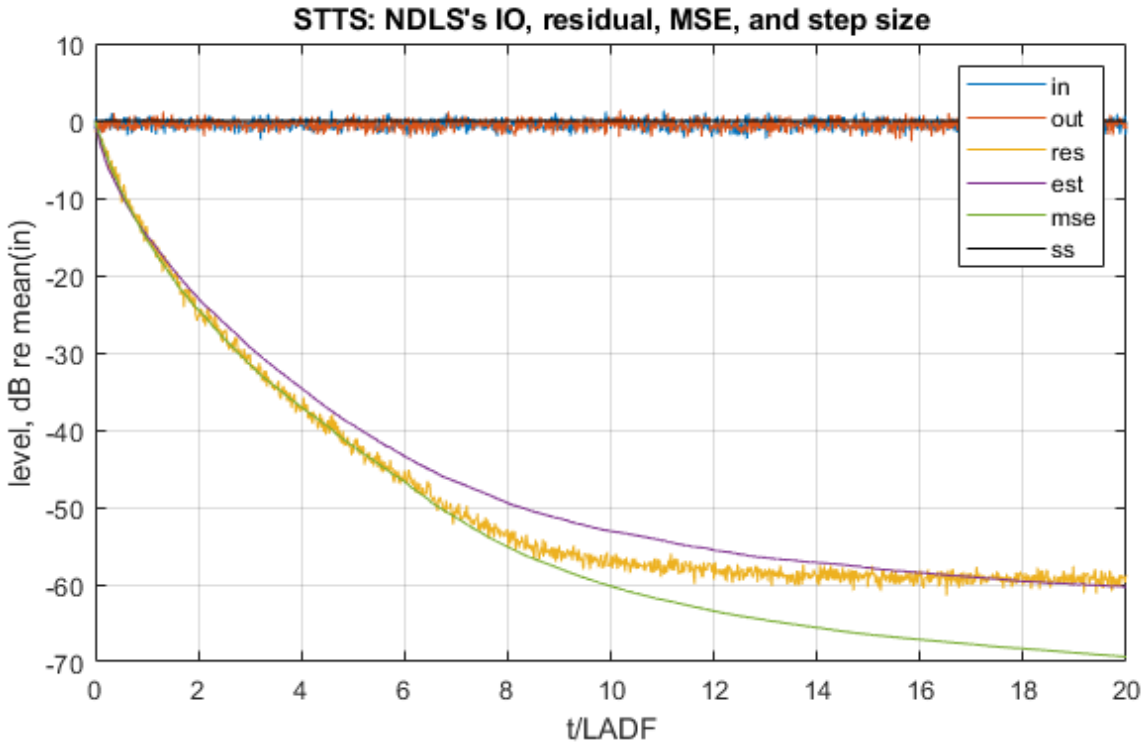
Depending on the processor and instruction set used, a straightforward implementation of NDLS shall take about 2x MIPS relative to LMS.

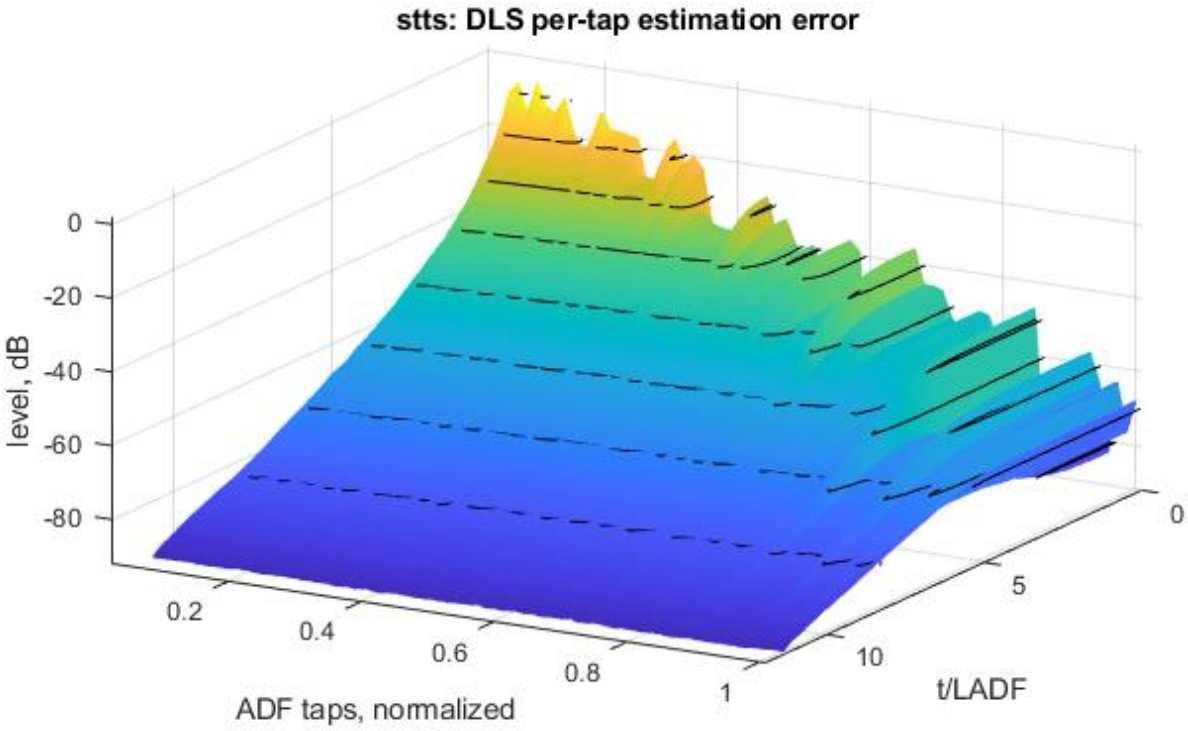
It’s easy to notice that NDLS resides between OSS LMS, WLMS, and ReRLS and in a certain sense performs WLMS gear-switching in a form of adjusting the effective adaptive filter length automatically. NDLS converges progressively slower than ReRLS: same speed in the very beginning, but to the -30dB threshold, about 5x slower than ReRLS - but 2.5x faster than the best of (W)LMS.



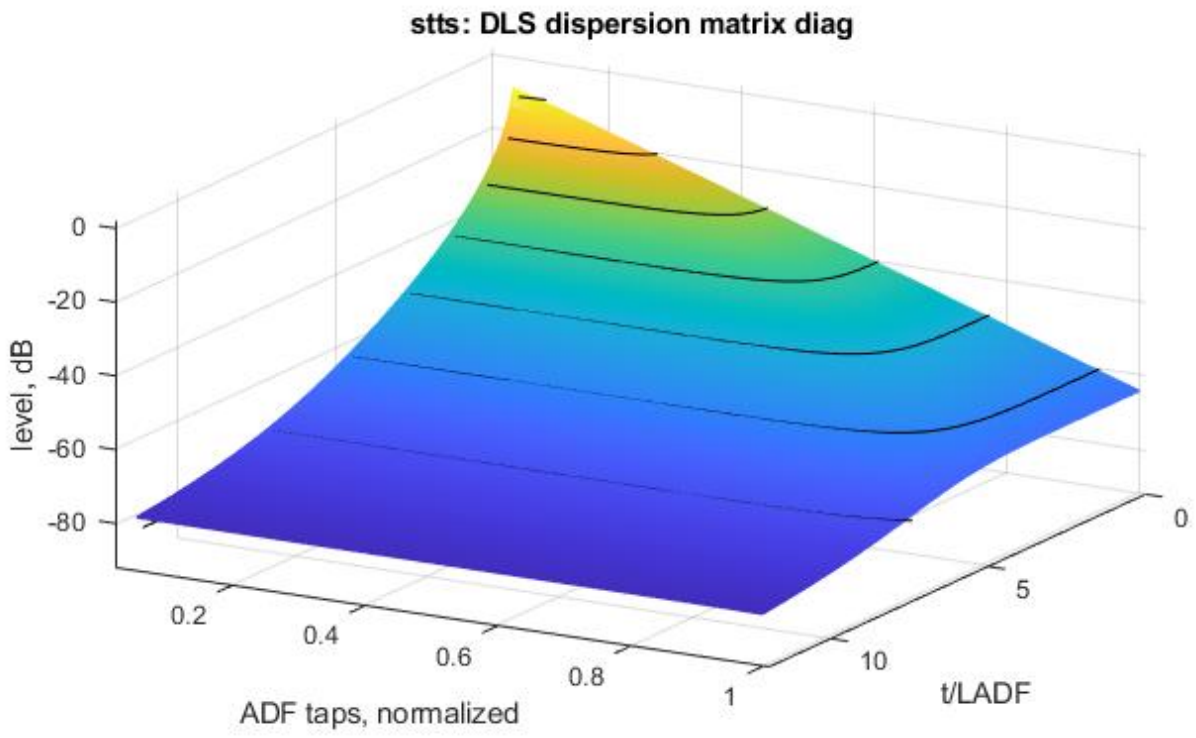
However, the main virtue of DLS lies in a completely different domain: DLS is capable of predicting the variance of residual error, and hence is capable of adjusting step-size automatically (and control post-filtering to suppress residual echo while passing low-level double talk) - what only [Re]RLS and OSS-LMS could have done, on the expense of huge MIPS and memory. This prediction is imperfect but it's much better than nothing.

The minimal value of d_{t+1} vector shall be limited by a carefully chosen d_{min}, which reflects under-modelling effect (a level related to RT_{60} and L_{ADF}, flat along time axis), and variability of the room due to real people present there, who breathe, nod, gesture etc (an exponential curve falling with RT_{60}).





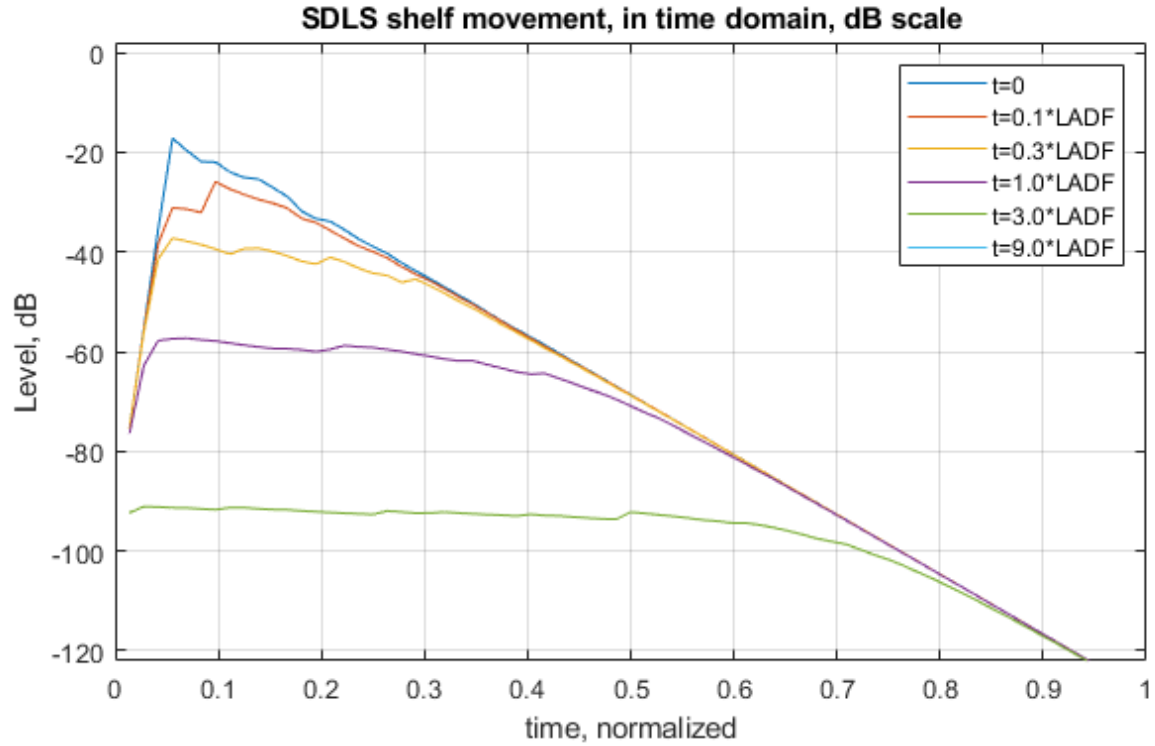
and here the estimated dispersion matrix (diag vector):



...which is the core of adaptive filtering because everything else can be derived from it.

4.4.3 Shelf DLS [208]

Shelf DLS (SDLS) models the estimation of dispersion matrix (of estimation errors) $E\{(h_t - h)(h_t - h)^H\}$ as a shelf moving [up or] down depending on the estimation of variance of residual error. The idea becomes clear if you look at the detailed picture for NDLS.



Suppose we have an initial estimation of [subband] RIR standard deviation:

$d0 = 10^{-\tau\, RT60/3}$; where $\tau = [0\ 1\ 2 \ldots LADF - 1]/FS$; but it does not have to be a simple exponent, it can contain a better description of the RIR.

$d_o = diag(d0/sum(d0))$; % if needed to normalize to 0dB IN-0>OUT gain

$s_0 = max(d_o)$;set shelf at the max value.

$d_t = min(d_o, s_t)$; % where scalar $s_t$defines a shelf moving down with convergence, and the tail of $d_t$below the shelf threshold is unaffected until the algorithm converges deep enough.

Then the same calculations applied:

$z_t = d_t . x_t$; % projection vector, d.*x

${\Sigma_t}^2 = {\sigma_{n,t}}^2 + {\sigma_{u,t}}^2 \quad + {\sigma_{a,t}}^2$; % the total noise on the output

${\nu_t}^2 = {x_t}^H z_t$; % estimation of residual error before the noise added

$\mu_t = {\nu_t}^2/({\nu_t}^2 + {\Sigma_t}^2)$; % Wiener optimal step size

$h_{t+1} = h_t + \mu_t z_t (y_t - {x_t}^H h_t)/{x_t}^H z_t$; % adaptation step

$u_t = max(shelf\ width(s_t, d_o), u_{min})$;in samples.

We do no specify how to find the shelf width and update it, it depends on $d_o$and unessential. Obviously, the convergence is inversely proportional to the shelf width, which usually grows linearly in db() domain.

$s_{t+1} = s_t . (1_v - \mu_t/u_t)$;

Computational complexity of SDLS is practically the same as for LMS because a flat shelf and an exponent can be realized with IIR filters.

4.4.4 Block DLS [209, 210]

Another quite obvious simplification is to represent RIR standard deviation as a blocked, piece-wise constant function.

- For block size c=1, BDLS becomes NDLS,
- for block size c= L_{ADF}, BDLS becomes flat LMS.
- For moderate c (like 8 or 16), computational complexity and memory of BDLS is comparable to the LMS.

In practice, the block size c would be most probably defined by SIMD vector size (like 4 for SSE, 8 for AVX, 16 for AVX2, etc).

First, sample-defined $d_o$shall be transformed into $b_o$of dimension (L_{ADF}/c), which is trivial. Of course, we assume that all blocks are of equal length, and L_{ADF} is divisible by c.

${\nu_t}^2 = {x_t}^H z_t$; % estimation of residual error before the noise added

$\eta_t = 1/({\nu_t}^2 + {\Sigma_t}^2)$; % normalizer of the Wiener optimal step size

for k=1:L_{ADF}/c

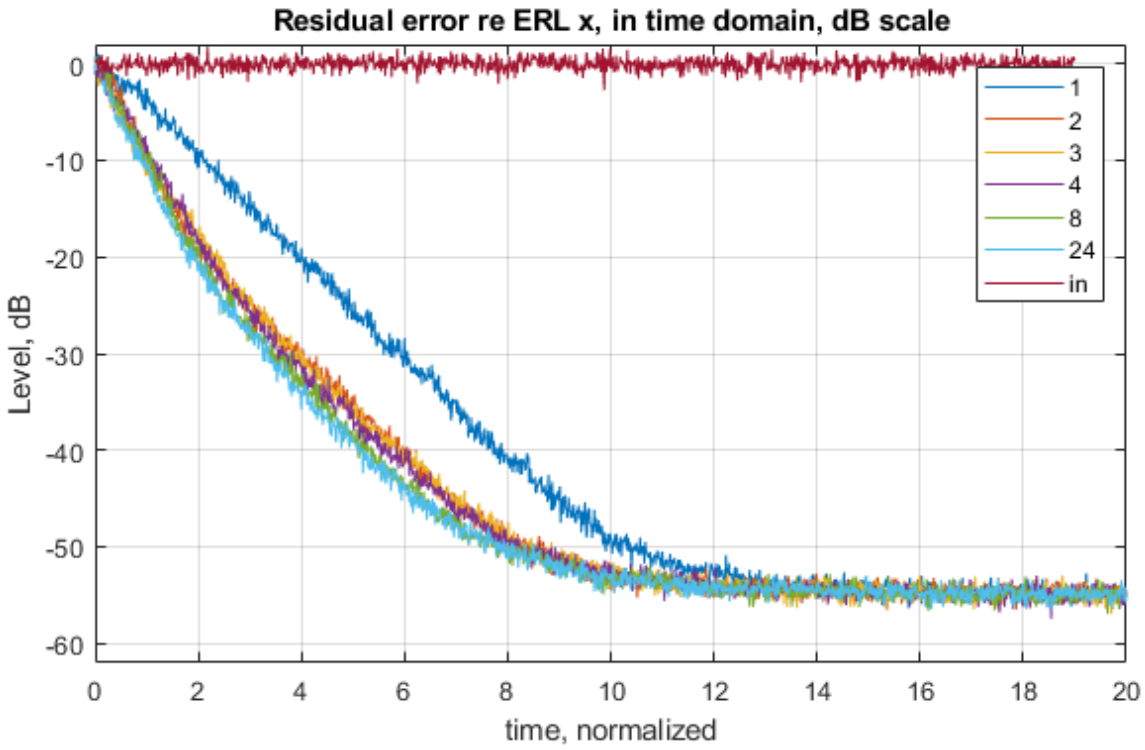
$X_t(k) = x_t(ck - c + 1 : ck)$; % corresponding block of excitation

$S_t(k) = X_t(k)^H X_t(k)/c;\ S_t = [S_t(1)\ S_t(2) \ldots S_t(blks)];$

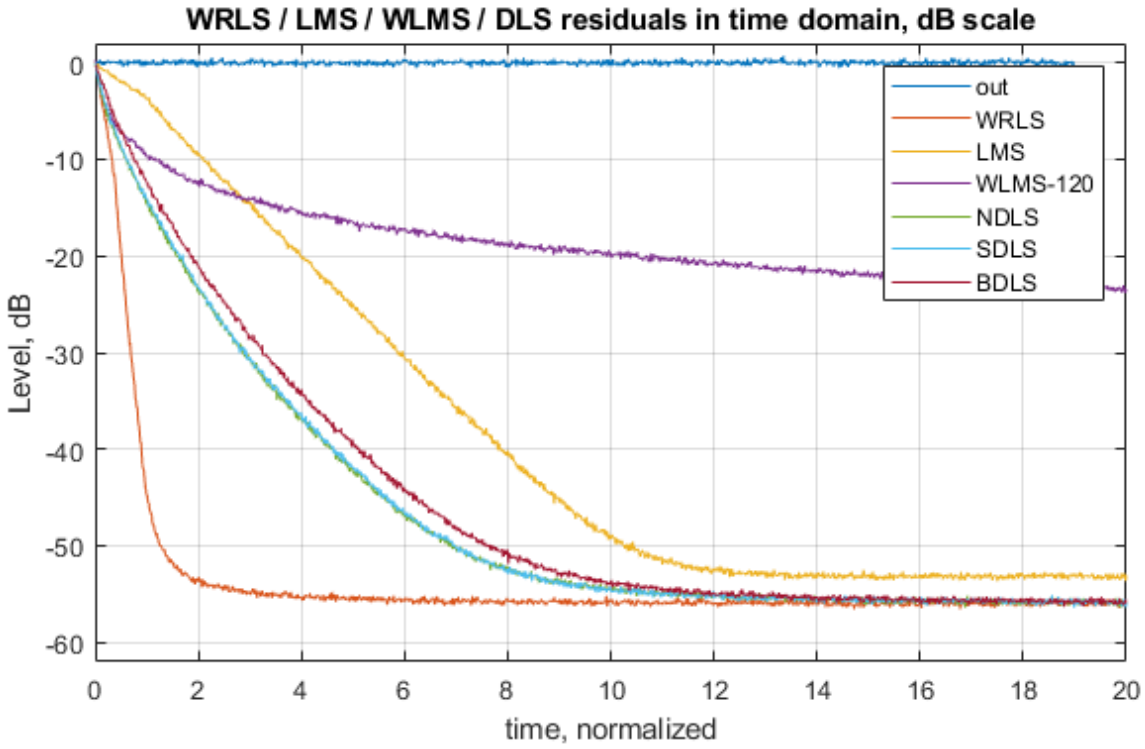
$b_{t+1}(k) = b_t(k)(1 - \eta_t b_t(k) S_t(k)\ \ );$

end

You don't need very many blocks to achieve reasonable performance:

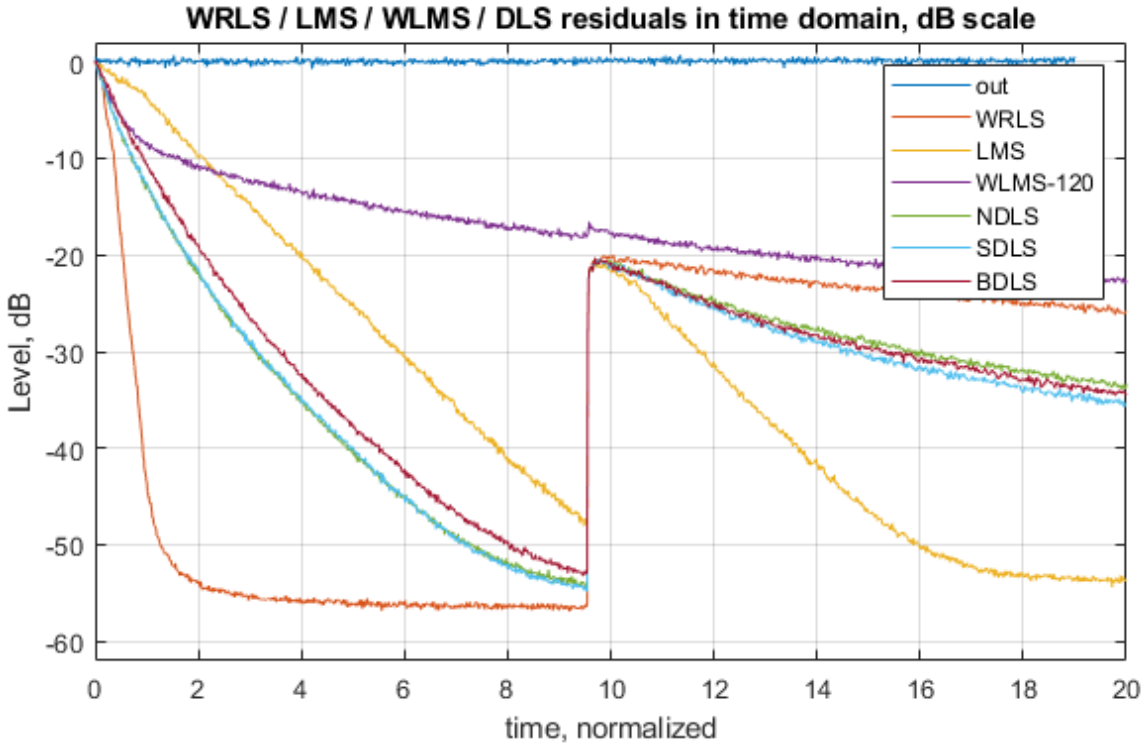


Now, let's look at how other variants of the DLS algorithm behave.



The SDLS and NDLS are practically indistinguishable, BDLS is slightly slower.

Of course, none of the algorithms is capable of dealing with a sudden RIR change adequately.



4.4.5 Relaxation

The expression for computation of d_{t+1} can (or should) be relaxed, especially if the excitation is autocorrelated:

$$d_{t+1} = d_t.(1_v - \mu_t x_t.d_t.conj(x_t) \ \ /{x_t}^H(d_t.x_t));$$

For that, we may modify

$$d_{t+1} = d_t.(1_v - \mu_0\mu_t x_t.d_t.conj(x_t) \ \ /{x_t}^H(d_t.x_t));$$

Where an empirical relaxation parameter $0 < \mu_0 < 1$ should be somewhat inversely proportional to the (average? Current? TBD) degree of excitation autocorrelation. The same consideration applies to all DLS variations.

4.4.6 Flat DLS

In the base we have BDLS with only one block, we get an approach to control the step size for LMS:

$${\nu_t}^2 = b_t {x_t}^H x_t$$

$$\eta_t = \frac{1}{{\nu_t}^2 + {\Sigma_t}^2}$$

$$b_{t+1} = b_t\left(1 - \eta_t {\nu_t}^2\right) = b_t(1 - \mu_0 \frac{{\nu_t}^2}{{\nu_t}^2 + {\Sigma_t}^2});$$

Where b_0 is chosen as the estimate of cumulative initial MSE *(trace(D0)).*

4.4.7 Relation to FTF and step response

Suppose that the input to [Re]RLS is a step function, with clean zeros before t=1. Then we should be able to find a solution of the first L_{ADF} equations, very easily:

[1	0	0	0	…	0]	*h(1)	$=y_1$
[1	1	0	0	…	0]	*h(2)	$=y_2$
[1	1	1	0	…	0]	*h(3)	$=y_3$
…							
[1	1	1	1	…	0]	*h(N-1)	$=y_{N-1}$
[1	1	1	1	…	1]	*h(N)	$=y_N$

…due to the lower-triangle shape of the matrix.

Does RLS do the same? The answer is assured "YES". Suppose that RLS is initialised to a scaled unity matrix, and σ_{nse}*->0.*

On the first step: $z_1=x_1$; $h_1(1)=y_1$; $D_1(1,1)=0$;

On the second step: $z_2(1)=0$; $z_2(2)=x_2(2)=1$; $h_2(2)=y_2-y_1$; $D_2(2,2)=0$;

On k^{th} step, $z_k(1:k-1)=0$; $z_k(k)=x_k(k)=1$; $h_k(k)=y_k-y_{k-1}-\ldots y_1$; $D_k(k,k)=0$;

The Kalman gain is a delta-function with *"1"* travelling along time axis. The dispersion matrix stays diagonal, with zeroing going on along time axis.

Thus, the NDLS shall operate more or less the same as ReRLS if the input is a sharp transition from silence to activity. The exact shape of activity waveform is unimportant, and σ_{nse} can be any and time-variant. Speech waveforms have lots of such sharp transitions, especially in high-frequency subbands.

Note that the assumption of zero input before *t=1* was the base of deriving FTF-class algorithms.

4.4.8 Summary

DLS is a straight-forward, direct, more than obvious descendant of ReRLS for the cases of more or less white noise excitation, which are not exactly uncommon, especially for subband adaptive filtering, which has been around since 1988. Why DLS is published so late? No idea.

4.5 Affine Projections (APA)

In essence, APA could be useful for low-frequency bands and can be combined with the DLS approach. The order of APA shall be equal to the number of harmonics of fundamental frequency in the sub-bands' passband and regularization must be heavy-handed to exclude close to band-edge harmonics. The matrix inversion does not need to be done on each sample.

The squeezing matrix ($I \ -\mu_t x_t \quad D_t {x_t}^H / {x_t}^H D_t x_t$)where D_t is P- diagonal matrix which is alike a mini-Fisher matrix and forms a "final-RLS-like" filter. This filter amplifies the weaker harmonics in square proportion, which isn't of un-debatable merit. A square-root APA, which dynamically equalizes the level of individual harmonics, would be more appropriate.

4.6 Summary

The domain of Adaptive Filtering is still very far from completed and thoroughly understood.

There are many more ReLS-based approaches to formalizing the audio-related problems. For example: typically, we form excitation vectors as $x_t = [in(t); x_{t-1}(1:L_{ADF}-1)]$. It does not have to be that way. An alternative RIR representation may be based on $x_t = [in(t); \lambda * x_{t-1}(1:L_{ADF}-1)]$ where $\lambda = 10^{(-6/RT_{60})} < 1$. Then, *h* will have uniform distribution along time axis, and the initial $diag(D_0)$ will be also flat. This approach solves quite a few critical problems with fixed-point implementations of adaptive filtering. Etc.

BTW, the Fast Kalman / FTF algorithms which are supposed to lower the RLS complexity into O(N) instead of $O(N^2)$ are unnecessary complications because FSAF (see the next part) lowers RLS complexity into $O(N^2/M^2)$.

5 Robustness of Adaptive Algorithms

5.1 Basics

Intuitively, a non-robust adaptive algorithm is an oxymoron. But… I wish it were so simple. Often, the degree of an algorithm's sensitivity to deviations from explicit and implicit assumptions is not obvious.

Adaptive control learned a while ago that "There is no algorithm worse than the optimal". Fast RLS is a good example of assumption taken too far. Generally, Gram-Schmidt orthogonalization is a non-robust algorithm because the residuals have very little in common with the theory even for mildly nonlinear systems, etc.

Here, we concentrate on volatility to the accuracy of the following estimates:

- noise level
- RT-60
- RIR variability

In FSAF, we need to be acutely aware of pollutants inserted by both [amplifier and loudspeaker] intermodulation products and inter-band aliasing, which shall be discussed in more detail in the following chapter. TBD.

5.2 NOISE

Acoustic noise level is a lesser problem (putting aside the discussion of what is noise and what is not). Most often, we can observe it at the beginning of conversation, and measure directly. HVAC turning on and off during the conversation, babble noise, traffic noise are all problems, however, solvable in reality thanks to the research efforts spent on noise reduction.

5.3 RT_{60}

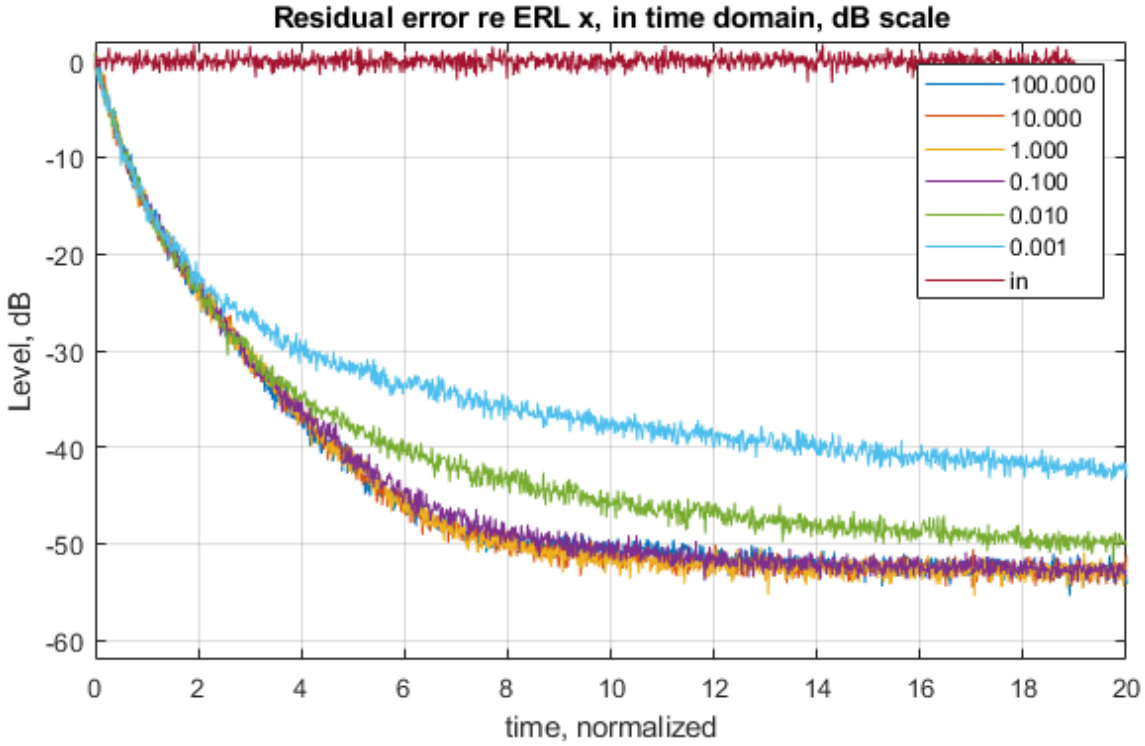
RT_{60} can be estimated by using a fully blown JTF-ReLS in a few very narrow subbands, as a hyperparameter of $arg(min|_{RT60}\{var(residual\_err(JTF\text{-}RELS(...)))\})$.

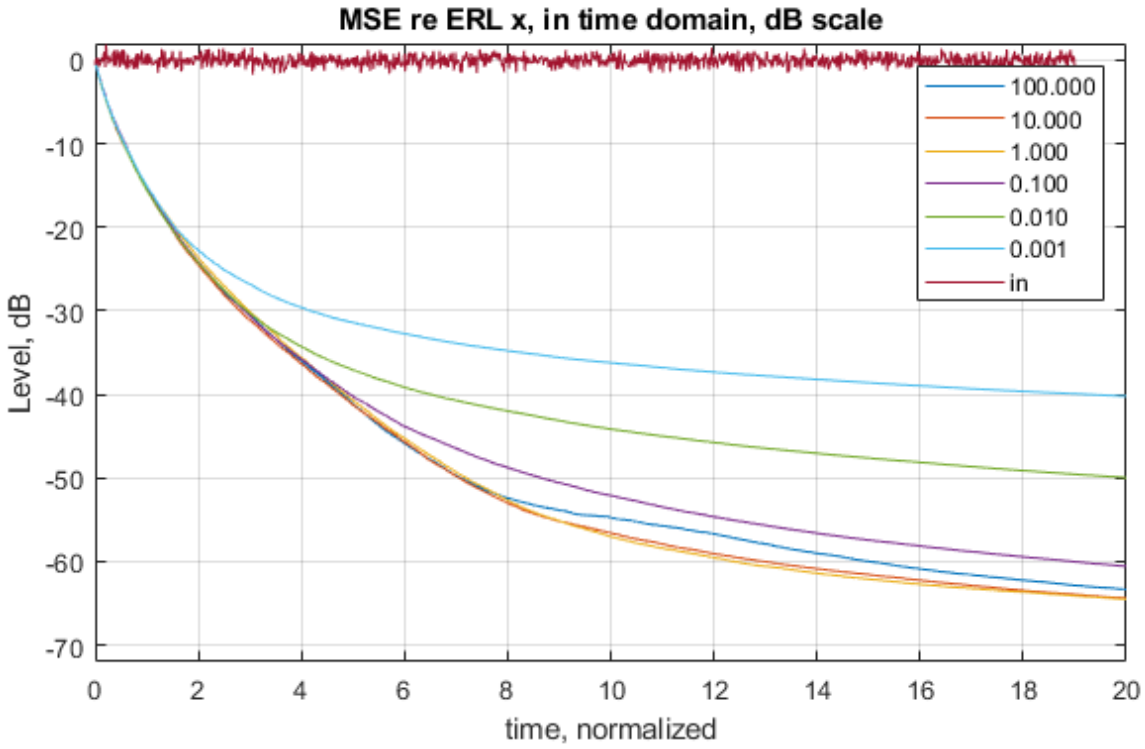
However, the room reverberation level (or it's variations) isn't easy to guess.

5.4 REVERBERATION LEVEL AND ITS VARIATIONS [209, 212]

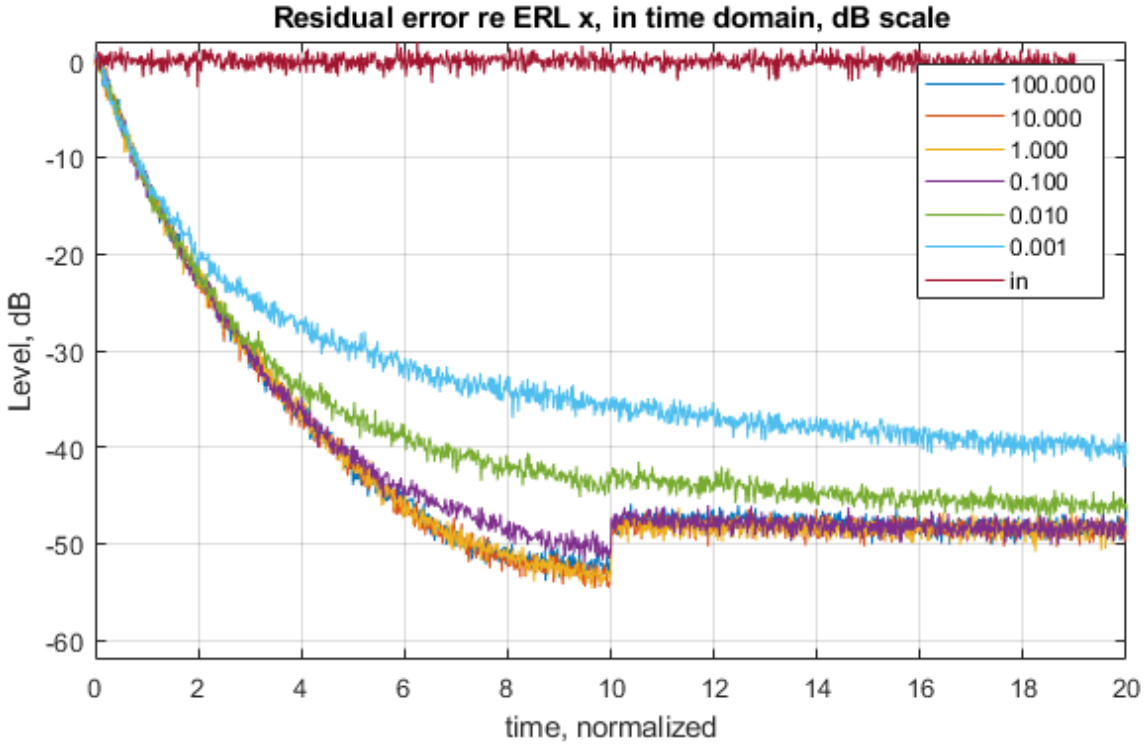
Some researchers assumed that estimations from above would work satisfactory i.e., if you start with estimated $est\{D_0\} > D_o \;\; if \;\; for \; any \; x, x^H est\{D_0\}x \; > x^H D_0 x$ then $est\{D_t\} > D_t \; for \; any \; t > 0$. In other words, we can err on a safe side. However, that assumption relied on a good knowledge of constant level noise, which is not true, and therefore there is no safe side to err.

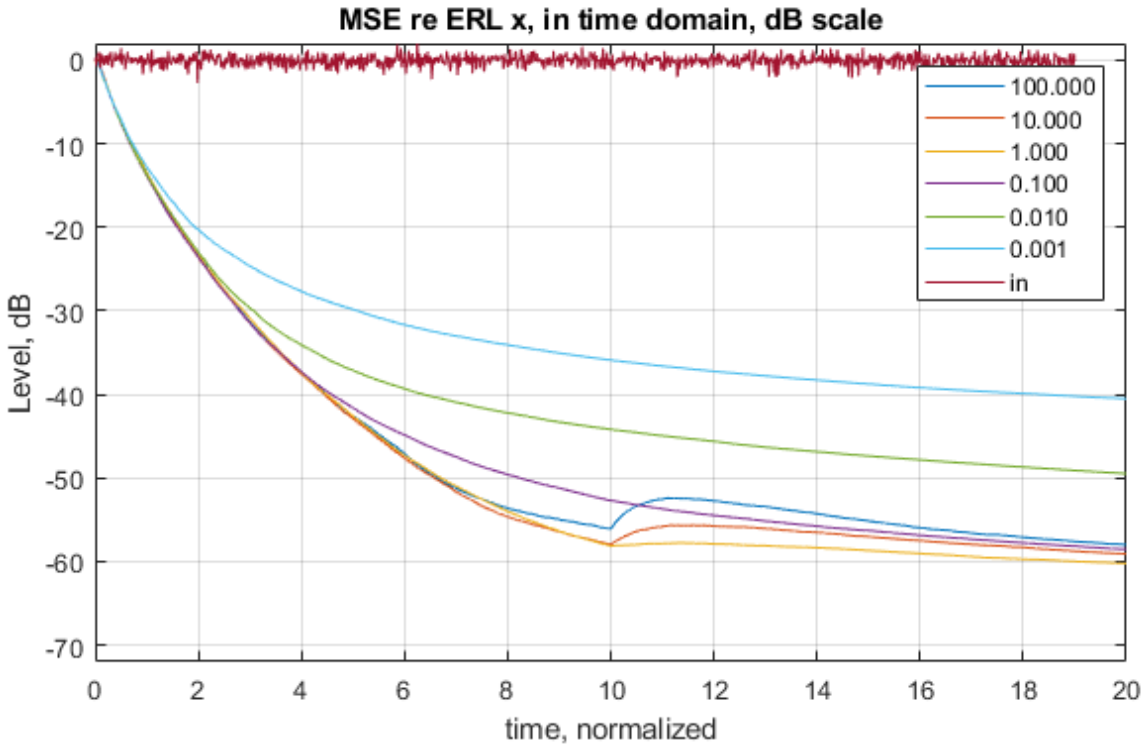
For a known constant level noise of -60dB re input, NDLS, depending on the ratio $est\{D_0\}/D_o$:

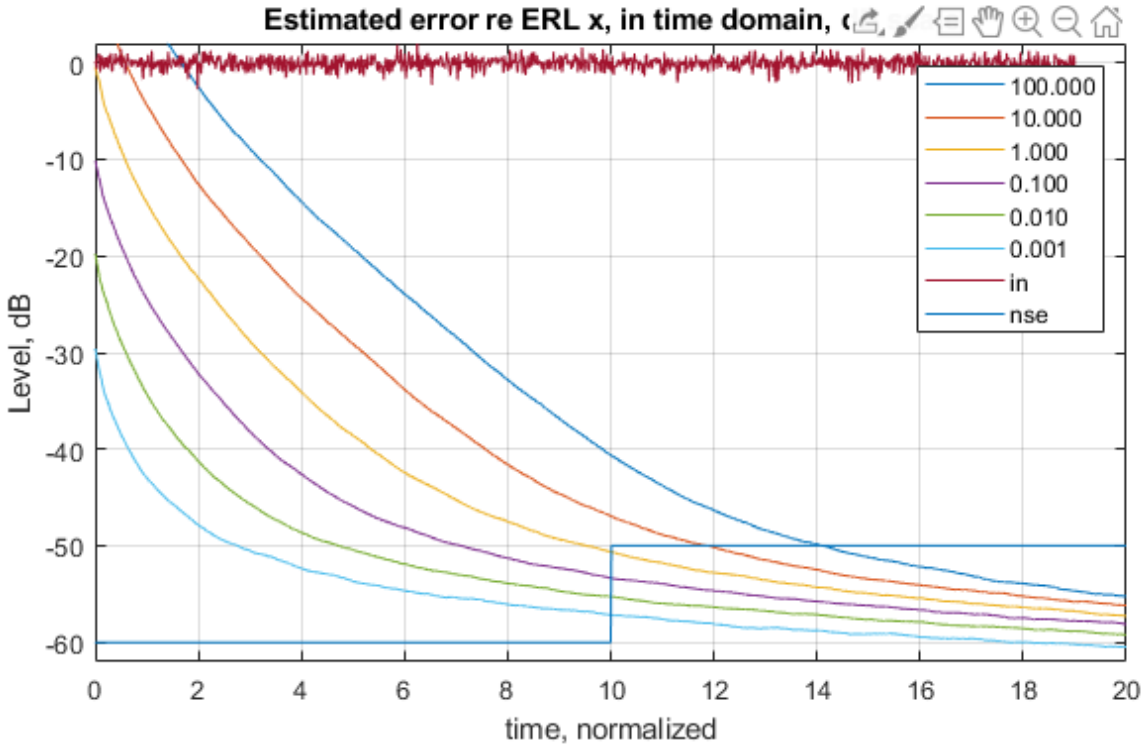




The case of $est\{D_0\} = D_o$ is still the best… but not by much, so you could have concluded that erring on a "safe" side is legit. However, that's not the whole story: if the noise level jumps up (even we know by how much):

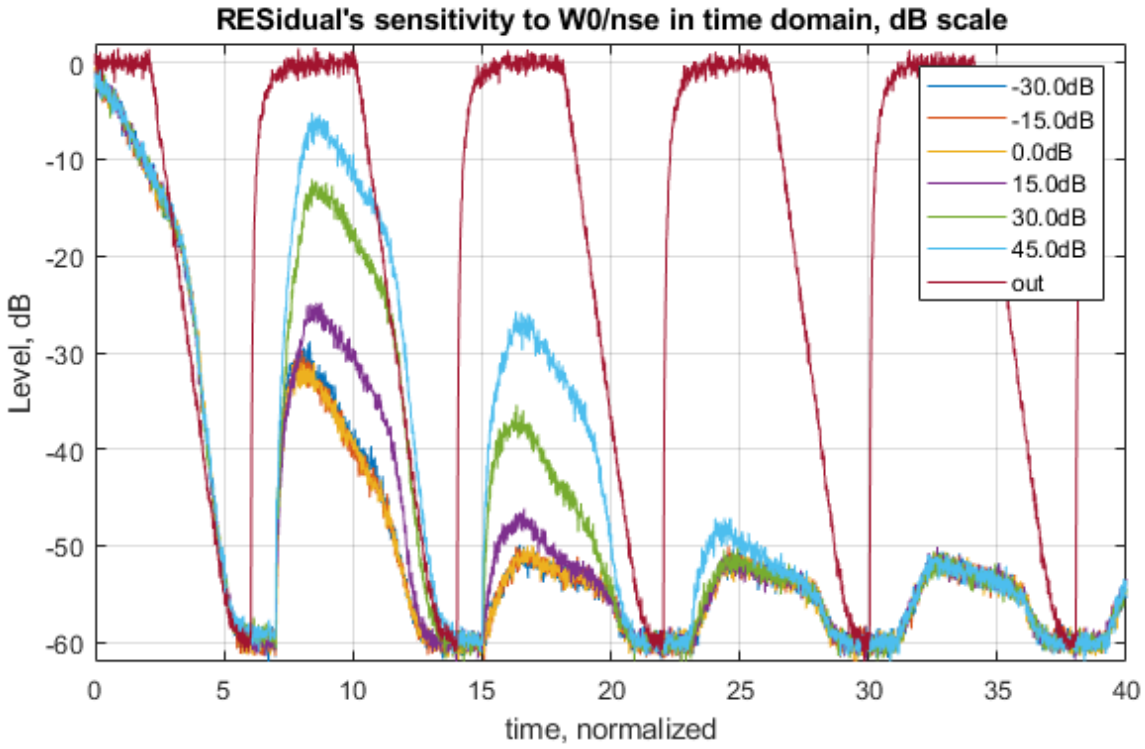


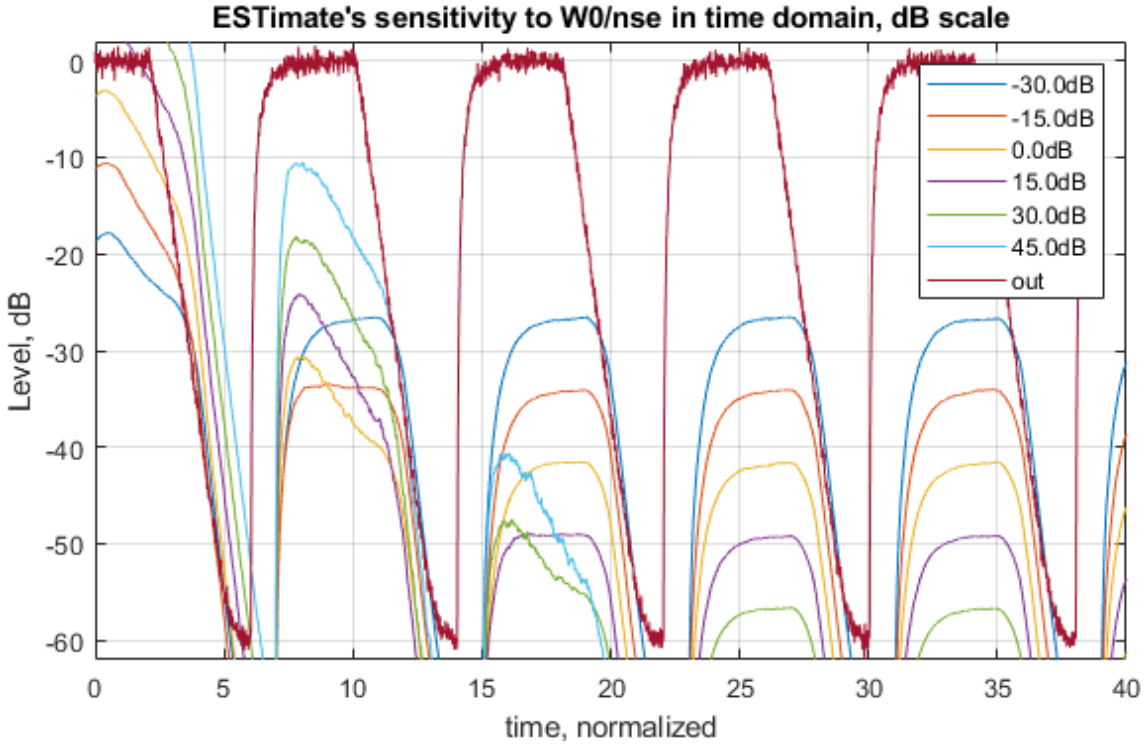


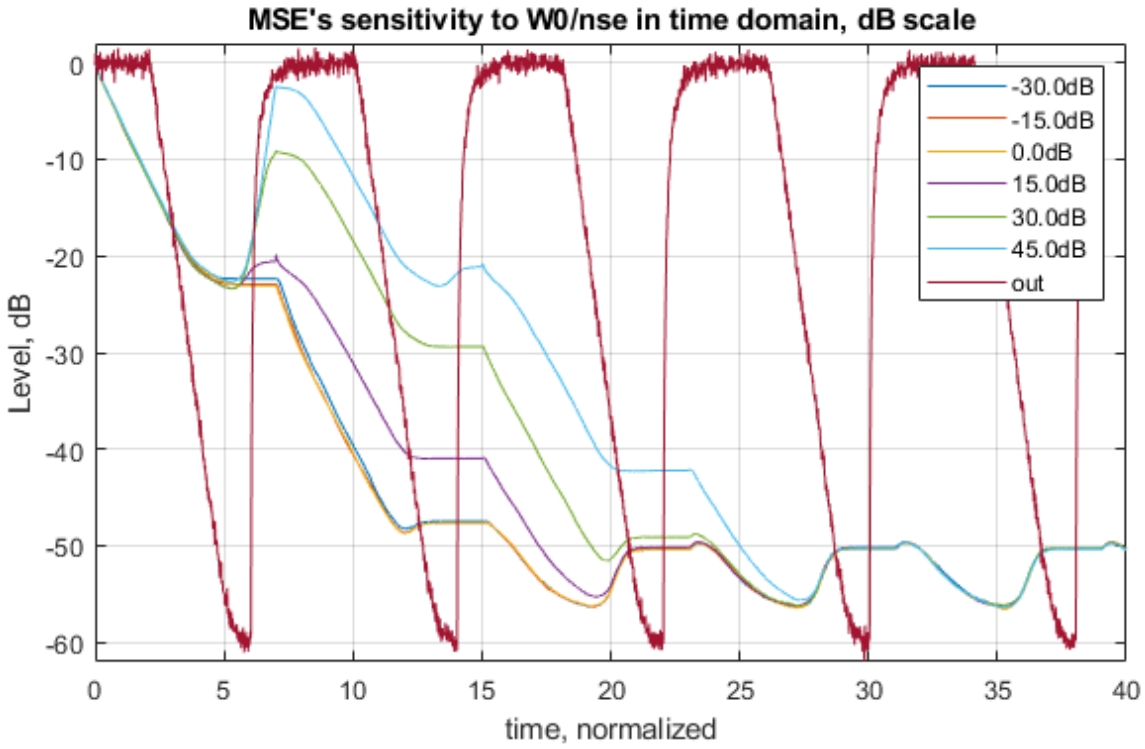


The algorithm should have dropped the step size, $\mu_t = {\nu_t}^2/({\nu_t}^2 + {\Sigma_t}^2)$; but due to overestimation of ${\nu_t}^2$it did not do that properly.

The effect of overestimation (even for constant noise level)$est\{D_0\} \gg D_o$ is much more pronounced if excitation is pulsed, as in normal speech, even if you use optimal step-size control.


RESidual's sensitivity to W0/nse in time domain, dB scale
Level, dB
time, normalized
-30.0dB
-15.0dB
0.0dB
15.0dB
30.0dB
45.0dB
out


ESTimate's sensitivity to W0/nse in time domain, dB scale
Level, dB
time, normalized
-30.0dB
-15.0dB
0.0dB
15.0dB
30.0dB
45.0dB
out



Summarising: the step size control algorithms, [pre] determined by excitation and initial estimations of noise and RIR variability, are not robust.

In other words, by using an adaptive approach, we can estimate IR permutations well enough… only if we have a good idea of their magnitude[6].

The existence of one magical algorithm which would work [well in all scenarios] is very questionable.

5.5 "HOME GROWN" STEP CONTROL

There is no such a thing as robustness of adaptive filtering to inadequate step-control.

A proper step-size control during activity transitions is quite tricky, and divergence / switch to half-duplex happens readily. As everyone learned from lockdowns' enforced teleconferencing, it's "headset-only" because commercially available AECs are poor half-duplex [in reality].

If the step control is not properly based on the well understood science of stochastic adaptive control and system identification… let us draw the curtain of charity over various improvisations on the topic.

The consequences of bad step control are even worse, pretty much fatal, for ARC/AFC.

5.6 SUMMARY

The real-life applications of adaptive filtering need to run unattended for indefinite time, limited only by UPS capacity. Converging is only half-a-work. Knowing how well an algorithm has converged is of the same importance, and is the key to the algorithm's robustness. However, any predetermined procedure of step size [program] control, which is not affected by the observed system output, is not robust.

[6] Which is essentially the same conclusion as in Ljung et al (2019) "$\Pi=\theta_0\theta_0^T$ (19) Not surprisingly the best regularization depends on the unknown system".

6 META ADAPTIVE APPROACH

6.1 BASICS

To overcome the deficiencies of predetermined program control, we may add another feedback loop on the top of the existing solution, and MM. For capable adaptive algorithms:

- we can derive the estimation of residual error variance, $v_t{}^2 = x_t{}^H D_t x_t,$ or in other ways
- we can observe the residual error variance.
- if estimation and observation are close, or below noise or within under-modelling error, do nothing.
- If estimation and observation are above noise and slowly diverging, we shall adjust something in the algorithm to make the match better… slowly, not to break the underlying adaptive loop.
- That something should be, IMHO, the algorithm's internal representations of the $E\{(h_t - h)(h_t - h)^H\}$ dispersion matrix.
- if the observed residual error variance jumps up well above estimation, we have either double talk or RIR variations. We can not distinguish between them readily but we may run two (or more) models in parallel:
  - (a) a conservative model shall freeze the adaptation
  - (b) a speculative model, on a rising front of excitation, add an exponent to the diagonal of D_t decaying with $2*RT_{60}$ and of the level corresponding to the observation.
  - control (a) and (b) models as described in Gustaffson's book.

Let's call adaptive algorithms with an explicit feedback loop on residual error variance "meta-adaptive"[7].

Alternatively, we can apply finely spaced (IMM, AFMM, etc) multiple model approach which is also capable of selecting the algorithm/model with the best dispersion matrix, but it would likely need many more models, and thus require higher MIPS and memory.

6.2 META-BDLS [213]

6.2.1 Conservative model

For each sample t, we can write a linear equation for ideal b_t in noise-less case:

$$v_t{}^2 = b_t S_t \; = \; |y_t \; - \; x_t{}^H h_t|^2 \; ;$$

With noise and non-idealities, there is an error

$$\varepsilon \; = \; |e_t|^2 \; - \; (v_t{}^2 \; + \; \Sigma_t{}^2);$$

and we can adapt our modelb_t- but slowly and if and only if this error does not indicate a radical RIR change. The basis (delta functions) is not physically meaningful. It can be improved by weighting the projection vector with b_0, or by rotating the basis so that the first ort is decaying exponent corresponding to the abrupt RIR change, or by other tricks.

$$Z_t = b_0 . S_t;$$

$$b'_t \; = max(\delta, b_t \; + \; \zeta_t Z_t \quad \varepsilon / S_t{}^H Z_t); \text{ where}$$

[7] Could there be meta-meta-meta-adaptive algorithms? In certain conditions… may be

$0 < \delta \approx w_0(end)/LADF$,

ζ_t- meta step size, can be chosen as a fraction of regular step size μ_t.

Another approach could be to control relaxation parameter μ_0 so that expected and observed variance of residual error agree.

The complexity of a particular approach chosen shall be determined by the researcher's understanding of applicable corner cases, ability to debug, and capacity to exhaustively test it.

6.2.2 Speculative model

Whenever RIR changes and $|e_t|^2 \gg (\nu_t{}^2 + \Sigma_t{}^2)$ in a sustained way, we need to add a scaled copy of $b_{0,scaled} = b_0 max(\varepsilon, 0)/b_0{}^H S_t$.

$$b_t{}' = b_t + b_0 max(\varepsilon, 0)/b_0{}^H S_t ;$$

The error $\varepsilon \gg \sigma^2$ can be thresholded, and /or passed through a procedure similar to decision-directed noise reduction to avoid acting on single outliers, and/or scaled down to reduce a potential for overreaction. Let's find scaling γ using Weighted LS (Weighted ReLS: TBD)

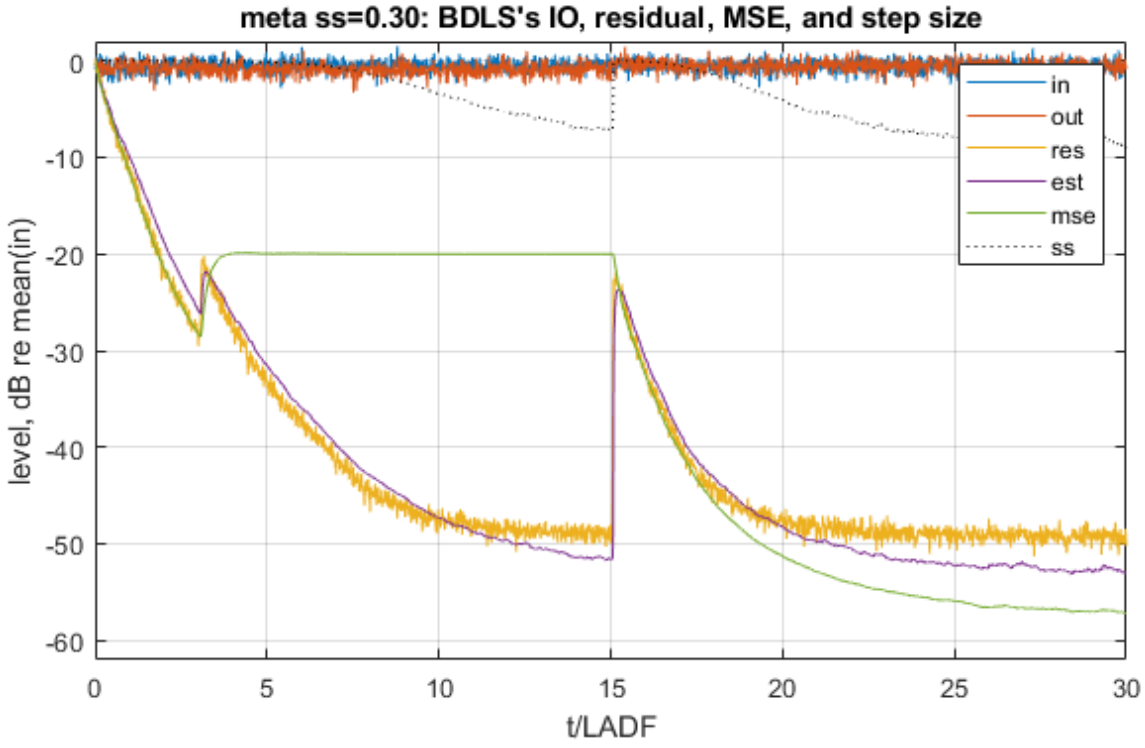
$b_0{}^H S_t\, \gamma = |e_t|^2 - (\nu_t{}^2 + \Sigma_t{}^2)$; Let's use N+1 of past equations and note:

$A_t = b_0{}^H S_t;\ A = [A_t A_{t-1} \cdots A_{t-N}]^T$

$B_t = |e_t|^2 - (\nu_t{}^2 + \Sigma_t{}^2);\ B = [B_t B_{t-1} \cdots B_{t-N}]^T$

$C_t = \nu_t{}^2 + \Sigma_t{}^2;\ C = diag\{[C_t C_{t-1} \cdots C_{t-N}]^T\}$

$\gamma_t = \frac{A^T C^{-1} B}{A^T C^{-1} A}$;



Here we simulate a change in RIR at 3.0 and return back at 15.0 seconds, at the level of -20dB relative to the main RIR. Noise is at -50dB relative to input.

You can see that meta-BDLS adapts to the RIR changes as fast as possible, at 15 "seconds" it adapts as fast as at t=0.

6.2.3 Combining models

$$J_{t+1} = db(e_{slow,t}) - db(e_{fast,t}) + J_t;$$

$$if\ -thr\ <\ J_{t+1}\ <\ thr, continue$$

$$elseif\ J_{t+1}\ >\ thr, h_{slow,t} = h_{fast,t};\ b_{slow,t} = b_{fast,t};\ J_{t+1} = 0;$$

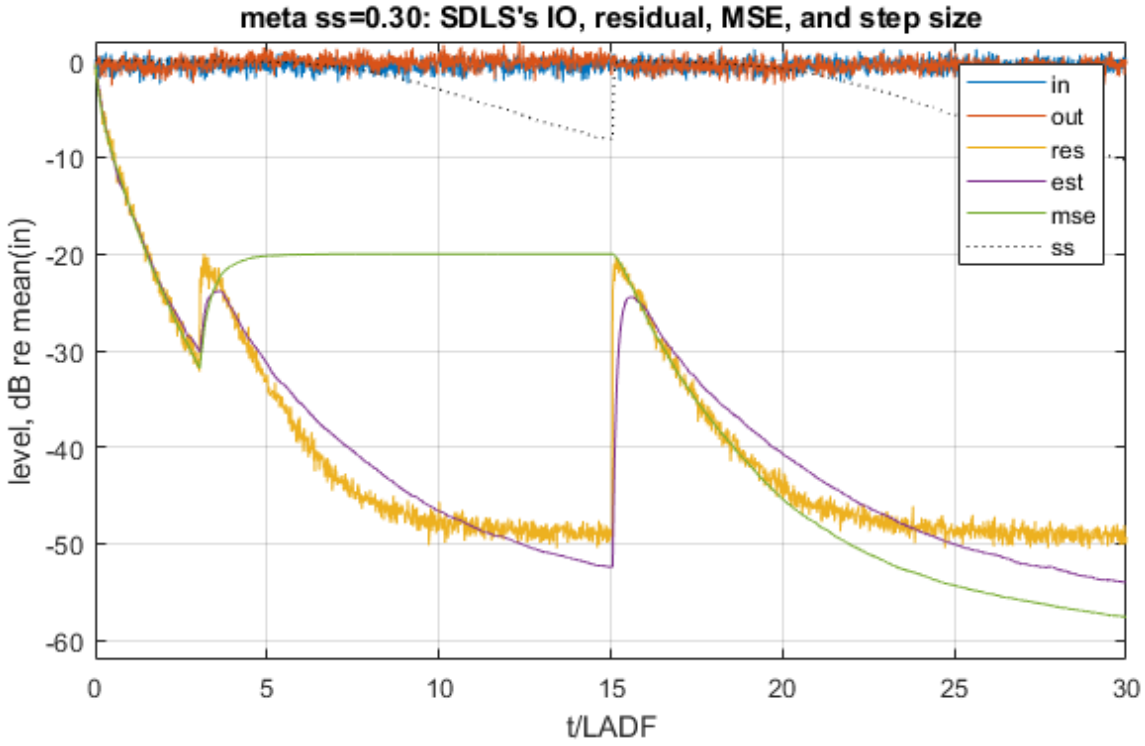
$$elseif\ J_{t+1}\ <\ -thr, h_{fast,t} = h_{slow,t};\ b_{fast,t} = b_{slow,t};\ J_{t+1} = 0;$$

$$e_{out,t} = min(e_{slow,t}, e_{fast,t});$$

6.3 META-SDLS

This algorithm is the simplest and is controlled by a single scalar, shelf level. When a RIR variation occurs, we need to add an exponent to the d_t... which we can not do within the SDLS approach. However, meta-SDLS can be used as a conservative model.

The performance of the SDLS depends on excitation spectra, so meta-SDLS can be corrected for that. The shelf level does not necessarily mean spectra-invariant MSE but can go up or down depending on how "novel" the current spectrum is. The step size for adapting the shelf position is the design choice.

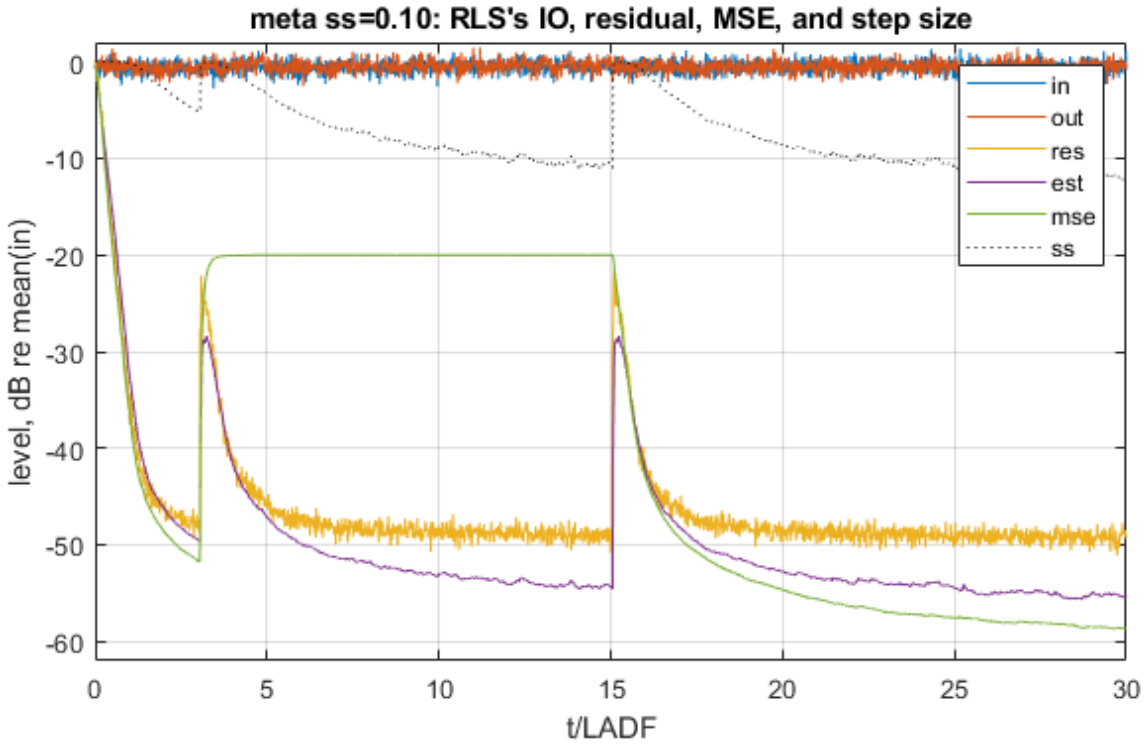


You can see that SDLS adapts to the RIR change as fast as it adapts on this level of MSE, it does not start "anew".

6.4 META-RERLS = REKALMAN-MM

No changes to the conservative model. You could potentially scale the dispersion matrix up or down … but it's unclear how to test that exhaustively.

Speculative model adds $diag(d_{0,scaled}\ )$ to the dispersion matrix in the manner similar to the Kalman filter. In subbands, you can add a matrix because you know the shape of the IN filter.



Quite close to as good as it gets – but so far on relatively low RIR variations only.

6.5 SUMMARY

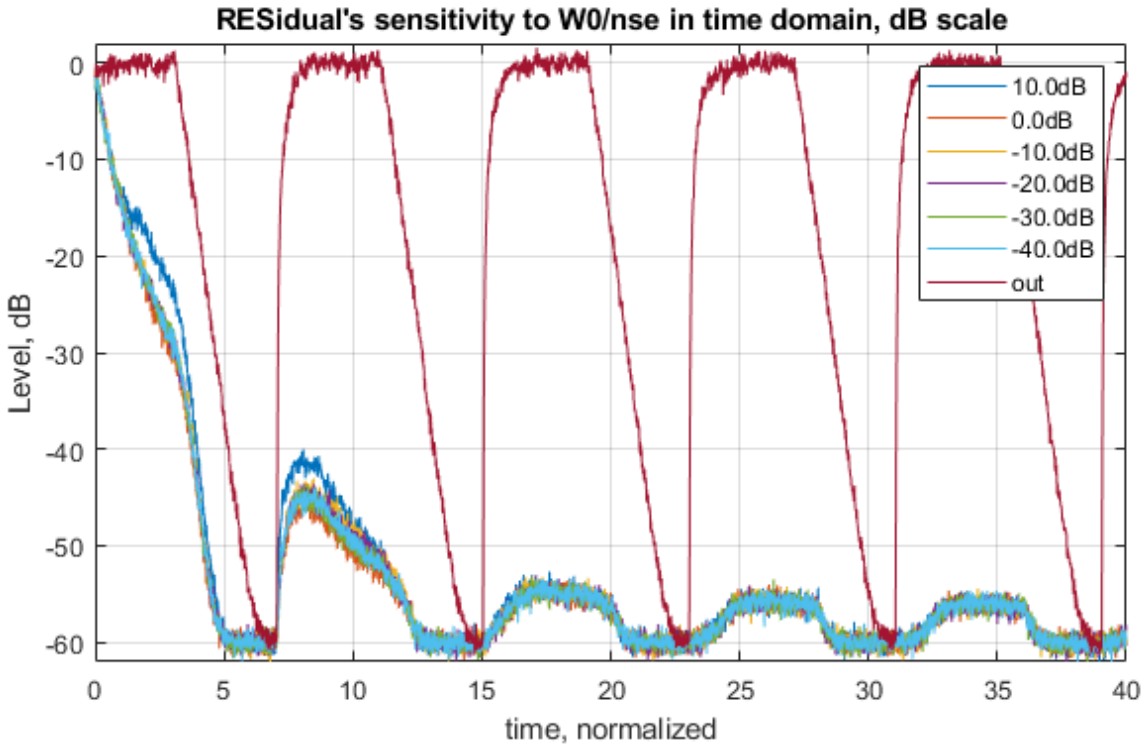
More TBD.

Sorry, it will take time because it's so awkward to translate thinking by images into a narrative.

7 ROBUSTNESS OF META ADAPTIVE ALGORITHMS

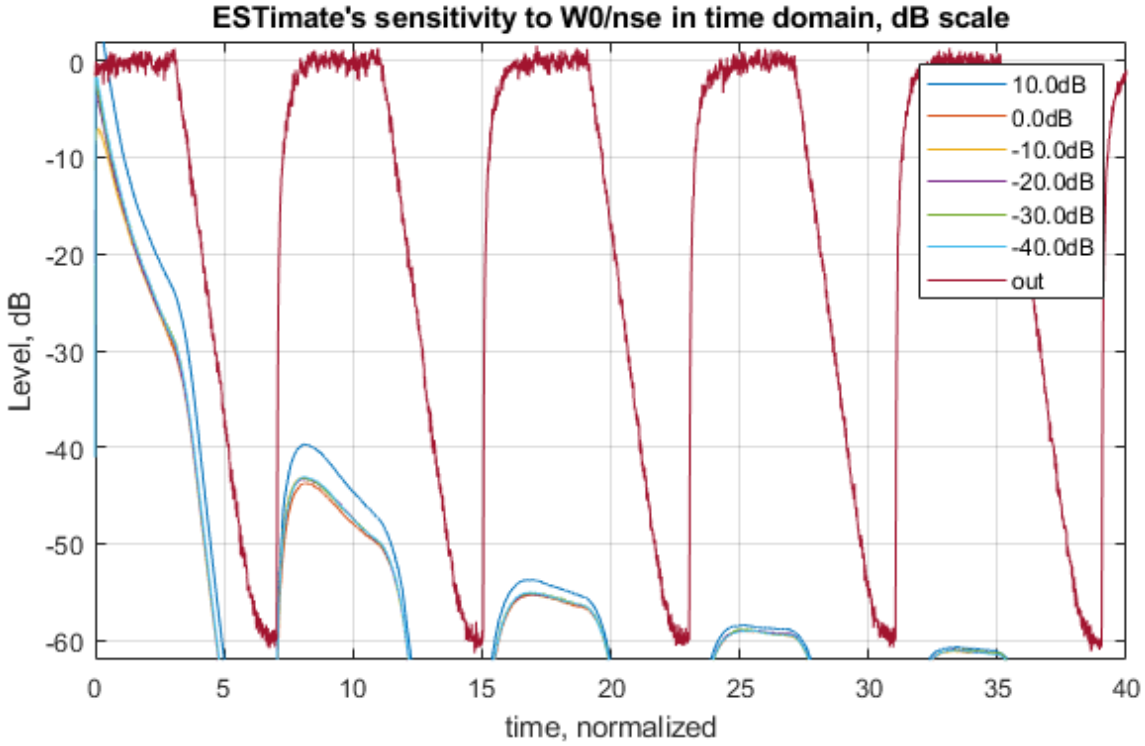
7.1 META-BDLS EXAMPLE [215]

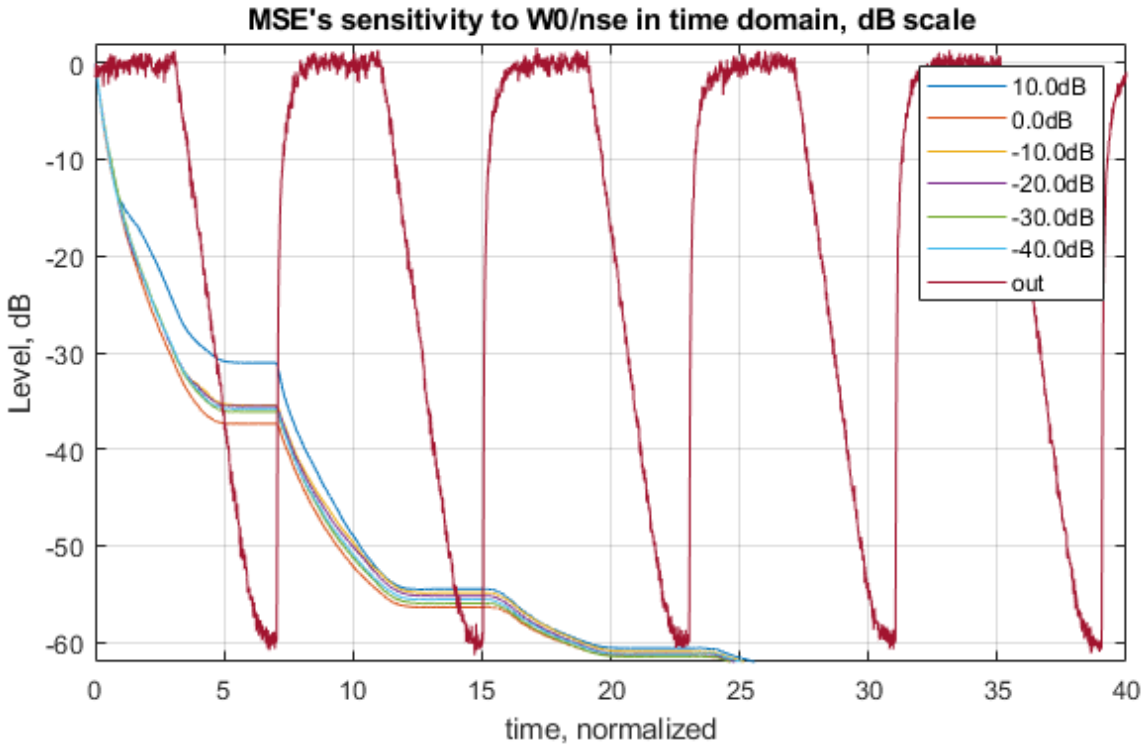
The robustness of meta-adaptive algorithms to errors in estimation of RIR and its variations can be demonstrated on meta-BDLS with the same pulsed input.



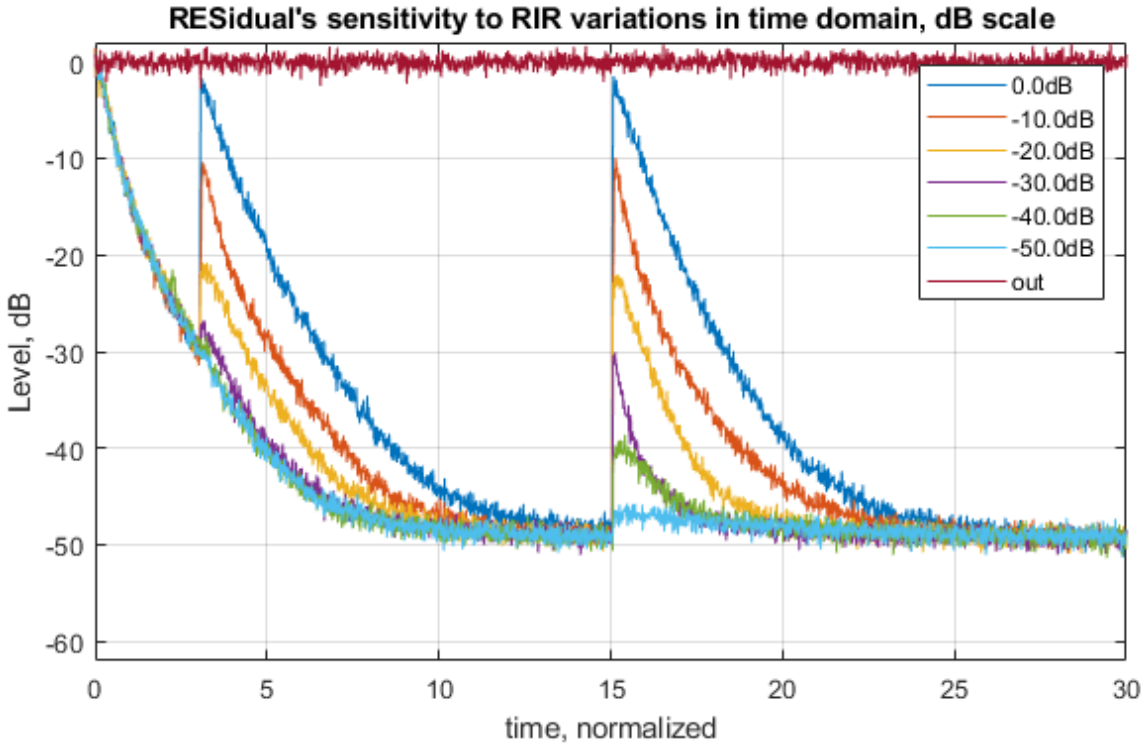
We can see that it's safe to err on the underestimate side. The algorithm will correct it, and correct it fast, using the knowledge of acoustics' basics and "known" RT_{60}.

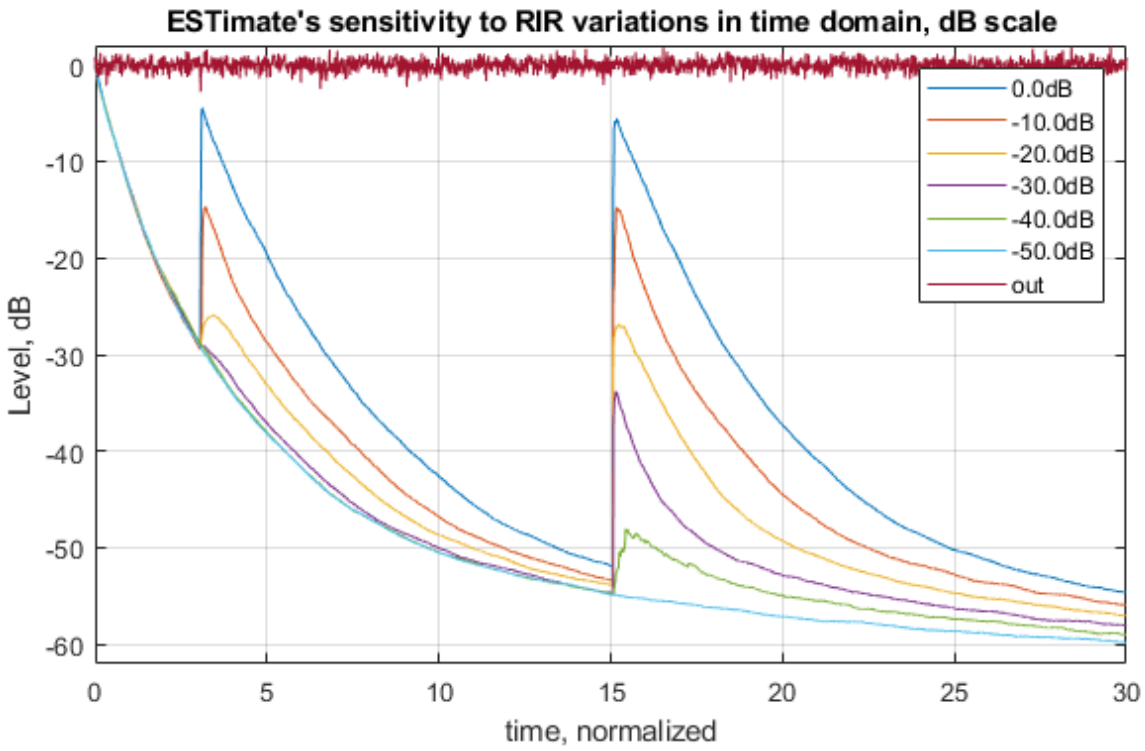
I am not aware of ways to perform the correction of RIR variability overestimation quickly.



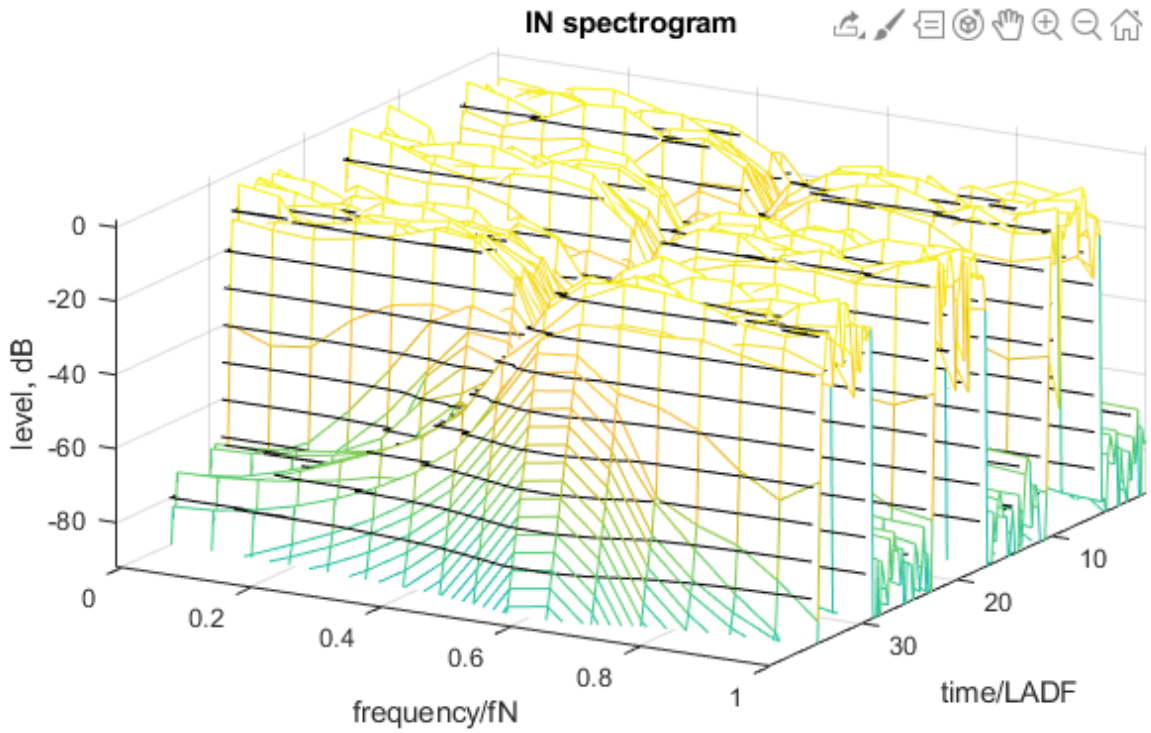


Let's also demonstrate how BDLS adapts to the varying level of RIR variations, as in 7.1.2:

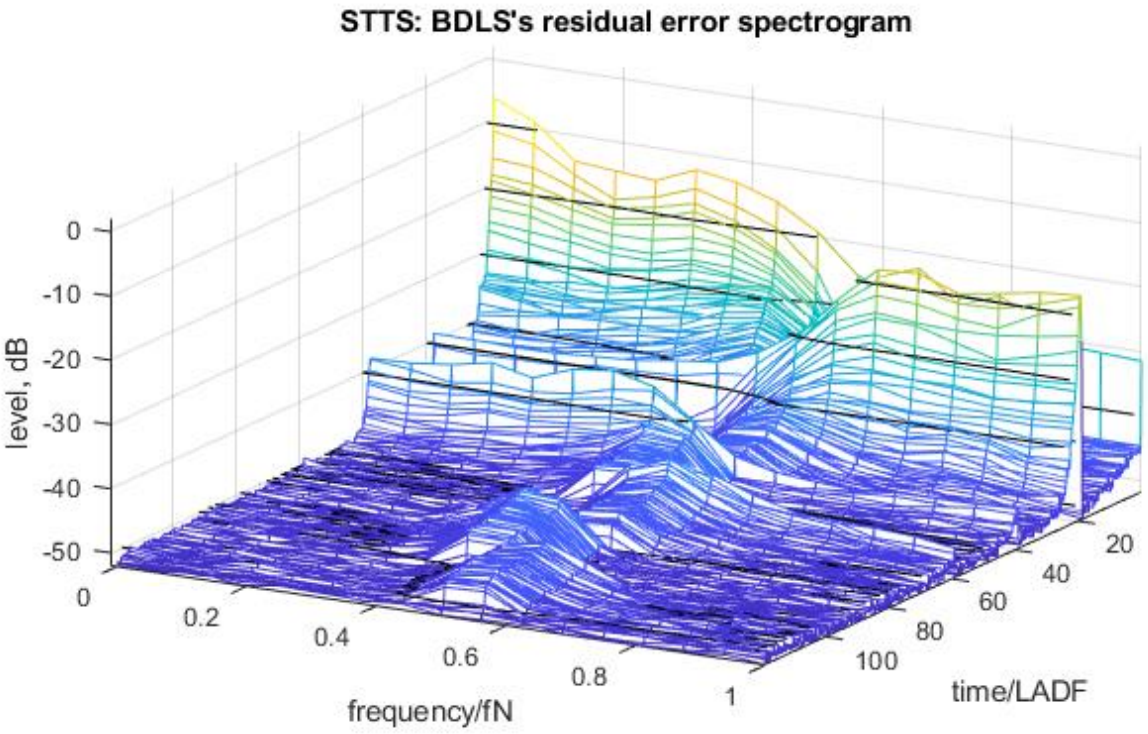


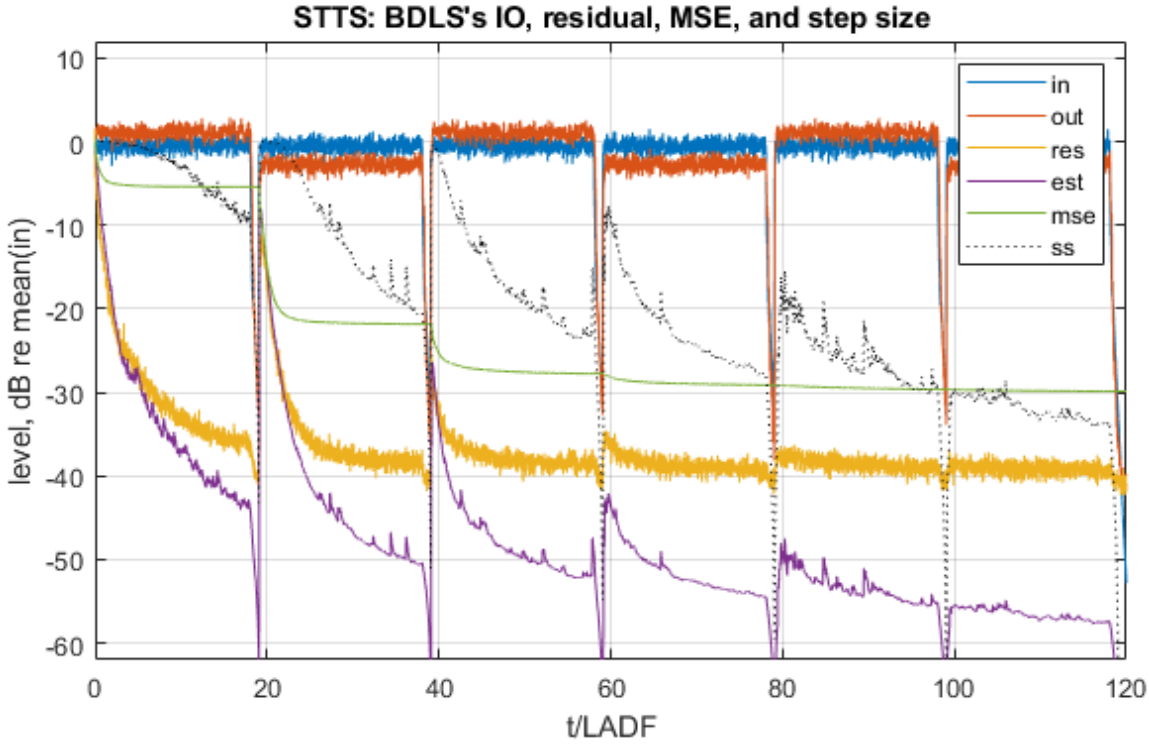


We also can demonstrate meta-BDLS robustness to non-white excitation, as in Chapter 3 (repeating):

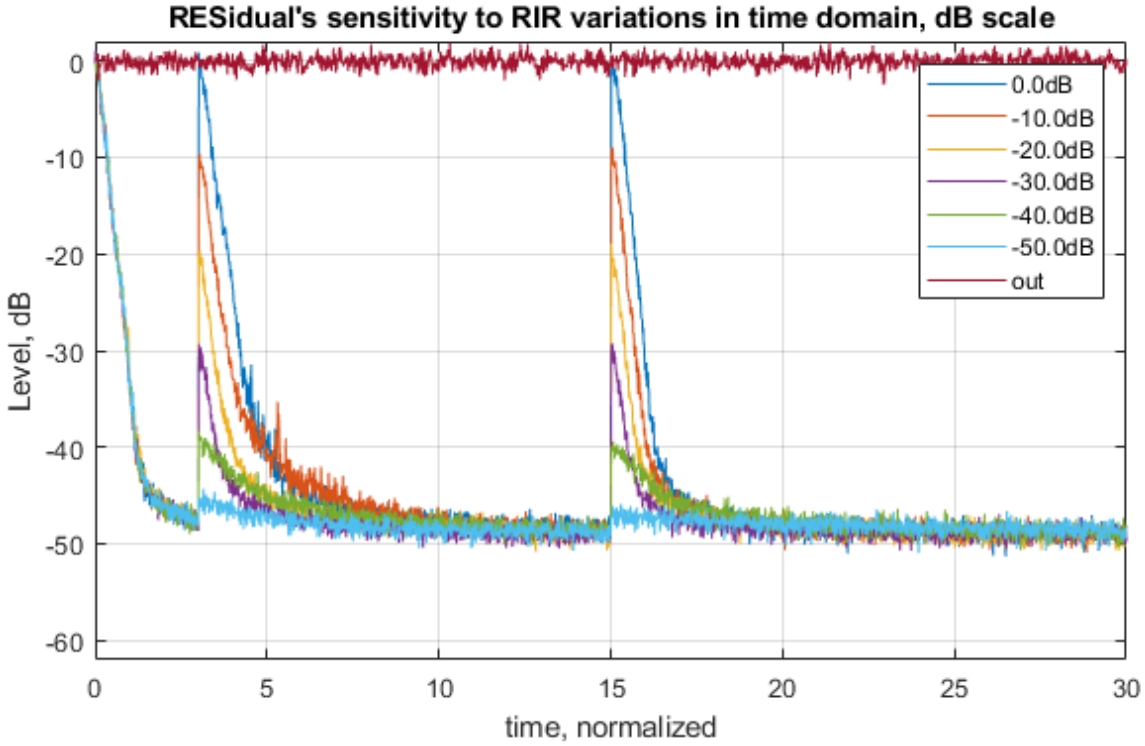


But the residual is completely different (because adding a feedback loop is a good thing):

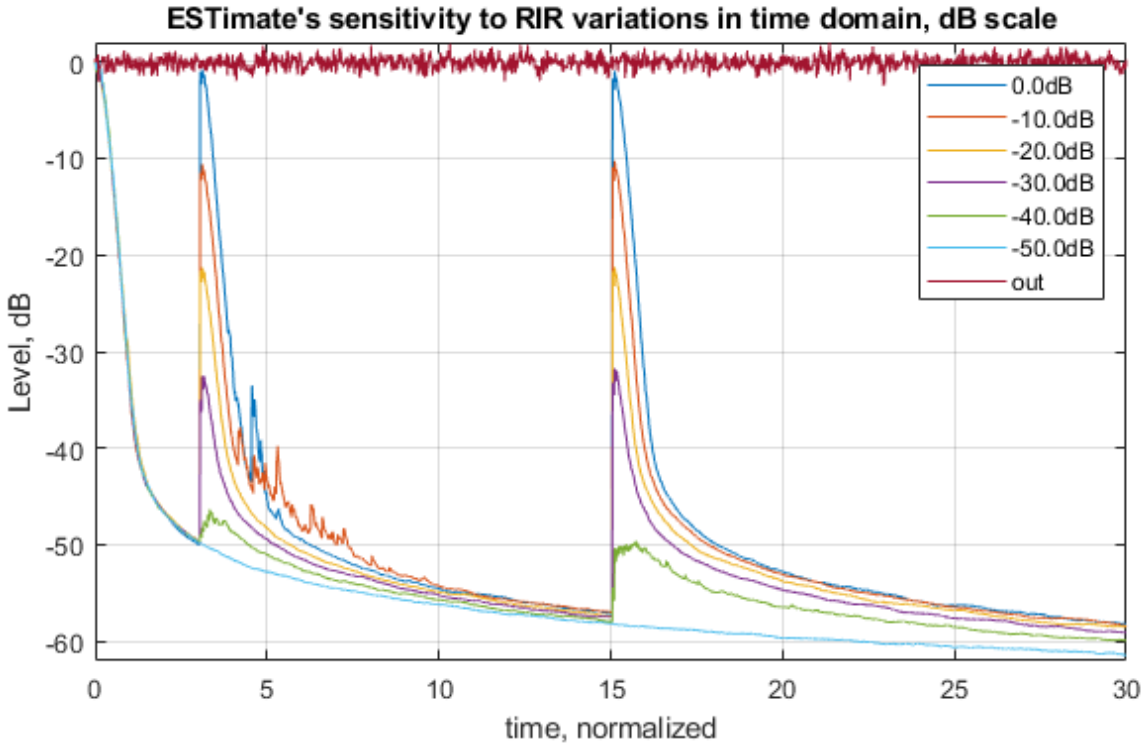

STTS: BDLS's residual error spectrogram
level, dB
0
-10
-20
-30
-40
-50
0
0.2
0.4
0.6
0.8
1
frequency/fN
20
40
60
80
100
time/LADF

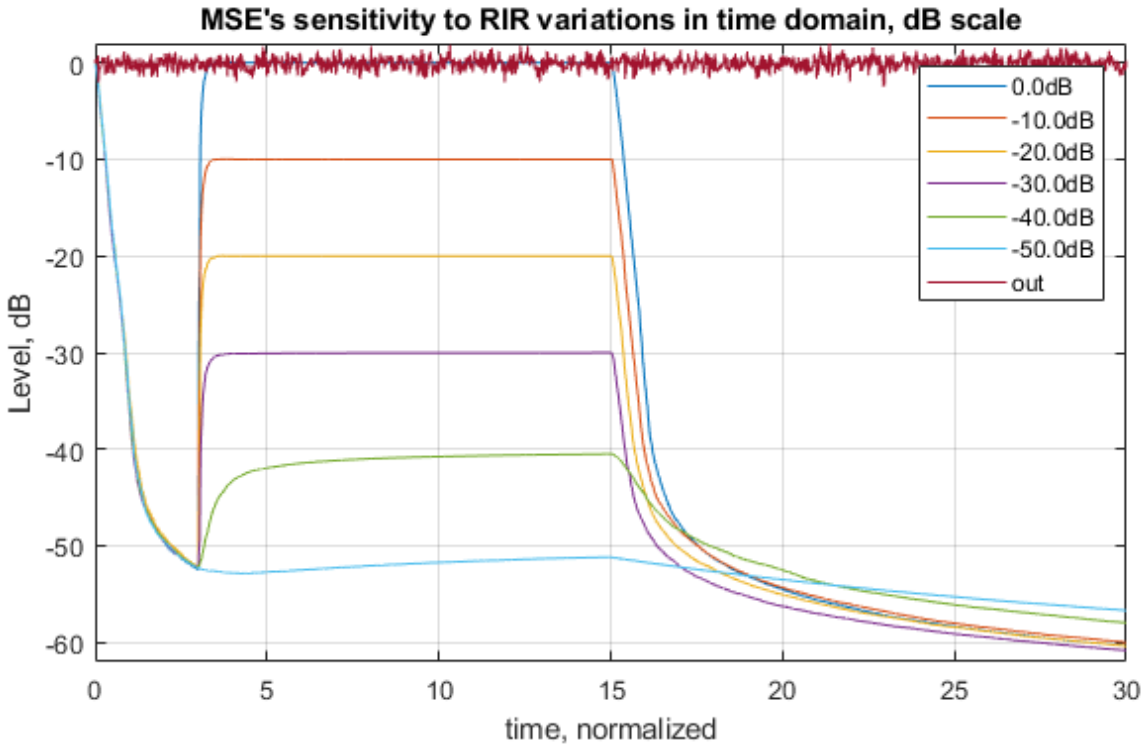

STTS: BDLS's IO, residual, MSE, and step size
level, dB re mean(in)
10
0
-10
-20
-30
-40
-50
-60
0
20
40
60
80
100
120
t/LADF
in
out
res
est
mse
ss

7.2 META-RERLS EXAMPLE [215]



The “beard” of extra corrections afterwards is strongly related to the insufficient level of primary correction, when RIR changes very radically. It’s a complex situation with stability of nested feedback loops in a strongly non-stationary (due to ReRLS) and non-linear system. It’s not clear yet how to deal with it.

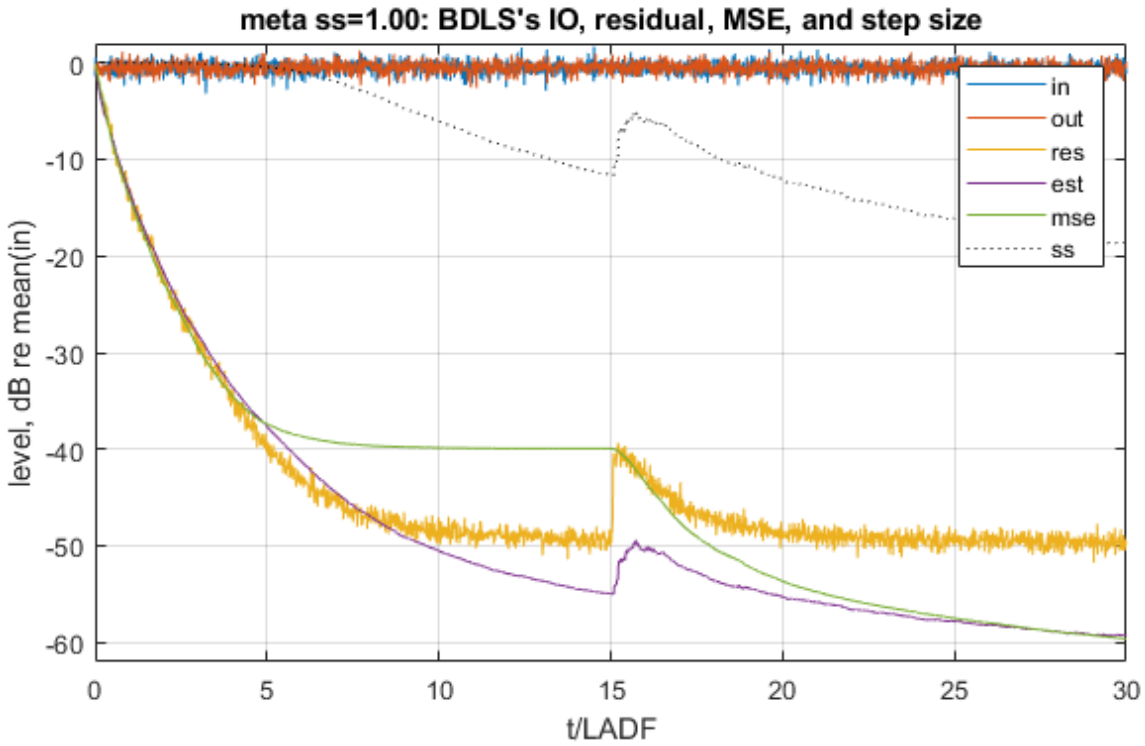




7.3 On detection of RIR variations [214]

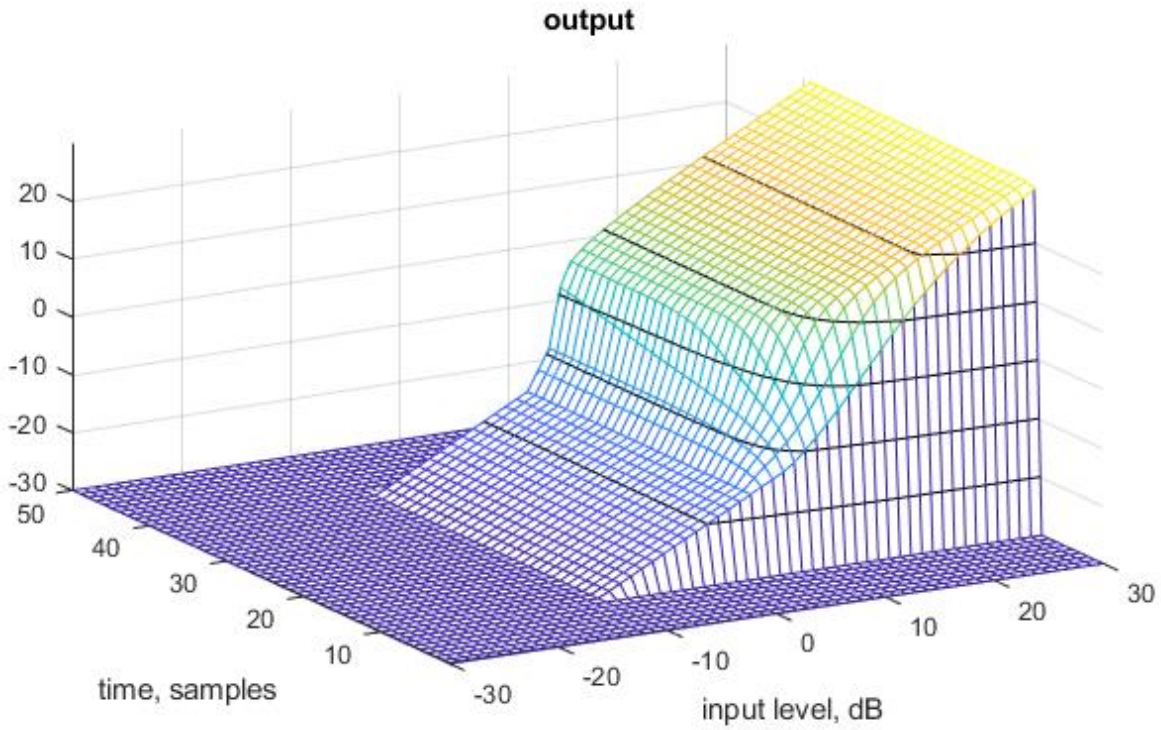
The meta-adaptive approach relies on detection of RIR variations, and for that you need to distinguish between normal and abnormal differences between observed residual and it's estimate. Normal difference is Gaussian, with relatively common single outliers but very rare 2 or 3 consecutive outliers.

Here, I use instant filtering using noise-reduction (NR) approach by Patrick J. Wolfe and Simon J. Godsill "Efficient alternatives to the Ephraim-Malah Suppression rule for audio signal enhancement", Feb 2003. This instant NR approach is not the best for detecting small variations:

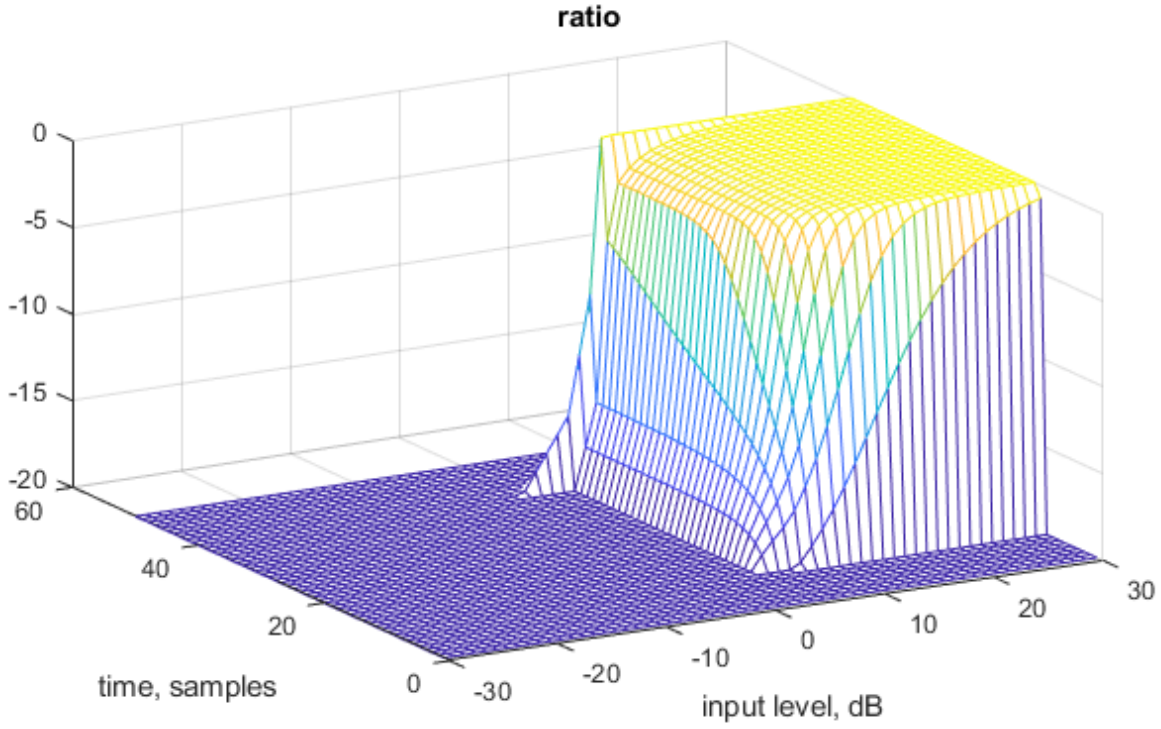


Actually, you need to accumulate the differences between observed residual and it's estimate over a longer time to exclude excessive false positives. Then you need to let adaptive filter converge, and only then apply correction again.

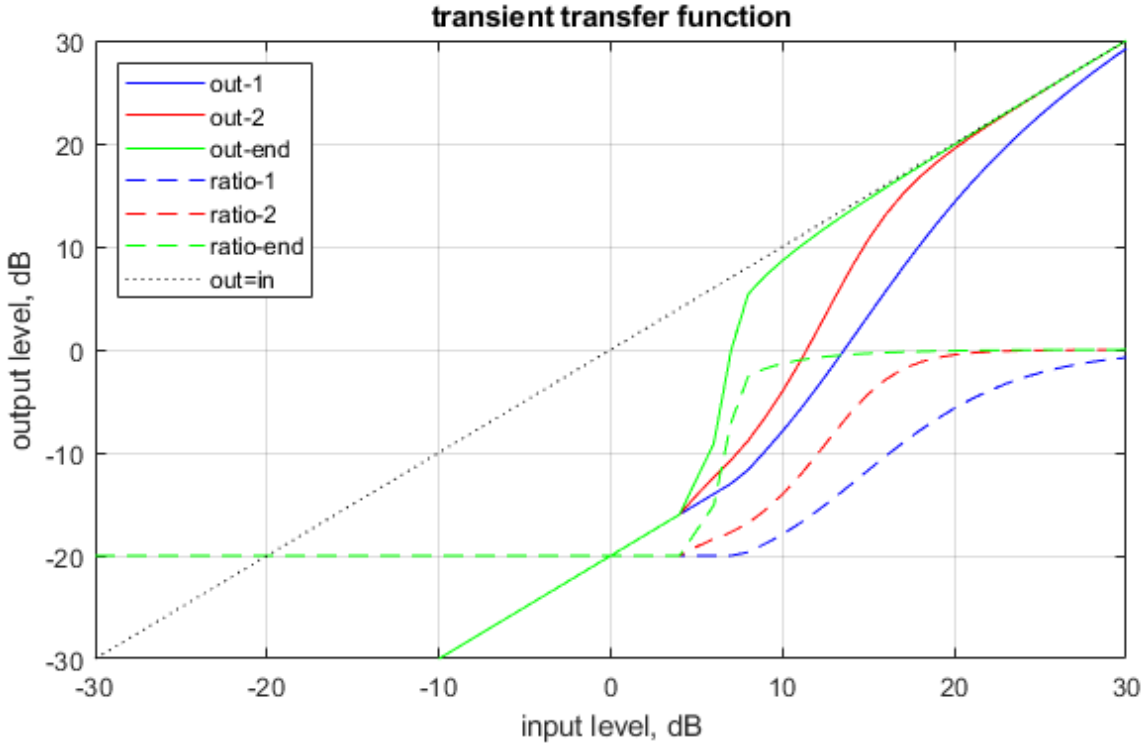
The instant NR principles can be demonstrated with input of a pulse of varying amplitude, starting at t=10, and ending at t=40:



You can see that the averaging filter τ depends on the pulse amplitude. The very first sample will be passed if and only if it's very strong, so a single noisy outlier will be almost always suppressed. But if such outliers are repeated, they will be passed through. The transfer ratio is funny on the falling edge, on transition between activity and silence. The very first "silence" is amplified, but then the ratio drops to zero.



Alternatively, you can see it in 2D, with transition point intentionally biased about +4dB:



There is no need to look for any speech-processing-specific meaning in that simple NR.