Audio Engineering Society

Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

A Novel Automatic Noise Removal Technique for Audio and Speech Signals

Harinarayanan.E.V[1], Deepen Sinha[2], Shamail Saeed[3], and Anibal Ferreira[4]

[1] ATC Labs, New Jersey, U.S.A
hari@atc-labs.com

[2] ATC Labs, New Jersey, U.S.A
sinha@atc-labs.com

[3] ATC Labs, New Jersey, U.S.A
shamail@atc-labs.com

[4] University of Porto, Portugal & ATC Labs, New Jersey, U.S.A
ajf@atc-labs.com

ABSTRACT

This paper introduces new ideas on wideband stationary/non-stationary noise removal for audio signals. Current noise reduction techniques have generally proven to be effective, yet these typically exhibit certain undesirable characteristics. Distortion and/or alteration of the audio characteristics of primary audio sound is a common problem. Also user intervention in identifying the noise profile is sometimes necessary. The proposed technique is centered on the classical Kalman filtering technique for noise removal but uses a novel architecture whereby advanced signal processing techniques are used to identify and preserve the richness of the audio spectrum. The paper also includes conceptual and derivative results on parameter estimation, a description of multi parameter Signal Activity Detector (SAD) and our new found improved results.

1. INTRODUCTION

In many commercial applications like broadcast and telecommunication field audio and/or speech signals are often degraded due to the presence of *noise* which can be either stationary or non-stationary. Persistence of background noise, like a hiss or a buzz, are examples of stationary noise. Presence of such noise for longer duration could severely lower the intelligibility and perceived quality of any speech/audio signal. On the other hand non-stationary

or transitory noise is short lived and in many cases occurs with a sudden boost in signal energy. Effects of transitory noise such as passing siren or background clapping sounds in a live commentary can also be disconcerting to the listener. In digital systems presence of noise has a secondary negative impact in that it may severely degrade the performance of low bit rate speech/audio coding schemes or speech recognition schemes which may be in use. Reduction and/or elimination of both stationary and non-stationary noises have therefore been found to be desirable in a whole range of applications. In the past, a number of different techniques [1],[3],[5],[6],[7] have been proposed to counter additive white noise for audio signals. These popular techniques fall under the broad classification of being either a time domain or a frequency domain based technique.

Most of the commercially available noise reduction tools use frequency domain based spectral subtraction technique. The theory of spectral subtraction involves noise spectrum estimation and subtraction on individual frequency bins on a frame by frame basis. In general, the inaccuracy involved in noise spectrum estimation and also, noise variance within a frame leaves prints of audible artifacts like annoying musical noise [3],[8] or short sinusoidal impulses [8] in the processed audio with varying frequencies for every frame. On the other hand, time domain noise filtering techniques use adaptive filters such as Kalman filter, Adaptive Weiner filter. Kalman filtering has often been investigated by authors for audio noise reduction with renditions [1],[6],[7]. The niche in using Kalman filter for noise reduction comes only with an accurate estimation of signal and noise parameters. These also often introduce unpleasant distortion in the signal such as increased *gargliness* in voices. In either class of algorithms the characteristics of the audio (such as perceived bandwidth and openness) is often adversely affected Furthermore, desirable level of noise reduction is only achieved with a user intervention in terms of selecting a suitable noise profile or identifying regions of the signal where noise is dominant, in effect these are two or more pass algorithms which are less suitable for real time applications. In this paper we introduce a novel noise reduction scheme that is designed around two core performance requirements. Firstly, the techniques achieves substantial noise reduction while preserving all the key audio characteristics of the primary signal sound and introduces minimal distortion to the primary sound. Secondly, identification of noise statistics is fully automatic and adaptive to any type of audio material. At the core of the proposed scheme is a perceptually optimized Adaptive Kalman Filtering algorithm. The first core performance requirement is achieved with an algorithm based on detailed frequency domain analysis, subtraction, and synthesis which helps preserve the richness of the original audio. The second performance requirement is achieved with the help of a new Signal Activity Detector and continuous update and validation of noise statistics. The rest of the paper is organized as follows: The second section introduces the architecture of the basic filtering model. The third section describes improved method for parameter estimation for Kalman filtering and we conclude by presenting results in the final section.

2. NOISE FILTERING

A high level architecture of the proposed generic noise filtering model is shown in Fig.1. The basic blocks of the architecture include a harmonic extraction block, a harmonic synthesis block, a Signal Activity Detector and a de-noising filter. The harmonic extraction along with harmonic synthesis preserves a subset of the harmonic structure of the original signal that is perceptually most significant. This technique helps in preserving the natural sound of the audio under interest. The noise removal algorithm works on the *residual* formed by this subtraction method. The processing of identifying perceptually significant subsets of harmonics is based on a detailed analysis of a *frame* of signal in both time and frequency domain. This analysis involves frequency domain perceptual modeling [5] and also an analysis of the *voiced/unvoiced* nature of the frame.

2.1. Choice of Adaptive Filtering Algorithm

The use of time domain filtering technique instead of frequency domain methods are for reasons stated under Section 1 like, artifacts distorting the original signal and also for the precision which is needed in the noise spectrum estimation. Among time domain techniques, Weiner filter is ideal for a short segment of 20-30 ms but, is not suited for non-stationary signals like audio. Though, an adaptive Wiener filter is optimum in the least-squares-sense, it does not

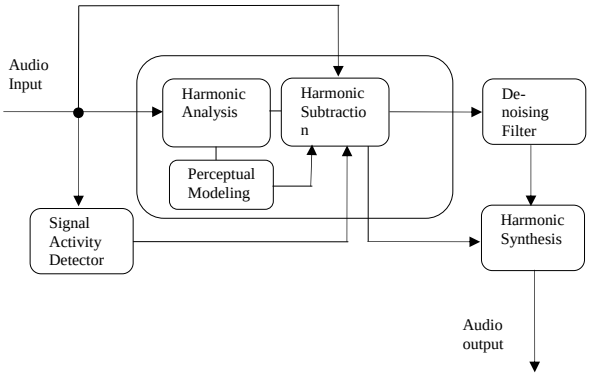

Figure 1: Architecture of Noise Reduction Module

exploit situations when additional knowledge about the signal such as the speech production process is available. Therefore Kalman filter for noise filtering is adopted in our scheme.

2.2. The Signal Model

The basic model of the clean (desired) audio and the additive noise measurements (available audio) are described by the following state space equations:

$$X_{k+1} = \Phi_k . X_k + Q_k \quad (1)$$

$$z_k = H_k X_k + r_k \quad (2)$$

Where,

X_k $(n \times 1)$ process state vector at time t_k

Φ_k $(n \times n)$ state transition matrix at time t_k relating X_k to X_{k+1}

Q_k $(n \times 1)$ process noise vector at time t_k

z_k Scalar measurement at time t_k

H_k $(1 \times n)$ measurement vector at time t_k relating X_k to z_k

r_k Scalar measurement noise at time t_k

The process noise vector Q_k and the scalar measurement noise r_k are assumed to be zero mean, white (time-uncorrelated) noise sequences with covariance structures given by $E[Q_k Q_i^T] = Q I_{\delta_{ki}}$ and $E[r_k r_i] = R\delta_{ki}$. Where, $E[.]$ denotes the expectation operator, the superscript T indicates a vector transpose, $I_{\delta_{ki}}$ is a $(n \times n)$ matrix with δ_{ik} along its diagonal with the rest of the values being zero, δ_{ik} is the Kronecker delta function and Q, R are scalar values. It is also assumed that Q_k and r_k are uncorrelated with initial state vector X_0.

2.3. Significance of the Harmonic Extraction Process

As shown in Fig.1 Harmonic Extraction block constitutes a harmonic analysis, perceptual modelling and a harmonic subtraction block. This block serves dual purpose by preserving the perceptually significant harmonics thereby preserving the main signal component and also this extraction process reduces the computational complexity by shrinking the LPC matrix order which would otherwise require exhaustive LPC matrix on account of long term LPC coefficients.

- Preserves the perceptually significant harmonic structure in the signal ensuring that it remains un-affected by the noise filtering process, thus ensuring that key natural characteristics of the audio are preserved.

- In the state space model, it is convenient and desirable to assume Φ_k to be a transition matrix corresponding to a Linear Predictive (LP) signal model [2]. This LP model in itself is generally not sufficient in ensuring that the resulting process noise, Qk, is uncorrelated white noise. For example a speech like signal typically involves both a long term as well as short term predictor to ensure suitable whitening. Similarly, for music signals, which can be quite rich in terms of their harmonic content, a very high order LP model may be necessary to ensure that the residual is sufficiently white. By introducing harmonic

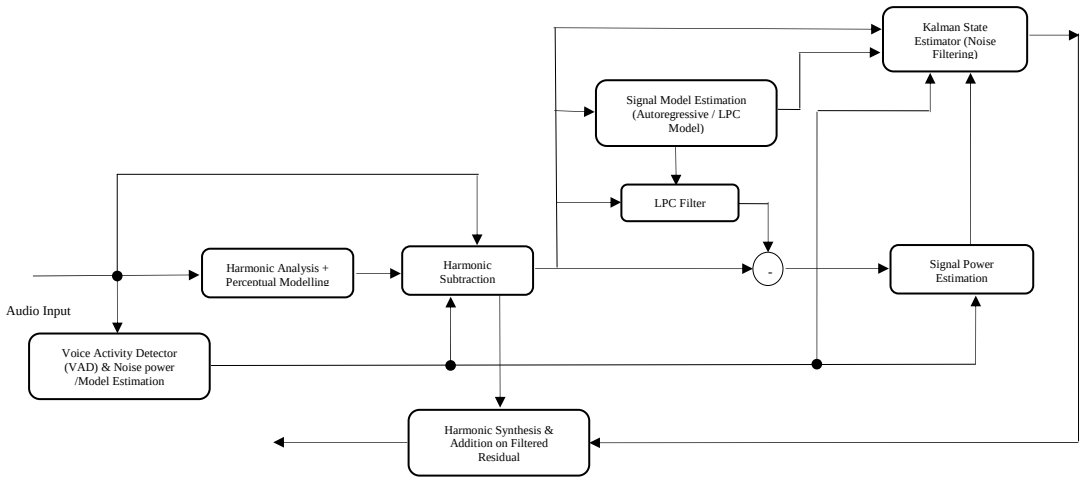

Figure 2: Kalman filter based noise reduction module

subtraction we can conveniently eliminate the need for including a long term predictor in the Φ_k model and also operate with a relatively small size LP model with its resulting numerical benefits.

2.4. Signal Activity Detector

The Signal Activity Detector (Fig.1) is a multi parameter based decision. The dependency of the algorithm on a multi parameter decision is to reduce the risk of a false alarm. From the architecture point of view signal activity decision plays a crucial role as the noise parameter (R) is updated only based on the decision of the detector, i.e, only for *noisy frames*. Fine tuning SAD is crucial as mentioned above, an aggressive SAD would be detrimental as the filter would filter even a signal as noise. On the other hand a conservative SAD decision would under filter the noise. The multi parameter decision of SAD discussed below has been fine tuned, in the implementation, with the goal of lowering the *false positive* duly because identifying signal as noise has a higher cost compared to a *miss*.

2.4.1. Correlation

Any Voice / Music signal exhibits characteristics of periodicity within a small window frame. Such signals with periodic nature are pinned by identifying periodic peaks on auto correlated input audio.

2.4.2. Fricative Detection

The auto-correlation based Signal Activity Detectors fair well except for those classes of genuine signals which are not periodic. Classic examples of such cases are *fricatives*, in the case of speech signals. From the studies of frequency domain envelope identification techniques [9] we employ a fricative detector to avoid filtering such fricatives.

2.4.3. Energy

The energy of the frame at analysis is expected to be higher than an approximate noise energy threshold. This threshold is set as a part of user setting – an approximate measure of expected noise. As this user parameter may not be an exact measure conservative decisions are taken based on the threshold input.

A noisy frame fails all the three tests, evaluated in the top down order listed above.

3. KALMAN FILTERING

The block diagram of the proposed noise reduction module is shown in Fig.2. In this section, we discuss estimation of parameters for Kalman filtering. The solution to Kalman filter equations requires knowledge of Φ_k, H_k, Q and R. As mentioned in section 2.4 we assume the transition matrix Φ_k (Eq.3)

to be based on a short term LP based signal production model. This matrix therefore has a simple sparse structure; the first row of the matrix is populated with N^{th} order LP filter coefficients. The second row through N^{th} row is populated by a shift operator whereby *(i-1)*th element of i^{th} row is 1 and remaining elements are zero. These LP coefficients are calculated by running a LP analysis on a short frame of noisy audio. The matrix H_k also follows a simple structure with a unit entry in its first row with the rest being zeros.

$$\Phi_k = \begin{bmatrix} \phi_{lpc}{}^{1} & \phi_{lpc}{}^{2} & . & . & . & \phi_{lpc}{}^{N} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ . & . & 1 & 0 & 0 & 0 \\ . & . & . & . & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}_{(NXN)} \quad (3)$$

3.1. Signal and Noise Parameter Estimation

Accuracy in the estimation of Q and R plays important roles in the stability of the Kalman filter. A wrong estimation of either Q or R could lead to an unstable filter. In our technique, we estimate the value of R during the portion of signal inactivity based on the decision from the VAD algorithm. In these frames the noise variance R_{new} is updated as per Eq.4.

$$R_{new} = \alpha R_{old} + (1-\alpha)\left(\sum_{i=1}^{M} z_i{}^{2}\right) / M \quad (4)$$

Where, α controls the influence of past variance on the updated value of noise variance and M is the frame size.

The value of Q is continuously updated during the active signal frames. The formula to estimate Q value is derived below.

The assumed H_k matrix structure has no influence on measurement values hence; it is dropped from Eq.2 at this stage of analysis. The measurement process (Eq.2) can be expanded as,

$$z_1 = \phi_{lpc}{}^{1}.x_0 + q_1 + r_1$$

$$z_2 = \phi_{lpc}{}^{1}.x_1 + \phi_{lpc}{}^{2}.x_0 + q_2 + r_2$$

$$z_n = \sum_{i=1}^{N} \phi_{lpc}^{i}.x_{n-i} + q_n + r_n \quad (5)$$

Where, ϕ_{lpc}^{i}s are the LPC coefficients, x_{n-i} is a scalar which is the ith element of the vector X_n and q_{n-i} is a scalar which is the ith element of the vector Q_n.

$$z_n = x_n + r_n \quad (6)$$

Eq.6 follows Eq.2. Substituting Eq.6 in Eq.5, Eq.5 can be recursively written in terms of z_{n-i} as,

$$z_n = \sum_{i=1}^{N} \phi_{lpc}^{i}.(z_{n-i} - r_{n-i}) + q_n + r_n \quad (7)$$

The above equation is a recursive reformulation of the basic audio model (Eq.1, 2).

LP analysis on the measurement values is given by the following equation,

$$\bar{z}_2 = \bar{\phi}_{lpc}^{1}.z_1$$

$$\bar{z}_3 = \bar{\phi}_{lpc}^{1}.z_2 + \bar{\phi}_{lpc}^{2}.z_1$$

$$\bar{z}_n = \bar{\phi}_{lpc}^{1}.z_{n-1} + \bar{\phi}_{lpc}^{2}.z_1 + \ldots + \bar{\phi}_{lpc}^{N}.z_{n-N} \quad (8)$$

Where, $\bar{z}_n$ is the linear prediction of z_n. Note, the LP coefficients in Eq.8 are not same as the coefficients in the signal model (Eq.7) because $\bar{\phi}_{lpc}^{i}$ are calculated after a LP analysis on the measurements and not from the noise free audio. But, at this point of analysis we approximate both of these coefficients for the sake of a simple formulation. The LP residual (v) is calculated by subtracting Eq.8 from Eq.7.

$$v = z_n - \bar{z}_n$$

$$= \left(\sum_{i=1}^{N} \phi_{lpc}^{i}.(z_{n-i} - r_{n-i}) + q_n + r_n\right) - \left(\sum_{i=1}^{N} \bar{\phi}_{lpc}^{i}.z_{n-i}\right)$$

$$v = q_n + r_n - \left(\sum_{i=1}^{N} \phi_{lpc}^{i} . r_{n-i} \right) \quad (9)$$

The measurement noise and process noise are uncorrelated and also, the measurement noise is assumed to be white. This reduces the variance of Eq.9 to

$$V_{\text{var}} = Q + R.(1 + \sum_{i=1}^{N} \phi_{lpc}^{i\ 2}) \quad (10)$$

Where, V_{var} is the variance of the random variable v. With the measurement of residual variance and the knowledge of R and ϕ_{lpc}^{i}, Eq.9 gives an estimate of Q.

4. RESULTS

The proposed algorithm was run on multiple stereo speech/music samples. We conducted a subjective measure on the database collection. The samples were acquired from various real life broadcast and communication setups with high level background noise with varying characteristics.

In our implementation of the overall scheme LP filter of order 15 (for the transition matrix in the Kalman filter) was found to be sufficient for the entire range of audio and a frame length of 1024 samples with input sampling rate of 32Kbps was used.

For subjective evaluation, a group of five *expert listeners* were asked to evaluate the samples against some of the commercially available noise removal modules. The other products that were put to test include Dart XP Pro V1.1.6p, Adobe Audition 1.5 and GoldWavev5.19. The rating was on a scale of 1 to 5 with 5 being the highest quality. The ratings were based on various criterion like effective noise suppression, any distortion introduced after processing, clarity/naturalness of the audio. The graph plotted shows the ratings of this experiment. Also, there were interesting feedbacks from the listener panel regarding these products which are being summarized here: GoldWave suppresses the noise completely but at the expense of distorting the main audio. Musical noise was noticed more in

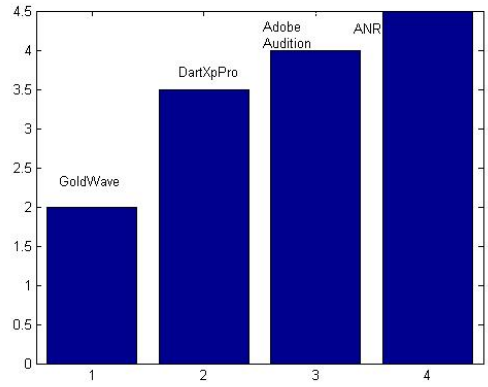

Figure 3 : Subjective results from different noise removal products

GoldWave compared to the other products. Dart Xp Pro's algorithm removes noise to a considerable extent but on the downside the algorithm seems to suppress the high frequency content of the signal or in other words the de-noised output has a low-pass filtered effect making it unfit for an audience expecting wideband audio. Adobe Audition 1.5's noise removal tool cleans up the noise considerably, the de-noised audio was wideband unlike Dart Xp Pro but, there were clear traces of musical noise, lesser than GoldWave. Also interestingly, the wideband speech signals exhibited signatures of comb filtering type distortions. Lastly, our method was unanimously appreciated for de-noising without distorting the main signal components of the audio. Also, the algorithm was appreciated for its wideband output. As a downside, occasional traces of noise were found to fade in and fade out of the audio in some of the samples. The audio samples used in this evaluation are available at http://www.atc-labs.com/anr.

5. ACKNOWLEDGEMENT

We acknowledge the support of Sirius Satellite Radio in facilitating this work. Constructive feedbacks and suggestions from Mr. Mark Kalman, Mr. Jim Tracey, and, Ms. Bhadresha Dedhia are appreciatively acknowledged.

6. REFERENCES

[1] K.K. Paliwal, A. Basu, A speech enhancement method based on Kalman filtering, Proceedings of ICASSP `87, Dallas, TX, USA, 1987, pp. 177-180.

[2] V.K. Jain and B.S. Atal, Robust LPC analysis of speech by extended correlation matching, Proc. ICASSP, 1985, pp. 473-476.

[3] O. Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Trans. Speech and Audio Processing, Vol.2, No. 1, pp. 345-349, April 1994.

[4] T.V. Ramabadran and D. Sinha. Speech Data Compression Through Sparse Coding of Innovations. IEEE Trans. Speech, Audio,. 2(2):274-284, 1994.

[5] Vijay K. Madisetti and Douglas B. Williams, *The Digita Signal Processing Handbook*, IEEE Press, 1998.

[6] M. Gabrea, Robust Adaptive Kalman Filtering-based Speech Enhancement Algorithm, in Proc. ICASSP'04, pp. 301-304.

[7] S. Gannot, D. Burshtein, and E. Weinstein, Iterative and Sequential Kalman Filter-Based Enhancement Algorithms, IEEE Trans. Speech and Audio Processing, vol.6, pp. 373-385, July 1998.

[8] M. Lorber and R. Holdrich, A Combined Approach for Broadband Noise Reduction, in Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 1997.

[9] Shrikanth Narayanan and Abeer Alwan, Noise Source Models for Fricative Consonants. IEEE Trans. Speech and Audio Processing, vol.8, No.2, March 2000.