Speech Enhancement Using Beamforming and Kalman Filter for In-Car Noisy Environment

G. Ramesh Babu and G. V. Sridhar

Abstract The effectiveness of the communication system gets seriously degraded in car by noises like engine sounds and ambient noise, thus decreasing the quality of speech. In modern cars, a lot of effort is put on reducing the background noise. In this paper, speech enhancement cascaded scheme named BEAM-KAL is developed to get the better intelligibility and quality of speech. For this, multichannel beamforming techniques are combined with single channel Kalman filter to get better quality of speech signals which suffer in-car noises. In beamforming, microphone arrays are used to extract the speech signal of interest from a specific desired direction, whereas signals contaminated with noises from various directions are attenuated. However, this technique does not appear to provide enough improvement by itself. Hence, the Kalman filter has been used for its further enhancement. Experiments are performed with real recordings taken while driving in a noisy automobile environment. The performance is investigated with SNR, PESQ and spectrograms and has been shown to produce a better quality of speech.

Keywords Speech enhancement · Kalman filter · Beamforming · BEAM-KAL · Car noise

1 Introduction

Nowadays, speech-based applications like automatic speech recognition systems (ASR) are becoming very important. But the performance of the ASR system in-cars is deteriorated very much by background noises and other various disturbances [1]. Hence, the essential step is to develop effective speech enhancement methods to suppress the background disturbing noises and improve the intelligibility of capturing speech under in-car noisy situations. In recent years, the devices which are activated

G. Ramesh Babu (✉) · G. V. Sridhar
Raghu Engineering College (A), Dakamarri, Visakhapatnam 531162, India
e-mail: drgramesh24@gmail.com

G. V. Sridhar
e-mail: sridhar1209@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

P. S. R. Chowdary et al. (eds.), *Microelectronics, Electromagnetics and Telecommunications*, Lecture Notes in Electrical Engineering 655,
https://doi.org/10.1007/978-981-15-3828-5_57

by human voice commands have been given much attention by car manufacturers. However, the performance of currently available commercial products is degrading substantially under real-world conditions. The noises originated from pumps, engines, audio equipment, wind, road and air-conditioning, radio and communication are usually non-stationary in nature and time-varying [1, 2]. So, some speech enhancement techniques dealing with in-car noises were developed in recent years to extract the original speech from the distorted noisy speech. The present scheme involves two main algorithms in which beamforming exploits the time correlation of speech signals captured by microphones and Kalman filter uses the different statistics in speech and noise signals to separate them. The purpose of combining the beamforming and Kalman filter is to exploit the strength of each individual method and to improve the quality of speech. When these two methods are combined, a strongly interfering source signal is first separated by beamforming and remaining noise signals are denoised by an adaptive Kalman filter [3]. Thus, they produce the best performance when worked jointly rather independently.

2 Microphone Array-Based Speech Enhancement System

The in-car acoustic ambient contains various sources of disturbances besides the speaker. In order to separate the speech signal of the speaker from these disturbances, multichannel speech enhancement system involving microphone arrays is being used. By processing the microphone array signal appropriately, we can achieve the direction-dependent sensitivity of the source. This technique is named as beamforming.

2.1 Beamforming

In beamforming, it isolates a source from a specific direction, while still maintaining some semblance of directionality of the receiving signal. The beamforming be able to represented by linearly arranged of the sensor outputs

$$R(l) = \sum_{i=1}^{M} w_i(l) Y_i(l) \tag{1}$$

where $w_i(l)$ are the weights associated to each ith sensors, and M represents the l number of sensors. The data is described by the vector,

$$\begin{aligned} R(l) &= w^T(l) Y(l) \\ w(l) &= [w_1(l) \ldots w_M(l)]^T \end{aligned} \tag{2}$$

Beamforming can be classified as conventional and adaptive where in conventional the weights across time are fixed, and in adaptive the weights are varied according to the acoustic surroundings of speech.

2.1.1 Fixed Beamforming

In conventional beamforming, [4] the weights $w_i(k)$ are fixed and determined by minimizing the signal power at the beamformer output and subjected to a constraint ensures that the desired signal is unvarnished [4], i.e., the optimal weights are the solution to

$$\min_{w(l)} w^*(l)\Psi_{yy}(l)w(l) \text{ subject to } w^*(l)1 = 1 \tag{3}$$

where * represents the complex conjugate transpose and whereas $\Psi_{yy}(l)$ is the power spectral density matrix of size $M \times M$ of the noisy speech signal with (i, j)th value which is $E\left[X_i(l)X_i^*(l)\right]$. The constraint of zero distortion in the desired direction is given by means of a vector of one's while we consider the array is pre-steered [5] towards the preferred signal direction. The solution is the minimum variance distortion less response beamformer with the constrained optimization [6]

$$w(l) = \frac{\Psi_{ww}^{-1}(l)1}{1^T\Psi_{ww}(l)1} \tag{4}$$

where $\Psi_{ww}(l)$ is the noise PSD matrix of size $M \times M$ whose $(i; j)$th entry is $E\left[W_i(l)W_i^*(l)\right]$. By considering the noise field as homogeneous, the solution is in terms of coherence matrix

$$w(l) = \frac{\Gamma_{ww}^{-1}(l)1}{1^{\mathrm{T}}\Gamma_{ww}(l)1} \tag{5}$$

The $(i; j)$th value of the coherence matrix $M \times M$ is

$$\Gamma_{ij}(l) = \frac{\Psi_{w_iw_j}(l)}{\sqrt{\Psi_{w_iw_i}(l)\Psi_{w_jw_j}(l)}} \tag{6}$$

$$= \frac{\Psi_{w_iw_j}(l)}{\Psi_{ww}(l)} \tag{7}$$

In the above equation, $\Psi_{w_iw_j}(l)$ is the cross spectral density between e ith and jth sensors and the noise signals. By the assumption of a homogeneous noise field, $\Psi_{w_iw_i}(l) = \Psi_{ww}(l)$ for i. The incoherence noise fields, $\Gamma_{ww} = \mathrm{I}$, $w = \frac{1}{M}$ 1 and the minimum variance distortion less response beamformer reduced to a delay-and-sum beamformer, in which first the sensor output speech signals are delayed and

then followed by average. The pre-steering corresponds to the delay and the speech signal components at various sensors added beneficially, and at the same time the noise components are get cancelled. In a delay-and-sum beamformer, the weights of the amplitude are fixed and the weights of phase introduce the delay. But both the amplitude and phase weights vary in filter-and-sum beamformer (FSB). These FSBs are used in designing beamformers with a specific pattern of direction for microphone arrays. Most of the noises fall into the category of noise fields, and the coherence function is in the form of

$$\Gamma_{ij}(l) = \sin c\left(\frac{2\Pi k}{l}\frac{D_{ij}}{c}\right) \tag{8}$$

where k is the frame length and c is the velocity of sound in air, $c = 339$ m/s. Where D_{ij} is the distance between the ith and jth sensors in the array and $\sin c(z) = \sin(z)/z$.

In the coherence matrix, if we use the equivalent expression for the resultant beamformer, it is known as a super directive beamformer (SDB). Although SDB is used in diffuse noise fields, it has a disadvantage of amplifying uncorrelated noises at low frequencies. It can be overcome by incorporating white noise gain restriction in the design.

2.1.2 Adaptive Beamforming

In this, the weights are changed according to the acoustic surroundings. The optimized weights are taken by means of minimizing the variance of the output signal. To make sure that the desired speech signal is not cancelled out or distorted, a distortion less constraint is forced on the desired signal. The generalized side lobe canceller (GSC) [6] is an efficient implementation of the linearly constrained minimum variance (LCMV) procedure, which converts the constrained optimized problem to an unconstrained one. This will give a better performance for the updated weights. The block diagram of GSC is shown in Fig. 1. The structure of GSC contains three modules a beamformer (BF), blocking matrix (BM) and noise canceller (NC). The BF is having a pre-steering module which aligns the desired speech components Y_{BF} by designing its coefficients. Blocking matrix is orthogonal to the beamformer and resulting outputs, called the noise reference signals by blocks the desired speech signal. By taking the differences between adjacent sensor signals, noise references will be formed. The noise cancellation in Fig. 1 removes any remaining noise residual in the speech reference that is correlated with the noise references.

3 Kalman Filter

When speech signal is corrupted with noise, the output y(l) is given as

$$y(l) = x(l) + v(l) \tag{9}$$

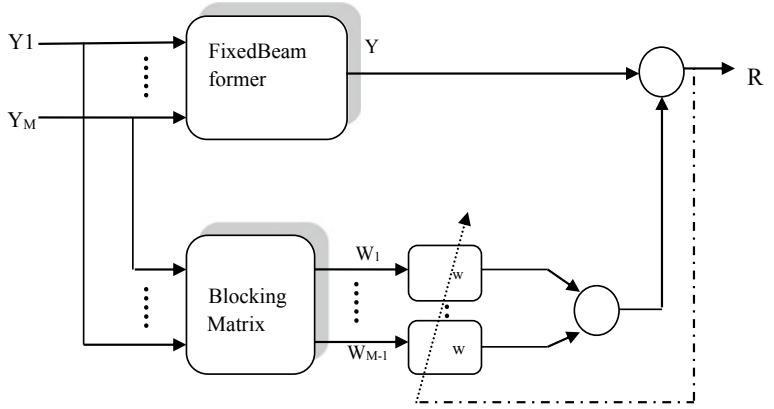

Fig. 1 Implementation of the generalized side lobe canceller in the frequency domain

$x(l)$ is the clean speech.

A qth order autoregressive AR predictor is used to model the speech signal.

Where $x(l)$, the present sample, depends on the linear combination of previous q samples added with a noise.

$$x(l) = \sum_{i=1}^{q} a_i x(l-i) + u(l) \tag{10}$$

where $x(l)$ is the lth sample of the clean signal, and $y(l)$ is the lth sample of the noisy speech, and $a_i(l)$ is ith autoregressive process parameter. This can be modelled by the following state-space expression. Where, the sequences $u(l)$ and $v(l)$ are uncorrelated Gaussian white noise sequences with the mean $\bar{u}$ and $\bar{v}$ and the variances σ_u^2 and σ_v^2.$x(l)$ is the $Q \times 1$ state vector.

$$X(l) = [s(l-q+1), \ldots, s(l), v(l-q+1), \ldots, v(l)]^T \tag{11}$$

The Kalman filter gives the updating state vector estimator equations

$$e(l) = y(l) - H\hat{X}(l/(l-1) \tag{12}$$

$$K(l) = q(l/(l-1))H \times \left[Hq\big(l/(l-1)H^T\right]^{-1} \tag{13}$$

$$\hat{X}(l/l) = \hat{X}(l/(l-1)) + k(l)e(l) \tag{14}$$

$$q(l/l) = [I - k(l)H]q(l/l-1) \tag{15}$$

$$\hat{\mathrm{x}}(l+1/l) = \mathrm{F}(l)\hat{\mathrm{x}}(l/l) + \mathrm{G}_{\bar{\mathrm{u}}} \tag{16}$$

$$Q(l+1/l) = F(l)P(l/l)F^T(l) + GG^T\sigma_u^2 \tag{17}$$

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

The predicted state error vector is $\hat{x}(l/l-1) = x(l) - \hat{x}(l/l-1)$.

$Q(l/l-1) = E[\tilde{x}(l/l-1)\tilde{x}^T(l/l-1)]$ is *predicted* state error correlation matrix where

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

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

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

$K(l)$ is the Kalman gain and $e(l)$ is the innovation sequence.

The estimated signal can be retrieved from the state vector estimator

$$\hat{s}(l) = H\hat{x}\left(\frac{l}{l}\right) \tag{18}$$

4 Experimental Results

The efficiency of the proposed system BEAM-KAL is tested under the speech signal corrupted with car interior noise at various inputs SNRs −6, 3 dB [7] which was taken from NOIZEUS database [7]. Perceptual evaluation of speech quality scores (PESQ) [8] for the proposed BEAM-KAL [9] are found to be consistently good at lower input SNR, i.e., −6 dB than individual filters shown in Fig. 2. The time domain graphs

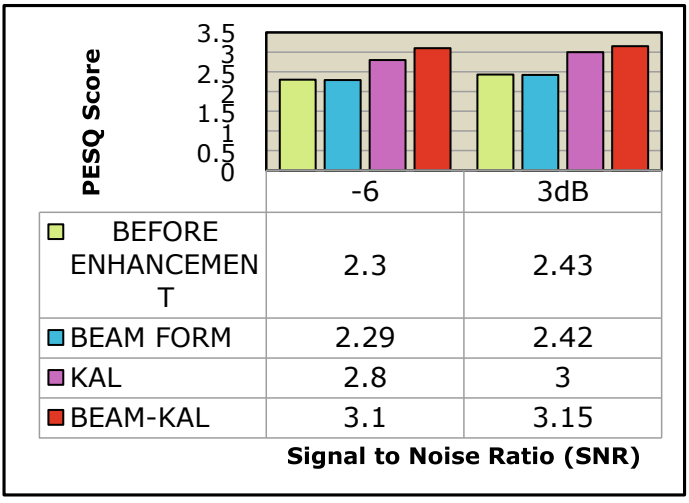

Fig. 2 PESQ scores for the proposed filter and with individual filters

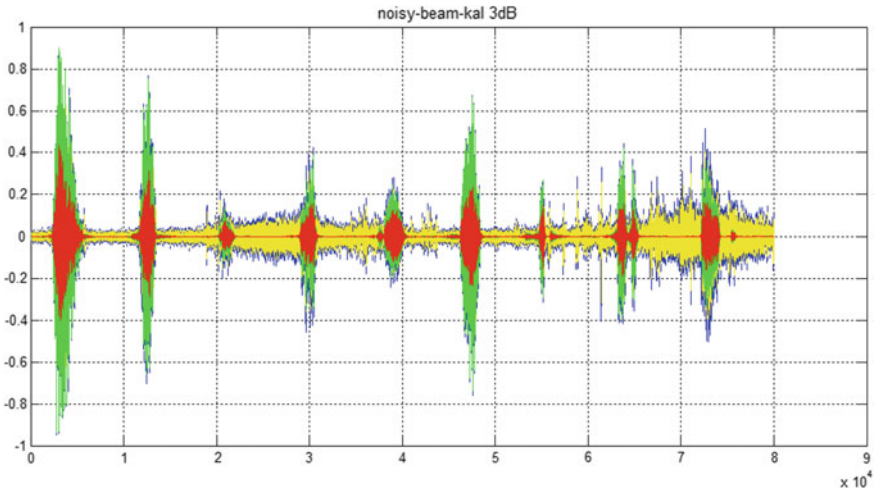

Fig. 3 Comparison of time domain plots at 3 dB. **a** Noisy signal (blue colour). **b** The enhanced beamform (yellow colour) **c**. The enhanced Kalman (green in colour). **d** Enhanced signal of the proposed BEAM-KAL (red colour) and the black circle shows the noise and it was removed

and spectrogram of noisy, beamform, Kalman and BEAM-KAL enhanced speech signals are done at 3 dB where the circles show the noise removal as in Figs. 3 and 4. From this, it is clear that BEAM-KAL combination is superior when compared to others in case of noise reduction. Speech enhancement must satisfy the listener by improving speech quality and hence subjective tests, like informal *A–B* testing, are used to evaluate the performance. In *A–B* testing, a group of people listening to a number of pairs of speech files (labelled *A* and *B*) are involved, and they decide which is better in each case. The obtained results show that the proposed system is extremely good at removing in-car noise, with excellent speech quality and high intelligibility even at low noise exhibits better performance than those obtained with the beamforming and Kalman filter (Table 1).

5 Conclusion

In this paper, a cascaded scheme BEAM-KAL based upon a combination of generalized side lobe canceller (GSC) beamformer and Kalman filter was proposed for the enhancement of speech signals corrupted with car noise. Simulation results were conducted, and the results of the proposed scheme are compared to the beamformer and Kalman filter individually at various input SNRs. The overall performance of the proposed method is shown to outperform beamforming and Kalman filter. Using an objective speech quality measure, spectrogram analysis, as well as formal subjective

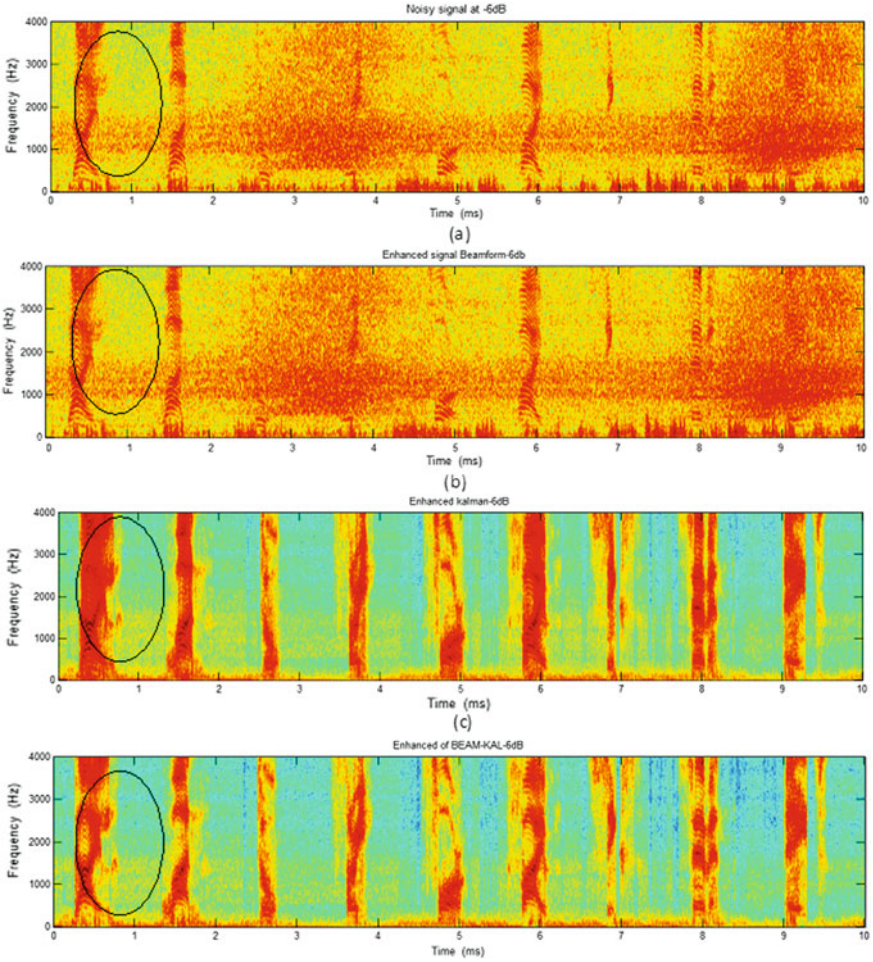

Fig. 4 Spectrogram of the speech sample sp01 from NOIZEUS data corrupted with car noise. **a** Corrupted speech with car noisy signal at −6 dB SNR. **b** Enhanced beamform the speech obtained signal at −6 dB. **c** The output of Kalman filter (enhanced signal at −6 dB). **d** Output signal of the proposed BEAM-KAL at input −6 dB SNR

Table 1 Subjective *A–B* test

Test	Listener preference		
	BEAM-KAL (%)	Other (%)	Undecided (%)
BEAM-KAL/noisy signal	100	0	0
BEAM-KAL/Beamform	96	4	0
BEAM-KAL/KAL	92	1	7

listening tests, showed that the proposed method is capable of reducing noise resulting in improved speech quality. Hence, this scheme provides a promising solution for real-time speech enhancement in noisy car environments.

References

1. Abut H, Hansen JHL, Takeda K (2005) DSP for in-vehicle and mobile systems. Springer
2. Poulat LD (2004) Robust speech recognition techniques evaluation for telephony server based in-car applications. In: ICASSP 2004, I-65–I-68
3. Paliwal K, Basu A (1987) A speech enhancement method based on Kalman filtering. In: Proceedings of IEEE international conference on acoustics speech and signal processing (ICASSP)
4. Van Veen Barry, Buckley Kevin M (1988) Beamforming: a versatile approach to spatial filtering. IEEE Sig Process Mag 5:4–24
5. Bitzer J, Simmer KU (2001) Superdirective microphone arrays. In: Brandstein MS, Ward DB (eds) Microphone arrays: signal processing techniques and applications. Springer, Berlin, pp 19–38 (Chapter 2)
6. Breed BR, Strauss J (2002) A short proof of the equivalence of LCMV and GSC beamforming. IEEE Sig Process Lett 9(6):168–169
7. Noizeus: a noisy speech corpus for evaluation of Speech enhancement algorithms, http://www.utdallas.edu/~loizou/speech/noizeus
8. ITU-T P.862 (2000) Perceptual evaluation of speech quality (PESQ) and objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs. ITU-T Recommendation, p 862
9. Ramesh Babu G, Rao R (2012) Combination of beamforming and Kalman filter techniques for speech enhancement. Int J Comput Sci Commun Netw 3(1):338–343