Combination of Beam forming and Kalman filter Techniques for Speech Enhancement

G. Ramesh Babu[1]
Department of E.C.E, Sri Sivani College of Engg.,
Chilakapalem, Srikakulam- 532401
Andhra Pradesh, India
E-mail: gramesh24@yahoo.co.in

Rameshwara Rao[2]
Professor, Department of ECE, Osmania Univeristy,
Hyderabad, India

Abstract

In all speech communication settings the quality and intelligibility of speech is of utmost importance for ease and accuracy of information exchange. Kalman filter is an adaptive least square error filter that provides an efficient computational recursive solution for estimating a signal in presence of noises. Beamforming is another possible method of speech enhancement, because, the beamformer minimizes the output signal power but maintains signals from the desired direction. Hence, an optimized cascaded scheme is implemented using the advantages of Kalman filter and Beamforming where the Kalman filter technique followed by Beamforming reduces stationary as well as residual noise. The proposed hybrid method gives better SNR and PESQ values as compared to that of individual techniques, thereby improving the quality of the speech.

Keywords: Kalman filter, Beamforming, Cascaded scheme

1. Introduction

When using hands-free speech communication systems, the speech signal acquisition is usually corrupted by reverberation and background noise which lead to a significant decrease in communication quality. For this reason, techniques for enhancing the desired speech signal are required which reduce the environmental noise. The objectives of speech enhancement are high quality and intelligibility of the output speech signal. Therefore, a noise reduction system is required which significantly attenuates the environmental noise without affecting the speech signal by additional distortions.

The scheme uses a linear microphone array to capture a speech signal that has been corrupted by babble noise, car noise and interference signals. Beamforming by itself, however, does not appear to provide enough improvement. Further, the performance of beamforming becomes worse if the noise source comes from many directions or the speech has strong reverberation. Kalman Filter [1] is an adaptive least square error filter that provides an efficient computational recursive solution for estimating a signal in presence of noises. It is an algorithm which makes optimal use of imprecise data on a linear (or nearly linear) system with errors to continuously update the best estimate of the system's current state. The present paper shows a combined technique using the advantages of Kalman filter and beamforming where the beamforming followed by Kalman filter reduces stationary as well as residual noise. Hence, the best performance is obtained when they work together rather than individually.

2. Multichannel speech enhancement system

Multi-channel enhancement algorithms [2], [3] and [4] exploit the spatial diversity. This diversity can be taken advantage of e.g., by steering a null towards the noise source and a beam towards the signal source. In this paper, a brief overview of one common multi-channel noise reduction technique known as beamforming technique is provided.

2.1 Beamforming

Beamforming is a means of performing spatial filtering [5]. In the frequency domain, beamforming can be viewed as a linear combination of the sensor outputs:

$$Z(k) = \sum_{i=1}^{M} b_i(k) Y_i(k) \qquad (1)$$

$b_i(k)$ is the beamformer weight corresponding to the i^{th} sensor, and M is the total number of sensors. In vector notation, we have

$$Z(k) = b^T (k)Y(k) \qquad (2)$$

where $b(k) = [b_1(k) \ldots b_M(k)]^T$

Beamforming can be classified into two categories - fixed, where the weights are fixed across time, and adaptive, where the weights vary in response to changes in the acoustic environment.

2.1. Fixed beamforming

In fixed beamforming, the weights $b_i(k)$ are fixed over time, and are determined by minimizing the power of the signal at the output of the beam former subject to a constraint that ensures that the desired signal is undistorted, i.e., the optimal weights are the solution to

$$\min_{b(k)} \mathbf{b}^*(k)\Phi_{yy}(k)\, \mathbf{b}(k) \text{ subject to } \mathbf{b}^*(k)\mathbf{1} = 1 \qquad (3)$$

where * refers to complex conjugate transpose and $\Phi_{yy}(k)$ is the M X M PSD matrix of the noisy input signals whose (i, j)th entry is $E[Y_i(k)Y_j{}^*(k)]$. Note that the constraint of zero distortion in the look direction is written using a vector of ones since we assume that the array shown in the fig. 1 has been presteered towards the desired signal direction. The solution to the constrained optimization problem is the well-known minimum variance distortionless response (MVDR) beamformer:

$$b(k) = \frac{\Phi_{ww}^{-1}(k)\mathbf{1}}{\mathbf{1}^T\Phi_{ww}(k)\mathbf{1}} \qquad (4)$$

where Φ_{ww} (k) is the M X M noise PSD matrix whose (i; j)th entry is $E[W_i(k)W_j{}^*(k)]$. Assuming a homogeneous noise field, the solution can be written in terms of the coherence matrix

$$b(k) = \frac{\Gamma_{ww}^{-1}(k)\mathbf{1}}{\mathbf{1}^T\Gamma_{ww}(k)\mathbf{1}} \qquad (5)$$

where the (i; j)th entry of the M X M coherence matrix is given by

$$\Gamma_{ij}(\mathbf{k}) = \frac{\Phi_{w_i w_j}(\mathbf{k})}{\sqrt{\Phi_{w_i w_i}(\mathbf{k})\Phi_{w_j w_j}(\mathbf{k})}} \qquad (6)$$

$$= \frac{\Phi_{w_i w_j}(\mathbf{k})}{\Phi_{ww}(\mathbf{k})} \qquad (7)$$

where $\Phi_{w_i w_j}(\mathrm{k})$ is the cross spectral density between the noise signals at the i^{th} and j^{th} sensors, and from the assumption of a homogeneous noise field, $\Phi_{w_i w_i}(k) = \Phi_{ww}(k)$ for all i.

For incoherent (or spatially white) noise fields, $\Gamma_{ww} = I$, $b = \frac{1}{M}$ 1 and the MVDR beam former reduces to a delay-and-sum beamformer (DSB),where the sensor signals are delayed and then averaged. The pre-steering corresponds to the delay and is such that the signal components at the different sensors sum up constructively while the noise components cancel each other. Incoherent noise fields are not common. An example of incoherent noise is electrical noise at the sensors, which is uncorrelated at the different sensors.

In a DSB, the amplitude weights are fixed across frequency (often equal) and the phase weights introduce the delay. A more general form is a filter-and-sum beam former (FSB), where both the amplitude and phase weights vary across frequency. FSBs are useful in designing beam formers with a specified directivity pattern for arbitrary microphone array configurations.

Many of the noise fields encountered in practice fall into the category of diffuse noise fields, whose coherence function has the form:

$$\Gamma_{ij}(k) = \mathrm{sinc}\left(\frac{2\Pi k}{k}\,\frac{d_{ij}}{c}\right) \qquad (8)$$

where sinc(x) = sin(x)/x, d_{ij} is the distance (in meters) between the i^{th} and j^{th} sensors, c = 340 m/s is the speed of sound in air and K is the frame length. If we use the corresponding expression for the coherence matrix, the resulting beamformer is called a super directive beamformer (SDB). While the SDB is useful in diffuse noise fields, its main disadvantage is an amplification of uncorrelated noise (e.g., sensor noise) at low frequencies. This problem is handled by incorporating a white noise gain constraint in the design.

2.2. Adaptive beamforming

In adaptive beamforming, the beamformer weights adapt to changes in the acoustic environment over time. The optimal weights are obtained by minimizing the variance of the output signal.

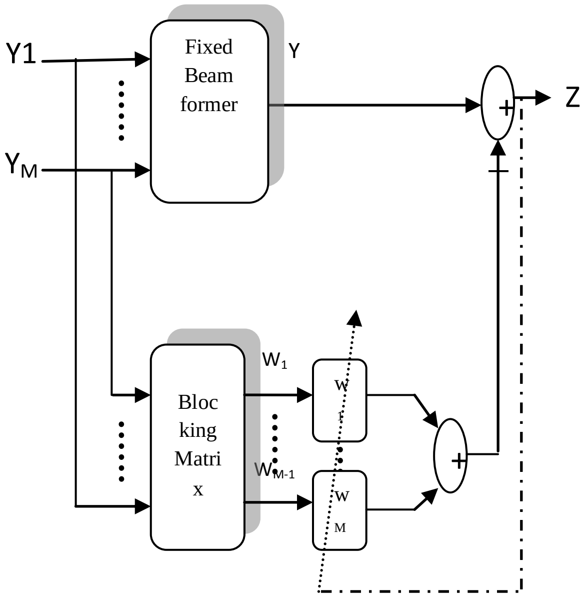


Figure 2. Frequency domain implementation of the Generalized Side lobe Canceller. The ANC is implemented by the adaptive filters $w_1 . . . .w_{M-1}$

To ensure that the speech signal is not cancelled out or distorted, a distortion less constraint is imposed on the desired signal. This results in the linearly constrained minimum variance (LCMV) beamformer, where the adaptive beamformer weights are obtained through a constrained minimization procedure.The generalized side lobe canceller (GSC) [6], is an efficient alternative implementation of Frost's LCMV approach, that converts the constrained optimization problem into an unconstrained one. This leads to an efficient implementation for the update of the beamformer weights.

The GSC consists of three parts - a fixed beamformer (FBF), a blocking matrix (BM) and an adaptive noise canceller (ANC) as shown in Fig.2. The FBF includes a pre-steering module and its weights are designed to produce a speech reference Y_{BF} with a specified gain and phase response. The FBF could either be a simple delay-and-sum beamformer, or a more advanced filter-and-sum or super directive beamformer. The BM is generally orthogonal to the FBF and produces M-1 outputs, called the noise references, by steering zeros towards the desired signal direction. One way to create the noise references is to take the difference between adjacent sensor signals. The ANC (implemented by the adaptive filters $w_1 . . . .w_{m-1)}$ in Fig. 2 removes any remaining correlation between the speech reference Y_{BF} and the noise references. Thus, any residual noise in the speech reference that is correlated to the noise references is removed. In practice, the noise references are not completely free of speech. As a consequence, the ANC results in some of the speech signal being cancelled. To minimize the effect of the speech leakage on the ANC, the noise-cancelling filters are adapted only during periods of speech absence. To reduce the amount of speech leakage, some variants of the GSC employ an adaptive blocking matrix.

3. Kalman filter

The speech signal s(n) is modeled as a P^{th} -order AR process
where

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \qquad (9)$$

$$y(n) = s(n) + v(n) \qquad (10)$$

Where, s(n) is the nth sample of the speech signal, y(n) is the nth sample of the observation, and $a_i(n)$ is the i^{th} AR parameter. This system can be represented by the following state-space Model. Where, the sequences u(n) & v(n) are uncorrelated Gaussian white noise sequences with the mean $\bar{u}$ and $\bar{v}$ and the variances σ_u^2 and σ_v^2.x(n) is the P x 1 state vector.

$$X(n)=[s(n-p+1),\ldots.,s(n),v(n-q+1),\ldots.,v(n)]^T \qquad (11)$$

F(n) is the P x P transition matrix

$$F(n)= \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 1 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & 1 \\ a_p & a_{p-1} & a_{p-2} & \ldots & a_1 \end{bmatrix}$$

G and H are, respectively, the P x 1 input vector and 1 x p observation row vector which is defined as follows

$$H = G^T = [0\ 0 \cdots\cdots 0\ 1] \qquad (12)$$

The standard Kalman filter [2] provides the updating state vector estimator equations

$$e(n) = y(n) - H\hat{x}(n/n-1) \qquad (13)$$

$$k(n) = p(n/(n-1))H \times [HP(n/(n-1))H^T]^{-1}$$
(14)

$\hat{x}(n/n)=\hat{x}(n/(n-1))+K(n)e(n)$ (15)

$P(n/n) = [I - K(n)H]\, P(n/n-1)$ (16)

$\hat{x}(n+1/n)=F(n)\hat{x}(n/n)+G\bar{u}$ (17)

$P(n+1/n)=F(n)P(n/n)\, F^{T}(n)+GG^{T}\sigma_{\bar{u}}^{2}$ (18)

Where, $\hat{x}(n+1/n)$ is the minimum mean square estimation of the state vector X(n) given the past n-1 observations y(1), ……. , y(n-1)

$\tilde{x}(n/n-1) = x(n) - \hat{x}(n/n-1)$ is the predicted state error vector.

$P(n/n-1) = E[\tilde{x}(n/n-1)\, \tilde{x}^{T}(n/n-1)]$ is predicted state error correlation matrix.

$\hat{x}(n/n)$ is the filtered estimation of the state vector.

$\tilde{x}(n/n) = x(n) - \hat{x}(n/n)$ is the filtered state error vector.

$P(n/n) = E[\tilde{x}(n/n-1)\, \tilde{x}^{T}(n/n)]$ is the filtered state error correlation vector.

e(n) is the innovation sequence.

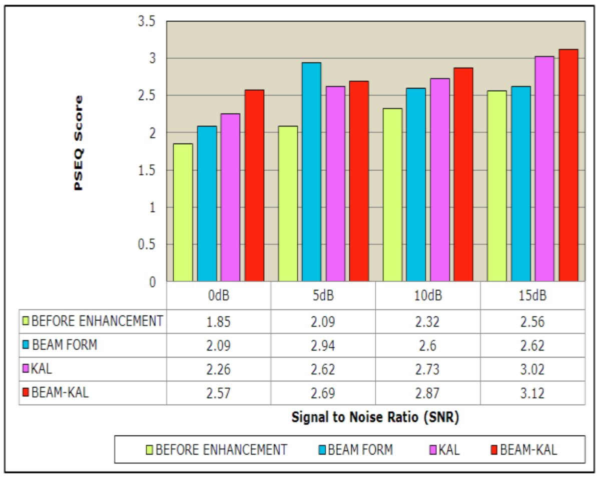
K(n) is the Kalman gain.

The estimated speech signal can be retrieved from the state-vector estimator

$$\hat{s}(n) = H\hat{x}\left(\frac{n}{n}\right) \quad (19)$$

4. Experimental Results

The experiment is carried by corrupting the speech signals using babble noise at various input SNRs 0 dB, 5 dB, 10 dB and 15 dB and performance was evaluated by cascading Beamform and Kalman filter (BEAM-KAL).



	0dB	5dB	10dB	15dB
□ BEFORE ENHANCEMENT	1.85	2.09	2.32	2.56
■ BEAM FORM	2.09	2.94	2.6	2.62
■ KAL	2.26	2.62	2.73	3.02
■ BEAM-KAL	2.57	2.69	2.87	3.12

Figure 3. Comparison of PESQ scores for the proposed BEAM-KAL filter

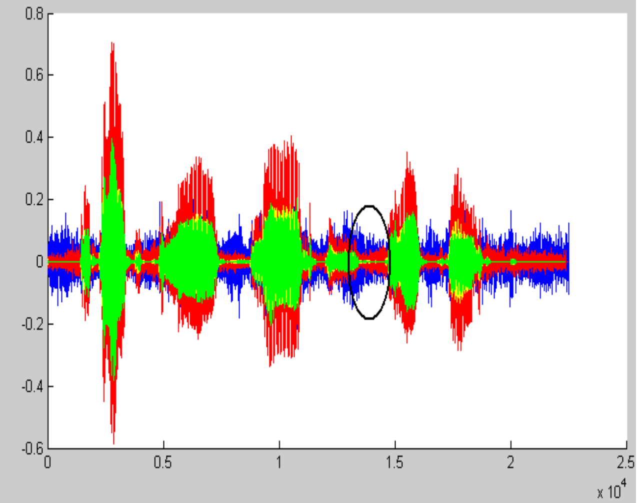


Figure 4. Time domain comparison of (a) Noisy input signal at 15 dB (blue colour) (b) Enhanced BEAMFORM signal at 15 dB (red colour) (c) Enhanced KALMAN signal at 15 dB (yellow colour) (d) Enhanced signal of the proposed BEAM-KAL at 15 dB (green colour) and the black circle shows how the noise was removed.

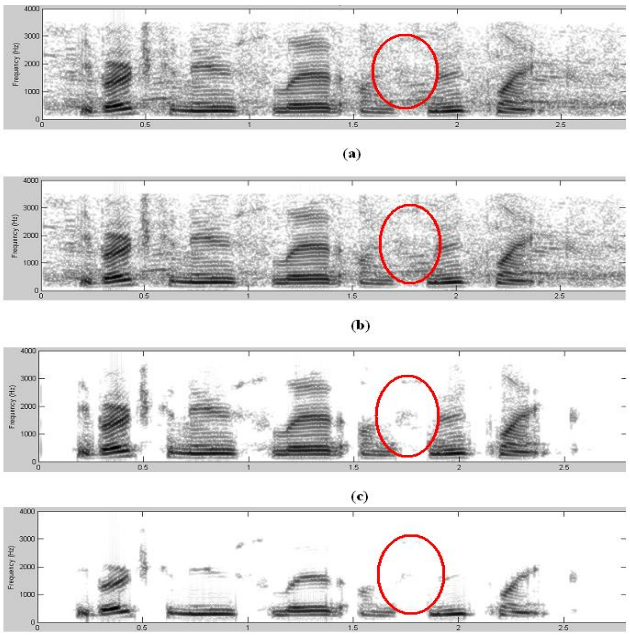


Figure 5. Spectrogram analysis of speech sample 'sp01' from NOIZEUS database corrupted with babble noise at 15 dB SNR. (a) Noisy signal at 15 dB (b) Enhanced BEAMFORM signal at 15 dB (c)Enhanced KALMAN signal at 15 dB (d) Enhanced signal of the proposed BEAM-KAL at 15 dB.

Perceptual Evaluation of Speech Quality (PESQ) scores [7] for the proposed BEAM-KAL was studied to be consistently good from low to high input SNR except at 5 dB where beamform was slightly better than this as shown in fig. 3. Fig. 4 & 5 shows the time domain and spectrogram comparison of noisy, Beamform, Kalman and BEAM-KAL enhanced speech signals at 15 dB. From this, it is clear that BEAM-KAL combination is superior to the others which can also be seen in the subjective A-B test shown in the table 1.

4.1. A-B Results for BEAM-KAL

In the same way second set of tests was conducted using a speech file corrupted with babble noise at 0 dB. The file was enhanced with the BEAM, KAL and BEAM-KAL. These tests show an almost unanimous preference for the proposed algorithm over all the others.

Table 1. Listener preferences for tests in babble noise at 0 dB

Test	Listener Preference		
	BEAM-KAL	Other	Undecided
BEAM-KAL/Noisy Signal	100%	0%	0%
BEAM-KAL/BEAM FORM	96%	4%	0%
BEAM-KAL/KAL	92%	1%	7%

5. Conclusion

The method presented here is based on a generalized side lobe canceller (GSC) adaptive beamformer combined with an Kalman filter. The objective test is conducted and the results of the above proposed system are compared to the beamformer and Kalman filter individually at various input SNRs. The enhanced signals of the proposed cascade BEAM-KAL and the individuals are compared to the unenhanced signal. It is worth mentioning that the improved system shows good performance. Because, the beamformer minimizes the output signal power but maintains signals from the desired direction.

References

[1] K. Paliwal, and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering", *Proceedings of IEEE Int.Conf. Acoust. Speech*, 1987.

[2] M. Drews, "Speaker localization and its application to time delay estimators for multi-microphone speech enhancement systems", *Proc. Eusipco*, 1996, pp. 483-486.

[3] M. Drews, "Construction of microphone arrays for the optimization of multi-channel speech enhancement systems", *Frequenz* 50, 1996, 223-227 (in German).

[4] Brandstein, M.S., D.B. Ward (Eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, 2001.

[5] B.D. Van Veen, and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, 1988, pp. 4-24.

[6] B. R. Breed, and J. Strauss, “A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Processing Lett*., vol. 9, no. 6, 2002, pp. 168-169.

[7] *Methods for Objective and Subjective Assessment of Quality, Recommendation* ITU-T P.830, International Telecommunication Union, Feb. 1996.