International Journal of Computer Applications (0975 - 8887)
Volume 180 - No.42, May 2018

Noisy Speech Recognition by Mel-LPC based AR-HMM with Power and Time Derivative Parameters

M. Babul Islam
Dept. of Applied Physics and Electronic Engineering
University of Rajshahi, Rajshahi 6205, Bangladesh

ABSTRACT

In this paper, AR-HMM on mel-scale with power and Mel-LPC based time derivative parameters has been presented for noisy speech recognition. The mel-scaled AR coefficients and mel-prediction coefficients for Mel-LPC have been calculated on the linear frequency scale from the speech signal without applying bilinear transformation. This has been done by using a first-order all-pass filter instead of unit delay. In addition, Mel-Wiener filter has been applied to the system to improve the recognition accuracy in presence of additive noise. The proposed system is evaluated on Aurora 2 database, and the overall recognition accuracy has been found to be 80.02% on the average.

Keywords

AR-HMM, Mel-LPC, Mel-Wiener filter, Aurora 2 database

1. INTRODUCTION

The speech recognition processes have been investigated by many researchers [1], [2], [3], [4], [5], [6] in the framework of autoregressive hidden Markov model (AR-HMM) [7]. In all of these works researchers have not tried to incorporate power and/or time derivative parameters of speech signal.

It has been found that the AR-HMM is a useful method to represent clean speech [1], [2]. In conventional AR-HMM, all the states are assumed to be stationary stochastic sequences whereas speech signal reveals the most notable nonstationary nature. Although the AR-HMM is suitable for LPC based front-end, it cannot deal with frequency dependent spectral variation and dynamic property of spectra. Consequently, AR-HMM based recognition system has inferior performance as compared to MFCC based system [6].

To overcome all these limitations, the pdf of Gaussian AR source for the static spectra and the Gaussian pdf for both energy and delta cepstrum have been combined in the proposed system. Furthermore, as in MFCC or PLP, auditory-like frequency resolution has been incorporated into AR-HMM by Mel-LPC [8], [9].

Though, the AR-HMM is suitable for clean speech as mentioned above, its performance severely degrades with noisy environments. Consequently, we have incorporated a filtering scheme along with the Mel-LPC based AR-HMM. Previously we have proposed a time-domain Mel-Wiener filter [10] on the linear frequency scale by using a first-order all-pass filter instead of unit delay. This Mel-Wiener filter has been efficiently applied in the autocorrelation domain to enhance mel-autocorrelation function of noisy speech.

The rest of the paper is organized as follows. Section 2 comprises of three subsections− subsection 2.1 introduces an overview of Mel-LPC analysis, subsection 2.2 deals with AR-HMM for bilinear transformed signal and subsection 2.3 describes the inclusion technique of power and time derivative parameters into autoregressive hidden Markov model. Section 3 gives a short description of previously proposed Mel-Wiener filter. System overview is introduced in section 4. Experimental setup for the proposed system is given in section 5. The performance of the proposed system is illustrated in section 6. Finally, conclusion is presented in section 7.

2. AR-HMM FOR MEL-LPC BASED FRONT-END

2.1 Mel-LPC Analysis

In the Mel-LPC analysis, the following all-pole model is defined for frequency warped speech signal $\tilde{x}[n]$ $(n = 0, 1, .., \infty)$ which is bilinear transformed [11] from a windowed input speech signal $x[n]$ $(n = 0, 1, .., N-1)$:

$$\tilde{H}_\alpha(\tilde{z}) = \frac{\tilde{\sigma}_e}{1 + \sum_{k=1}^{p} \tilde{a}_k \tilde{z}^{-k}} \tag{1}$$

where $\tilde{z}^{-1}$ is a first-order all-pass filter,

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \tag{2}$$

The phase response of $\tilde{z}^{-1}$ is given by

$$\tilde{\lambda} = \lambda + 2 \cdot \tan^{-1}\left\{\frac{\alpha \sin\lambda}{1 - \alpha\cos\lambda}\right\} \tag{3}$$

This phase function determines a frequency mapping. As shown in Figure 1, $\alpha = 0.35$ and $\alpha = 0.40$ can approximate the mel-scale and bark-scale at the sampling frequency of 8 kHz respectively.

Actually, in Mel-LPC analysis, the spectral envelope of $\tilde{X}(\tilde{z})\tilde{W}(\tilde{z})$ is approximated by the all-pole model given in Equation (1), where the frequency weighting function $\tilde{W}(\tilde{z})$ is given by

$$\tilde{W}(\tilde{z}) = \frac{\sqrt{1-\alpha^2}}{1 + \alpha \cdot \tilde{z}^{-1}} \tag{4}$$

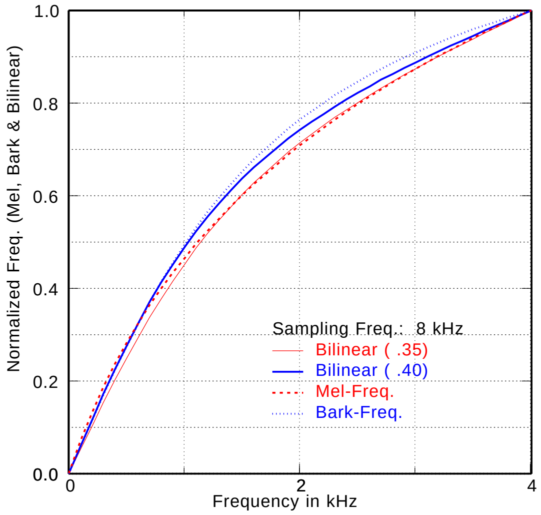

Fig. 1. Frequency mapping by bilinear transformation.

which is derived from

$$\frac{d\lambda}{d\tilde{\lambda}} = \left|\tilde{W}(\tilde{z})\right|^2 \tag{5}$$

The mel-prediction coefficients $\{\tilde{a}_k\}$ can be calculated directly from the input speech signal without applying bilinear transformation as shown in [8], [9].

2.2 AR-HMM for Bilinear Transformed Signal

Let $\tilde{x}[n]$ be the bilinear transformed and gain normalized signal of $x[n]$, that is, in the LPC terminology this is equivalent to the normalization by the square root of average residual energy. It is assumed that $\tilde{x}[n]$ is generated by an Mth order zero mean autoregressive process. Therefore

$$\tilde{e}_n = \sum_{i=0}^{M} \tilde{a}_i \tilde{x}[n-i] \tag{6}$$

where $\tilde{e}[n]$ are Gaussian i.i.d. random variables with zero mean and unity variance, and $\{\tilde{a}_i\}$ are the mel-scaled AR coefficients with $\tilde{a}_0 = 1$. It should be noted that the mel-scaled AR coefficients can be calculated from the gain normalized signal of $x[n]$ in the same way as mel-prediction coefficients are calculated.

Now, for large N, the probability density function for x can be approximated by [7]

$$f_a(\boldsymbol{x}) \approx (2\pi)^{-N} \exp\{-\frac{1}{2}\delta(\boldsymbol{x};\tilde{\boldsymbol{a}})\} \tag{7}$$

where

$$\delta(\boldsymbol{x};\tilde{\boldsymbol{a}}) = R_{\tilde{a}}[0]\tilde{r}_x[0] + 2\sum_{i=1}^{M} R_{\tilde{a}}[i]\tilde{r}_x[i] \tag{8}$$

$R_{\tilde{a}}[i]$ is the autocorrelation function of AR coefficients and $\tilde{r}_x[i]$ is the mel-autocorrelation function [8], [9] of x.

2.3 Inclusion of Energy and Time Derivative Parameters

Energy and time derivative parameters of Mel-LPC can be included by estimating the joint probability of AR coefficients, energy and time derivative Mel-LPC parameters, which is given by

$$f(x,y) = f_a(x)f_g(y) \tag{9}$$

where $f_a(x)$ is the autoregressive probability density function, given by Equation 7, and $f_g(y)$ is the Gaussian probability density function, given by

$$f_g(y) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} \exp\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\} \tag{10}$$

and $y = [c_0, \Delta c_0, \Delta c_1, ..., \Delta c_p]$.

3. MEL-WIENER FILTER

It has been found form the recognition experiment that the performance of AR-HMM is satisfactory for clean speech only. On the contrary, its performance severely degrades against SNRs irrespective of noise types. Consequently, an effective enhancement system is required to improve the performance of AR-HMM in different SNR conditions.

To improve the performance of the proposed system, we have used the previously implemented Mel-Wiener filter (MWF) [10] which was formulated on the linear frequency scale by using a first-order all-pass filter instead of unit delay and a remarkable improvement has been obtained.

The transfer function of the Mel-Wiener filter on z domain is defined as

$$\tilde{H}_w(\tilde{z}) = \sum_{n=0}^{p-1} \tilde{h}_w[n]\tilde{z}^{-n} \tag{11}$$

Now, the estimated clean speech $\hat{s}_w[n]$ based on filter $\tilde{H}_w(\tilde{z})$ is given by

$$\hat{s}_w[n] = \sum_{k=0}^{p-1} \tilde{h}_w[k]x_k[n] \tag{12}$$

where $x_k[n]$ is the output signal of k cascaded all pass filter $\tilde{z}^{-k}$.

It should be noted that filtering is done in the autocorrelation domain as follows:

$$\hat{\tilde{r}}_s[m] = \sum_{k=-p+1}^{p-1} r_{\tilde{h}}[k]\tilde{r}_x[m-k] \tag{13}$$

where $r_{\tilde{h}}[m]$ is the autocorrelation function of filter coefficients $\{\tilde{h}_w[m]\}$.

4. SYSTEM OVERVIEW

The block diagram of the proposed system is shown in Figure 2. The filtering is done in autocorrelation domain to obtain the mel-autocorrelation function of the enhanced speech. From the estimated mel-autocorrelation function the energy and time derivative mel-cepstra have been calculated. The mel-scaled AR coefficients have been obtained by using gain normalized mel-autocorrelation

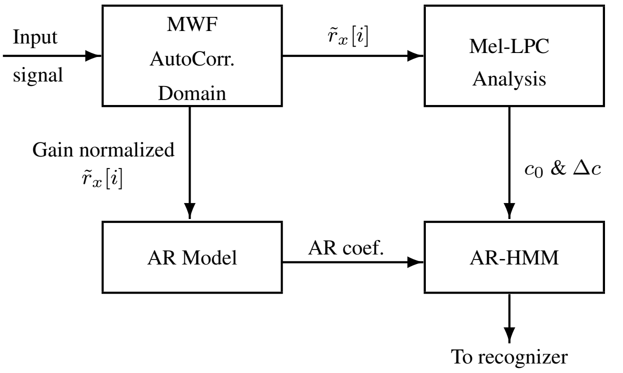

Fig. 2. Mel-LPC based AR-HMM system with energy and time derivative parameters.

function. Finally, the autoregressive hidden markov model (AR-HMM) for each word has been created by using the joint probability of AR coefficients, energy and time derivative Mel-LPC parameters which is given by Equation (9).

5. EXPERIMENTAL SETUP

The order of AR process and Mel-LPC analysis were set to 12. The speech signal was windowed using Hamming window of length 20 ms with 10 ms frame period. As the frequency weighting function $\tilde{W}(\tilde{z})$ defined by Equation (4) acts like a preemphasis, the speech signal was not preemphasized. The warping factor was set to 0.4. Each feature vector consists of 13 AR coefficients, one energy term and 14 delta mel-cepstral coefficients.

The reference recognizer was based on HTK (Hidden Markov Model Toolkit) software package. The HMM was trained on clean condition. The digits are modeled as whole word HMMs with 16 states per word and a mixture of 3 to 10 Gaussians per state using left-to-right models. In addition, two pause models 'sil' and 'sp' are defined. The 'sil' model consists of 3 states. This HMM models the pauses before and after the utterance. A mixture of 6 Gaussians models each state. The second pause model 'sp' is used to model pauses between words. It consists of a single state, which is tied with the middle state of the 'sil' model.

Table 1. Performance of AR-HMM on word accuracy as a function of mixture components for subway noise in set A.

No. of mixture	SNR [dB]		
	clean	20	15
3	90.3	74.2	54.5
6	91.3	76.2	56.2
10	91.3	75.0	55.0

Table 2. Effect of power term and delta parameters on recognition accuracy for subway noise in test set A.

Parameters	SNR [dB]		
	clean	20	15
with c_0	94.2	41.7	17.8
with c_0 & Δc	99.2	81.5	67.5

Table 3. Recognition accuracy for AR-HMM with c_0 and Δc_0 for test set A.

Noise	SNR [dB]		
	clean	20	15
Subway	99.2	81.5	67.5
Babble	99.1	81.4	52.1
Car	98.9	80.3	57.2
Exhibition	99.3	83.0	65.9

Table 4. Recognition accuracy for Mel-LPC with c_0 and Δc_0 for test set A.

Noise	SNR [dB]		
	clean	20	15
Subway	98.9	97.3	93.8
Babble	98.8	91.1	64.9
Car	98.6	97.6	87.3
Exhibition	99.1	97.3	94.4

6. PERFORMANCE EVALUATION

The performance of the proposed system was evaluated on Aurora 2 database [12], which is a subset of TIDigits database [13] contaminated by additive noises and channel effects.

At first, the effect of mixture components on recognition accuracy has been examined for the AR-HMM system using test set A and noise kind subway. As shown in Table 1, the increase of mixture components does not improve the recognition accuracy as in [6] rather the optimum accuracy is obtained for 6 mixture components. So, in the subsequent recognition experiment the mixture components were set to 6.

Effect of power and time derivative parameters is shown in Table 2. Incorporation of power term improves the accuracy for clean speech and degraded performance is observed for noisy speech as expected. On the contrary, inclusion of power and delta parameters improves the accuracy for all cases. As shown in Table 3 and 4, for clean speech, AR-HMM with power and delta parameters attains slightly better performance than that of Mel-LPC cepstral parameters. However, for noisy speech, performance is very weak against SNRs.

As the AR-HMM with power and delta parameters outperforms Mel-LPC for clean speech only, an enhancement scheme is required to obtain the better performance in noisy conditions. Therefore, from this viewpoint, Mel-Wiener filter (MWF) has been used to speech signal before estimating the AR-HMM and the overall recognition performance for all three test sets A, B and C of Aurora 2 database is presented in Table 5. In this table it has been shown that a remarkable improvement has been achieved for all three sets and on the average, recognition accuracy has been improved from 44.66% to 80.02%.

7. CONCLUSION

In this work, auditory-like frequency resolution has been achieved for AR-HMM by estimating mel-AR coefficients. Since AR-HMM cannot deal with power and time derivative parameters, the power and Mel-LPC based delta parameters have been effectively incorporated. Thus the significant improvement is obtained in word accuracy. Furthermore, Mel-Wiener filter has been applied to the system to improve the recognition accuracy in different noisy conditions. The overall recognition accuracy has been obtained with the proposed system is 80.02%.

8. REFERENCES

[1] Juang, B. and Rabiner, L. R. 1986. Mixture autoregressive hidden Markov models for speech signals. IEEE Trans. Acoust., Speech, Signal Processing, 33: 1404-1413.

[2] Ephraim, Y. 1992. Gain adapted hidden Markov models for recognition of clean and noisy speech. IEEE Trans. Signal Processing, 40(6): 1303-1316.

[3] Ruske, G. and Lee, K. Y. 1999. Speech recognition and enhancement by a nonstationary AR HMM with gain adaptation under unknown noise. Proceedings ICASSP'99.

[4] Deng, L. 1992. A generalized hidden Markov model with state conditioned trend functions of time for speech signal. Signal Processing, 27: 65-72.

[5] Lee, K. Y. and Lee, J. 2001. Recognition of noisy speech by a nonstationary AR HMM with gain adaptation under unknown noise. IEEE Trans. Speech and Audio Processing, 1 9(7): 741-746.

[6] Logan, B. T. and Robinson, A. J. 1997. Improving autoregressive hidden Markov model recognition accuracy using a nonlinear frequency scale with application to speech enhancement. Proc. of EUROSPEECH, 2103-2106.

[7] Juang, B. 1984. On the hidden Markov model and dynamic time warping for speech recognition - a unified view. AT&T Bell Lab. Tec. Journal, 63(7): 1213-1243.

[8] Strube, H. W. 1980. Linear prediction on a warped frequency scale. J. Acoust. Soc. America, 68(4): 1071-1076.

[9] Matsumoto, H., et al. 1998. An efficient Mel- LPC analysis method for speech recognition. Proc. of ICSLP98: 1051-1054.

[10] Islam, M. B., et al. 2007. Mel-Wiener filter for Mel-LPC based speech recognition. IEICE Transactions on Information and Systems, E90-D (6): 935-942.

[11] Oppenheim, A. V. and Johnson, D. H. 1972. Discrete representation of signals. IEEE Proc., 60(6): 681-691.

[12] Hirsch, H. G. and Pearce, D. 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ITRW ASR2000: 181-188.

[13] Leonard, R. G. 1984. A database for speaker independent digit recognition. ICASSP84, 3: 42.11.1-42.11.4.

Table 5. Overall recognition accuracy for Mel-LPC based AR-HMM (including c_0 & Δc_0) w/ and w/o MWF.

SNR	Set A		Set B		Set C		Average	
	w/o MWF	w/ MWF	w/o MWF	w/ MWF	w/o MWF	w/ MWF	w/o MWF	w/ MWF
clean	99.19	98.32	99.19	98.32	99.33	97.87	99.30	98.30
20 dB	84.95	97.04	84.66	96.65	85.77	96.08	85.00	96.70
15 dB	67.53	95.46	64.54	94.02	73.88	93.88	67.60	94.60
10 dB	42.52	90.62	39.86	87.43	55.37	87.57	44.10	88.80
5 dB	18.80	77.20	15.09	70.58	31.65	73.72	19.90	73.90
0 dB	7.12	49.08	2.71	41.46	13.39	49.25	6.70	46.10
-5 dB	5.38	16.90	2.42	12.29	8.81	23.51	4.90	16.40
Average	44.19	81.88	41.38	78.03	52.02	80.10	44.66	80.02