Speech Enhancement using Hidden Markov Models and Mel-Frequency

Priyanka T B[1], Sindhu[2], Uma D[3], Varshitha S M[4], Manjula G[5]
B.E Students[1234] (ECE) at GSSS Institute of Engineering & Technology for women, Mysore, Karnataka, India.
Asst.prof.[5]Dept of ECE, GSSSIETW, Mysore, Karnataka, India.
Visvesvaraya Technology University, Belagavi, Karnataka, India

***Abstract* - Speech enhancement using Hidden Markov model (HMM) - based minimum mean square error in Mel-frequency domain is mainly focused and to estimate clean speech waveform from a noisy signal, an inversion from the Mel- frequency domain to the spectral domain is required which introduces distortion artifacts in the spectrum estimation and filtering. To obtain a more accurate hidden Markov model (HMM) of noisy speech using the Vector Taylor Series (VTS) which is used to estimate the mean vectors and covariance matrices of HMM for noisy speech. To reduce the distortion derived from the inversion operation, a parallel Mel-frequency and log magnitude (PMLM) modeling is used. Experimental results show that, in comparison with the reference methods, the proposed method can get better performance for different noise environments and input SNRs.**

Keywords - HMM-based speech enhancement, vector Taylor series, Parallel Mel-frequency and log magnitude modeling.

INTRODUCTION

Speech signals need to be enhanced in many applications for various purposes such as boosting overall speech quality, increasing intelligibility or improving the performance of speech coding and speech recognition systems. Although speech enhancement has been studied in various aspects, single microphone speech enhancement is of wide interest due to the extent and variety of its applications. Furthermore, it has remained as a challenging topic over the years and numerous researchers have tried to suggest solutions for this problem.

Speech enhancement using HMM is originally proposed by Ephraim [1], then the speech enhancement based on minimum mean square error (MMSE) criterion using HMM is proposed [2]. Later, some scholars made many improvements based on HMM [3-5], including the changes of the training features. Zhen-Zhen gao [6] discuss the HMM-based speech enhancement using VTS and PMLM modeling in Me frequency domain. The VTS is used to estimate the mean vectors and covariance matrices of noisy speech HMM more accurately, and the PMLM modeling is used to reduce the spectral distortion resulted by the inversion from Mel frequency domain to spectral domain.

Speech enhancement aims to improve speech quality by using various algorithms. The objective of enhancement is improvement in intelligibility and/or overall perceptual quality of degraded speech signal using audio signal processing techniques. The algorithms of speech enhancement for noise reduction can be categorized into three fundamental classes: filtering techniques, spectral restoration, and model-based methods. This project uses filtering method.

Speech enhancement is required in many applications such cellular phones, hands-free communication, teleconferencing, hearing aids, speech recognition and audio fingerprinting among others. If the speech signal is corrupted by additive noise the goal of speech enhancement is to find an optimal estimate ~s(n) of the clean speech signal s(n) , given a noisy observation x(n):

x(n) = s(n) + v(n) where v(n) is the additive noise

Conventional speech enhancement methods, such as spectral subtraction, wiener filter, short-time spectral amplitude estimator rely on a separate noise estimation algorithm, while the noise estimation is not accurate under non-stationary conditions. However, the speech enhancement based on hidden Markov model (HMM) can overcome the deficiencies. Hidden Markov Model (HMM) of noisy speech and each individual filter is estimated using the HMM of clean speech and noise. So, the more accurate the HMM of noisy speech and these filters are estimated, the better performance will be obtained.

METHODOLOGY

From Fig. 1, we can see that,

The proposed speech enhancement method contains two processes: training process and enhancement process. In the training process, the HMMs of clean speech and noise are trained using PMLM modeling method, and the HMM of noisy speech is constructed by VTS. In the enhancement process, the wiener filter is estimated using the HMMs and noisy speech. estimate clean speech waveform from a noisy signal, an inversion from the Mel-frequency domain to the spectral domain is required which introduces distortion artifacts in the spectrum estimation and the filtering. To reduce the corrupting effects of the inversion, the PMLM modeling is proposed. This method performs concurrent modeling in both cepstral and magnitude spectral domains.

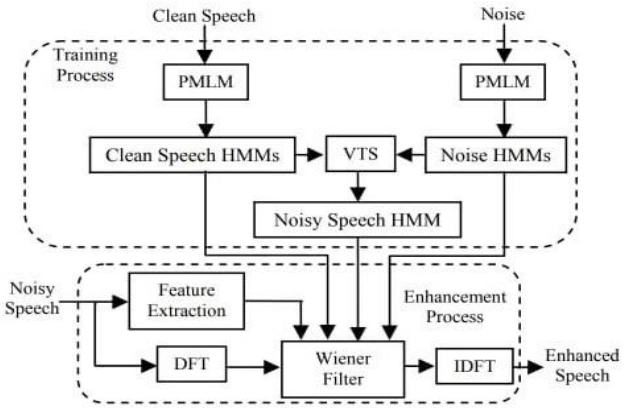

Fig1:The block diagram of the proposed speech enhancement method

HMM-based speech enhancement in mfs domain

The noisy speech vector y_n in the nth frame can be expressed as follows:

$$y_n=s_n+d_n \qquad (1)$$

where s_n and d_n denote the clean speech and additive noise vectors in time domain, respectively. s_n, d_n and y_n are defined as the MFS feature vectors of clean speech, noise and noisy speech in the nth frame, respectively. A general MFS feature vector x_n is defined as:

$$\log=(\mathrm{MEL}(|\mathrm{DFT}(\mathrm{win}(x_n))|)) \qquad (2)$$

where the symbol, win, DFT,|.|, MEL and log denote the operations of windowing, discrete Fourier transform, magnitude calculation, Mel-filtering, logarithmic, respectively. Similar to x_n which denotes the MFS domain feature vector for x_n, the superscript notations spc, mag, log-mag, and Mel indicate the DFT, magnitude of DFT, log-magnitude spectra and MEL coefficients of speech signal in the following parts, respectively. The HMMs of clean speech, noise and noisy speech in MFS domain are denoted as λs, λd and λy, respectively. In this paper, it is assumed that the feature vectors of clean speech, noise and noisy speech in MFS domain have a Gaussian distribution, and each HMM is defined as :

$$\lambda= \{ N\ M\ \pi\ a\ c\ \mu\ \Sigma \} \qquad (3)$$

where N and M are the number of states and mixtures, respectively, π={πβ} is the set of initial state probabilities, a={aαβ} is the set of state transition probabilities, c={cγ|β } is the set of mixture weights, μ={μ(γ|β)} and Σ={Σ(γ|β)} are the sets of Gaussian density parameters for mean vectors and covariance matrices, respectively, and the ranges for α, β and γ are defined as 1≤α, β≤N and 1≤γ≤M, respectively. To estimate the parameters of λs and λd, the standard maximum likelihood hidden Markov model with Baum re-estimation is used.

The type of the HMM in this paper is ergodic and the transition from one state to another is assumed to have first order Markovian property. Given the noisy speech $y_{0:n}=[y_0, y_1,\ldots,y_n]$, the MMSE criterion is used to estimate clean speech $s\wedge_n^{spc}$.

$$\hat{\mathbf{s}}_n^{spc} = \underset{\hat{\mathbf{s}}_n^{spc}}{\arg\min}\left[\mathbf{E}\left\|\hat{\mathbf{s}}_n^{spc}-\mathbf{s}_n^{spc}\right\|^2\right]=\mathbf{E}\left(\mathbf{s}_n^{spc} \mid \mathbf{y}_{0:n}^{spc}\right)$$

$$=\sum_{\beta_s=1}^{N_s}\sum_{\gamma_s=1}^{M_s}\sum_{\beta_d=1}^{N_d}\sum_{\gamma_d=1}^{M_d} P_{\lambda_y}\left(\beta_s,\gamma_s,\beta_d,\gamma_d \mid \mathbf{y}_{0:n}^{mfs}\right)$$

$$\cdot\mathbf{E}\left(\mathbf{s}_n^{spc} \mid \mathbf{y}_n^{spc},\beta_s,\gamma_s,\beta_d,\gamma_d\right) \qquad (4)$$

where Ns and Ms are the number of states and mixtures of clean speech HMM, respectively. Nd and Md are the number of states and mixtures of noise HMM, respectively. y Pλ (ßs, γs, ßd, γd|y0:n mfs) is the conditional probability of clean speech state βs and mixture γs, and noise state βd and mixture γd given the noisy speech y0:n mfs. E($s_n$spc|y_n spc, ßs, γs, ßd, γd) is the conditional expectation of clean speech.

Weiner Filtering

For discrete-time signals, the Wiener filter is a linear shift-invariant (LSI) filter whose output is an estimate of the desired signal when its input is the noisy signal x(n). It minimizes the mean squared error (MSE) between the estimated signal and the desired signal.

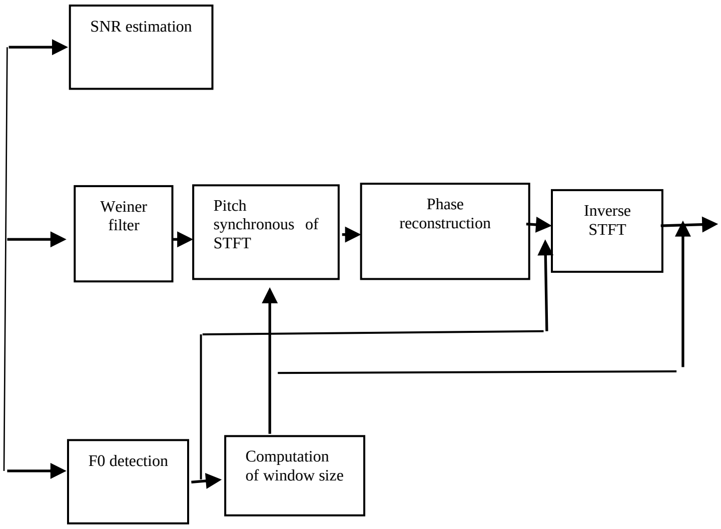

Fig 2: Block diagram of Weiner filtering

In Eq. 4, the conditional probability is calculated by the ‘forward-backward’ algorithm. Due to the Gaussian mixtures, the conditional expectation can be obtained by the following Wiener filtering

$$E(\ s_n^{spc}\ \big|\ y_n^{spc}, \beta_s, \gamma_s, \beta_d, \gamma_d)$$

$$= H(\ \gamma_s\ \gamma_d\ \big|\ \beta_s\ \beta_d).\ y_n^{spc} \qquad (5)$$

Where

$$H(\ \gamma_s\ \gamma_d\ \big|\ \beta_s\ \beta_d) = \frac{(\mathscr{M}_s^{mag}\ (\gamma_s\ \big|\ \beta_s))^2}{(\mathscr{M}_s(\gamma_s+\beta_s))^2 + (\mathscr{M}_d\ (\gamma_d+\beta_d))^2} \qquad (6)$$

If the impulse response of the Wiener filter is h[n], then:

$$\hat{s}(n) = \sum_{m \in B} h(m)x(n-m) \tag{7}$$

The MSE is

$$\varepsilon = E\left[\left(s(n) - \hat{s}(n)\right)^2\right] \tag{8}$$

In order to minimize the MSE, we differentiate it with respect to h[i] where i = 0…B. The result is the following set of equations:

$$\sum_{m \in B} h(m)R_{xx}(i-m) = R_{sx}(i) \tag{9}$$

where $R_{sx}(i)$ is the cross-correlation between s(n) and x(n) at lag i .These equations are called the Wiener-Hopf equations.

If the filter is not restricted to be causal, we can obtain the transfer function by taking the DTFT of equation and rearranging the terms.

$$H(e^{j\omega}) = \frac{S_{sx}(e^{j\omega})}{S_{xx}(e^{j\omega})} \tag{10}$$

If we assume that the underlying speech signal and

the noise Are uncorrelated, then

$$R_{xx}(i) = R_{ss}(i) + R_{vv}(i) \tag{11}$$

Hence we get

$$H(e^{j\omega}) = \frac{S_{ss}(e^{j\omega})}{S_{ss}(e^{j\omega}) + S_{vv}(e^{j\omega})} \tag{12}$$

$$H(e^{j\omega}) = \frac{SNR(e^{j\omega})}{SNR(e^{j\omega}) + 1} \tag{13}$$

Parallel frequency and log - magnitude method

In this paper, the feature vectors of MFS domain obey Gaussian distribution, and taking into account the linear relationship between the feature vectors of MFS domain and the feature vectors of LOG-MAG domain, so the feature vectors of LOG-MAG domain should also obey Gaussian distribution. In order to not only utilize the benefits of the feature vectors of MFS domain in the HMM-based speech enhancement system but also eliminate the distortion effect of the inversion process in calculating the filters, we propose a parallel Mel-frequency and log-magnitude (PMLM) modeling approach to train the HMMs, as shown in Fig. 3.

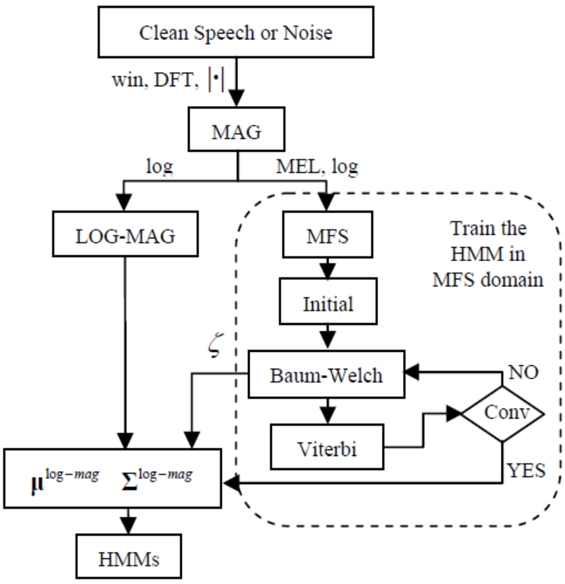

Fig3:Training process of parallel Mel frequency and log magnitude modeling

From Fig. 3 we can see that, the HMM in MFS domain is trained using the conventional HMM formulation given , and only mean vectors and covariance matrices of the HMM in LOG-MAG domain are obtained. In order to create the concurrent HMM in LOG-MAG domain, it is assumed that the alignment of the frames and states in the MFS and LOG-MAG domains are identical. The mean vectors and covariance matrices of HMM in LOG-MAG domain are given by Eq. 14 and Eq. 15 respectively, where x_n log-mag is the observation vector in the LOG-MAG domain corresponding to x_n mfs, ζ_n mfs(β,γ) is the probability of being in mixture γ of state β given the observation sequence x0:n mfs in the MFS domain, N indicates the total number of frames.

$$\mathcal{M}^{\text{log-mag}}(\beta,\gamma) = \frac{\sum_{n=1}^{N} \zeta_n^{\text{mfs}}(\beta,\gamma).x_n^{\text{log-mag}}}{\sum_{n=1}^{N} \zeta_n^{\text{mfs}}(\beta,\gamma)} \tag{14}$$

$$\Sigma^{\text{log-mag}}(\beta,\gamma) = \frac{\sum_{n=1}^{N} \zeta_n^{\text{mfs}}(\beta,\gamma)\ B.B^T}{\sum_{n=1}^{N} \zeta_n^{\text{mfs}}(\beta,\gamma)} \tag{15}$$

Where

$$B = x_n^{\text{log-mag}} - \mathcal{M}^{\text{log-mag}}(\beta,\gamma) \tag{16}$$

Now, the mean vectors and covariance matrices of HMM in LOG-MAG domain are known, the mean vectors in MAG domain can be obtained by

$$\mathcal{M}_i^{\text{mag}} = \exp(\mathcal{M}_i^{\text{log-mag}} + \Sigma_{ii}^{\text{log-mag}}/2) \tag{17}$$

RESULTS

In this section we will compare the results of the input speech signal and after applying HMM algorithm. The original Speech of the signal and the HMM approaches are shown.

The results have been obtained in the form of various performance parameters. At very first the original signal has been obtained both before and after enhancement. The signal enhancement is the technique which is used to improve the quality of the speech signal. The speech signal enhancement has been performed by using the of the hidden Markov model. The PMLM method has returned the feature extracted from the speech signal which is further used as the core sample to improve the quality of the speech signal using the HMM model. The speech signal has been obtained before and after the enhancement. There are the minute visible changes in the signal after the speech signal enhancement and so in the case of signal noise. The signal noise has been also calculated before and after the speech signal enhancement.

Noisy Speech is enhanced and noiseless speech is generated by filtering. Obtained a more accurate hidden Markov models (HMM) of noisy speech using the vector Taylor series (VTS) which is used to estimate the mean vectors and covariance matrices of HMM for noisy speech.

Reduced the distortion derived from inversion operation, a parallel Mel-frequency and log-magnitude (PMLM) modeling. Estimated clean speech waveform from a noisy signal, an inversion from the Mel-frequency domain to the spectral domain is required which introduces distortion artifacts in the spectrum estimation and the filtering.

Input audio is considered and the enhancement is played.

Following figure shows log magnitude plot of the original speech.

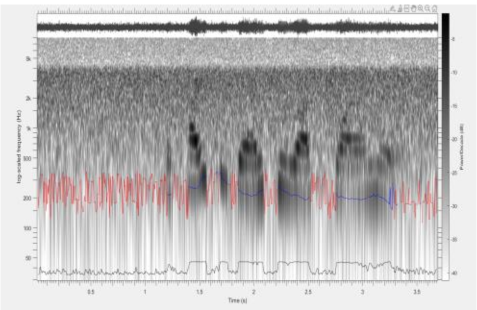

Fig 4: Log magnitude plot of the original speech

Following are the spectrograms of noisy and filtered speech.

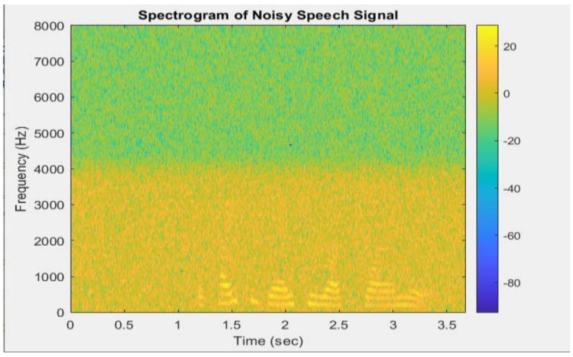

Fig5: Spectrogram of noisy and filtered speech

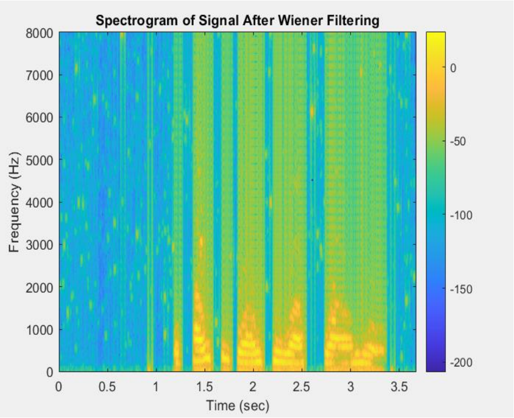

Fig6: Spectrogram of signal after Weiner filter

CONCLUSION

In this paper, we present a HMM-based speech enhancement using VTS and PMLM modeling in Mel frequency domain. The VTS is used to estimate the mean vectors and covariance matrices of noisy speech HMM more accurately, and the PMLM modeling is used to reduce the spectral distortion resulted by the inversion from Mel frequency domain to spectral domain. As obviously shown in the evaluation results, the performance of proposed method is better than the reference methods.
Demonstrating that the proposed methods reduce the annoying non-stationary noise associated with the enhanced speech.
Generating the spectrograms and calculating the mean and co-variance vectors.

ACKNOWLEDGMENT

We are really thankful to our guide without which the accomplishment of the task would have never been possible. we thankful to all other helpful people for providing us relevant information and necessary clarifications.

REFERENCES

[1] Y. Ephraim, D. Melah and B. Juang, “On the Application of Hidden Markov Models for Enhancing Noisy Speech,” Acoustics, Speech and Signal Processing, IEEE Transactions on, 1989, vol. 37, pp.1846-1856.

[2] Y. Ephraim, “A minimum mean square error approach for speech enhancement,” Acoustics, Speech and Signal Processing (ICASSP), 1990 International Conference on. IEEE, 1990, pp.829-832

[3] H. Sameti, H. Sheikhzadeh, L. Deng and L. B. Robert, “HMMbased strategies for enhancement of speech signals embedded in nonstationary noise,” Speech and Audio Processing, IEEE Transactions on, 1998, vol. 6, pp.445-455.

[4] M. Nilsson, M. Dahl and I. Claesson, “HMM-based speech enhancement applied in non-stationary noise using cepstral features and log-normal approximation,” Proc. DSPCS, 2003, pp.82-86.

[5] H. Veisi and H. Sameti, “A Parallel Cepstral and Spectral Modeling for HMM-based Speech Enhancement,” Digital Signal Processing (DSP), 2011 17th International Conference on. IEEE, 2011, pp.1-6.

[6] Zhen-zhen Gao, Chang-chun Bao, Feng Bao, Mao-shen, “HMM based speech enhancement using vector Taylor series and parallel modeling in Mel frequency domain,” IEEE ,2014

[7] Berdugo, B. and Cohen, I., “Noise estimation by minima controlled recursive averaging for robust speech enhancement”, IEEE Signal Proc. Letters, vol. 9, no. 1, pp. 12-15, Jan. 2002.

[8] R. Martin, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimation”, in Proc. IEEE Int. Conf. Acoust Speech, Signal Processing, pp. 504—512, 2002.

[9] Cohen, “Noise spectrum estimation in adverse environments Improved minima controlled recursive averaging,” IEEE Trans. on speech and audio processing, vol. 11, no. 5, pp. 466-475, Sept. 2003.

[10] Barinder pal singh, Dr shahi Bhushan, Mr. Karan Mahajan,“speech enhancement using Hidden Markov model,” IJCSMC, Vol. 4, Issue. 7, July 2015.

[11] Hadi veisi,Hossein sameti, “speech enhancement using hidden Markov models in Mel-frequency domain,” speech communication,vol.55,Issue 2,feb 2013.

[12] Daniel Dzibela, Armin Sehr “Hidden Markov Model Based Speech Enhancement”, 2017 IEEE.

[13] Akhirao Kato and Ben Milner “Using Hidden Markov Models for Speech Enhancement” 2014 ICSA

[14] Przybyla Dymarski “Hidden Markov Models, Theory and Applications” 2011.