On the use of Kalman Filter for Enhancing Speech Corrupted by Colored Noise

BOUBAKIR CHABANE [1,2], BERKANI DAOUED [2]
[1]LAMEL, Faculté des Sciences de l'Ingénieur
Université de Jijel.
BP. 98, Ouled Aissa, 18000, Jijel, Algérie.
ALGERIE
[2]Laboratoire Signal et Communications.
Ecole Nationale Polytechnique.
10, av. Hassen Badi, El Harrach, Alger, Algérie.
ALGERIE
chboubakir@yahoo.fr, dberkani@hotmail.com

Abstract: - Kalman filtering is a powerful technique for the estimation of the speech signal observed in additive background noise. This paper presents a contribution in the enhancement of noisy speech with white and colored noise assumption. Some tests were performed with ideal filter parameters, others using the Expectation Maximization (EM) algorithm to iteratively estimate the spectral parameters of the speech and noise. Simulation results show that the application has the best performance evaluated with objective quality scores, observation of the waveforms, as well as informal listening tests in the case of Noizeus database.

Key-Words: - Speech enhancement, Kalman filtering, colored noise, EM algorithm.

1 Introduction

Speech enhancement has been a hot research area in recent years with the fast development of mobile communications systems and other applications. In the presence of additive continuous broadband noise, enhancement of speech remains a challenging task, especially in moderate to high noise levels (SNRs -5 to 10 dB). In many cases, background noise can be deemed as stationary process, whereas speech is short-time stationary. Moreover, real world noise, for instance, vehicle engine noise or radio channel noise tend to be colored, which yields more challenge to speech enhancement than that under the assumption of white noise. Here we investigate into both cases.

A speech enhancement algorithm can be viewed as successful if it suppresses perceivable background noise, and preserves or enhances perceived signal quality. This paper focuses on a single microphone system and the aim is to minimize the effect of noise to improve the speech quality. Many approaches have been investigated in that way. Many of these methods are based on spectral subtraction like power spectral subtraction [1][2], parametric spectral subtraction [3], multi-band subtraction [4][5].Other methods are based on Bayesian approach [6]-[13] like Wiener filtering [6], soft-decision estimation [7] and Minimum Mean Square Error estimation [8][9]. Moreover, other methods have been developed from the state-space approach in which a state equation models the dynamics of the signal generation process and an observation equation models the noisy and distorted observation signal (Kalman filter theory).

The Kaman Filter is a general estimation technique that has been widely used in many areas from tracking to speech enhancement.

The use of Kalman filtering for speech enhancement was first proposed in [14] and later extended to the colored noise case in [15]. The Kalman filter is best suitable for reduction of white noise to comply with Kalman assumption. In deriving Kalman equations it is normally assumed that the process noise is uncorrelated and has a normal distribution. This assumption leads to whiteness character of this noise. There are, however, different methods developed to fit the Kalman approach to colored noise.

A variety of Kalman filter implementations have been proposed for speech enhancement, some

concerned with the speech model [15]-[19], some with parameters estimation schemes [20]. This paper investigates some of theses tools and technologies related to Kalman speech enhancement. Some of the tests are performed with ideal filter parameters of the speech and noise models so that we can obtain a more reliable order q for the different noises of the database. However, for single channel noise suppression these parameters are not available and have to be estimated from the noisy observations. Therefore, additional tests were performed with the EM algorithm.

This paper is organized as follows. We present in section 2 the noisy speech model and Kalman filtering. The section 3 is concerned with the EM algorithm for the parameters estimation of linear Gaussian state-space models, and the different steps of the EM iterative procedure. The section 4 presents the different objectives measures used in the speech quality assessment. In section 5, we provide the experiment results in the case of white and colored noise. Conclusions are drawn in the last section.

2 Noisy Speech Model and Kalman Filtering

Let the noisy signal measured by the microphone be given by:

$$y(n) = s(n) + v(n) \tag{1}$$

Where s(n) represents the sampled speech signal, and v(n) represents additive background noise uncorrelated with the speech signal. It is assumed that speech signal is stationary during each frame, that is, the AR model of speech remains the same across the segment. In order to apply the Kalman filter, we model the speech and noise as autoregressive processes of model order p and q respectively [14]-[19]:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + \sqrt{g_s}\,u(n) \tag{2}$$

$$v(n) = \sum_{j=1}^{q} b_j v(n-j) + \sqrt{g_v}\,w(n) \tag{3}$$

Where u(n) and w(n) are uncorrelated Gaussian normalized white noise sequences with (zero means and unit variances), a_i is the ith AR speech model parameter and b_j is the jth AR noise model parameter. The system of equations (1-3) can be represented in a state-space form:

$$\begin{cases} x(n) = \Phi x(n-1) + Gd(n) \\ y(n) = Hx(n) \end{cases} \tag{4}$$

$x(n) = [s(n-p+1),...,s(n),v(n-q+1),...,v(n)]^T$ is the (p+q)×1 state vector and $d(n) = [0,...,0,u(n),0,...,0,w(n)]^T$.

The explicit expressions for the transition matrix Φ, G and observation row vector H are given below:

$$\Phi = \begin{bmatrix} \Phi_s & 0_{p,q} \\ 0_{p,q} & \Phi_v \end{bmatrix} \tag{5}$$

$$\Phi_s = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{6}$$

$$\Phi_v = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ b_q & b_{q-1} & b_{q-2} & \cdots & b_1 \end{bmatrix} \tag{7}$$

$$G = \begin{bmatrix} g_s & 0_{p,q} \\ 0_{p,q} & g_v \end{bmatrix} \tag{8}$$

Where g_s and g_v are the following p and q dimensional vectors: $g_s^T = [0,...,0,\sqrt{g_s}]$, $g_v^T = [0,...,0,\sqrt{g_v}]$

$$H = \left[\underbrace{0,...,0,1}_{p}, \underbrace{0,...,0,1}_{q} \right] \tag{9}$$

The Kalman algorithm provides a method to compute recursively the minimum mean-squared error estimate $\hat{s}(n)$ from the available noisy observations, which can be retrieved as the pth component of the state-vector estimator $\hat{x}(n/n)$.
The covariance matrices Q=G*G';

3 Parameters Estimation

The Kalman filter is a recursive estimator. This means that only the estimated state from the

previous time step and the current measurement are needed to compute the estimate for the current state. The Kalman filter has two distinct phases: Predict and Update. The predict phase uses the estimate from the previous time step to produce an estimate of the current state. In the update phase, measurement information from the current time step is used to refine this prediction to arrive at a more accurate estimate.

Many approaches using Kalman filtering have been referenced in the literature [14]-[20]. They usually operate in two steps: first, the noise and the signal parameters are estimated, and second, the speech signal is estimated by using Kalman filtering. These approaches differ essentially one from the other by the choice of the algorithm used to estimate the parameters of such model, the models adopted for the speech signal and the additive noise.

There are several methods for extraction of linear prediction (LP) model parameters from noisy observations [21]. In this work, some of the tests are performed with ideal filter parameters so that we can assess the potential of Kalman filter for speech enhancement without worrying about the extraction of theses parameters and the effect of this error on the system. However, for single channel noise suppression these parameters are not available and have to be estimated from the noisy observations. Other methods try to calculate the LP model parameters first and then use them for denoising the speech signal or iteratively estimate and correct these values and enhance the speech (EM algorithm).

The Expectation Maximization (EM) algorithm is used for estimating the parameters of linear dynamic system [20][22][23]. Linear time-invariant dynamical systems, also known as linear Gaussian state-space models, can be described by the following two equations:

$$\begin{cases} x(n) = \Phi x(n-1) + w(n), & \text{n} = 1,\ldots,\text{N} \\ y(n) = Hx(n) + v(n), & \text{n} = 1,\ldots,\text{N} \end{cases} \quad (10)$$

In this paper, the parameters of the system are estimated on a frame by frame basis, reference [22] provides the EM iterative procedure in the case of a sequence of N output vectors $(y_1, y_2, \ldots, y_N)$.

3.1 The Expectation Step

From the initial estimators $\mu(0), \Phi(0), Q(0)$ and $R(0)$ calculate $x_n^N = E(x_n / y_1, \ldots, y_N)$ and $P_n^N = \text{cov}(x_n / y_1, \ldots, y_N)$, with the following Kalman filter forward recursions, for n = 1,…,N:

Predict

Predicted state

$$x_n^{n-1} = \Phi_n x_{n-1}^{n-1} \quad (11)$$

Predicted estimate covariance

$$P_n^{n-1} = \Phi_n P_{n-1}^{n-1} \Phi_n^T + Q_n \quad (12)$$

Update

Optimal Kalman gain

$$K_n = P_n^{n-1} H^T \left(H P_n^{n-1} H^T + R_n \right)^{-1} \quad (13)$$

Update state estimate

$$x_n^n = x_n^{n-1} + K_n \left(y_n - H x_n^{n-1} \right) \quad (14)$$

Update estimate covariance

$$P_n^n = P_n^{n-1} - K_n H P_n^{n-1} \quad (15)$$

Where we take $x_0^0 = \mu$ and $P_0^0 = \Sigma$.

Since the speech signal is often assumed stationary during an analysed frame (20-30 ms), the Kalman smoother can be carried out and provides better estimates of the state since it is based on a higher number of observations.
In order to calculate x_n^N and P_n^N one performs the set of backward recursions (smoothing) [23]: For n = N,N-1,…,1 on the equations

$$J_{n-1} = P_{n-1}^{n-1} \Phi_n^T \left(P_n^{n-1} \right)^{-1} \quad (16)$$

$$x_{n-1}^N = x_{n-1}^{n-1} + J_{n-1} \left(x_n^N - \Phi_n x_{n-1}^{n-1} \right) \quad (17)$$

$$P_{n-1}^N = P_{n-1}^{n-1} + J_{n-1} \left(P_n^N - P_n^{n-1} \right) J_{n-1}^T \quad (18)$$

We also require the covariance $P_{n,n-1}^N = \text{cov}(x_n, x_{n-1} / y_1, \ldots, y_N)$ which can be obtained through the backward recursions.

For n = N,N-1,…,2

$$P_{n-1,n-2}^N = P_{n-1}^{n-1} J_{n-2}^T + J_{n-1} \left(P_{n,n-1}^N - \Phi_n P_{n-1}^{n-1} \right) J_{n-2}^T \quad (19)$$

Which is initialized

$$P_{N,N-1}^{N} = (I - K_N H)\Phi_N P_{N-1}^{N-1} \quad (20)$$

3.2 The Maximization Step

Estimate $\mu_0(1) = x_0^N$ and fix the value of Σ at some reasonable baseline level. Get $\Phi(1), Q(1),$ and $R(1)$ respectively with the equations:

$$\Phi(r+1) = BA^{-1} \quad (21)$$

$$Q(r+1) = \frac{1}{N}\left(C - BA^{-1}B^T\right) \quad (22)$$

And

$$R(r+1) = \frac{1}{N}\sum_{n=1}^{N}[(y_n - Hx_n^N)(y_n - Hx_n^N)^T + HP_n^N H^T] \quad (23)$$

$$A = \sum_{n=1}^{N}\left(P_{n-1}^N + x_{n-1}^N x_{n-1}^{N\,T}\right) \quad (24)$$

$$B = \sum_{n=1}^{N}\left(P_{n,n-1}^N + x_n^N x_{n-1}^{N\,T}\right) \quad (25)$$

$$C = \sum_{n=1}^{N}\left(P_n^N + x_n^N x_n^{N\,T}\right) \quad (26)$$

3- Repeat 1 and 2 above until the estimates and the log likelihood function are stable.

$$LL = -\frac{1}{2}\sum_{n=1}^{N}\left(y_n - Hx_n^{n-1}\right)^T\left(HP_n^{n-1}H^T + R_n\right)^{-1}\left(y_n - Hx_n^{n-1}\right) - \frac{1}{2}\sum_{n=1}^{N}\log\left|HP_n^{n-1}H^T + R_n\right| \quad (27)$$

In this case of the extended model and from the equations (4) and (10), we can observe that $w(n) \equiv G.d(n)$ and $v(n) \equiv 0$, so we need to calculate only Φ and Q, but R=0.

4 Objective Quality Results

To measure quality of the enhanced signal, we have used the segmental SNR, the Log-Likelihood Ratio measure (LLR), the Weighted Spectral Slope measure (WSS) [24] and the Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862) [25]. All the measures show high correlation with subjective quality.

The WSS measure is based on an auditory model and finds a weighted difference between the spectral slopes in each band. The magnitude of each weight reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum. A per-frame measure in decibels is found as,

$$d_{wss}(j) = K_{spl}\left(K - \hat{K}\right) + \sum_{k=1}^{36} w_a(k)\left(S(k) - \hat{S}(k)\right)^2 \quad (28)$$

Where $K, \hat{K}$ are related to overall sound pressure level of the original and enhanced speech, and K_{spl} is a parameter which can be varied to increase overall performance.

The LLR measure is given by:

$$d_{LLR}\left(\vec{a}_d, \vec{a}_\phi\right) = \log\left(\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right) \quad (29)$$

Where $\vec{a}_\phi$ and $\vec{a}_d$ represent the linear prediction (LP) coefficient vectors for the clean and processed speech frame respectively, and R_ϕ is the autocorrelation matrix of the clean speech signal. The LLR is a spectral distance measure which mainly models the mismatch between the formants of the original and enhanced signals. The mean LLR value was obtained by averaging the individual frame LLR values across the sentence.
The highest 5 % of the LLR and WSS measures values were discarded, as suggested in [24], to exclude unrealistically high spectral distance values. The lower the LLR and WSS measures for an enhanced speech, the better are its perceived quality.

Segmental SNR is based on the classical SNR and it is one of the most widely used methods for testing enhancement algorithms. Segmental SNR is more correlated to any subjective attribute of speech quality than classical SNR. This can be explained that speech is time varying and classical SNR weights all time domain errors in the speech signal equally.
Segmental SNR is measured over short frames and final result is obtained by averaging the value of each frame over all the segments. The corresponding segmental SNR can be formulated as [24],

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{t=Nm}^{Nm+N-1} s_{\phi}^{2}(n)}{\sum_{t=Nm}^{Nm+N-1}\left[s_{d}(n)-s_{\phi}(n)\right]^{2}} \quad (30)$$

where $s_{\phi}(n)$ is the clean speech signal, $s_{d}(n)$ is the enhanced signal after performing noise reduction algorithm on noisy speech signal, *M* is the number of segments and *N* is the segment length. The lower and upper thresholds are selected to be -10 dB and +35 dB, respectively.

The PESQ (Perceptual Evaluation of Speech Quality) algorithm is an objective method to predict the results of subjective mean-opinion score (MOS) tests, designed purposely for handset telephony speech codecs. Although PESQ scores were not designed for speech enhancement algorithms evaluation, they are still found to provide a meaningful indication of performance and they are frequently used by researchers for this purpose.

The PESQ algorithm compares the original clean speech signal to the output of the enhancement algorithm, and penalizes the final score based on measures of the distortion. The PESQ is perceptual in the sense that the amount of distortion is measured in the context of a model for the human auditory system. In the P.862 standard, the lowest PESQ score is -0.5 and the highest score is 4.5. High scores stand for good speech quality.

5 Experiment Results

For the experiment, the Noizeus database [26] was used. The noisy database contains 30 IEEE sentences [27] produced by three male and three female speakers (5 sentences/speaker), and was corrupted by eight different real-world noises at different SNRs 0dB, 5dB, 10dB and 15dB. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz.
Noise signals were taken from the AURORA database [28] and included the following recordings from different places: babble (crowd of people), car, exhibition hall, restaurant, street, airport, train station, and train.

To simulate the receiving frequency characteristics of telephone handsets, the speech and noise signals were filtered by the modified Intermediate Reference System (IRS) filters used in ITU-T P.862 for evaluation of the PESQ measure.

We added computer generated telephone–bandwidth white Gaussian noise as an extra noise source, since it is not present in the database. The frame size is 20 ms, with an overlap of 50 %, although small changes in the frame size did not degrade the performance.

In this section we describe the results of simulations of the presented system, with both ideal and estimated filter parameters. The overall SNR is computed according to ITU P.56 standard [29]. The tool we use is MATLAB. The performance evaluation is based on objective measures using SNR, LLR, WSS and PESQ as well as subjective listening. Though the measures have been observed over a wide range of SNR 0dB, 5dB, 10dB and 15dB, only few are tabulated due to limitation of space in tables 1 to 9.

The iterative scheme proposed in [22] and presented in section 3 was used to estimate the system parameters in some results. The algorithm needs three iterations or plus to get the highest SNR gain.

5.1 Model for Speech with White Noise

In the case of simple Kalman filter (model for speech with white noise), a Matlab code was developed, where a simple Kalman filter was applied on all 30 files of the Noizeus database.

In this first approach, only the speech signal is modeled by an AR model of order P=10. The 10 AR coefficients were updated for every analysis frame of 20 ms duration, using the linear prediction analysis method (LPC), which is directly applied to the clean signal. The additive measurement noise is assumed to be a white noise, even in the case of real noises of the database.

The test results of the measures (LLR, SNRseg, WSS, PESQ) in this case are presented in table 1. The rows (DWT and EWT) of table 1 give the measures of the degraded signal in the case of a white noise (DWT), and that of the enhanced signal (EWT). The following rows (DCR and ECR), (DTN and ETN), (DBB and EBB), (DRT and ERT), (DSN and ESN), (DAP and EAP), (DEX and EEX) and (DSTR and ESTR) corresponds respectively to the results in the case of car, train, babble, restaurant, station, airport, exhibition and street noises.

Each value of the table corresponds to an average of thirty different measures carried out on the database sentences with the same characteristics (the same noise, the same SNR).

Table 1. Objective Quality Scores under White Noise

	0 dB				5 dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
DWT	1.80	-5.08	82.69	1.539	1.54	-2.33	69.97	1.79
EWT	0.70	3.17	68.79	2.345	0.58	4.86	57.47	2.625
DCR	1.01	-4.96	66.94	1.634	0.79	-2.17	54.08	1.891
ECR	0.88	1.96	67.24	2 .073	0.75	3.62	54.39	2 .350
DTN	1.18	-4.50	60.25	1.60	0.99	-1.69	48.12	1.859
ETN	0.74	2.45	61.49	2.20	0.61	4.36	48.39	2 .515
DBB	0.89	-4.63	70.35	1.705	0.71	-1.78	56.02	2.006
EBB	1.03	1.58	69.57	2 .068	0.86	3.34	55.45	2 .352
DRT	0.84	-4.19	66.42	1.754	0.68	-1.39	53.60	2.001
ERT	1.01	1.86	65.96	2.093	0.84	3.62	52.64	2.354
DSN	0.94	-4.71	69.05	1.665	0.73	-1.89	54.67	1.958
ESN	0.98	1.86	67.94	2.062	0.84	3.47	54.19	2.366
DAP	0.86	-4.41	71.52	1.726	0.69	-1.67	56.05	2.021
EAP	1.06	1.85	69.09	2.051	0.89	3.52	54.73	2.338
DEX	1.20	-4.67	63.52	1.585	0.94	-1.84	51.93	1.882
EEX	0.84	2.61	62.75	2.194	0.71	4.34	50.72	2.463
DSTR	0.99	-4.26	63.34	1.563	0.80	-1.58	50.04	1.904
ESTR	0.90	2.54	62.07	2.194	0.73	4.06	49.77	2.448

According to the results of table 1, we note an improvement in terms of (LLR, SNRseg, WSS, PESQ) measures for the two values of signal to noise ratio SNR = 0dB and SNR = 5dB. The listening tests and waveforms obtained for each sentence confirm that there is a better quality of the enhanced speech, a low level of residual noise while preserving intelligibility and natural sound. Besides, the obtained measurements in the case of white noise are significantly higher than those obtained in the case of real noises, because Kalman filter design is based on the assumption that the additive noise is white noise. For real colored noise, in addition to speech modeling, noise modeling is more than necessary.

As an example, Fig.1 shows the plots of the clean speech, the noisy speech with a white Gaussian noise at SNR equal to 5 dB and the enhanced speech using the above model.

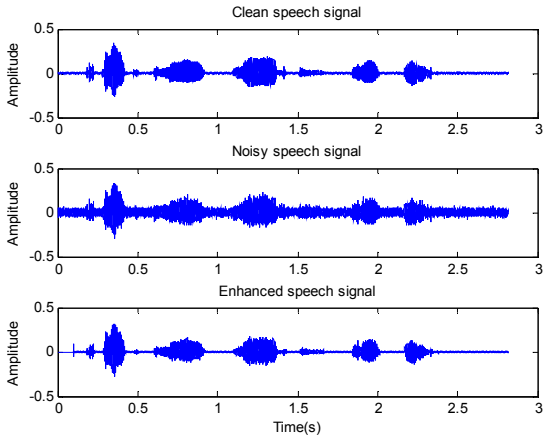

Fig.1 Clean, noisy and enhanced speech waveforms in the case of white noise.

5.2 Model for Speech with Colored Noise

With noise model, test results (LLR, SNRseg, WSS, PESQ) are presented in the following tables, in the case of a car noise (Table 2), train noise (Table 3), station noise (Table 4), street noise (Table 5), airport noise (Table 6), restaurant noise (Table 7), exhibition noise (Table 8) and babble noise (Table 9). In this case, all these results are obtained with a speech signal modelling with order P = 10, an analysis frame of 20 ms duration and real noise modeling with a variable order q = 2, 4, 6, 8, 10.

Table 2. Objective Quality Scores under Car Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	1.01	-4.96	66.94	1.634	0.79	-2.17	54.08	1.891
2	0.61	2.26	66.60	2.130	0.51	3.88	53.80	2.398
4	0.56	2.35	65.03	2.162	0.47	3.99	52.71	2.435
6	0.55	2.34	64.39	2.154	0.46	3.99	52.06	2.431
8	0.53	2.30	64.01	2.149	0.45	3.96	51.59	2.430
10	0.53	2.28	64.36	2.146	0.44	3.93	51.78	2.425

Table 3. Objective Quality Scores under Train Noise

	0 dB				5 dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	1.18	-4.50	60.25	1.605	0.99	-1.69	48.12	1.859
2	0.62	2.14	61.92	2.194	0.52	4.34	48.64	2.502
4	0.54	2.60	61.15	2.229	0.45	4.48	48.32	2.534
6	0.54	2.66	59.73	2.209	0.46	4.53	47.29	2.517
8	0.53	2.67	58.85	2.205	0.45	4.54	46.45	2.517
10	0.53	2.67	58.70	2.201	0.44	4.53	46.24	2.514

Table 4. Objective Quality Scores under Station Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	0.94	-4.71	69.05	1.665	0.73	-1.89	54.67	1.958
2	0.56	2.34	65.90	2.166	0.46	3.98	53.01	2.437
4	0.54	2.35	65.06	2.174	0.44	4.01	52.23	2.453
6	0.54	2.35	64.78	2.170	0.44	4.01	51.74	2.443
8	0.53	2.34	64.03	2.169	0.43	3.99	51.17	2.438
10	0.52	2.32	63.84	2.165	0.43	3.97	51.11	2.437

Table 5. Objective Quality Scores under Street Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	0.99	-4.26	63.34	1.563	0.80	-1.58	50.04	1.904
2	0.57	2.81	60.96	2.220	0.46	4.36	49.19	2.483
4	0.54	2.86	59.54	2.242	0.45	4.38	48.61	2.493
6	0.54	2.85	59.63	2.239	0.45	4.36	48.47	2.485
8	0.52	2.84	58.83	2.237	0.44	4.34	47.73	2.482
10	0.51	2.83	58.56	2.235	0.44	4.32	47.72	2.478

The two measures that are more correlated with subjective tests and listening tests are PESQ and SNRseg. From the tables 2 to 5, we note that the different noises (car, train, station and street) can be well modeled by an AR model of order q = 4. With this order value the best results are obtained. In this case, the slowly time-varying short-term spectrum can be modeled adequately by an AR process of order q = 4.

Table 6. Objective Quality Scores under Airport Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	0.86	-4.41	71.52	1.726	0.69	-1.67	56.05	2.021
2	0.57	2.58	65.25	2.177	0.46	4.19	52.49	2.457
4	0.52	2.63	63.84	2.196	0.44	4.26	51.35	2.473
6	0.52	2.64	63.24	2.195	0.43	4.27	50.62	2.472
8	0.51	2.63	62.63	2.196	0.42	4.26	50.09	2.469
10	0.49	2.63	62.14	2.197	0.41	4 .25	49.89	2.471

According to Table 6 an order q = 6 gives the best results in the case of airport noise.

Table 7. Objective Quality Scores under Restaurant Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	0.84	-4.19	66.42	1.754	0.68	-1.39	53.60	2 .001
2	0.56	2.55	63.62	2.197	0.47	4.31	51.05	2.471
4	0.54	2.63	62.31	2.222	0.45	4.38	50.11	2.488
6	0.53	2.69	60.72	2.229	0.44	4.44	48.89	2.494
8	0.52	2.70	59.83	2.238	0.43	4.45	48.25	2.496
10	0.51	2.69	59.51	2.235	0.43	4.44	48.04	2.491

Table 8. Objective Quality Scores under Exhibition Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	1.20	-4.67	63.52	1.585	0.94	-1.84	51.93	1.882
2	0.60	2.65	62.05	2.208	0.51	4.41	50.19	2.479
4	0.56	2.75	60.68	2.245	0.48	4.52	49.05	2.512
6	0.56	2.82	59.67	2.252	0.47	4.59	48.14	2.522
8	0.54	2.87	58.04	2.262	0.46	4.62	46.63	2.530
10	0.53	2.88	57.73	2.262	0.45	4.62	46.45	2.528

Table 9. Objective Quality Scores under Babble Noise

	0dB				5dB			
	LLR	SNR	WSS	PESQ	LLR	SNR	WSS	PESQ
Deg	0.89	-4.63	70.35	1.705	0.72	-1.78	56.02	2.006
2	0.58	2.17	67.16	2.150	0.47	3.90	54.15	2.432
4	0.57	2.24	65.77	2.164	0.45	3.98	52.91	2.449
6	0.56	2.30	64.58	2.165	0.44	4.04	51.66	2.457
8	0.54	2.31	63.43	2.166	0.44	4.04	50.98	2.460
10	0.53	2.29	63.47	2.164	0.43	4.02	50.79	2.455

For restaurant noise, exhibition noise and babble noise (Tables 7 to 9), a higher order for modeling (q=8), with respect to the other noises is justified by the fact that this noise has nearly the same waveform of the speech signal.

As seen in the results of Table 2 to Table 9, the noise modeling for real colored noise in addition to speech modeling, provides better performance than the standard assumption (Table 1).

An example of the considered model in the case of babble noise with SNR = 5 dB and parameter q=8 is presented in Fig.2. The comparison of noisy and enhanced speech waveforms reveals dramatic noise suppression with negligible residual noise and only slight speech distortion in the output signal.

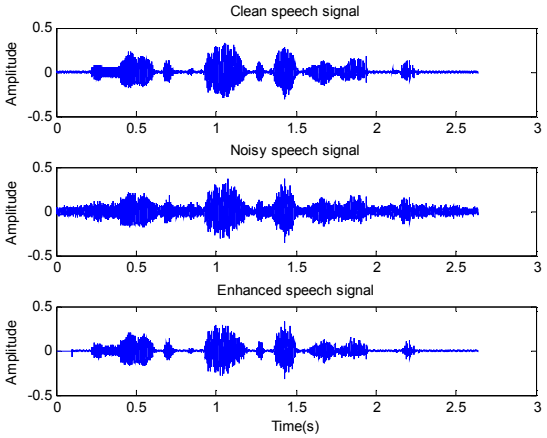

Fig.2 Clean, noisy and enhanced speech waveforms in the case of babble noise.

5.3 Results of the EM Algorithm

Further tests were conducted on the database speech files, directly using the EM algorithm for the AR parameters estimation from the noisy speech files. Table 10 shows the results of the measures (LLR, SNRseg, WSS, PESQ) of a single sentence in the case of a white noise with an SNR = 5 dB. We remark that three iterations can achieve acceptable results.

TABLE 10. Test results of the EM algorithm in the case of White Noise

	LLR	SNR	WSS	PESQ
Degraded	1.6019	-1.3279	33.9343	1.6872
Iteration 1	0.8118	2.8780	33.4534	2.1588
Iteration 2	0.8211	3.7545	33.7064	2.3100
Iteration 3	0.8146	4.1097	35.3840	2.3135
Iteration 4	0.8793	4.1062	38.9420	2.3552

7 Conclusions

We have presented in this paper a contribution to the use of the Kalman filter in the speech enhancement in the case of white and real colored noises. The performance evaluation based on objective quality measures, observation of the waveforms, as well as informal listening tests, show clearly that the Kalman filter provides a lower signal distortion and a higher noise reduction in the two cases.

Furthermore, for real colored noise, in addition to speech modeling, noise modeling with q=4 (car, train, station, street), q=6 (airport) and q=8 (restaurant, exhibition, babble) provides better performance.

Finally, an EM Kalman approach with Kalman smoother was applied to iteratively estimate the speech and noise parameters directly from the noisy observation. Simulation results have shown that the idea leads to very promising results.

References:

[1] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing, April 1979, pp. 208-211.

[2] C. Boubakir and D. Berkani, "Approche mono voie de réduction de bruit basée sur la soustraction spectrale et un modèle statistique de la VAD," TAIMA'05, Hammamet , Tunisia, 26 September 2005.

[3] B.L. Sim, Y.C. Tong, J.S. Chang and C.T. Tan, "A Parametric formulation of the

generalized spectral subtraction method," IEEE Trans. Speech and Audio Processing, Vol.6, No.4, July 1998, pp. 328-337.

[4] S. Kamath, P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in Proc. of ICASSP-2002, Orlando, FL, May 2002.

[5] C. Boubakir, D. Berkani and F. Grenez, "A frequency-dependent speech enhancement methods," Third International Summer School on Signal Processing and its Applications (I3SPA'06), JIJEL, ALGERIA, July 2006.

[6] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. of the IEEE, Vol.67, No.12, December 1979, pp.1586-1604.

[7] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.28, No.2, December 1980, pp. 137–145.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.32, No.6, December 1984, pp. 1109-1121.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.33, No.2, April 1985, pp. 443-445.

[10] R. Martin, "Spectral subtraction based on minimum statistics," in Proc. Euro. Signal Processing Conf. (EUSIPCO), 1994, pp. 1182-1185.

[11] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," in Proc. 21st IEEE Int. Conf. Acoustics, Speech and Signal Processing, Atlanta, GA, May 1996, pp. 629-632.

[12] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed priors," IEEE Proc. Intern. Conf. on Acoustics, Speech and Signal Processing, May 2002, pp. 504-512.

[13] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," IEEE Signal Processing Letters, Vol.11, No.9, September 2004, pp. 725-728.

[14] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in Proc. Int. Conf. Acoust., Speech, Signal Processing, 1987, pp. 177–180.

[15] J.D. Gibson, B. Koo, and S.D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Processing*, Vol.39, Aug. 1991, pp. 1732-1742.

[16] N.Ma, M. Bouchard and R.A. Goubran, "Perceptual Kalman filtering for speech enhancement in colored noise," in Proc. ICASSP'04, 2004, pp. 717-720.

[17] M. Gabrea, "Adaptive Kalman filtering-based speech enhancement algorithm," in Proc. of Canadian Conference on Electrical and Computer Engineering, Federicton, New-Brunswick, vol.1, 2001, pp. 521-526.

[18] M. Gabrea, "Robust adaptive Kalman filtering-based speech enhancement algorithm," in Proc. ICASSP'04, 2004, pp. 301-304.

[19] V. Grancharov, J. Samuelsson, and W.B. Kleijn, "Improved Kalman filtering for speech enhancement," in Proc. ICASSP'05, 2005, pp. 1109-1112.

[20] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," IEEE Trans. Speech Audio Process. Vol.6, No.4, 1998, pp. 373-385.

[21] J.S. Lim and A.V. Oppenheim, "All-pole modeling of degraded speech," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.26, No.3, June 1978, pp. 197-210.

[22] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, Vol.3, No.4, 1982, pp. 253-264.

[23] H.E. Rauch, F.Tung, and C.T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, Vol.3, 1965, pp. 1445-1450.

[24] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, "Objective measures of speech quality," Englewood Cliffs, NJ: Prentice-Hall, 1988.

[25] A.Rix, J.Beerends, M.Hollier and A.Hekstra, "Perceptual evaluation of speech quality (PESQ) -- A new method for speech quality assessement of telephone networks and codecs," in Proc.IEEE Int. Conf. Acoustics, Speech and Signal Processing, Salt Lake City, UT, Vol.2, 2001, pp. 749-752.

[26] "Noizeus: A noisy speech corpus for evaluation of speech enhancement algorithms," http://www.utdallas.edu/~loizou/speech/noizeus/.

[27] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio and Electroacoustics, 1969, pp. 225–246.

[28] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in ISCA ITRW ASR2000, Sept. 2000, Paris, France.

[29] "Objective measurement of active speech level," ITU-T Recommendation P.56, March 1993.