A Modified Oesophageal Speech Enhancement Using Ephraim-Malah Filter For Robust Speech Recognition

C.GANESH BABU
Department of ECE
Bannari Amman Institute of Technology,
Sathyamangalam
INDIA
bits_babu@yahoo.co.in

P. T. VANATHI
Department of ECE
PSGCT,
Peelamedu
INDIA
ptvani@yahoo.com

JIBBY PETER DCRUZ
Department of ECE
Bannari Amman Institute of Technology,
Sathyamangalam
INDIA
dr_dcruz@yahoo.co.in

Abstract: - This paper presents a modified Oesophageal Single Channel Speech Enhancement using Ephraim-Malah Filter for Robust Speech Recognition. An Oesophageal voice is due to the laryngectomy undergone by those persons with larynx cancer and it has extremely low intelligibility. This work was already proposed with a method of Kalman Filtering technique to improve the Speech Quality. A Novel Approach to Enhance the Speech Quality with Ephraim-Malah filter for Robust Speech Recognition is presented in this paper where we present a speech-to-text system using isolated word recognition with voice samples in English (for the words Eight and Nine) and statistical modeling (Hidden Markov Model - HMM) for machine Speech Recognition.

Key-Words: - Oesophageal Voice, Pulse Code Modulation, Robust Speech Recognition, Speech Enhancement, Ephraim-Malah Filter, Hidden Markov Model.

1 Introduction

Patients who have undergone a laryngectomy as a result of larynx cancer have exceptionally low intelligibility [1]. This is due to the removal of their vocal folds, which forces them to use the air flowing through the oesophagus, known as oesophageal speech. Low intelligibility is the main problem in both oral and telephone communications with other people [2]. In addition, the noise of this kind of speech signal is especially high.

Single channel Speech Enhancement Algorithms process a noisy, monaural speech signal and estimate of what the signal would have been in a less noisy environment. Because signal processing cannot create information, the output signal cannot contain more information about what was said than existed in the noisy input. It can only have less noise. For human perception, the goals are to make the speech more intelligible and to improve the perceived quality of the speech. These two goals are often conflicting, because as more noise is removed more speech is often removed as well. For Automatic Speech Recognition, the main goal is increased recognition accuracy.

In the training phase, the uttered digits are recorded using 8-bit Pulse Code Modulation (PCM) with a sampling rate of 8 KHz and saved as a wave file using sound recorder software. One hundred different voice samples are considered. The system performs speech analysis using the Linear Predictive Coding (LPC) method of degree. From the LPC coefficients, the weighted cepstral coefficients and cepstral time derivatives are derived. From these variables the feature vector for a frame is arrived. Then, the system performs Vector Quantization (VQ) utilizing a vector codebook which result vectors form of the observation sequence. For the given word, the system builds an HMM model and trains the model during the training phase. The proposed Robust Speech Recognition System is shown in Figure 1.

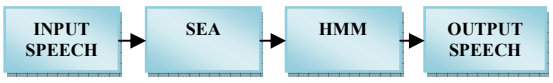

Fig.1 Proposed Robust Speech Recognition System

2 Overview of Speech Enhancement

Speech Enhancement in the past decades has focused on the suppression of additive background noise. From a signal processing point of view additive noise

is easier to deal with than convolutive noise or nonlinear disturbances. Moreover, due to the bursty nature of speech, it is possible to observe the noise by itself during speech pauses, which can be of great value.

$$y(t) = s(t) + n(t) \quad (1)$$

Speech Enhancement aims to improve speech quality by using various algorithms. By the word quality, it can be at least

- clarity and intelligibility
- pleasantness
- compatibility with some other method in speech processing and The goal of speech enhancement is to find an optimal estimate (i.e., preferred by a human listener)ŝ(t), given a noisy measurement

3 Methodology

The general idea of the algorithm presented in this work is to filter the noisy speech signal to obtain a less noisily corrupted one.

3.1 Kalman Filter

The Kalman Filter (KF) is an efficient recursive filter that estimates the state of a linear dynamic system from a series of noisy measurements. The Kalman Filter is a recursive estimator. This means that only the estimated state from the previous time step and the current measurement are needed to compute the estimate for the current state. In contrast to batch estimation techniques, no history of observations and/or estimates is required. The state of the filter is represented by two variables:

- The a posteriori state estimate at time k given observations up to and including at time k
- The a posteriori error covariance matrix (a measure of the estimated accuracy of the state estimate).

The Kalman Filter has two distinct phases:

3.1.1 Predict

The predict phase uses the state estimate from the previous time step to produce an estimate of the state at the current time step. This predicted state estimate is also known as the a priori state estimate because, although it is an estimate of the state at the current time step, it does not include observation information from the current time step.

3.1.2 Update

In the update phase, the current a priori prediction is combined with current observation information to refine the state estimate. This improved estimate is termed the a posteriori state estimate.

Given the past and present observations, Kalman Filtering is able to obtain the optimum estimate of the state, due to its recursive method. When using the KF, speech and noise are usually modeled as an Autoregressive (AR) approach.

3.2 Autoregressive Model

Voice must be characterized in order to obtain the speech parameters. This fundamental task is performed with the help of the autoregressive model approach. This model is used to obtain the parameters of both signals, that is, the speech and additional noise. The model equation is given below where v(t) is a unit-variance zero-mean colored noise, the system poles a_i and the zero b:

$$y(t) = \sum_{i=1}^{n} ai \,.\, y(t-i) \; + \; b \,.\, v(t), \; b \geq 0 \quad (2)$$

3.3 Voice Activity Detection

An improved Voice Activity Detection (VAD) algorithm employing long-term signal processing and maximum spectral component tracking. It improves the speech/non-speech discriminability and speech recognition performance in noisy environments. Two problems are solved using VAD. The first one is performance of VAD in low noise condition and the second is with noisy environment.

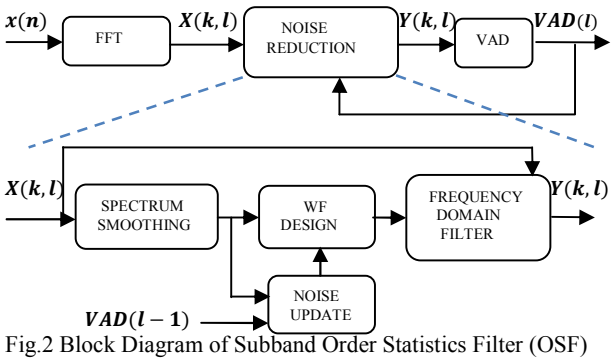

Fig.2 Block Diagram of Subband Order Statistics Filter (OSF) based VAD

The subband based VAD uses two order statistics filters for the multi-band quantile (MBQ) SNR estimation [3]. The implementation of both OSF is based on a sequence of 2N+1 log-energy values {E(m − N,k), . . . , E(m,k), . . . , E(m + N,k)} around the frame to be analyzed. The block diagram of the subband based VAD is shown in the Figure 2. This algorithm operates on the subband log-energies. Noise reduction is performed first and the VAD decision is formulated on the de-noised signal. The noisy speech signal is decomposed into 25-ms frames with a 10-ms window shift. Let X(m,l) be the spectrum magnitude for the mth band at frame l .The design of the noise reduction block is based on Wiener Filter (WF) theory whereby the attenuation is a function of the signal-to-noise ratio (SNR) of the input signal. The VAD decision is formulated in terms of the de-noised signal, being the subband log-energies processed by means of order statistics filters.

The noise reduction block consists of four stages.
i) Spectrum smoothing: The power spectrum is averaged over two consecutive frames and two adjacent spectral bands.
ii) Noise estimation: The noise spectrum Ne(m,l) is updated by means of a 1st order IIR filter on the smoothed spectrum Xs(m,l),

$$Ne(m,l)=\lambda\, Ne(m,l-1)+(1-\lambda)\, Xs(m,l) \quad (3)$$

where $\lambda=0.99$ and $m=0,1,\ldots,NFFT/2$
iii) Wiener Filter design: First, the clean signal S(m,l) is estimated by combining smoothing and spectral subtraction

$$S(m,l)=\gamma S'(m,l-1)+(1-\gamma)\max(Xs(m,l-Ne(m,l),0) \quad (4)$$

where $\gamma=0.98$

Then, the Wiener Filter H(m,l) is designed as

$$H(m,l)=\frac{\eta(m,l)}{1+\eta(m,l)} \quad (5)$$

$$\text{Where } \eta(m,l)=\max\left[\frac{S(m,l)}{Ne(m,l)},\eta\min\right] \quad (6)$$

ηmin is selected so that the filter yields a 20 dB maximum attenuation. S'(m,l), the spectrum of the cleaned speech signal, is assumed to be zero at the beginning of the process and is used for designing the Wiener Filter through Equation 3 to Equation 5. It is given by

$$S'(m,l)=H(m,l)X(m,l) \quad (7)$$

The filter H(m,l) is smoothed in order to eliminate rapid changes between neighbor frequencies that may often cause musical noise. Thus, the variance of the residual noise is reduced and consequently, the robustness when detecting non-speech is enhanced. The smoothing is performed by truncating the impulse response of the corresponding causal FIR filter to 17 taps using a Hanning window. With this operation performed in the time domain, the frequency response of the Wiener filter is smoothed and the performance of the VAD is improved.
iv) Frequency domain filtering: The smoothed filter is applied in the frequency domain to obtain the denoised spectrum

$$Y(m,l)=Hs(m,l)X(m,l) \quad (8)$$

3.4 Minimum Mean Square Error Approach To Speech Enhancement

In these systems the Short Time Spectral Amplitude (STSA) of the speech signal is estimated, and combined with the short-time phase of the degraded speech, for constructing the enhanced signal [4]. To derive the MMSE STSA estimator, a priori information of the speech and noise spectrum is needed. Since in practice they are unknown, one can think of measuring each probability distribution or, alternatively, assume a reasonable statistical model. The MMSE STSA estimator depends on the parameters of the statistical model it is based on and consists of two parts namely, the Decision-Directed method estimating the a priori speech spectrum, and the MMSE Short-Time Spectral Amplitude (STSA) estimator.

4 Hidden Markov Model Approach

As mentioned above the technique used to implement speech recognition is Hidden Markov Model (HMM). The HMM [5] is used to represent the utterance of the word and to calculate the probability of that the model which created the sequence of vectors. There are some fundamental problems in designing of HMM for the analysis of speech signal. The present hidden Markov Model is represented by

$$\lambda=(\pi, A, B) \quad (9)$$

π = initial state distribution vector.
A= State transition probability matrix.

B=continuous observation probability density function matrix.

Given appropriate values of A, B and π, the HMM can be used as a generator to give an observation sequence
$O=O_1 O_2 \ldots\ldots O_T$ (10)

Where each observation O_t is one of the symbols from the observation symbol V and T is the number of observation in the sequence as follows:
(i)Choose an initial state $q_1=S_i$ according to the initial state distributionπ.
(ii) Set t=1
(iii)Choose $O_t=v_k$ according to the symbol probability distribution in state S_i.
(iv)Transit to a new state $q_{t+1}=S_j$ according to the state transition probability distribution for state S_i.
(v)Set t=t+1(return to step3) if t<T; otherwise terminate the procedure.
The above procedure can be used as both a generator of observations, and as a model for how a given observation sequence was generated by an appropriate HMM.
After re-estimating the parameters, the model is represented with the following denotation

$\lambda = (A, \mu, \Sigma)$ (11)

The model is saved to represent that specific observation sequences, i.e. an isolated word. The basic theoretical strength of the HMM is that it combines modeling of stationary stochastic processes and the temporal relationship among the processes together in a well-defined probability space. This allows us to study these two separate aspects of modeling a dynamic process using one consistent framework. Also, HMM is relatively easy and straightforward to train a model from a given set of labeled training data.

4.1 Linear Predictive Coding Analysis

One way to obtain observation vectors O from speech samples is to perform a front end spectral analysis. The type of spectral analysis that is often used is linear predictive coding [6].The steps in the processing as shown in Figure 3 are as follows:
(i) Pre-emphasis: The digitized speech signal is processed by a first-order digital network in order to spectrally flatten the signal.
(ii) Block into Frames: Sections of N_A consecutive speech samples are used as a single frame. Consecutive frames are spaced M_A samples apart.
(iii) Frame Windowing: Each frame multiplied by an N_A sample window(Hamming Window) w(n) so as to minimize the adverse effects of chopping an N_A samples section out of the running speech signal.
(iv) Auto Correlation Analysis: Each windowed set of speech sample is auto-correlated to give a set of (p+1) coefficient, where p is order of the desired LPC analysis.
(v) LPC / Cepstral Analysis: A Vector of LPC coefficients is computed from the autocorrelation vector using a Levinson or a Durbin recursion method. An LPC derived cepstral vector is then computed up to the Qth component.
(vi) Cepstral Weighting: The Q-coefficient cepstral vector $c_l(m)$ at time frame l is weighted by a window $W_c(m)$[6,7].

$W_c(m)=1+[(Q/2)(\sin(\pi m/Q))],n1\leq m\leq Q$ (12)

To give $\hat{c}_l(m)=c_l(m).W_c(m)$ (13)
(vii) Delta Cepstrum: The time derivative of the sequence of weighted cepstral vectors is approximated by a first-order orthogonal polynomial over a finite length window of frames centered around the current vector [8].

$\Delta\hat{c}_l(m)=[\sum_{k=-K}^{K} k\ \hat{c}_{l-k}(m)].G$ (14)

Where G is the gain term to make the variance of $\hat{c}_l(m)$ and $\Delta\hat{c}_l(m)$ equal.

$Q_l(m)= \{\hat{c}_l(m), \Delta\hat{c}_l(m)\}$ (15)

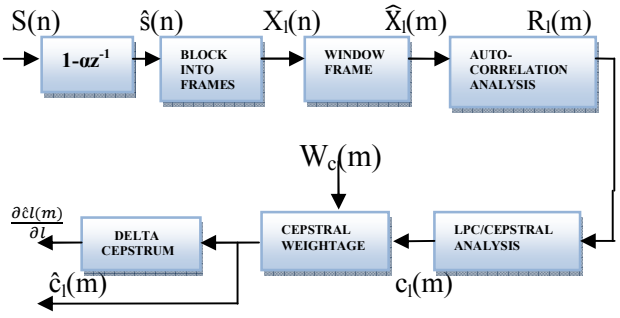

Fig.3 Linear Predictive Coding

Analysis:

$\hat{s}(n)=s(n)-\alpha s(n-1)$ (16)
$X_l(n)= \hat{s}(ml+n),\ o\leq n\leq N-1\ ;\ \ 0\leq l\leq L-1$ (17)

$\tilde{X}_l(n)=x_l(n).w(n),\ 0\le n\le N-1$ (18)

$R_l(m)=\sum_{n=0}^{N-m}\tilde{X}_l(n)\tilde{X}_l(n+m),\ 0\le m\le p$ (19)

$a_l(m)$=Lpc Coefficients, $0\le m\le p$ (20)

$c_l(m)$=Cepstral Coefficient, $1\le m\le Q$ (21)

$\hat{c}_l(m)=c_l(m).w_c(m),\ 1\le m\le Q$ (22)

$\Delta\hat{c}_l(m)=\partial\hat{c}_l(m)/\partial l,\ 1\le m\le Q$ (23)

4.2 Vector Quantization

To use HMM with discrete observation symbol density, a Vector Quantizer (VQ) is required to map each continuous observation vector in to a discrete code book index [9]. The procedure basically partitions the training vector in to M disjoin sets. The distortion steadily decreases as M increases. Hence HMM with codebook size of from M=32 to 256 vectors has been used in speech recognition experiments using HMMs.

During the training phase the system trains the HMM for each digit in the vocabulary. The same weighted cepstrum matrices for various samples and digits are compared with the code book and their corresponding nearest codebook vector indices is sent to the Baum-Welch algorithm to train a model for the input index sequence. After training we have three models for each digit that corresponds to the three samples in our vocabulary set. Then we find the average of A, B and π matrices over the samples to generalize the models.

The input speech sample is preprocessed to extract the feature vector. Then, the nearest codebook vector index for each frame is sent to the digit models. The system chooses the model that has the maximum probability of a match.

5 Results and Discussion

Table 1 Performance of KF and EM Filter for digit '8' for various 0DB noise sources

Noise	KF	EM	%Improvement
Airport	3	12	75
Babble	1	4	75
Exhibition	1	5	80
Street	1	10	90
Restaurant	1	2	50
Station	0	0	0
Car	1	1	0

Table 2 Performance of KF and EM Filter for digit '9' for various 0DB noise sources

Noise	KF	EM	%Improvement
Airport	5	19	73.68
Babble	3	3	0
Exhibition	0	0	0
Street	3	33	90.9
Restaurant	1	2	50
Station	3	3	0
Car	3	7	57.14

Table 3 Performance of KF and EM Filter for digit '8' for various 5DB noise sources

Noise	KF	EM	%Improvement
Airport	1	21	95.23
Babble	13	14	7.14
Exhibition	1	4	75
Street	1	1	0
Restaurant	20	26	23.07
Station	1	9	88.8
Car	1	1	0

Table 4 Performance of KF and EM Filter for digit '9' for various 5DB noise sources

Noise	KF	EM	%Improvement
Airport	4	23	82.61
Babble	1	22	95.45
Exhibition	1	14	92.86
Street	1	18	94.44
Restaurant	12	24	50.00
Station	7	24	70.83
Car	15	20	25.00

Table 5 Performance of KF and EM Filter for digit '8' for various 10DB noise sources

Noise	KF	EM	%Improvement
Airport	1	21	95.23
Babble	28	37	24.3
Exhibition	6	17	64.7
Street	9	23	60.86
Restaurant	23	42	95.23
Station	5	9	44.4
Car	3	24	87.5

Table 6 Performance of KF and EM Filter for digit '9' for various 10DB noise sources

Noise	KF	EM	%Improvement
Airport	9	18	50
Babble	11	27	59.25
Exhibition	4	14	71.42
Street	11	21	45.83
Restaurant	13	24	45.83
Station	12	16	25
Car	13	23	43.47

Table 7 Performance of KF and EM Filter for digit '8' for various 15DB noise sources

Noise	KF	EM	%Improvement
Airport	32	46	30.43
Babble	24	38	36.84
Exhibition	10	33	69.69
Street	1	10	90
Restaurant	24	35	31
Station	23	39	41
Car	2	38	94.7

Table 8 Performance of KF and EM Filter for digit '9' for various 15DB noise sources

Noise	KF	EM	%Improvement
Airport	19	36	47.22
Babble	19	26	26.92
Exhibition	17	33	48.48
Street	2	24	91.66
Restaurant	11	22	50
Station	12	28	57.14
Car	18	24	25

6 Conclusion

From the tabulated results shown in Table [1-8], we can easily identify that speech samples of eight and nine by Ephraim-Malah filtering yielded comparatively better Speech Recognition Accuracy in the presence of the noises considered. For example, Nine has an accuracy improvement of 57.14% for 0dB in the presence of car noise. The highest speech recognition accuracy improvement of 95.45% is seen for the digit Nine in the presence of 5dB Babble noise. No improvement was seen for the digits Eight and Nine in the presence of 0dB Station noise. Also, it was found that the Ephraim-Malah filtering not only results in better noise reduction but also increased the signal strength of the speech samples compared to Kalman Filtering.

References:

[1]B. Garcia, I. Ruiz," *Oesophageal Speech Enhancement Using Poles Stabilization and Kalman Filtering*", IEEE 2008, pg 93-97.

[2]Ronan Flynn and Edward Jones," *Robust Distributed Speech Recognition Using Speech Enhancement*" Members IEEE, May 19, 2008, pg 1267-1273.

[3] Javier Ramírez, José C. Segura*, Senior Member, IEEE*, Carmen Benítez, Ángel de la Torre and Antonio Rubio," An Effective Subband OSF-Based VAD With Noise Reduction for Robust Speech Recognition".

[4]Yariv Ephraim and David Malah, "*Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator",* IEEE transactions on acoustics, speech, and signal processing, vol. assp-32, no. 6**,** December 1984.

[5Nick Bardici,Björn Skarin:"*Speech Recognition Using HMM*".

[6]J. Makhoul,"*Linear Prediction a Tutorial View*,"

[7]J.D.Markel and A.H.Gray Jr., "*Linear Prediction of Speech"*, New York, NY:springer-Verilag,1976.

[8]Y.Tokhura,"*A Weighted Cepstral Distance Measure for Speech Recognition*," IEEE Trans.Acoust.speech signal processing, vol.ASSP-35,no.10.pp.1414-1422 ,oct.1987.

[9]J. Makhoul, S.Roucos and H.Gish, "*Vector Quantization In Speech Coding*", Proc. IEEE, vol.73, no.11, pp. 1551-1558, Nov 1985.