CCIT Report #402 October 2002

Sequential-Joint Estimation of Signal and Parameters Using the Unscented Kalman Filter with Application to Single- and Multi-Microphone Speech Enhancement

Sharon Gannot

Department of Electrical Engineering, Technion — Israel Institute of Technology,
Technion City, Haifa 32000, Israel
E-mail: gannot@siglab.technion.ac.il;
Tel.: +972 4 8294756; Fax: +972 4 8323041.

Marc Moonen

Department of Electrical Engineering, ESAT-SISTA
K.U.Leuven,
Kasteelpark Arenberg 10,
3001 Leuven-Heverlee, Belgium
E-mail: marc.moonen@esat.kuleuven.ac.be

I. Introduction

The problem of estimating both signals and parameters arises in many applications, such as,

- Sensor array direction finding.
- Multi path distinction.
- Single microphone signal enhancement.
- Multi microphone signal enhancement.

The most commonly used procedure for solving this estimation problem, when a statistical model with unknown parameters is given, is the *estimate-maximize* (EM) method [1]. This method is essentially an iterative solution for the *maximum-likelihood* (ML) parameter estimation. The signal (or its statistics) estimation is usually a by-product of the algorithm. The ML estimator looks for the parameters which explain the observation in the best way,

$$\max_{\boldsymbol{\theta}} \log f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\theta}) \rightarrow \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$$

where, $\boldsymbol{z}$ is the *Observed data* (measurements). The iterative solution works with the *complete data* notation. Let,

$$\boldsymbol{z} = \mathcal{H}(\boldsymbol{y})$$

where, $\boldsymbol{y}$ is the *Complete data* and $\mathcal{H}$ is some arbitrary non-invertible transform. Then, instead of solving the original problem, we might solve the following problem.

$$\max_{\boldsymbol{\theta}} \log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) \rightarrow \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$$

This is done by applying the following iterations

For $l = 0, 1, 2, \ldots$ iterate between:

Estimation:

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}) = E\{\log f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta}) | \boldsymbol{z}, \hat{\boldsymbol{\theta}}^{(l)}\}$$

Maximization:

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}) = 0 \Rightarrow \hat{\boldsymbol{\theta}}^{(l+1)}$$

Thus, In the E-step the signals' statistics is estimated using the current parameters values, and in the M-step the parameters are estimated using the current statistics. The idea is also summarized in Figure 1.

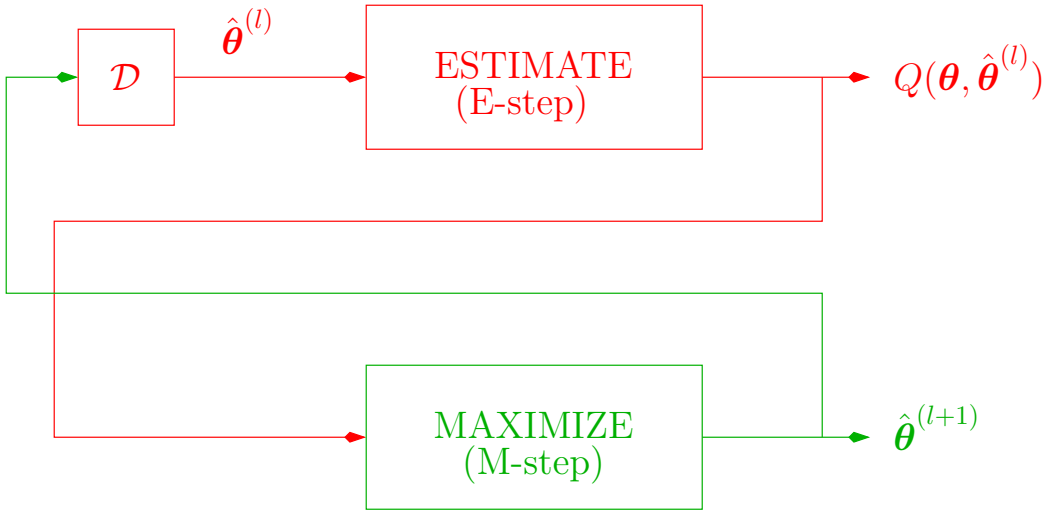


Fig. 1. The EM algorithm.

It can be shown that the EM procedure have the following properties:

- $\hat{\boldsymbol{\theta}}^{(l)} \overset{l \to \infty}{\longrightarrow} \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$. The likelihood is increasing at each iteration.
- Expectation stage used for signal estimation.
- Initialization: local minimum may be reached.
- *Complete data* controls complexity and convergence rate.

Usually, this batch in nature procedure is computationally too complicated and a sequential approximation is searched for (see for instance [2],[3]). In this report we will address such sequential solution for the problem in hand. We will concentrate on the newly suggested Unscented Kalman filter and suggest some application in the field of speech processing.

The structure of this report is as follows. In Section II the unscented transform and its application to non-linear Kalman filter is presented. Sections III,IV and V discuss application of the method to the problems of single microphone speech enhancement, two microphone speech enhancement and two microphone speech dereverberation, respectively. We draw some conclusions and discuss some further directions in Section VI.

II. Preliminaries

A. The Unscented Transform

Let $\boldsymbol{x}$ be an L-dimensional random vector with mean $\bar{\boldsymbol{x}}$ and covariance matrix P_{xx}.
Let,

$$\boldsymbol{y} = f(\boldsymbol{x})$$

be a nonlinear transformation from the random vector $\boldsymbol{x}$ to another random vector $\boldsymbol{y}$. The first and second order statistics of the vector $\boldsymbol{y}$ is to be calculated.

The unscented transform is a method for calculating the statistics of a random variable which undergoes a nonlinear transformation and it was first suggested by Julier and Uhlmann [4],[5],[6] and was further extended by Wan *et al.* [7],[8],[9],[9],[10],[11],[12],[13],[14],[15],[16]. Matlab© code by van der Merwe *et al.* [17] and by Gannot [18] and a preliminary toolbox by Julier [19] are available.

We will briefly summarize the method. The mean and covariance of $\boldsymbol{x}$ are represented by the following $2L+1$ points and weights

$$\begin{aligned}
\mathcal{X}_0 &= \bar{\boldsymbol{x}} \\
\mathcal{X}_l &= \bar{\boldsymbol{x}} + \left(\sqrt{(L+\lambda)P_{xx}}\right)_l ;\ l = 1,\ldots,L \\
\mathcal{X}_{l+L} &= \bar{\boldsymbol{x}} - \left(\sqrt{(L+\lambda)P_{xx}}\right)_l ;\ l = 1,\ldots,L \\
W_0^{(0)} &= \frac{\lambda}{L+\lambda} \\
W_0^{(c)} &= \frac{\lambda}{L+\lambda} + (1-\alpha^2+\beta) \\
W_l^{(m)} &= W_l^{(c)} = \frac{1}{2(L+\lambda)};\ l = 1,2,\ldots,2L
\end{aligned}$$

where, $\left(\sqrt{(L+\lambda)P_{xx}}\right)_l$ is the l-th row or column of the corresponding matrix square root, and $\lambda = \alpha^2(L+\kappa) - L$. α determines the spread of the sigma points, we used in our simulations $\alpha = 1$. κ is a secondary scaling parameter, if $\kappa = 3 - L$ the kurtosis of a Gaussian vector is maintained. Through our simulations κ is set to 0. β is used to incorporate prior knowledge of the distribution ($\beta = 2$ for Gaussian distributions). A proper choice of these parameters and its influence on the achieved performance is still an open topic.

The mean and covariance of the vector $\boldsymbol{y}$ are calculated using the following procedure,

1. Construct the sigma points $\mathcal{X}_l,\ l = 0,\ldots,2L$.
2. Transform each point: $\mathcal{Y}_l = f\left(\mathcal{X}_l\right),\ l = 0,\ldots,2L$.
3. Estimate the mean by weighted averaging: $\bar{\boldsymbol{y}} \approx \sum_{l=0}^{2L} W_l^{(m)} \mathcal{Y}_l$.
4. Estimate the covariance by weighted outer product: $P_{yy} \approx \sum_{l=0}^{2L} W_l^{(c)} \left(\mathcal{Y}_l - \bar{\boldsymbol{y}}\right)\left(\mathcal{Y}_l - \bar{\boldsymbol{y}}\right)^T$.

The advantage of using the unscented transform is evident from the following figures. In Figure 2 a Monte-Carlo method for estimating the mean and covariance of random vector is presented. Random number generator is generating many points. Each of them is transformed by the (known) non-linear transformation. Sample mean and covariance are calculated using the transformed points. An

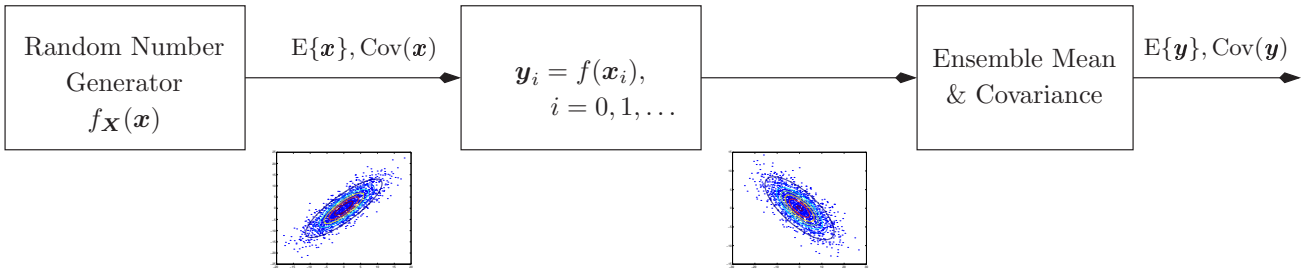


Fig. 2. Monte Carlo

alternative and more efficient method based on carefully chosen *sigma points* is presented in Figure 3.

Only $2L+1$ points are needed, if the points are scattered around the mean vector in accordance to the square root of the covariance matrix.

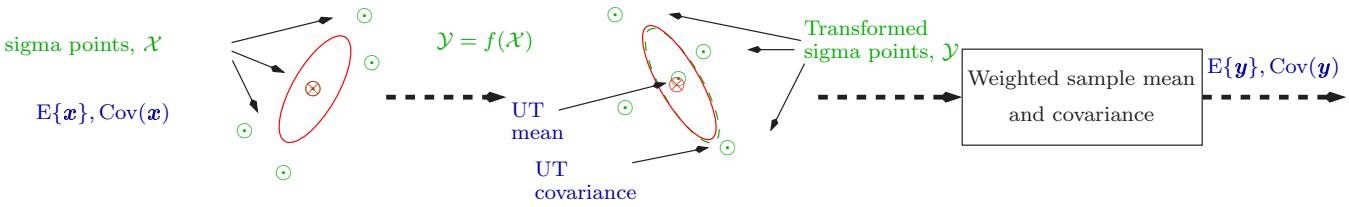


Fig. 3. The unscented transform

B. *The application of the Unscented transform to the non-linear Kalman filtering problem*

The Kalman smoother is a recursive solution for the MMSE linear estimator. An optimal casual solution is given by the Kalman filter. The Kalman equations are formulated with the state-space notation. As depicted in Figure 4 the Kalman filter constitutes of two stages. *Propagation* stage in

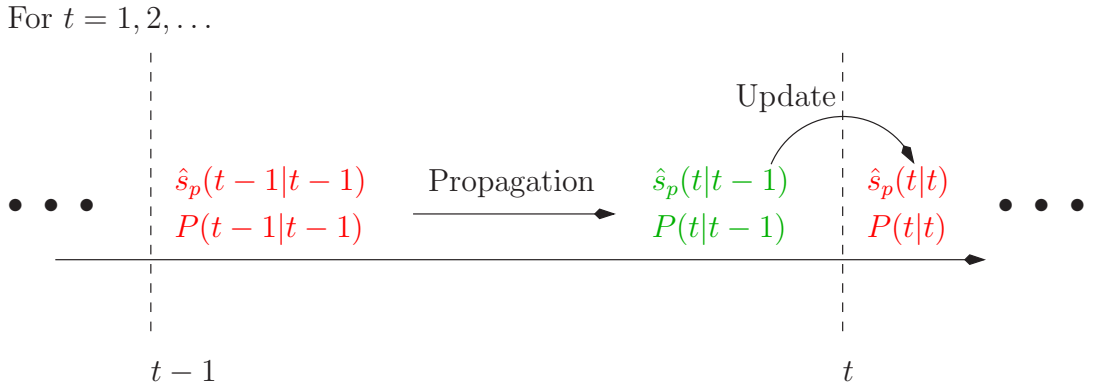


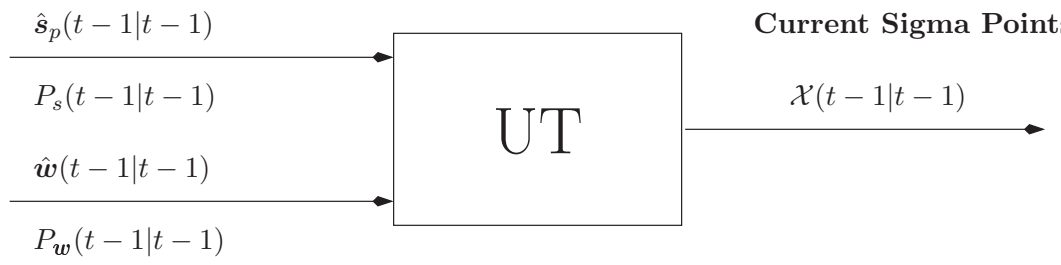
Fig. 4. Propagation and update stages of the Kalman filter

which the state is predicted based on the system dynamics and the previous time instance estimate, and an *update* stage in which this prediction is optimally weighted with the new measurement. The error covariance, interpreted as the amount of confidence we have in the estimate is propagated in a similar fashion.

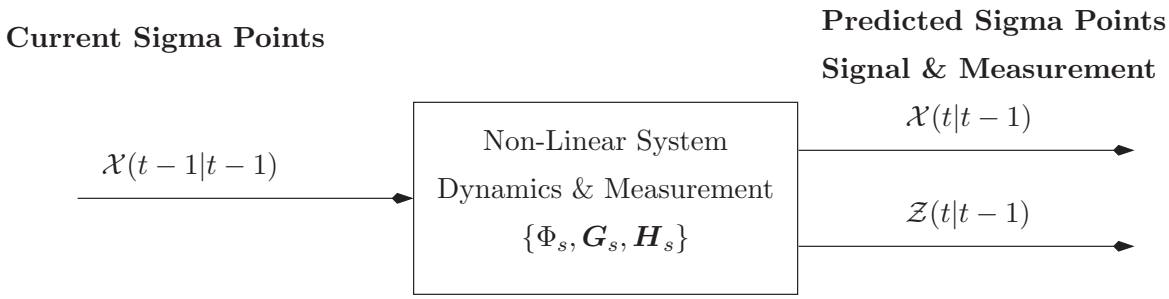
When the system dynamics and the measurement equation are linear, all the calculations involved are straight forward. The situation is more complex when the involved equations are non-linear. In this case a method for propagating mean and covariance through non-linearities is needed. In the past the *extended Kalman filter* (EKF), based on linearizing of the equations, was used. This method might be quite complex, as it involves the calculation of derivatives, but yet not accurate enough, as only first-order approximation is applied.

A better method, suggested by [4], is to use the Unscented transform in order to propagate the mean and covariance through the non-linearities. Figure 5 shows the steps involved in this non-linear Kalman filter. The method consists of calculating the mean and covariance of the state vectors undergoing a non-linear transform (that might be either speech or parameters or both) by virtue of the unscented transform. The complexity of the suggested method is quite low since only increase of dimensions by $2L+1$ is needed.

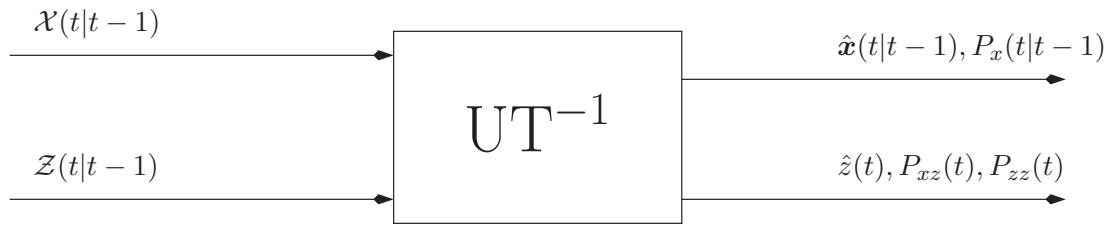
Claims for optimality in ML or MAP sense are presented in [4].



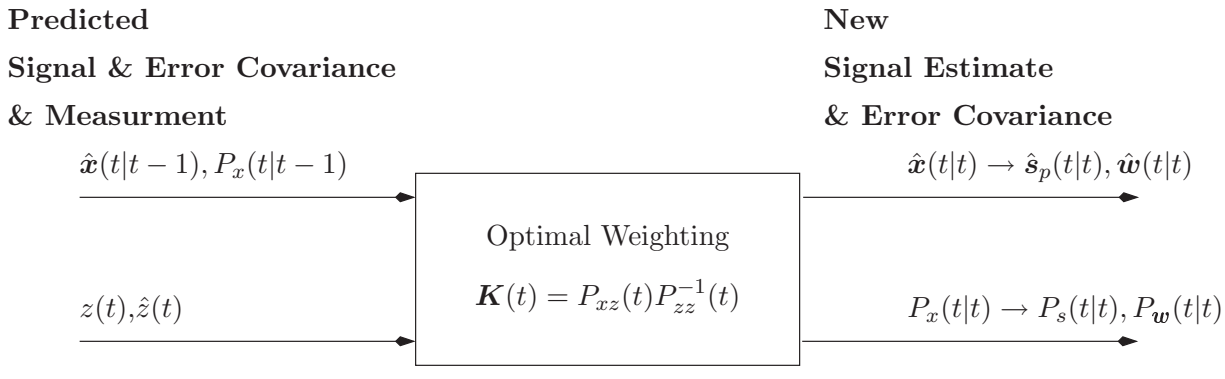
(a)



(b)



(c)



(d)

Fig. 5. Unscented Kalman filter: (a) Unscented transform. (b) Propagation equations. (c) Inverse unscented transform. (d) Update equations.

C. Application to speech processing

In many model-based problems in speech processing (e.g. single microphone speech enhancement, multi-microphone speech enhancement and dereverberation) a problem of estimating both the speech signal and various parameters arises. These problems can be addressed in two ways. In the first, referred to as *Dual estimation*, a two step approach is taken. In each time instance a Kalman filtering step for the signal is applied based on the current estimate of the parameters. In parallel a parameter estimate step is applied based on the current signal state estimate. The parameter estimation might be conducted using recursive methods such as RLS or LMS. Other option is to give the parameters a dynamic model and to use the Kalman filter. This approach will be used through out this report.

The *Dual estimation* method can be seen as a sequential variant of the EM procedure, but no claims of optimality are valid. Discussion on the subject can be found in [3]. The method is summarized in Figure 6.

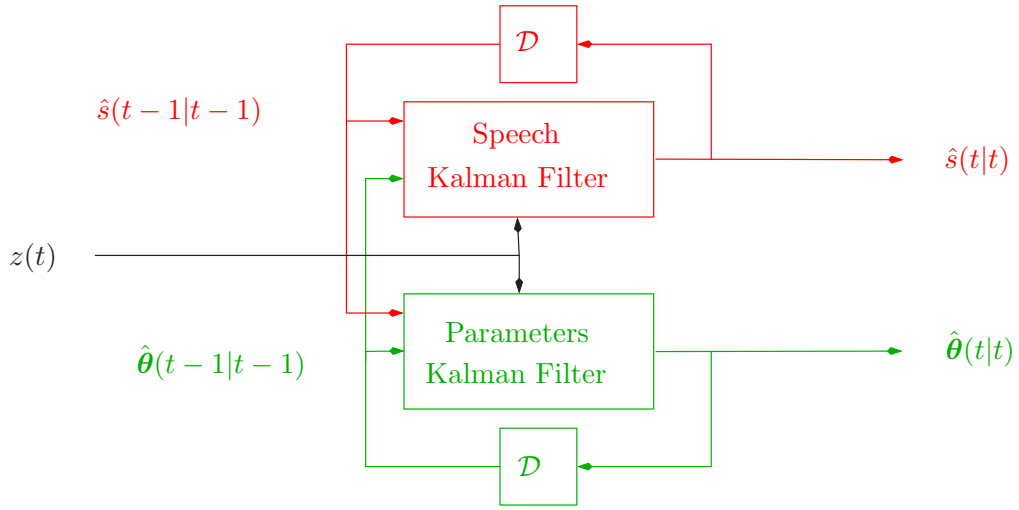


Fig. 6. Dual estimation procedure

The same problem can be reformulated. Note that most operations involve parameter and state vector multiplications. Thus, the problem of *Joint estimation* of the speech and the parameters becomes non-linear if both are modeled as stochastic processes. We remark that as this non-linearity is *separable* this formulation might lead to the same performance as in the Dual scheme. This subject is still under investigation. The approach of jointly estimating speech signal and its parameters is summarized in Figure 7.

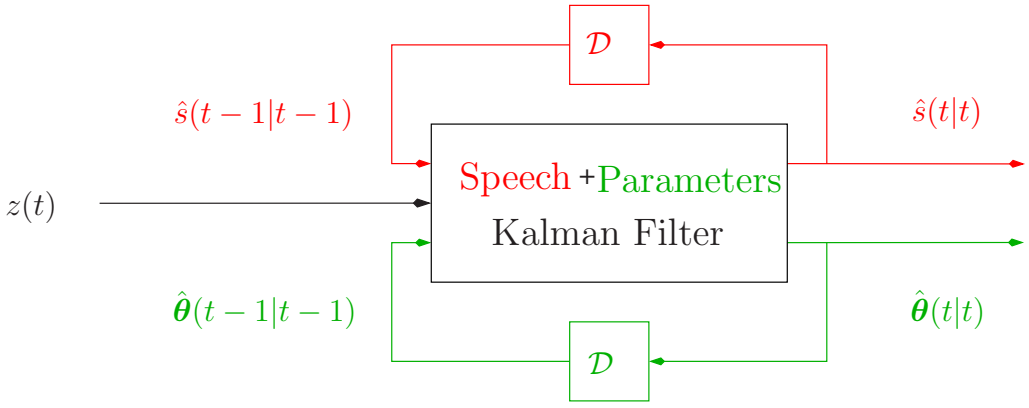


Fig. 7. Joint estimation procedure

III. Single Microphone Speech Enhancement

The problem of single-microphone speech enhancement was extensively studied. Specifically, the use of Kalman filter for estimating both signal and parameters. Two versions of an algorithm are presented by Gannot *et al.* [3]. The first is batch in nature, i.e. the noisy signal is divided into frames, EM iterations are performed in each frame iterating between Kalman filtering, using the current parameters, and a modified version of *Yule-Walker* equations for parameter estimation, using the current signal estimate. A computationally more efficient method is constructed by applying a sequential solution for the parameters estimation using RLS or LMS type methods. Another variant of the sequential solution uses a separate Kalman filter for the parameter estimation. We will concentrate on the latter. These estimation methods are within the framework *Dual* estimation. The problem remains linear as, in each time instance, the signal Kalman filter assumes the AR parameters are known and the parameter estimation assumes the signal is known.

A. Signal and parameter models

A.0.a Speech model. Let the signal measured by the microphone be given by:

$$z(t) = s(t) + v(t) \tag{1}$$

where $s(t)$ represents the sampled speech signal and $v(t)$ represents an additive background noise.

We shall assume time varying LPC model for the speech signal, i.e.

$$s(t) = -\sum_{k=1}^{p} \alpha_k(t) s(t-k) + g_s(t) u_s(t) \tag{2}$$

where the excitation $u_s(t)$ is a normalized (zero mean unit variance) white noise. $g_s(t)$ represents the innovation gain, and $\alpha_1(t), \alpha_2(t), \ldots, \alpha_p$ are the AR coefficients. The additive noise $v(t)$ is also assumed to be a realization from a zero mean white Gaussian stochastic AR process with variance g_v^2.

Eq. 1 and Eq. 2 may be represented in a state-space form:

$$\begin{aligned} \boldsymbol{s}_p(t) &= \Phi_s(t)\boldsymbol{s}_p(t-1) + \boldsymbol{g}_s(t)u_s(t) \\ z(t) &= \boldsymbol{h}_s^T \boldsymbol{s}_p(t) + v(t) \end{aligned} \tag{3}$$

where

$$\boldsymbol{s}_p^T(t) = \begin{bmatrix} s(t) & s(t-1) & \ldots & s(t-p) \end{bmatrix}$$

The state transition matrix $\Phi_s(t)$ is given by:

$$\Phi_s(t) = \begin{bmatrix} -\alpha_1(t) & -\alpha_2(t) & \cdots & \cdots & -\alpha_p(t) & 0 \\ 1 & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 & 0 \end{bmatrix}$$

$\boldsymbol{g}_s(t)$ is

$$\boldsymbol{g}_s^T(t) = \begin{bmatrix} g_s(t) & \ldots & 0 & 0 \end{bmatrix}$$

and $\boldsymbol{h}_s$ is

$$\boldsymbol{h}_s^T = \begin{bmatrix} 1 & \ldots & 0 & 0 \end{bmatrix}$$

A.0.b Parameter model.

$$\begin{aligned} \boldsymbol{\alpha}(t) &= \Phi_{\boldsymbol{\alpha}}\boldsymbol{\alpha}(t-1) + u_{\boldsymbol{\alpha}}(t) \\ z(t) &= \boldsymbol{h}_{\boldsymbol{\alpha}}^T(t)\boldsymbol{\alpha}(t) + \boldsymbol{g}_s(t)u_s(t) + v(t) \end{aligned} \tag{4}$$

where,

$$\boldsymbol{h}_{\boldsymbol{\alpha}}^T(t) = \begin{bmatrix} s(t-1) & s(t-2) & \dots & s(t-p) \end{bmatrix}$$

and $\Phi_{\boldsymbol{\alpha}} = I_{p\times p}$ or very close to it. Define also,

$$\boldsymbol{g}_{\boldsymbol{\alpha}}^T = \underbrace{\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}}_{p}$$

B. Dual Scheme

In one hand, assuming that all the signal and noise parameters are known, which implies that $\Phi_s(t)$, $\boldsymbol{h}_s$ and $\boldsymbol{g}_s(t)$ are known, the optimal casual minimum mean square error (MMSE) linear state estimate, which includes the desired speech signal $s(t)$, is obtained using the Kalman filtering equations. And the other hand, assuming the speech signal is known, i.e. $\boldsymbol{h}_{\boldsymbol{\alpha}}^T(t)$ is known, a Kalman filter for the parameter estimate is applied.

Since both signal and parameters are not known, the dual scheme presented in Figure 6 may be applied. In each time instance the AR parameters are estimated using the estimated speech signal and the speech signal is estimated using the current parameter estimate.

B.1 Speech Kalman filtering

B.1.a Propagation equations:.

$$\begin{aligned} \hat{\boldsymbol{s}}_p(t|t-1) &= \Phi_s\hat{\boldsymbol{s}}_p(t-1|t-1) \\ P(t|t-1) &= \Phi_s P(t-1|t-1)\Phi_s^T + \boldsymbol{g}_s\boldsymbol{g}_s^T \end{aligned} \tag{5}$$

B.1.b Kalman gain:.

$$\boldsymbol{k}(t) = \frac{P(t|t-1)\boldsymbol{h}_s}{\boldsymbol{h}_s^T P(t|t-1)\boldsymbol{h}_s + g_v^2} \tag{6}$$

B.1.c Update equations:.

$$\begin{aligned} \hat{\boldsymbol{s}}_p(t|t) &= \hat{\boldsymbol{s}}_p(t|t-1) + \boldsymbol{k}(t)\left[z(t) - \boldsymbol{h}_s^T\hat{\boldsymbol{s}}_p(t|t-1)\right] \\ P(t|t) &= P(t|t-1) - \boldsymbol{k}(t)\left[\boldsymbol{h}_s^T P(t|t-1)\boldsymbol{h}_s + g_v^2\right]\boldsymbol{k}^T(t) \end{aligned} \tag{7}$$

B.2 Parameters Kalman filtering

B.2.a Propagation equations:.

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(t|t-1) &= \Phi_{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}(t-1|t-1) \\ P_{\boldsymbol{\alpha}}(t|t-1) &= \Phi_{\boldsymbol{\alpha}}P_{\boldsymbol{\alpha}}(t-1|t-1)\Phi_{\boldsymbol{\alpha}}^T + \boldsymbol{g}_{\boldsymbol{\alpha}}\boldsymbol{g}_{\boldsymbol{\alpha}}^T \end{aligned} \tag{8}$$

B.2.b Kalman gain:.

$$\boldsymbol{k}_{\boldsymbol{\alpha}}(t) = \frac{P_{\boldsymbol{\alpha}}(t|t-1)H_{\boldsymbol{\alpha}}}{\boldsymbol{h}_{\boldsymbol{\alpha}}^T P_{\boldsymbol{\alpha}}(t|t-1)\boldsymbol{h}_{\boldsymbol{\alpha}} + g_u^2 + g_v^2} \tag{9}$$

B.2.c Update equations:.

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(t|t) &= \hat{\boldsymbol{\alpha}}(t|t-1) + \boldsymbol{k}_{\boldsymbol{\alpha}}(t)\left[z(t) - \boldsymbol{h}_{\boldsymbol{\alpha}}^T\hat{\boldsymbol{\alpha}}(t|t-1)\right] \\ P_{\boldsymbol{\alpha}}(t|t) &= P_{\boldsymbol{\alpha}}(t|t-1) - \boldsymbol{k}_{\boldsymbol{\alpha}}(t)\left[\boldsymbol{h}_{\boldsymbol{\alpha}}^T P_{\boldsymbol{\alpha}}(t|t-1)\boldsymbol{h}_{\boldsymbol{\alpha}} + g_u^2 + g_v^2\right]\boldsymbol{k}_{\boldsymbol{\alpha}}^T(t) \end{aligned} \tag{10}$$

The dual scheme suggested in Figure 6 is then used.

C. Joint scheme

An augmented state vector of the speech and the parameters is constructed.

$$\boldsymbol{x}^T(t) = \left[\begin{array}{cc} \boldsymbol{s}_p(t) & \boldsymbol{\alpha}(t) \end{array} \right]$$

Then,

$$\begin{aligned} \boldsymbol{x}(t) &= \underbrace{\Phi \boldsymbol{x}(t-1)}_{\text{non-linearity}} + \left[\begin{array}{c} \boldsymbol{g}_s(t) u_s(t) \\ \boldsymbol{g}_{\boldsymbol{\alpha}} \boldsymbol{\alpha}(t) \end{array} \right] \\ z(t) &= \underbrace{\left[\begin{array}{cc} \boldsymbol{h}_s^T & \boldsymbol{h}_{\boldsymbol{\alpha}}^T(t) \end{array} \right] \boldsymbol{x}(t)}_{\text{non-linearity}} + v(t) \end{aligned} \tag{11}$$

This set of equation is non-linear since it involves a multiplication of the speech state space and the transition matrix comprised of the parameters process. So, the joint scheme suggested in Figures 5,7 can be used.

D. Results

Time varying Gaussian AR process (4 coefficients) embedded in white Gaussian noise with input SNR level of about 20dB is processed by the joint Kalman scheme. The noise level is estimated during non-signal portions of the noisy signal. The tracking ability of the parameters is presented in Figure 8.

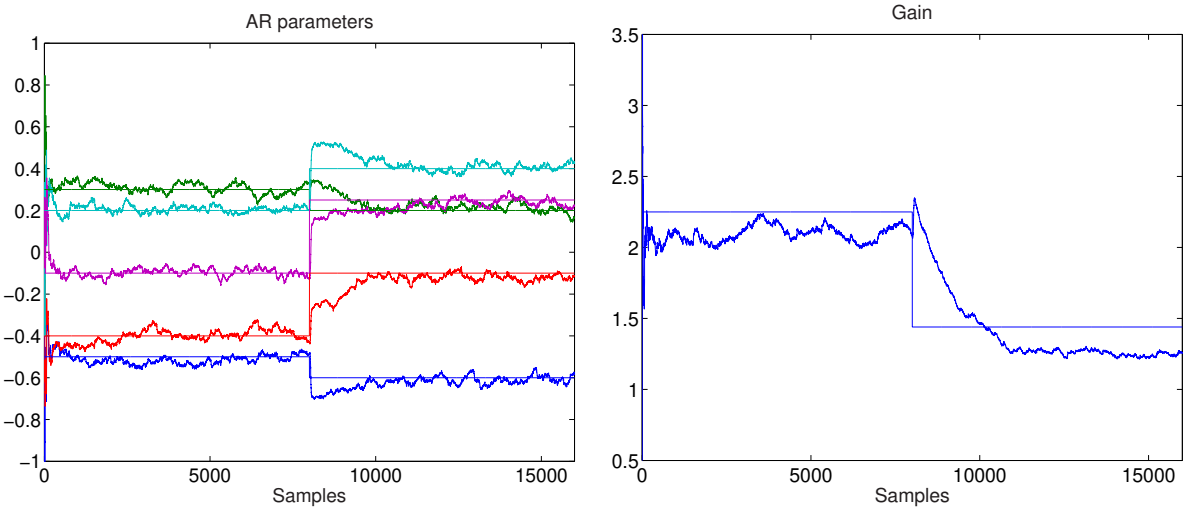


Fig. 8. Results

IV. Two Microphones Speech Enhancement

The basic system of interest consists of a desired (speech) signal source and a noise source both present in the same acoustic environment, say a living room or an office. We install two microphones. The speech and the noise are both coupled into each microphone by the acoustic field in the environment. An extra low-level white Gaussian noise signal is picked by each microphone representing some sensor noise.

A. Speech and parameters model

A.0.d Speech model. The mathematical model for the received signals, assuming FIR model for the room acoustics, is given by, (see [2]):

$$\begin{aligned} z_1(t) &= s(t) + \sum_{k=0}^{n_b-1} b_k v(t-k) + e_1(t) \\ z_2(t) &= \sum_{k=0}^{n_a-1} a_k s(t-k) + v(t) + e_2(t) \end{aligned} \tag{12}$$

Speech and noise signals are both modeled as AR processes:

$$\begin{aligned} s(t) &= -\sum_{k=1}^{p} \alpha_k s(t-k) + g_s u_s(t) \\ v(t) &= -\sum_{k=1}^{q} \beta_k v(t-k) + g_v u_v(t) \end{aligned} \tag{13}$$

$e_1(t)$ and $e_2(t)$ are both low-level white noise processes with levels g_1 and g_2, respectively. The speech and noise state-vectors are given by,

$$\begin{aligned} \mathbf{s}_{n_a}^T(t) &= \left[\begin{array}{cccc} s(t) & s(t-1) & \cdots & s(t-n_a+1) \end{array}\right] \\ \mathbf{v}_{n_b}^T(t) &= \left[\begin{array}{cccc} v(t) & v(t-1) & \cdots & v(t-n_b+1) \end{array}\right] \end{aligned} \tag{14}$$

Denote by $\boldsymbol{\theta}(t)$ the vector of unknown parameters:

$$\boldsymbol{\theta} = \left[\begin{array}{cccccccc} \boldsymbol{\alpha}(t) & g_s(t) & \boldsymbol{\beta}(t) & g_v(t) & \boldsymbol{a}(t) & \boldsymbol{b}(t) & g_1 & g_2 \end{array}\right]^T \tag{15}$$

Define a state-space form of the equations (assuming $n_a > p$ and $n_b > q$):

$$\begin{aligned} \left[\begin{array}{c} \boldsymbol{s}_{n_a}(t) \\ \boldsymbol{v}_{n_b}(t) \end{array}\right] &= \left[\begin{array}{cc} \Phi_s(t) & \\ & \Phi_v \end{array}\right] \left[\begin{array}{c} \boldsymbol{s}_{n_a}(t-1) \\ \boldsymbol{v}_{n_b}(t-1) \end{array}\right] + G(t) \left[\begin{array}{c} u_s(t) \\ u_v(t) \end{array}\right] \\ \left[\begin{array}{c} z_1(t) \\ z_2(t) \end{array}\right] &= H^T(t) \left[\begin{array}{c} \boldsymbol{s}_{n_a}(t) \\ \boldsymbol{v}_{n_b}(t) \end{array}\right] + \left[\begin{array}{c} e_1(t) \\ e_2(t) \end{array}\right] \end{aligned} \tag{16}$$

$\Phi_s(t)$ is the $n_a \times n_a$ speech transition matrix and $\Phi_v(t)$ is the $n_b \times n_b$ speech transition matrix:

$$\Phi_s(t) = \underbrace{\left[\begin{array}{cccccccc} -\alpha_1(t) & -\alpha_2(t) & \cdots & -\alpha_p(t) & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ \vdots & & & & & 1 & 0 \end{array}\right]}_{n_a} \quad \Phi_v(t) = \underbrace{\left[\begin{array}{cccccccc} -\beta_1(t) & -\beta_2(t) & \cdots & -\beta_q(t) & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ \vdots & & & & & 1 & 0 \end{array}\right]}_{n_b}$$

G is,

$$G^T(t) = \left[\begin{array}{cccccccc} g_s(t) & 0 & \cdots & 0 & \overset{n_a}{\overset{\downarrow}{0}} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & g_v(t) & 0 & \cdots & 0 \end{array}\right] \tag{17}$$

and H is,

$$H^T(t) = \begin{bmatrix} 1 & 0 & \cdots & 0 & \overset{n_a}{\downarrow} & & & \\ a_1(t) & a_2(t) & \cdots & a_{n_a-1}(t) & b_1(t) & b_2(t) & \cdots & b_{n_b-1}(t) \end{bmatrix} \tag{18}$$

A.0.e Parameters model.

$$\begin{aligned} \boldsymbol{\alpha}(t) &= \Phi_{\boldsymbol{\alpha}}\boldsymbol{\alpha}(t-1) + u_{\boldsymbol{\alpha}}(t) \\ \boldsymbol{a}(t) &= \Phi_{\boldsymbol{a}}\boldsymbol{a}(t-1) + u_{\boldsymbol{a}}(t) \\ \boldsymbol{b}(t) &= \Phi_{\boldsymbol{b}}\boldsymbol{\alpha}(t-1) + u_{\boldsymbol{b}}(t) \end{aligned} \tag{19}$$

$$\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \begin{bmatrix} H^T(t) & H_{\boldsymbol{\alpha}}^T & H_{\boldsymbol{a}}^T & H_{\boldsymbol{b}}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_{n_a}(t) \\ \boldsymbol{s}_{n_b}(t) \\ \boldsymbol{\alpha}(t) \\ \boldsymbol{a}(t) \\ \boldsymbol{b}(t) \end{bmatrix} + G(t) \begin{bmatrix} u_s(t) \\ u_v(t) \end{bmatrix} + \begin{bmatrix} e_1(t) \\ e_2(t) \end{bmatrix}$$

$H_{\boldsymbol{\alpha}}$, $H_{\boldsymbol{a}}$ and $H_{\boldsymbol{b}}$ are all-zeros matrices. It is straight forward to plug this into the framework of both dual and joint Kalman filtering. The joint scheme is again non-linear due to the multiplication operations.

B. Short discussion

The two microphone speech enhancement scenario was tested with both the dual and the joint schemes. Preliminary results suggest that the algorithm is very sensitive to initialization. There are two many parameters to adjust simultaneously and whenever there is no good starting the algorithm tends to converge to an improper solution. A-prior knowledge of one of the filters $\boldsymbol{a}(t)$ or $\boldsymbol{b}(t)$ may be required. Such initialization may be available from a portion of the noisy signals in which only one source is active.

V. Two microphone speech Dereverberation

In the two channel dereverberation problems a speech signal, modeled as an AR process is filtered by an *acoustical transfer function* (ATF), modeled as an FIR filters. Noise is then added to the output constructing the noisy and reverberated speech signals, as depicted in Figure 9.

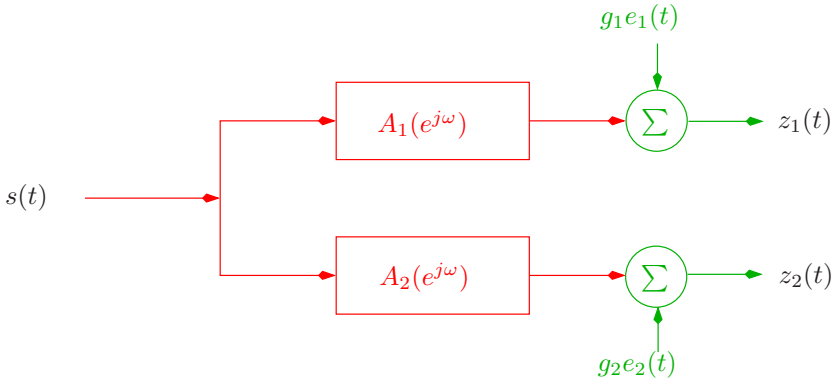


Fig. 9. Two channel dereverberation problem.

A. Signals' model Model

The reverberated and noisy signals in Figure 9 are given by the following model,

$$\begin{aligned} s(t) &= -\sum_{k=1}^{p} \alpha_k s(t-k) + g_{u_s} u_s(t) \\ z_1(t) &= \sum_{k=0}^{n_a-1} a_1(k) s(t-k) + g_1 e_1(t) \\ z_2(t) &= \sum_{k=0}^{n_a-1} a_2(k) s(t-k) + g_2 e_2(t) \end{aligned} \tag{20}$$

Thus, we have again the problem of estimating both speech signal and the following parameters,

$$\boldsymbol{\theta}^T(t) = \left[\begin{array}{cccccc} \boldsymbol{\alpha}(t) & g_{u_s}(t) & \boldsymbol{a}_1(t) & \boldsymbol{a}_2(t) & g_1 & g_2 \end{array} \right]$$

B. Joint Speech and Parameters estimation

Define a state vector

$$\boldsymbol{s}_{n_a}^T(t) = \left[\begin{array}{cccc} s(t) & s(t-1) & \cdots & s(t-n_a+1) \end{array} \right]$$

and the vectors,

$$\begin{aligned} \boldsymbol{g}_s^T(t) &= \underbrace{\left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \end{array} \right]}_{n_a} \\ \boldsymbol{g}_{\boldsymbol{\alpha}}^T(t) &= \underbrace{\left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array} \right]}_{p} \\ \boldsymbol{g}_{\boldsymbol{a}_1}^T(t) &= \underbrace{\left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array} \right]}_{n_a} \\ \boldsymbol{g}_{\boldsymbol{a}_2}^T(t) &= \underbrace{\left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array} \right]}_{n_a} \end{aligned}$$

and the transition matrix,

$$\Phi_s(t) = \underbrace{\left[\begin{array}{cccccccc} -\alpha_1(t) & -\alpha_2(t) & \cdots & -\alpha_p(t) & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ \vdots & & & & & 1 & 0 \end{array} \right]}_{n_a}$$

Then, the augmented transition measurement equations can be written as,

$$\begin{bmatrix} \boldsymbol{s}_{n_a}(t) \\ \boldsymbol{\alpha}(t) \\ \boldsymbol{a}_1(t) \\ \boldsymbol{a}_2(t) \end{bmatrix} = \underbrace{\begin{bmatrix} \Phi_s(t) & 0 & 0 & 0 \\ 0 & I_{p\times p} & 0 & 0 \\ 0 & 0 & I_{n_a\times n_a} & 0 \\ 0 & 0 & 0 & I_{n_a\times n_a} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_{n_a}(t-1) \\ \boldsymbol{\alpha}(t-1) \\ \boldsymbol{a}_1(t-1) \\ \boldsymbol{a}_2(t-1) \end{bmatrix}}_{\text{non-linearity}} + \begin{bmatrix} \boldsymbol{g}_s u_s(t) \\ \boldsymbol{g}_\alpha u_\alpha(t) \\ \boldsymbol{g}_{a_1} u_{a_1}(t) \\ \boldsymbol{g}_{a_2} u_{a_2}(t) \end{bmatrix} \tag{21}$$

$$\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{a}_1(t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{a}_2(t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_{n_a}(t) \\ \boldsymbol{\alpha}(t) \\ \boldsymbol{a}_1(t) \\ \boldsymbol{a}_2(t) \end{bmatrix}}_{\text{non-linearity}} + \begin{bmatrix} g_1 e_1(t) \\ g_2 e_2(t) \end{bmatrix}$$

C. Preliminary results

For a low level white noise signal, which variance is estimated from signal free segments, the tracking ability of the algorithm is presented in Figure 10. It is worth mentioning that the presented problem is a very simple one, the order of the AR process is 1 and the filters $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are 3 taps long. The SNR value is very high. Even in this simple case convergence is not guaranteed.

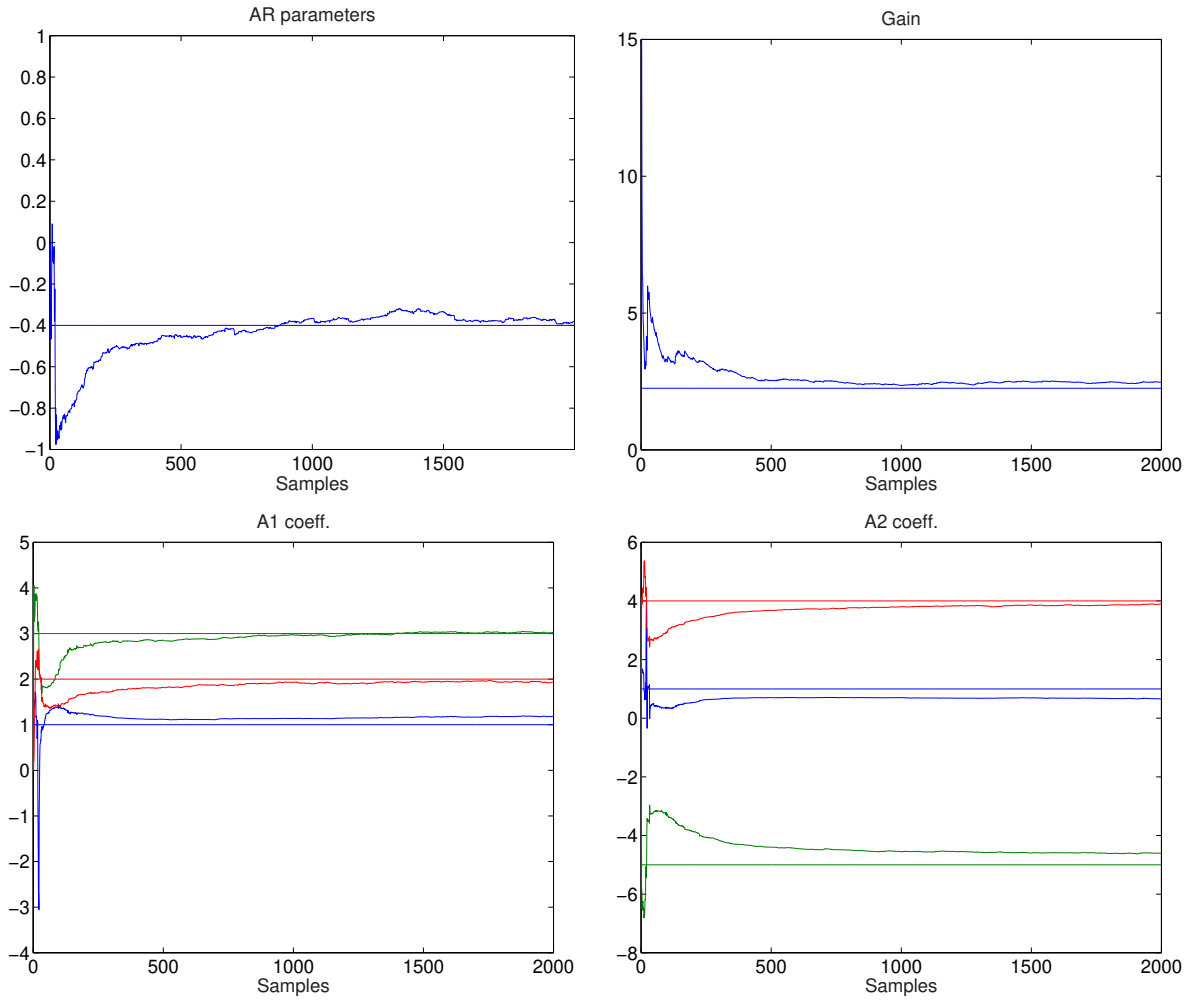


Fig. 10. Tracking ability of the parameters of the the dereverberation problem.

VI. Discussion

In this report we propose to use the newly proposed *unscented Kalman filter* (UKF) for several speech enhancement problems. Results show that although the method is applicable to the problems in hand, it can only solve at this stage simple problems (low dimensional and high SNR). Several issues are under current research,

- The affect of the unscented transform parameters on the convergence rate and the achievable performance.
- Is the behavior of the *dual* and the *joint* solutions different for the separable problems, i.e. problems in which the non-linearity is due to a multiplication of the parameters and the signal.
- The affect of initialization.
- In what sense is the UKF solution optimal.

References

[1] A.P. Dempster N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. Ser. 3g, pp. 1–38, 1977.

[2] M. Feder E. Weinstein, A. V. Oppenheim and J. R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. Signal Processing*, vol. 42, pp. pp. 846–859, 1994.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[4] S.J. Julier and J.K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Tech. Rep. RRG, Dept. of Engineering Science, University of Oxford, England, 1996.

[5] S.J. Julier and J.K. Uhlmann, "A new extension of the kalman to nonlinear systems," in *The 11th International Symposium on AeroSense/Defence Sensing, Simulation and Controls*, Orlando, Florida, USA, 1997, vol. Multi Sensor Fusion, Tracking and Resource Management II.

[6] S. J. Julier, "The scaled unscented transformation," Preprint submitted to Elsevier Preprint (Rev. 3), Jan. 1999.

[7] E.A. Wan, "Homepage," `http://www.ece.ogi.edu/~ericwan/`.

[8] R. van der Merwe and E. A. Wan, "Efficient Derivative-Free Kalman Filters for Online Learning," in *European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, Apr. 2001.

[9] R. van der Merwe and E. A. Wan, "The Square-Root Unscented Kalman Filter for State and Parameter-Estimation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, May 2001.

[10] E. A. Wan, R. van der Merwe and A. T. Nelson, "Dual Estimation and the Unscented Transformation," in *Advances in Neural Information Processing Systems*, S.A. Solla, T.K. Leen and K.-R. Muller, Ed., vol. 12, pp. 666–672. MIT Press, Nov. 2000.

[11] E. A. Wan and R. van der Merwe, "The Unscented Kalman Filter for Nonlinear Estimation," in *Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC)*, Lake Louise, Alberta, Canada, Oct. 2000, IEEE.

[12] A. T. Nelson and E. A. Wan, "A Two-Observation Kalman Framework for Maximum-Likelihood Modeling of Noisy Time Series," in *International Joint Conference on Neural Networks (INNS)*. IEEE, May 1998.

[13] E. A. Wan and A. T. Nelson, "Removal of noise from speech using the dual EKF algorithm," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1998.

[14] A. T. Nelson and E. A. Wan, "Neural Speech Enhancement Using Dual Extended Kalman Filtering," in *International Conference on Neural Networks*, Jun. 1997, pp. 2171–2175.

[15] E. A. Wan and A. T. Nelson, *Kalman Filtering and Neural Networks*, chapter Dual EKF Methods, Simon Haykin. Wiley, 2001.

[16] E. A. Wan and R. van der Merwe, *Kalman Filtering and Neural Networks*, chapter The Unscented Kalman Filter, Simon Haykin. Wiley, 2001.

[17] R. van der Merwe, "Matlab©code," `/users/sista/sgannot/matlab/Ukf_W/`, May 2001.

[18] S. Gannot, "Matlab© code," `/users/sista/sgannot/matlab/Ukf_W/`, Aug. 2001.

[19] S.J. Julier, "Matlab© toolbox," `/users/sista/sgannot/matlab/Ukf_J/`, Aug. 2001.