Autoregressive Hidden Markov Model based Speech Enhancement using Sparsity

A. Gayathri
PG Scholar, Department of ECE, DVR and Dr. HS MIC College of Technology
Kanchikacherla, A.P, India.

G. Chenchamma, PhD
Professor, Department of ECE, DVR and Dr. HS MIC College of Technology
Kanchikacherla, A.P, India.

K. V. V. Kumar
Assistant Prof., Department of ECE, DVR and Dr. HS MIC College of Technology
Kanchikacherla, A.P, India.

ABSTRACT
Speech enhancement is required to enhance the quality of speech corrupted by the background noise and can be used in many applications such as hearing aids, mobile communication etc. In this paper a speech enhancement method is presented in which first Autoregressive (AR) model is applied for the noisy speech signal to find the speech parameters and then Hidden Markov model is applied to model those parameters. Later, the sparsity is encouraged into the model by adding the regularization parameter. The objective results for the proposed method and Wiener filter are compared. Speech quality in non-stationary noise conditions is observed through listening. The average log-likelihood score is obtained for different noises and observed that the performance is improved compared to the reference methods.

Keywords
Speech enhancement, non-stationary noise, sparse autoregressive hidden markov model (SARHMM).

1. INTRODUCTION
In general, speech enhancement is an efficient approach to improve the speech signal corrupted by background noise. The speech enhancement is trade-off between the reduction of noise level and speech distortion. Therefore, the trade-off is chosen according to the particular application for the best speech quality. The aim of speech enhancement is to improve the quality and intelligibility of degraded speech. The characteristics of the speech signal depends on the nature and characteristics of the noise signal and changes with the time when it is corrupted by the noise. The design of algorithm differs from the application to application and so the performance of the algorithm can also differ for each application. The nature of the noise is a predominant factor which decides the speech enhancement method. Therefore, a good noise model is important to know the performance of speech enhancement system and also to analyze how well a speech enhancement method works with different types of noise. Noise may be different based on various statistical, spectral or spatial properties. A noise is said to be non-stationary if it's spectral properties change with time. It is difficult to analyze the non-stationary noises as noise changes with the surrounding environment. Over the years, many speech enhancement algorithms have been proposed. But still the research has been continued to improve the quality of the speech signal. The spectral subtraction algorithm does not require prior information and is very simple to implement. The Wiener filtering algorithm derives the enhanced signal in an optimal way but it requires prior information of speech and noise. The above two algorithms do not perform well in non-stationary environments. Later Autoregressive Hidden Markov models (ARHMM) and codebooks have been proposed and used successfully to model the statistics of speech and noise for speech enhancement in non-stationary environments. These methods modeled the speech and noise signals as AR processes and their spectral characteristics are modeled by considering signal gain as a deterministic parameter instead of random variable. The gain variances of the speech and noise are modeled accurately which can play an important role in speech enhancement for non-stationary noise environments.

ARHMM is further improved and the speech and noise gains are considered as random processes that describe the power levels of speech and noise [9]. By learning the speech and noise characteristics on-line, prior information of the gains can be obtained the more accurately. The combinations of speech and noise spectral shapes results in ambiguity to distinguish speech spectral shapes and noise spectral shapes. Therefore, it is difficult to separate the speech and the noise components. This problem is known as ambiguity problem. The ambiguity problem increases with the number of states of HMM and the Gaussian mixture components per state of the HMM. This problem will be less if only few states are considered and number of Gaussian mixture components per state is low. But using only small number of states and restricting the number of noise environments do not perform well for speech enhancement [8]. Thus, the ambiguity problem limits the overall performance of the ARHMM model. Besides this problem ARHMM has another problem known as inherent problem that does not model the spectral fine structure of speech. The model parameters of the ARHMM are estimated by maximizing the likelihood. The solution to ambiguity problem is to introduce sparsity to ARHMM model. In this paper, sparse ARHMM is used to improve the performance which is limited by ARHMM.

This paper is organized as follows. In Section 2, we present Autoregressive Hidden Markov Model. Signal model of ARHMM for speech and noise is provided in Section 3. In Section 4, we present Parameter Estimation. In Section 5, we present Speech and Noise Estimation of ARHMM using Sparsity. Simulation results are provided in Section 6. Conclusion is provided in Section 7.

2. AUTOREGRESSIVE HIDDEN MARKOV MODEL
Autoregressive Hidden Markov model (ARHMM) is a class of Hidden Markov model which is particularly applicable for speech processing. For ARHMM, the observation vectors are drawn from an autoregression process [4], [5].

Consider the observation vector,

$$O = (x_0, x_1, x_2, \dots x_{K-1}) \quad (1)$$

The elements x_i could be the speech waveform samples.

The components of O are assumed to be from an autoregressive Gaussian source.

$$x_k = -\sum_{i=1}^{p} a_i x_{k-i} + e_k \quad (2)$$

where e_k, $k = 0,1,2, \dots K-1$ are Gaussian, independent, identically distributed random variables with zero mean and variance σ_e^2 and a_i, $i = 1,2, \dots p$, are autoregression or prediction coefficients. For large value of K, the density function is given as follows:

$$f(0) = (2\pi\sigma_e^2)^{-k/2} \exp\left\{-\frac{1}{2\sigma_e^2}\delta(0,a)\right\} \quad (3)$$

where

$$a = \left[1, a_1, a_2, \dots a_p\right]', \quad (a_0 = 1) \quad (4)$$

If $r_a(i)$ is the autocorrelation of the autoregressive coefficients,

$$r_a(i) = \sum_{n=0}^{p-i} a_n a_{n+i}, \qquad 1 \le i \le p \quad (5)$$

If $r(i)$ is the autocorrelation of the observation samples,

$$r(i) = \sum_{n=0}^{k-i-1} x_n x_{n+i}, \qquad 0 \le i \le p \quad (6)$$

$\delta(0,a)$ is a residual energy resulting from inverse filtering the data x_i with an all-zero filter defined by a.

To separate the signal level from the spectral shape, gain normalization is used.

$$\hat{O} = O/\sigma_{e0} \quad (7)$$

where σ_{e0}^2 is the minimum linear prediction residual energy per sample.

The elements of $\hat{O}$, $\hat{x}_k = x_k/\sigma_{e0}$, still satisfy the autoregressive relationship

$$\hat{x}_k = -\sum_{i=1}^{p} a_i \hat{x}_{k-i} + \hat{e}_k \quad (8)$$

where variance of $\hat{e}_k$ is unity.

The probability density function for the output of an all-pole system defined by a, with zero mean and unit variance Gaussian independent and identically distributed (i.i.d) sequence can be given as follows:

$$f(\hat{O}) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\delta(\hat{O},a)\right\} \quad (9)$$

This type of pdf is often referred to as a 'gain -independent' pdf.

Assume a mixture density of the form

$$b_j(0) = \sum_{k=1}^{M} C_{jk}\, b_{jk}(0) \quad (10)$$

where

$$b_{jk}(0) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\delta(O,a_{jk})\right\} \quad (11)$$

a_{jk} is the autoregression vector.

Autocorrelation sequence for the jth state, kth mixture component is of the form,

$$\bar{\mathbf{r}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \mathbf{r}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \quad (12)$$

where $\mathbf{r}_t = [\mathbf{r}_t(0), \mathbf{r}_t(1), \dots, \mathbf{r}_t(p)]'$ is the autocorrelation vector for tth frame.

$\gamma_t(j,k)$ is defined as the probability of being in state j at the time t and using mixture component k.

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\right]\left[\frac{C_{jk}\, b_{jk}(O_t)}{\sum_{k=1}^{M} C_{jk}\, b_{jk}(O_t)}\right] \quad (13)$$

Note that $\bar{\mathbf{r}}_{jk}$ is a weighted sum of the normalized autocorrelation of the frames in the observation sequence. A set of normal equations can be solved from $\bar{\mathbf{r}}_{jk}$, to obtain the corresponding autoregressive coefficient vector $\bar{\mathbf{a}}_{jk}$ for kth mixture of state j.

3. SIGNAL MODEL OF ARHMM

Assuming that the clean speech signal is corrupted by an uncorrelated additive noise, the noisy speech signal can be modeled as follows:

$$Y_t = X_t + W_t \quad (14)$$

where Y_t, X_t and W_t are random vectors which represents frame segments of the noisy speech, clean speech and noise signals respectively denoting t as the index of the speech frame.

Consider, each frame contains K signal samples, i.e., $\mathbf{y}_t = \{y_t(0), y_t(1), \dots y_t(K-1)\}$, $\mathbf{x}_t = \{x_t(0), x_t(1), \dots x_t(K-1)\}$, and $\mathbf{w}_t = \{w_t(0), w_t(1), \dots w_t(K-1)\}$.

3.1 Speech Model

Let, $x_0^{T-1} = \{x_0, x_1, \dots x_{T-1}\}$ denote a T-frame clean speech sequence from frame 0 to $T-1$ frame and $p(x_0^{T-1})$ be the probability density function of the model for the clean speech sequence x_0^{T-1}.

The statistics of the clean speech frame sequence x_0^{T-1} are modeled by an $\bar{N}$ state ARHMM model and is given as follows:

$$p(x_0^{T-1}) = \sum_{\bar{s}_0^{T-1}} \prod_{t=0}^{T-1} \bar{a}_{\bar{s}_{t-1}\bar{s}_t} p_{\bar{s}_t}(x_t) \quad (15)$$

where $\bar{s}_0^{T-1} = \{\bar{s}_t\}_{t=0,1,\dots,T-1}$ denotes a sequence of speech ARHMM states and $\bar{s}_t \in \{1,2,\dots,\bar{N}\}$ denotes the state of speech at frame t. $\bar{a}_{\bar{s}_{t-1}\bar{s}_t}$ is the state transition probability from state $\bar{s}_{t-1}$ at frame $T-1$ to $\bar{s}_t$ at frame t and $\bar{a}_{\bar{s}_{-1}\bar{s}_0}$ is probability of the initial state $\bar{s}_0$. $p_{\bar{s}_t}(x_t)$ is probability density function of the clean speech frame x_t for given state $\bar{s}_t$.

$$p_{\bar{s}_t}(x_t) = \int_{-\infty}^{\infty} p_{\bar{s}_t}(\bar{g}_t')\, p_{\bar{s}_t}(x_t|\bar{g}_t')\, d\bar{g}_t' \quad (16)$$

where $\bar{g}_t' = \log(\bar{g}_t)$ and $\bar{g}_t$ denotes the linear speech gain, which is the variance of the prediction error of AR model.

Now consider, the speech gain $\bar{g}_t$ as a stochastic process and model the probability density function (pdf) $p_{\bar{s}_t}(\bar{g}_t^{'})$ of $\bar{g}_t$ as a state-dependent log-normal distribution:

$$p_{\bar{s}_t}(\bar{g}_t^{'}) = \frac{1}{\sqrt{2\pi\bar{\sigma}_{\bar{s}_t}^2}} \exp\left(-\frac{\left[\bar{g}_t^{'} - \left(\bar{u}_{\bar{s}_t} + \bar{q}_t\right)\right]^2}{2\bar{\sigma}_{\bar{s}_t}^2}\right) \quad (17)$$

where $\bar{u}_{\bar{s}_t} + \bar{q}_t$ denotes a mean value composed of a global average $\bar{u}_{\bar{s}_t}$ and a local bias $\bar{q}_t$, $\bar{\sigma}_{\bar{s}_t}^2$ denotes variance.

The parameters $\bar{u}_{\bar{s}_t}$ and $\bar{\sigma}_{\bar{s}_t}^2$ are time-invariant and can be estimated off-line together with the other speech ARHMM model parameters using training of speech data. The parameter $\bar{q}_t$ is used to compensate for the speech-gain bias, which can be estimated and updated on-line.

$p_{\bar{s}_t}(x_t|\bar{g}_t^{'})$ is the pdf of the clean speech vector x_t, given the speech gain $\bar{g}_t^{'}$.

Assume speech to be a zero-mean $\bar{p}$th order Gaussian AR processes and the conditional probability density function can be described as follows:

$$p_{\bar{s}_t}(x_t|\bar{g}_t^{'}) = \frac{\exp\left(-\frac{1}{2\bar{g}_t} x_t^{\#} \bar{\boldsymbol{D}}_{\bar{s}_t}^{-1} x_t\right)}{(2\pi\bar{g}_t)^{K/2}\left|\bar{\boldsymbol{D}}_{\bar{s}_t}\right|^{1/2}} \quad (18)$$

where # denotes Hermitian transposition, $\bar{\boldsymbol{D}}_{\bar{s}_t} = \left(\bar{\boldsymbol{A}}_{\bar{s}_t}^{\#} \bar{\boldsymbol{A}}_{\bar{s}_t}\right)^{-1}$ is the covariance matrix of the AR process, $\bar{\boldsymbol{A}}_{\bar{s}_t}$ is a $K \times K$ lower triangular Toeplitz matrix in which the first column is $\left[\bar{\alpha}_0, \bar{\alpha}_1, \ldots, \bar{\alpha}_{\bar{p}}, 0, \ldots 0\right]^T$, where $\bar{\alpha}_1, \ldots, \bar{\alpha}_{\bar{p}}$ constitute the speech AR coefficients and $\bar{\alpha}_0 = 1$.

3.2 Noise Model

Now the noise is modeled similar to the speech to capture the high diversity and variability of acoustical noises in a non-stationary environment. Thus, an ARHMM is used for the noise that is nearly identical to the ARHMM for the speech. The noise model parameters are labelled as '..', in contrast to the overbar '-' for speech model parameters. The probability density function of the noise frame ω_t for given state $\ddot{s}_t$ is $p_{\ddot{s}_t}(\omega_t)$ and is given as follows:

$$p_{\ddot{s}_t}(\omega_t) = \int_{-\infty}^{\infty} p(\ddot{g}_t^{'})\, p_{\ddot{s}_t}(\omega_t|\ddot{g}_t^{'}) d\ddot{g}_t^{'} \quad (19)$$

Where $\ddot{g}_t^{'} = \log(\bar{g}_t)$ and $\ddot{g}_t$ denotes the linear noise gain, which is the variance of the prediction error of AR model. The noise gain can be modeled as follows:

$$p(\ddot{g}_t^{'}) = \frac{1}{\sqrt{2\pi\ddot{\sigma}^2}} \exp\left(-\frac{(\ddot{g}_t^{'} - \ddot{u}_t)^2}{2\ddot{\sigma}_{\ddot{s}_t}^2}\right) \quad (20)$$

The mean value $\ddot{u}_t$ is a time-varying parameter that can be estimated and updated on-line together with the parameter $\bar{q}_t$.The conditional probability density of the noise, is defined similarly to that of conditional probability density function of the speech.

3.3 Noisy Speech Model

The probability density function of the noisy speech sequence y_0^{T-1} can be derived based on the speech and noise ARHMM models as follows:

$$p(y_0^{T-1}) = \sum_{s_0^{T-1}} \prod_{t=0}^{T-1} \bar{a}_{s_{t-1}s_t} p_{s_t}(y_t) \quad (21)$$

where $s_0^{T-1} = \{s_t\}_{t=0,1,\ldots,T-1}$ denotes sequence of noisy speech states and $s_t = (\bar{s}_t, \ddot{s}_t)$ denotes the noisy speech state for frame t, which is a composite state of speech and noise. $\bar{a}_{s_{t-1}s_t} = \bar{a}_{\bar{s}_{t-1}\bar{s}_t} \ddot{a}_{\ddot{s}_{t-1}\ddot{s}_t}$ is the transition probability from the composite state s_{t-1} at frame $t-1$ to the composite state s_t at frame t, and $p_{s_t}(y_t)$ denotes the pdf of the noisy speech y_t for a given composite state y_t. Note that there are $\bar{N} \times \ddot{N}$ states in the noisy speech model.

The joint speech and noise density $p_{s_t}(y_t)$ can be written as

$$p_{s_t}(y_t) = \iint p_{s_t}(y_t, \bar{g}_t^{'}, \ddot{g}_t^{'})\, d\bar{g}_t^{'}\, d\ddot{g}_t^{'} \quad (22)$$

$$= \iint p_{\bar{s}_t}(\bar{g}_t^{'}) p(\ddot{g}_t^{'}) p_{s_t}(y_t|\bar{g}_t^{'}, \ddot{g}_t^{'})\, d\bar{g}_t^{'}\, d\ddot{g}_t^{'}$$

where the pdf $p_{s_t}(y_t|\bar{g}_t^{'}, \ddot{g}_t^{'})$ is a Gaussian distribution with zero mean and covariance $\boldsymbol{D}_{s_t} = \bar{g}_t \bar{\boldsymbol{D}}_{\bar{s}_t} + \ddot{g}_t \ddot{\boldsymbol{D}}_{\ddot{s}_t}$ that can be given as follows:

$$p_{s_t}(y_t|\bar{g}_t^{'}, \ddot{g}_t^{'}) = \frac{\exp\left(-\frac{1}{2} y_t^{\#} \boldsymbol{D}_{s_t}^{-1} y_t\right)}{(2\pi)^{K/2}\left|\boldsymbol{D}_{s_t}\right|^{1/2}} \quad (23)$$

The pdf in equation (22) can be approximated as follows:

$$p_{s_t}(y_t, \bar{g}_t^{'}, \ddot{g}_t^{'}) \approx p_{s_t}(y_t, \bar{g}_t^{'}, \ddot{g}_t^{'}) \delta(\bar{g}_t^{'} - \hat{\bar{g}}_t^{'}) \delta(\ddot{g}_t^{'} - \hat{\ddot{g}}_t^{'}) \quad (24)$$

where $\delta(.)$ is a Dirac delta function and $\{\hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}\}$ are the optimal speech and noise gains.

$$\{\hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}\} = \arg\max_{\bar{g}_t, \ddot{g}_t} \log p_{s_t}(y_t, \bar{g}_t^{'}, \ddot{g}_t^{'}) \quad (25)$$

Therefore, according to the obtained optimal speech and noise gain pair $\{\hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}\}$, the pdf $p_{s_t}(y_t)$ of noisy speech in (22) can be approximated as

$$p_{s_t}(y_t) \approx p_{s_t}(y_t, \bar{g}_t^{'}, \ddot{g}_t^{'}) \quad (26)$$

4. PARAMETER ESTIMATION

The time-invariant parameters are estimated off-line and time-variant parameters are estimated and updated online.

4.1 Off-line Parameter Estimation

ARHMM parameters are commonly estimated using the Baum-Welch approach that is based on the expectation maximization (EM) algorithm. The EM algorithm iterates between the expectation (E) step and the maximization (M) step of the Q function. The Q function can be split into separate terms for the three types of model parameters $\theta = (\pi, A, B)$:

1. The initial distribution of states (π).

2. The state transition probability matrix (A).

3. The observation probability matrix (B).

If the speech model is taken as an example, then the function Q can be given as follows:

$$Q(\theta, \theta^{'}) = \sum_{\bar{s}_t} \log p(\bar{s}_t|O, \theta) p(\bar{s}_t \backslash O, \theta^{'}) \quad (27)$$

$$= \sum_{\bar{s}_t} \log \pi_{\bar{s}_0} p(\bar{s}_t|O,\theta') + \sum_{\bar{s}_t} \sum_{t=0}^{T-1} \log \bar{a}_{\bar{s}_{t-1}\bar{s}_t} p(\bar{s}_t|O,\theta') + \sum_{\bar{s}_t} \sum_{t=0}^{T-1} \log p_{\bar{s}_t}(x_t) p(\bar{s}_t|O,\theta')$$

where $\bar{s}_t$ denotes state of speech and $O$is observation sequence, $O = \{x_t\}_{t=0}^{T-1}$ and θ represents model parameters and θ' previous estimation of model parameters θ.

As the three types of model parameters can be optimized independently, we can encourage sparsity for the transition probability term and the observation probability term to derive the SARHMM.

First, the sparsity is induced to the transition probabilities and update equation can be derived for the transition probabilities. The sparsity for the transition probabilities $\bar{a}_{\bar{s}_{t-1}\bar{s}_t}$ of the ARHMM can be encouraged by introducing the l_p norm $H(A)$ to the second term of equation (27), and then maximizing

$$\sum_{\bar{s}_t} \sum_{t=0}^{T-1} log\, \bar{a}_{\bar{s}_{t-1}\bar{s}_t} p(\bar{s}_t|O,\theta') - \eta_1 H(A) \tag{28}$$

$$= \sum_{i}^{\bar{N}} \sum_{j}^{\bar{N}} \sum_{t=0}^{T-1} \log \bar{a}_{ij}\, p(\bar{s}_{t-1} = i, \bar{s}_t = j|O,\theta') - \eta_1 H(A)$$

where θ' represents previous parameters estimation of SARHMM, A is the matrix of transition probabilities, $H(A) = \|A\|_{1,p_1} = \sum_i \left(\sum_j \bar{a}_{ij}^{p_1}\right)^{1/p_1}$ is l_p regularization norm.

Although the l_1 norm encourages sparsity, we cannot directly use the l_1 norm, because the transition probability A and the observation probability B are stochastic matrices. Their entries are non-negative and the summation of each row must be 1, thus l_1 regularization norm is meaningless. The p_1 in the l_1 norm is a regularization parameter, which encourages sparsity for $0 \le p_1 \le 1$. The equation of transition probabilities of SARHMM can be updated by setting the derivation of equation (28) to zero and satisfying the constraints $\sum_j^{\bar{N}} \bar{a}_{ij} = 1$ and $\bar{a}_{ij} \ge 0$ for each state i.

$$\bar{a}_{ij} = \frac{\max\left(\sum_{t=0}^{T-1} p(\bar{s}_{t-1} = i, \bar{s}_t = j|O,\theta') - \eta_1 \bar{A}_{ij}, 0\right)}{\sum_h^{\bar{N}} \max\left(\sum_{t=0}^{T-1} p(\bar{s}_{t-1} = i, \bar{s}_t = h|O,\theta') - \eta_1 \bar{A}_{ih}, 0\right)} \tag{29}$$

where the maximization operation max (.,0) is added to make sure that all the transition probabilities are greater than zero and $\bar{A}_{ij}$ is the regularization term for the transition probability and can be given as

$$\bar{A}_{ij} = \bar{a}_{ij} \nabla_{\bar{a}_{ij}} H(A) = \bar{a}_{ij} \left[\bar{a}_{ij} \Big/ \left(\sum_h \bar{a}_{ih}^{p_1} \right)^{1/p_1} \right]^{p_1 - 1} \tag{30}$$

where ∇ is a differential operator.

From the equation (29) it can be observed that the low transition probabilities are rapidly driven to zero and strong transition probabilities are enforced. Thus, it is ensured that only a few states have a significant probability for transition. Second, the sparsity can also be encouraged to the observation probability $p_{\bar{s}_t}(x_t)$ of speech ARHMM by introducing the l_p norm $H(B)$ to the third term of equation (14).

$$\sum_{\bar{s}_t} \sum_{t=0}^{T-1} log\, p_{\bar{s}_t}(x_t) p(\bar{s}_t \backslash O,\theta') - \eta_2 H(B) \tag{31}$$

where $H(B) = \|B\|_{1,p_2} = \sum_{\bar{s}_t} \left[\sum_t p_{\bar{s}_t}^{p_2}(x_t) \right]^{1/p_2}$, B is the observation probability matrix and p_2 is regularization parameter. A regularization term $\bar{B}_{\bar{s}_t,x_t}$ for the observation probability of speech SARHMM can be given as

$$\bar{B}_{\bar{s}_t,x_t} = p_{\bar{s}_t}(x_t) \nabla_{p_{\bar{s}_t}(x_t)} H(B) \tag{32}$$

$$= p_{\bar{s}_t}(x_t) \left[\frac{p_{\bar{s}_t}(x_t)}{\left(\sum_t p_{\bar{s}_t}^{p_2}(x_t) \right)^{1/p_2}} \right]^{p_2 - 1}$$

The update equations of the training parameters of the observation probability of SARHMM can be derived by using the regularization term $\bar{B}_{\bar{s}_t,x_t}$ of equation (32).The parameters of the observation probability term in the equation (27) are $\bar{\theta} = \{\bar{\mu}_{\bar{s}}, \sigma_{\bar{s}}^2, \bar{\alpha}_{\bar{s}}, \bar{q}_r\}$, which represent the mean of the speech-gain model, the variance of the speech-gain model, the AR coefficients of speech-gain model and the gain bias of speech model, respectively. The speech gain bias is assumed to be a constant for each speech training utterances. Thus $\bar{q}_r$ denotes the speech gain bias of the rth utterance. Therefore, the third term in the equation (27) can be written as the auxiliary function $\Theta\left(\bar{\theta}|\hat{\bar{\theta}}^{(j-1)}\right)$.

$$\Theta\left(\bar{\theta}|\hat{\bar{\theta}}^{(j-1)}\right) = \sum_r \sum_{\bar{s}} \sum_t \bar{\omega}(\bar{s}_t) \int p_{\bar{s}_t}\left(\bar{g}_t'|x_t, \hat{\bar{\theta}}^{(j-1)}\right) \left[\log p_{\bar{s}_t}(x_t|\bar{g}_t', \theta') + \log p_{\bar{s}_t}(\bar{g}_t'|\bar{\theta})\right] d\bar{g}_t' \tag{33}$$

where $\bar{\omega}(\bar{s}_t)$ represents the posterior state probability.

In the above equation j denotes the iteration index, r denotes the index of the speech utterance in the database, $\{\bar{s}_0^{T-1}, \bar{g}_0^{T-1}\}$ denote the missing data of the EM algorithm, that are the sequence of the underlying states and speech gains.

$$\bar{\omega}(\bar{s}_t) = p\left(\bar{s}_t|x_0^{T-1}, \hat{\bar{\theta}}^{(j-1)}\right) \tag{34}$$

The posterior state probability $\bar{\omega}(\bar{s}_t)$ can be estimated by the forward-backward algorithm which is used for HMMs.

The sparsity can be encouraged to the observation probabilities by applying the regularization term $\bar{B}_{\bar{s}_t,x_t}$ to posterior state probability $\bar{\omega}(\bar{s}_t)$. The auxiliary function $\Theta\left(\bar{\theta}|\hat{\bar{\theta}}^{(j-1)}\right)$ then becomes auxiliary function $\tilde{\Theta}\left(\bar{\theta}|\hat{\bar{\theta}}^{(j-1)}\right)$.

$$\tilde{\Theta}\left(\bar{\theta}|\hat{\bar{\theta}}^{(j-1)}\right) = \sum_r \sum_{\bar{s}} \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t,x_t}, 0\right) \int p_{\bar{s}_t}\left(\bar{g}_t'|x_t, \hat{\bar{\theta}}^{(j-1)}\right) \left[\log p_{\bar{s}_t}(x_t|\bar{g}_t', \theta') + \log p_{\bar{s}_t}(\bar{g}_t'|\bar{\theta})\right] d\bar{g}_t' \tag{35}$$

where η_2 is a regularization parameter.

The max operation is used in the above equation in order to ensure that the posterior state probabilities are nonnegative. The update equations of the parameters for the jth iteration can be obtained by differentiating equation (35) with respect to the model parameters and setting the derivative to zero.

The update equation of mean of the speech-gain model is,

$$\bar{\mu}_{\bar{s}}^{(j)} = \frac{\sum_r \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right) \int \bar{g}_t' p_{\bar{s}_t}\left(\bar{g}_t' | x_t, \hat{\bar{\theta}}^{(j-1)}\right) d\bar{g}_t' - \bar{q}_r}{\sum_r \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right)} \quad (36)$$

The update equation of variance of the speech-gain model is,

$$\sigma_{\bar{s}}^{2(j)} = \frac{\sum_r \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right) \int \left(\bar{g}_t' - \bar{\mu}_{\bar{s}}^{(j)} - \bar{q}_r\right) p_{\bar{s}_t}\left(\bar{g}_t' | x_t, \hat{\bar{\theta}}^{(j-1)}\right) d\bar{g}_t'}{\sum_r \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right)} \quad (37)$$

The autocorrelation sequence of the speech is estimated to learn the AR coefficients $\bar{\alpha}_{\bar{s}}$ of each state. The update equation of gain bias of speech model and the update equation of autocorrelation sequence is given in the equation (38) and (39) respectively.

$$\bar{\mu}_{\bar{s}}^{(j)} = \frac{\sum_r \sum_t \frac{\max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right)}{\sigma_{\bar{s}}^{2(j)}} \int \left(\bar{g}_t' - \bar{\mu}_{\bar{s}}^{(j)}\right) p_{\bar{s}_t}\left(\bar{g}_t' | x_t, \hat{\bar{\theta}}^{(j-1)}\right) d\bar{g}_t'}{\sum_r \sum_t \frac{\max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right)}{\sigma_{\bar{s}}^{2(j)}}} \quad (38)$$

$$\bar{r}_{\bar{\alpha}_{\bar{s}}^{(j)}[i]} = \frac{\sum_r \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right) \bar{r}_{x_t}[i] \int \left(\bar{g}_t'\right)^{-1} p_{\bar{s}_t}\left(\bar{g}_t' | x_t, \hat{\bar{\theta}}^{(j-1)}\right) d\bar{g}_t'}{\sum_r \sum_t \max\left(\bar{\omega}(\bar{s}_t) - \eta_2 \bar{B}_{\bar{s}_t, x_t}, 0\right)} \quad (39)$$

where $\bar{r}_{x_t}[i]$ denotes the autocorrelation sequence of the speech observations x_t and i denotes the index of the autocorrelation sequence.

Now, Levinson-Durbin recursion algorithm is applied to obtain the AR coefficients $\bar{\alpha}_{\bar{s}}$.

The noise SARHMM can be obtained by encouraging the sparsity to transition probabilities and observation probabilities in a manner similar to that for the speech SARHMM. For the training of the noise SARHMM, independence is assumed between the noise gain and spectral shape. The training parameters are $\ddot{\theta} = \{\ddot{a}_{\ddot{s}}, \sigma_{\ddot{s}}^2, \ddot{\alpha}_{\ddot{s}}\}$, which are transition probability, the variance of noise gain model and the AR coefficients of noise gain model, respectively. The noise training data set is normalized by the long-term averaged noise gain, and then the transition probability can be optimized using the standard Baum Welch algorithm. The noise gain variance $\sigma_{\ddot{s}}^2$ can be estimated as the sample variance of the logarithm of the excitation variances after the normalization, and the estimation and update process for the noise AR coefficients $\ddot{\alpha}_{\ddot{s}}$ are similar to speech AR coefficients $\bar{\alpha}_{\bar{s}}$: first autocorrelation sequence of the noise is estimated, and then the Levinson-Durbin recursion algorithm is applied to update the noise AR coefficients $\ddot{\alpha}_{\ddot{s}}$.

4.2 Online Parameter Estimation

The time-varying parameters $\{\bar{q}_t, \ddot{u}_t\}$ as defined in (17) and (20) are to be estimated on-line using the observed noisy speech. The recursive EM algorithm is applied to perform the on-line parameter update. That is, the parameters are updated recursively for each observed noisy speech segment and the likelihood score is improved on average. The update equations for $\{\bar{q}_t, \ddot{u}_t\}$ can be given as follows:

$$\hat{\bar{q}}_t = \hat{\bar{q}}_{t-1} + \frac{1}{\Xi_t'} \sum_{s_t} \frac{\omega(s_t)}{\Omega_t \bar{\sigma}_{\bar{s}_t}^2} \left(\hat{\bar{g}}_t' - \bar{u}_{\bar{s}_t} - \hat{\bar{q}}_{t-1}\right) \quad (40)$$

$$\hat{\ddot{\mu}}_t = \hat{\ddot{\mu}}_{t-1} + \frac{1}{\Xi_t} \sum_{s_t} \frac{\omega(s_t)}{\Omega_t} \left(\hat{\ddot{g}}_t' - \hat{\ddot{\mu}}_{\bar{s}_t}\right) \quad (41)$$

where Ξ_t and Ξ_t' are two nondecreasing normalization factors that control the impact of the previous noisy segments to one new noisy segment, which is because the parameters are considered time-varying. Therefore, the normalization factors are calculated by recursive summation of the past values:

$$\Xi_t' = p_{\bar{q}} \Xi_{t-1}' + \sum_{s_t} \frac{\omega(s_t)}{\Omega_t \bar{\sigma}_{\bar{s}_t}^2} \quad (42)$$

$$\Xi_t = p_{\ddot{\mu}} \Xi_{t-1} + 1$$

where $0 \le p_{\bar{q}}$ and $p_{\ddot{\mu}} \le 1$ are forgetting factors. Thus, $p_{\bar{q}} = p_{\ddot{\mu}} = 1$ implies there is no forgetting, and the equations of $\omega(s_t)$ and Ω_t are defined later.

5. SPARSE ARHMM SPEECH ENHANCEMENT

The traditional ARHMM method estimates the speech directly. This results in inherent problem besides the ambiguity problem. Inherent problem is that the spectral fine structure of voiced speech is not obtained. Due to this problem, the noise cannot be removed between the speech harmonics. Therefore, the perceptual quality of the enhanced speech signal in voiced segments will be poor. In sparse ARHMM (SARHMM) speech enhancement, the speech is not directly estimated to solve the inherent problem. First we estimated the noise power spectrum using SARHMM approach [8]. Then the clean speech spectrum is estimated from the noisy speech using a Bayesian estimator. Finally, the enhanced speech can be obtained by applying the inverse Fourier transform to the estimated speech spectrum.

5.1 Noise Estimation

Given noisy observations y_0^t , minimizing the expected value of $\left\|\hat{f}(\omega_t) - f(\omega_t)\right\|^2$ results in the minimum mean square error (MMSE) estimator of function $f(\omega_t)$.

$$\hat{f}(\omega_t) = E\{f(\omega_t)|y_0^t\} = \int f(\omega_t) p(\omega_t|y_0^t) d\omega_t \quad (43)$$

where $\|.\|$ denotes the vector norm, $E\{.\}$ indicates the statistical expectation. $p(\omega_t|y_0^t)$ is the posterior noise pdf given the noisy observations, and can be given as follows:

$$p(\omega_t|y_0^t) = \frac{p(\omega_t, y_0^t)}{p(y_0^t)} = \frac{p\left(\omega_t, y_t|y_0^{t-1}\right)}{p(y_t|y_0^{t-1})} \quad (44)$$

The numerator of equation (44) can be expressed by speech and noise gains given a composite state s_t using the Markov assumption.

$$p\left(\omega_t, y_t|y_0^{t-1}\right) = \sum_{s_t} \gamma(s_t) p_{s_t}\left(\omega_t, y_t\right) \quad (45)$$

$$= \sum_{s_t} \gamma(s_t) \iint p_{s_t}(y_t, \bar{g}_t', \ddot{g}_t') p_{s_t}(\omega_t|y_t, \bar{g}_t', \ddot{g}_t') d\bar{g}_t' d\ddot{g}_t'$$

where $\gamma(s_t)$ denotes the probability of being in the composite state s_t given all past noisy observation up to frame $t-1$, which can be defined by

$$\gamma(s_t) = p(s_t|y_0^{t-1}) = \sum_{s_{t-1}} p(s_{t-1}|y_0^{t-1}) a_{s_{t-1}s_t} \quad (46)$$

where $p(s_{t-1}|y_0^{t-1})$ is the forward probability at frame $t-1$, which can be obtained by the forward algorithm of HMM.

Based on the equation (24), the equation (45) can be rewritten as follows:

$$p(\omega_t, y_t|y_0^{t-1}) \approx \sum_{s_t} \gamma(s_t)\, p_{s_t}(y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t')\, p_{s_t}(\omega_t|y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t') \quad (47)$$

The denominator of equation (44) can be written as follows:

$$p(y_t|y_0^{t-1}) = \int p(\omega_t, y_t|y_0^{t-1}) d\omega_t, \quad (48)$$

$$\approx \int \sum_{s_t} \gamma(s_t)\, p_{s_t}(y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t')\, p_{s_t}(\omega_t|y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t')\, d\omega_t,$$

$$= \textstyle\sum_{s_t} \gamma(s_t) p_{s_t}(y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t').$$

Substituting the equations (47) and (48) in (44),

$$p(\omega_t|y_0^t) = \frac{\sum_{s_t} \gamma(s_t)\, p_{s_t}(y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t')\, p_{s_t}(\omega_t|y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t')}{\sum_{s_t} \gamma(s_t) p_{s_t}(y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t')} \quad (49)$$

The equations of $\omega(s_t)$ and Ω_t are defined as follows:

$$\omega(s_t) = \gamma(s_t) p_{s_t}(y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t') \quad (50)$$

$$\Omega_t = p(y_t|y_0^{t-1}) = \sum_{s_t} \omega(s_t) \quad (51)$$

Now, substituting the equations (50) and (51) in (49), the posterior noise pdf $p(\omega_t|y_0^t)$ can be rewrite as follows:

$$p(\omega_t|y_0^t) = \frac{1}{\Omega_t} \sum_{s_t} \omega(s_t)\, p_{s_t}(\omega_t|y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t') \quad (52)$$

From the above equation, the equation (43) can be written as

$$\hat{f}(\omega_t) = \int f(\omega_t) p(\omega_t|y_0^t) d\omega_t \quad (53)$$

$$= \frac{1}{\Omega_t} \sum_{s_t} \omega(s_t) \int f(\omega_t) p_{s_t}(\omega_t|y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t') d\omega_t$$

$$= \frac{1}{\Omega_t} \sum_{s_t} \omega(s_t)\, E_{s_t}[f(\omega_t)|y_t, \hat{\bar{g}}_t', \hat{\ddot{g}}_t']$$

Let $W_t(k)$ denote the kth spectral magnitude of the noise ω_t. Using $W_t^2(k) = f(\omega_t)$, the power spectrum $\hat{\ddot{\lambda}}_t(k)$ of the noise ω_t can be estimated as follows:

$$\hat{\ddot{\lambda}}_t(k) = \frac{1}{\Omega_t} \sum_{s_t} \omega(s_t)\, E_{s_t}[W_t^2(k)|Y_t(k), \hat{\bar{g}}_t', \hat{\ddot{g}}_t'] \quad (54)$$

where k is the index of the frequency bins, $W_t(k)$ and $Y_t(k)$ are the kth spectral amplitude of noise ω_t and noisy speech y_t respectively.

According to the ARHMM signal model, the MMSE estimation of noise power spectrum for composite state s_t is given by

$$E_{s_t}[W_t^2(k)|Y_t(k), \hat{\bar{g}}_t', \hat{\ddot{g}}_t'] = \left[\left(1 - H_{s_t}(k)\right) Y_t(k)\right]^2 + H_{s_t}(k) \ddot{\lambda}_{\ddot{s}_t}(k) \quad (55)$$

where,

$$\bar{\lambda}_{\bar{s}_t}(k) = \frac{\hat{\bar{g}}_t}{\left|\sum_{j=0}^{\bar{p}} (j)\, \bar{\alpha}_{\bar{s}_t}(j) \exp(-2\pi jk)/k\right|^2} \quad (56)$$

$$\ddot{\lambda}_{\ddot{s}_t}(k) = \frac{\hat{\ddot{g}}_t}{\left|\sum_{j=0}^{\ddot{q}} \ddot{\alpha}_{\ddot{s}_t}(j) \exp(-2\pi jk)/k\right|^2} \quad (57)$$

$\bar{\alpha}_{\bar{s}_t}(j)$ and $\ddot{\alpha}_{\ddot{s}_t}(j)$ are the jth AR coefficients of speech and noise models, respectively; $\bar{p}$ and $\ddot{q}$ are the order of speech and noise AR coefficients, respectively.

Now, substituting equation (55) in equation (54), the final estimation of noise power spectrum $\hat{\ddot{\lambda}}_t(k)$can be obtained.

$$\hat{\ddot{\lambda}}_t(k) = \frac{1}{\Omega_t} \sum_{s_t} \omega(s_t)\, E_{s_t}[W_t^2(k)|Y_t(k), \hat{\bar{g}}_t', \hat{\ddot{g}}_t'] \quad (58)$$

$$= \frac{1}{\Omega_t} \sum_{s_t} \omega(s_t) \left\{\left[\left(1 - H_{s_t}(k)\right) Y_t(k)\right]^2 + H_{s_t}(k) \ddot{\lambda}_{\ddot{s}_t}(k)\right\}$$

5.2 Speech Estimation

In traditional ARHMM the spectral fine structure of voiced speech is not obtained. This results in the presence of clearly audible noise in the voiced segments of the estimated speech. Therefore, the perceptual quality of the speech will be reduced if the speech is directly estimated. In order to solve this problem, Forward-backward algorithm is used in which speech is indirectly estimated.

All the above mathematical analysis can be represented in terms of block diagram as shown as in the Fig 1.

The noisy speech signal is taken as the input and is divided into blocks of frames having the frame length of 32ms where the sampling frequency is 8KHz. Now each frame is multiplied by the sampling window. Here, hamming window is used with the 50% overlap in order to avoid the loss of speech information in between the frames.

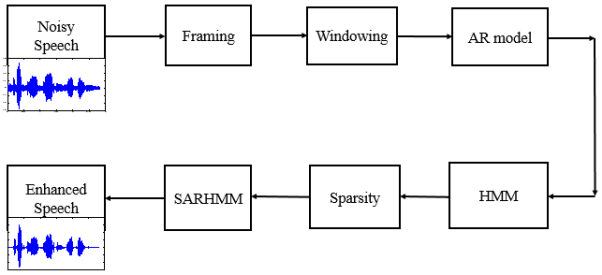


Fig 1: Block diagram of ARHMM based speech enhancement using sparsity

Now, each windowed set of speech samples is auto-correlated to give a set of (p + 1) coefficients, where p is the order of the desired LPC analysis. This is referred as AR model from which the speech parameters are obtained and then Hidden Markov model is applied to model those parameters. The above two steps constitute to form ARHMM model. Later, the sparsity is encouraged into the model by adding the regularization parameter to form SARHMM. Finally, the enhanced speech signal is obtained.

6. SIMULATION RESULTS

In this section, the simulation results for the proposed speech enhancement method is compared with the Wiener filter method. The objective results are obtained. Log-likelihood Scores for different noise signals is obtained.

The objective quality measures are given as follows:

6.1 Signal to Noise Ratio (SNR)

Signal to noise ratio is defined as follows:

$$SNR = 10\log_{10}\left(\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N}[x(n)-\hat{x}(n)]^2}\right) \quad (59)$$

where $x(n)$ is the original speech signal and $\hat{x}(n)$ is enhanced speech signal [1].

6.2 Segmental SNR ($\mathbf{SNR_{seg}}$)

Segmental Signal to Noise ratio can be evaluated either in time or frequency domain. But time domain measure is the simplest objective measure used to evaluate speech enhancement algorithms [6]. It can be defined as follows:

$$SNR_{seg} = \frac{1}{L}\sum_{l=0}^{L-1} 10.\log_{10}\left(\frac{\sum_{n=Nl}^{Nl+N-1} x^2(n)}{\sum_{n=Nl}^{Nl+N-1}[x(n)-\hat{x}(n)]^2}\right) \quad (60)$$

where $x(n)$ is the original speech signal and $\hat{x}(n)$ is the enhanced speech signal. N is the frame length, M is the number of frames.

6.3 Log-likelihood Ratio (LLR)

Log-likelihood ratio for the speech segment is based on the assumption that over the short time intervals speech can be represented by an all-pole linear predictive coding model of the form [6],

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + Gu(n) \quad (61)$$

where, a_k $(a_1, a_2, \ldots, a_p)$ are the coefficients of all-pole filter, G is the filter gain and $u(n)$ is a unit variance white noise excitation.

The LLR measure is defined as follows:

$$d_{LLR}(a_x, a_{\hat{x}}) = \log\frac{a_{\hat{x}}^T R_x a_{\hat{x}}}{a_x^T R_x a_x} \quad (62)$$

where a_x^T denotes the vector with the LPC coefficients of clean speech signal, $a_{\hat{x}}^T$ denotes the vector with the coefficients of the enhanced speech signal, R_x is the auto correlation matrix of the clean speech signal. The Log-likelihood ratio values were limited to the range of [0, 2].

6.4 Perceptual Evaluation of Speech Quality (PESQ)

PESQ measure is currently the most reliable measure for assessment of overall quality of speech processed by noise-reduction algorithms [6]. The final PESQ sore is computed as a linear combination of the average disturbance value d_{sym} and the average asymmetrical disturbance value d_{asym} [6]as follows:

$$PESQ = 4.5 - 0.1.d_{sym} - 0.0309 d_{asym} \quad (63)$$

The range of the PESQ score is -0.5 to 4.5, although for most cases the output range will be a MOS-like score, i.e., a score between 1.0 and 4.5.

6.5 Spectrograms

The time varying spectral characteristics of the speech signal can be graphically displayed through the spectrograms. The spectrogram is a two dimensional graphical pattern in which the vertical dimension corresponds to the frequency and horizontal dimension corresponds to the time. The darkness of the pattern indicates the energy present in the speech signal.

6.6 Mean Square Error (MSE)

The mean square error is defined as the mean square value of the error signal which is the difference between the signal implied and the true but unknown signal. The value of the mean square error should be minimum [7].

$$MSE = E\left\{\left(x(n)-\hat{x}(n)\right)^2\right\} \quad (64)$$

where $x(n)$ is the input signal and $\hat{x}(n)$ is the estimated signal.

6.7 Log-likelihood (LL) Score

The log-likelihood score of the estimated speech and noise models is evaluated using the true speech and noisy signals in order to evaluate the modeling accuracy [10]. The LL score of the estimated speech model for the nth block is defined as follows:

$$LL(x_n) = \log\left(\frac{1}{\Omega_n}\sum_{s}\omega_n(s) f_{\bar{s}}(x_n|\hat{\bar{g}}_n)\right) \quad (65)$$

where $\omega_n(s)$ is the state probability given the observations y_0^n and $f_{\bar{s}}(x_n|\hat{\bar{g}}_n)$ is the density function evaluated using the estimated speech gain $\hat{\bar{g}}_n$.

The signal to noise ratio (SNR) for different noise signals car, train and airport noises with input SNR of 0dB, 5dB,10dB and 15dB is obtained as shown in Fig 2.

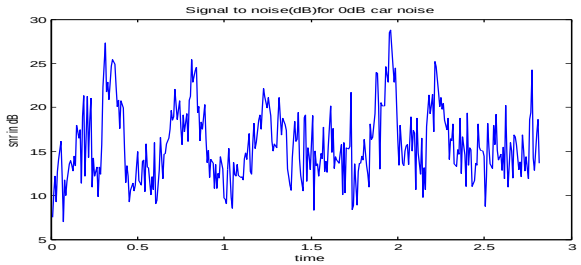


Fig 2: Signal to noise ratio for speech degraded with 0dB car noise

The Mean squared error (MSE) for different noise signals car, train and airport noises with input SNR of 0dB, 5dB,10dB and 15dB is obtained as shown in Fig 3.

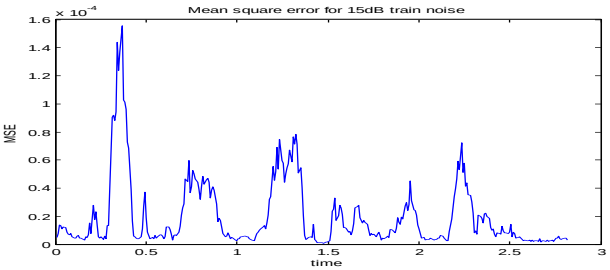


Fig 3: Mean square error for speech degraded with 15dB train noise

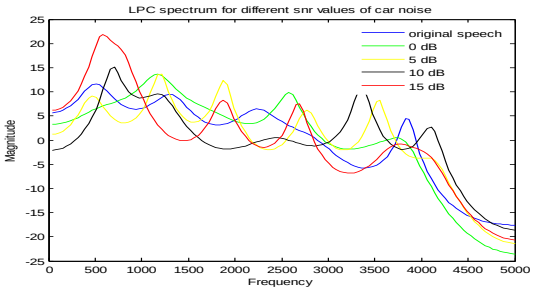


Fig 4: LPC Spectrums for different SNR values of Car noise

The Timing diagrams comparison for Wiener filter and LPC can be shown in Fig 5.

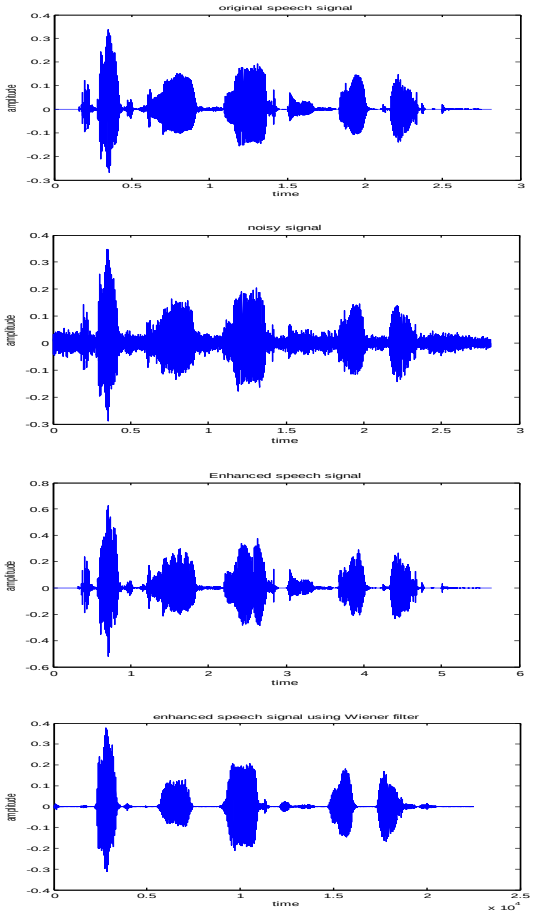


Fig 5: Timing diagrams comparison for LPC and Wiener filter.

The comparison of Wiener filter and LPC for the speech signal degraded with different noises in terms of segmental SNR can be shown in Table 1.

Table 1. Comparison of Wiener filter and LPC in terms of Segmental SNR (dB)

Noise (dB)	Wiener Filter	Linear Prediction Coding (LPC)
Airport-0	-1.508	25.420
Airport-5	-0.035	26.915
Airport-10	1.552	27.118
Airport-15	2.182	33.712
Car-0	-0.634	15.794
Car-5	0.348	22.102
Car-10	0.524	20.461
Car-15	2.755	25.630
Train-0	0.524	14.150
Train-5	0.044	16.258
Train-10	1.590	18.574
Train-15	2.202	21.703

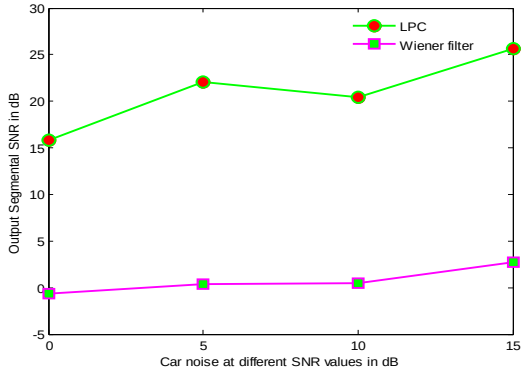
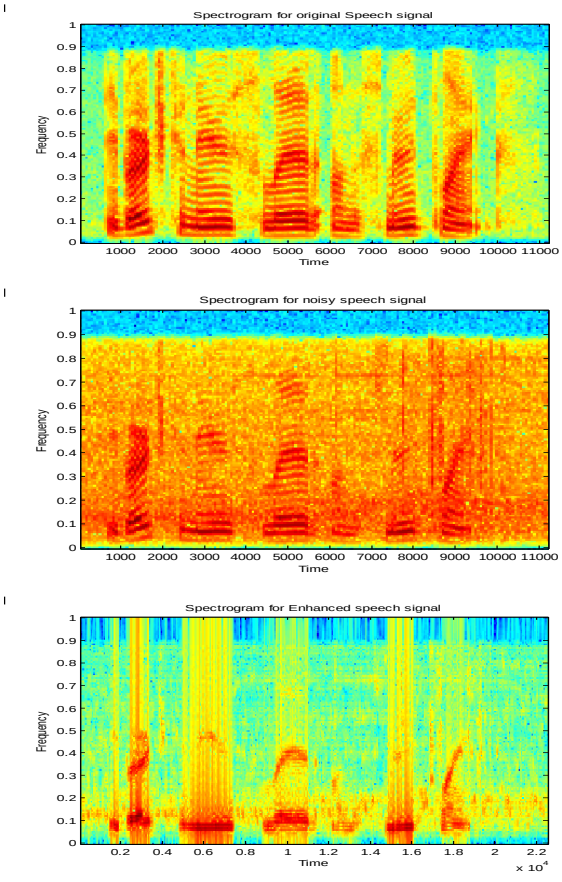


Fig 6: Segmental SNR(dB) Comparison of LPC and Wiener filter for the speech degraded with Car noise.

The Spectrograms comparison for Wiener filter and LPC can be shown in Fig 7.



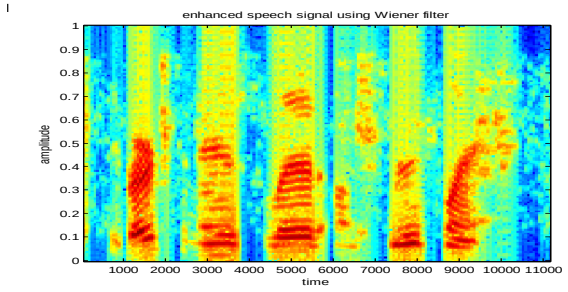


Fig 7: Spectrograms comparison for LPC and Wiener filter

The comparison of Wiener filter and LPC for the speech signal degraded with different noises in terms of Log-likelihood ratio LLR (dB) can be shown in Table 2.

The comparison of Wiener filter and LPC in terms of Perceptual evaluation of speech quality PESQ (MOS) can be shown in Table 3.

Table 2. LLR (dB) Comparison for Wiener filter and LPC

Noise (dB)	Wiener Filter	Linear Prediction Coding (LPC)
Airport-0	1.446	1.550
Airport-5	1.304	1.498
Airport-10	0.932	1.233
Airport-15	0.994	1.254
Car-0	1.504	1.668
Car-5	1.185	1.338
Car-10	1.343	1.550
Car-15	1.018	1.293
Train-0	1.343	1.421
Train-5	1.578	1.638
Train-10	1.317	1.467
Train-15	1.347	1.428

Table 3. PESQ (MOS) Comparison for Wiener filter and LPC

Noise (dB)	Wiener Filter	Linear Prediction Coding (LPC)
Airport-0	1.472	1.238
Airport-5	1.492	2.039
Airport-10	2.025	1.112
Airport-15	2.249	1.029
Car-0	1.165	1.980
Car-5	1.694	1.644
Car-10	1.921	1.377
Car-15	2.265	1.106
Train-0	1.921	1.263
Train-5	1.680	1.254
Train-10	2.008	1.103
Train-15	2.004	1.056

Log-likelihood scores for the speech signal degraded with the car noise and airport noise with input SNR values of 0dB, 5dB, 10dB and 15dB for Gaussian mixtures 2, 4, 8 and 16 with states 2 and state 3 can be shown in Table 4 and Table 5 respectively.

Log-likelihood scores for the speech signal degraded with the car and airport noises with input SNR values of 0dB, 5dB, 10dB and 15dB for Gaussian mixtures 2, 4, 8 and 16 with states 2 and state 3 can be shown in Fig 8, Fig 9 and Fig 10, Fig 11 respectively.

Table 4. Log-likelihood Scores for state 2 and state 3 for the speech degraded with Car noise

Noise (dB)	Gaussian mixture	State 2	State 3
Car-0	2	2047.8	2080.8
	4	2118.0	2155.3
	8	2140.8	2164.8
	16	2230.9	2254.9
Car-5	2	1887.7	1992.5
	4	1958.4	2026.8
	8	2070.8	2126.5
	16	2167.6	2187.2
Car-10	2	1614.7	1585.2
	4	1669.7	1797.0
	8	1965.3	1884.4
	16	1982.8	2006.4
Car-15	2	1388.8	1506.2
	4	1562.0	1571.8
	8	1681.2	1690.2
	16	1761.7	1723.9

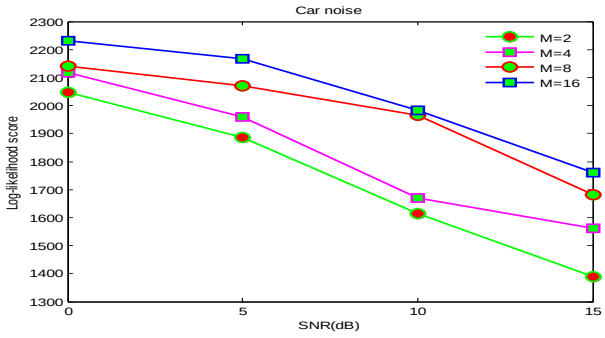


Fig 8: Log-likelihood scores for speech degraded with Car noise for State 2.

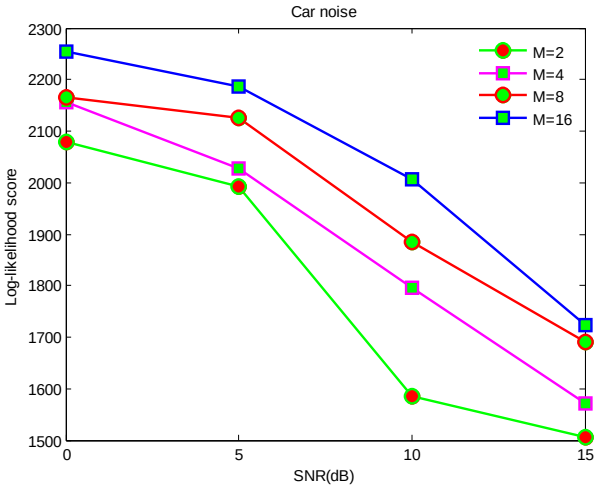


Fig 9: Log-likelihood scores for speech degraded with Car noise for State 3.

Table 5. Log-likelihood Scores for state 2 and state 3 for the speech degraded with Airport noise.

Noise (dB)	Gaussian mixture	State 2	State 3
Airport-0	2	1881.5	1958.1
	4	1894.6	2039.0
	8	2063.0	2109.7
	16	2117.9	2153.7
Airport-5	2	1583.0	1560.5
	4	1617.9	1705.3
	8	1764.2	1720.2
	16	1902.9	1918.6
Airport-10	2	1451.8	1497.1
	4	1602.4	1582.1
	8	1790.5	1714.2
	16	1920.6	1908.2
Airport-15	2	1239.8	1252.2
	4	1351.5	1256.0
	8	1569.5	1642.4
	16	1591.3	1653.6

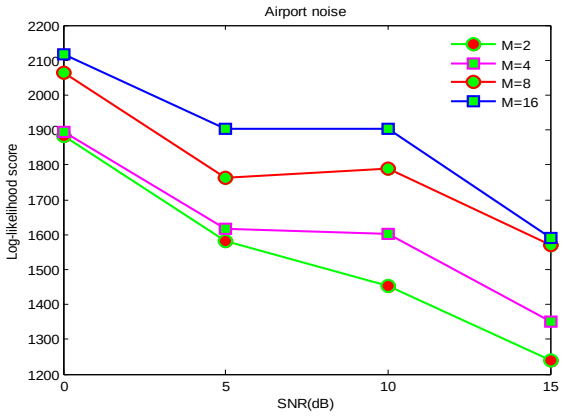


Fig 10: Log-likelihood scores for speech degraded with Airport noise for State 2.

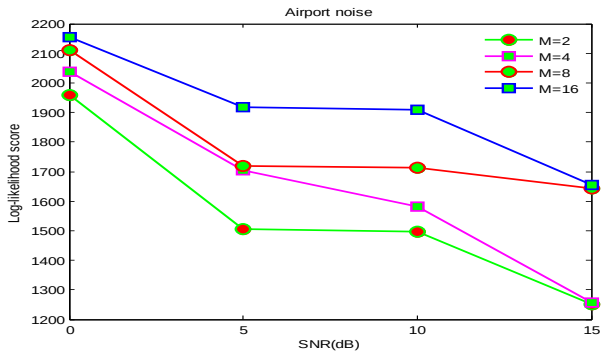


Fig 11: Log-likelihood scores for speech degraded with Airport noise for State 3.

It is observed that the log-likelihood score is decreased as the SNR value of the noise increased.

7. CONCLUSION

This work has presented a speech enhancement method in which first AR model is applied for the noisy speech signal to find the speech parameters and then Hidden Markov model is applied to model those parameters to form ARHMM model. Later, the sparsity is encouraged into the model by adding the regularization parameter to form SARHMM to overcome the ambiguity and inherent problems in ARHMM. The objective results for the proposed method and Wiener filter are compared. Speech quality in non-stationary noise conditions is observed through listening. The average log-likelihood score is obtained for different noises and observed that the performance is improved compared to the reference methods. The present work can be used for different signals and can be used in real time applications like hearing aids.

8. REFERENCES

[1] Lawrence R. Rabiner and Ronald W. Schafer, Digital Processing of Speech Signals. *Prentice-Hall, Inc.*, Englewood Cliffs, New Jersey 07632.

[2] Lawrence R. Rabiner and Ronald W. Schafer, Introduction to Digital Speech Processing.

[3] Thomas F. Quatieri, Discrete-Time Speech Processing, Principles and Practice.

[4] Lawrence R. Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall, Signal Processing Series.

[5] L. Rabiner, "A tutorial on Hidden Markov models and Selected Applications in Speech Recognition," *proc. IEEE*, vol. 77, no. 2, Feb. 1989.

[6] Philipos C. Loizou, Speech Enhancement: Theory and practice, second edition, CRC press.

[7] Simon Haykin, Adaptive Filter Theory, third edition, Prentice-Hall, Information and System Sciences Series.

[8] Feng Deng, Changchun Bao, and W. Bastiaan Kleijn, Sparse Hidden Markov Models for speech Enhancement in Non-Stationary Noise Environments, *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 23, no. 11, Nov. 2015.

[9] Feng Deng, Changchun Bao, and W. Bastiaan Kleijn, "Sparse HMM-based Speech Enhancement method for Stationary and Non-Stationary Noise Environments," in *proc. IEEE International conf. on Acoustics, Speech and signal Processing (ICASSP)*, 2015.

[10] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 15, no. 3, Mar. 2007.