Performance Evaluation of Blind Equalization for Mel-LPC based Speech Recognition under Different Noisy Conditions

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

Md. Mahfuzur Rahman
Department of Computer Science and Engineering,
Comilla University, Comilla, Bangladesh

ABSTRACT
This study is aimed to develop a noise robust distributed speech recognizer (DSR) for real-world applications by employing Blind Equalization (BEQ) for robust feature extraction. The main focus of the work is to cope with different noisy environments in recognition phase. To realize this objective, Mel-LP based speech analysis has been used in speech coding on the linear frequency scale by applying a first-order all-pass filter instead of a unit delay. Mismatch between training and test phases is reduced through robust feature extraction and this is achieved by applying BEQ on Mel-LP cepstral coefficients as an effort to reduce additive noise and channel distortion. The performance of the proposed system has been evaluated on test set A and set C of Aurora-2 database. The baseline performance, that is, for Mel-LPC the average word accuracy has found to be 59.05% and 63.99% for sets A and C, respectively. By applying the BEQ on Mel-LP cepstral coefficients, the performance has been improved to 65.66% and 64.65% for sets A and C, respectively.

Keywords
Mel-LPC, bilinear transformation, BEQ, Aurora 2 database

1. INTRODUCTION
The accuracy of Automatic Speech Recognizers (ASRs) has reached to a satisfactory level under controlled and matched training and recognition conditions. However, the performance of ASR severely degrades when there is a mismatch between training and test phases, caused by additive noise and channel effect. Environmental noises as well as channel effects contaminate the speech signal and change the data vectors representing the speech, for instance, reduce the dynamic range, or variance of feature parameters within the frame[1][2]. Consequently, a serious mismatch is occurred between training and recognition conditions, resulting in degradation in recognition accuracy.

Noise robust ASR can be achieved in many ways, such as, enhancement of speech signal either in time domain [3] or in frequency domain[4][5][6][7][8], enhancement in cepstral domain [9][10][11][12], that is, feature parameter compensation, and acoustic model compensation or model adaptation[13][14][15].

In HMM based recognizer, the model adaptation approaches have been shown to be very effective to remove the mismatch between training and test environments. However, for a distributed speech recognition system, speech enhancement and parameter compensation approaches are suitable than the model adaptation approach. Because the acoustic model resides at a server, so adaptation or compensation of model from the front-end is not feasible. Therefore, this paper deals with the design of front-end with parameter compensation, such as BEQ.

Since the human ear resolves frequencies non-linearly across the speech spectrum, designing a front-end incorporating auditory-like frequency resolution improves recognition accuracy[16][17][18]. In nonparametric spectral analysis, Mel-frequency Cepstral Coefficient (MFCC)[16] is one of the most popular spectral features in ASR. This parameter takes account of the nonlinear frequency resolution like the human speech perception system.

In parametric spectral analysis, the linear prediction coding (LPC) analysis [19][20] based on an all-pole model is widely used because of its computational simplicity and efficiency. While the all-pole model enhances the formant peaks as an auditory perception, other perceptually relevant characteristics are not incorporated into the model unlike MFCC. To alleviate this inconsistency between the LPC and the auditory analysis, several auditory spectra have been simulated before the all-pole modeling [17][[21][22][23].

In contrast to the different spectral modification, Strube [24] proposed an all-pole modeling to a frequency warped signal which is mapped onto a warped frequency scale by means of the bilinear transformation [25], and investigated several computational procedures. However, the methods proposed by Oppenheim and Johnson [25] to estimate warped all-pole model have rarely been used in automatic speech recognition. Recently, as an LP-based front-end, a simple and efficient time domain technique to estimate all-pole model is proposed by Matsumoto *et al*.[26], which is referred to as a “Mel-LPC” analysis. In this method, the all-pole model has been estimated directly from the input signal without applying bilinear transformation. Hence, the prediction coefficients can be estimated without any approximation by minimizing the prediction error power at a two-fold computational cost over the standard LPC analysis.

In our previous work [12] effect of CMN has been examined on test set A of Aurora 2 database. As an extension of that study this paper deals with the design of Mel-LP based front-end with BEQ as parameter compensation. As compared to previous work this research uses test sets A and C of Aurora 2 database in the evaluation phase. It should be noted that the filtering characteristics for sets A and C are different which is mentioned in subsection 4.1.

2. MEL-LP ANALYSIS
The frequency-warped signal $\tilde{x}[n]$ $(n = 0,\dots,\infty)$ obtained by the bilinear transformation [25] of a finite length windowed signal $x[n]\,(n = 0,1,\dots,N-1)$ is defined by

$$\tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} = X(z) = \sum_{n=0}^{N-1} x[n] z^{-n} \tag{1}$$

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

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

where $0 < \alpha < 1$ is treated as frequency warping factor.

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 Fig. 1, $\alpha = 0.35$ and $\alpha = 0.40$ can approximate the mel-scale and bark-scale [27][28] at the sampling frequency of 8 kHz respectively.

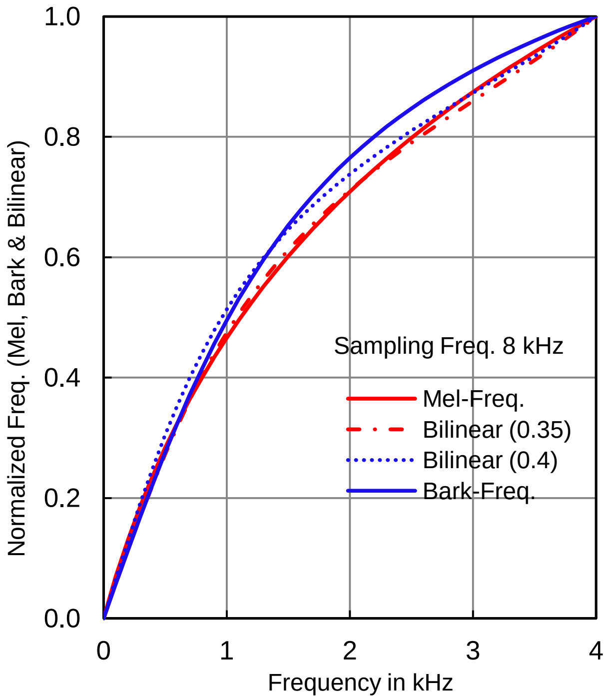

Fig. 1: The frequency mapping functions by bilinear transformation.

Now, the all-pole model on the warped frequency scale is defined as

$$\tilde{H}(\tilde{z}) = \frac{\tilde{\sigma}_e}{1 + \sum_{k=1}^{p} \tilde{a}_k \tilde{z}^{-k}} \tag{4}$$

where $\tilde{a}_k$ is the k-th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy[24].

On the basis of minimum prediction error energy for $\tilde{x}[n]$ over the infinite time span, $\tilde{a}_k$ and $\tilde{\sigma}_e$ are obtained by Durbin's algorithm from the autocorrelation coefficients $\tilde{r}[m]$ of $\tilde{x}[n]$ defined by

$$\tilde{r}[m] = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{x}[n-m] \tag{5}$$

which is referred to as mel-autocorrelation function.

The mel-autocorrelation coefficients can easily be calculated from the input speech signal $x[n]$ via the following two steps[26][29]. First, the generalized autocorrelation coefficients are calculated as

$$\tilde{r}_\alpha[m] = \sum_{n=0}^{N-1} x[n] x_m[n] \tag{6}$$

where $x_m[n]$ is the output signal of an m-th order all pass filter $\tilde{z}^{-m}$ excited by $x_0[n] = x[n]$. That is, $\tilde{r}_\alpha[m]$ is defined by replacing the unit delay z^{-1} with the first order all-pass filter $\tilde{z}(z)^{-1}$ in the definition of conventional autocorrelation function as shown in Fig. 2.

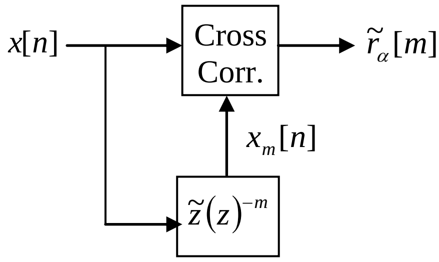

Fig. 2: Generalized autocorrelation function.

Due to the frequency warping, $\tilde{r}_\alpha[m]$ includes the frequency weighting $\tilde{W}(e^{j\tilde{\lambda}})$ defined by

$$\tilde{W}(\tilde{z}) = \frac{\sqrt{1 - \alpha^2}}{1 + \alpha \tilde{z}^{-1}} \tag{7}$$

which is derived from

$$\frac{d\lambda}{d\tilde{\lambda}} = \left|\tilde{W}(e^{j\tilde{\lambda}})\right|^2 \tag{8}$$

Thus, in the second step, the weighting is removed by inverse filtering in the autocorrelation domain using $\{\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})\}^{-1}$.

As feature parameters for recognition, the Mel-LP cepstral coefficients can be expressed as:

$$\log \tilde{\mathrm{H}}(\tilde{z}) = \sum_{n=0}^{\infty} c_k \tilde{z}^{-n} \tag{9}$$

where $\{c_k\}$ are the mel-cepstral coefficients.

The mel-cepstral coefficients can also be calculated directly from mel-prediction coefficients $\{\tilde{a}_k\}$[30] using the following recursion:

$$c_k = -\tilde{a}_k - \frac{1}{k}\sum_{j=1}^{k-1}(k-j)\tilde{a}_k c_{k-j} \tag{10}$$

It should be noted that the number of cepstral coefficients need not be the same as the number of prediction coefficients.

3. ENHANCEMENT OF MEL-LP CEPSTRUM

3.1 Blind Equalization

Blind equalization is a technique effective for minimizing the channel distortion which is caused by the differences in the input devices' frequency characteristics. It uses adaptive filtering technique to reduce these effects. It can be applied both in spectral domain as well as in cepstral domain [31][32]. But in the cepstral domain it is easier to implement, and it requires less operations than in the spectral domain. This technique is based on the least mean square (LMS) algorithm, which minimizes the mean square error computed as a difference between the current and reference cepstrum.

In this study, the same algorithm is used as that implemented in [3] with same values of different parameters, and the long-term cepstrum of training clean speech is used as reference cepstrum.

The algorithm is as follows:

$$w = Min(1, Max(0, \ln E - 4.75)$$

$$step = 0.008 * w$$

$$c_{eq}[i] = c[i] - bias[i], \qquad 0 \le i \le 13$$

$$bias[i] = bias[i] + step * (c_{eq}[i] - C_{Ref}[i]), \quad 0 \le i \le 13$$

where w is the weighting parameter, $\ln E$ indicates the log energy of the current frame, $bias[i]$ is initialized on 0.0 $(0 \le i \le 13)$ and $C_{Ref}[i]$ is the reference cepstrum.

4. EVALUATION ON AURORA 2 DATABASE

4.1 Experimental Setup

The proposed system was evaluated on Aurora-2 database [33], which is a subset of TIDigits database [34] contaminated by additive noises and channel effects. This database contains the recordings of male and female American adults speaking isolated digits and sequences up to 7 digits. In this database, the original 20 kHz data have been down sampled to 8 kHz with an ideal low-pass filter extracting the spectrum between 0 and 4 kHz. These data are considered as clean data. Noises are artificially added with SNR ranges from 20 to -5 dB at an interval of 5 dB.

To consider the realistic frequency characteristics of terminals and equipment in the telecommunication area an additional filtering is applied to the database. Two standard frequency characteristics G.712 and MIRS are used which have been defined by ITU [35]. The frequency responses of G.712 and MIRS filters have been shown in Fig. 3 and Fig. 4, respectively.

It should be noted that the whole Aurora 2 database was not used in this experiment rather a subset of this database was used as shown in Table 1.

Table 1: Definition of training and test data.

	Filter	Data set	Noise Type	SNR [dB]
Training	G.712	Clean	–	∞
Test	G.712	Set A	subway, babble, car, exhibition	clean, 20, 15, 10, 5, 0, -5
	MIRS	Set C	subway, street	clean, 20, 15, 10, 5, 0, -5

The recognition experiments were conducted with a 12th order prediction model of Mel-LPC analysis. The pre-emphasized speech signal with a pre-emphasis factor of 0.95 was windowed using Hamming window of length 20 ms with 10 ms frame period. The frequency warping factor was set to 0.35. As front-end, 14 cepstral coefficients and their delta coefficients including 0th terms were used. Thus, each feature vector size is 28.

The reference recognizer was based on HTK (Hidden Markov Model Toolkit, version 3.4) 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 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 which is illustrated in Fig. 5. This HMM shall model 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.

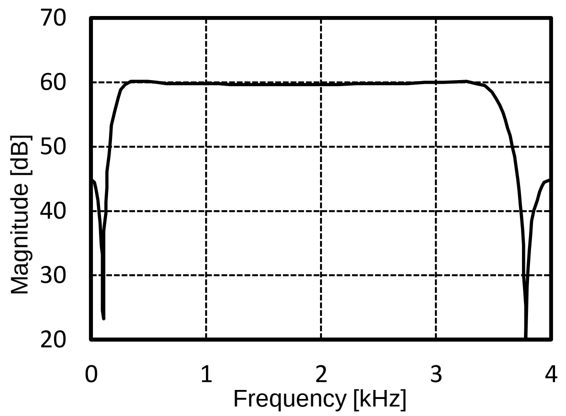

Fig. 3: Frequency response of G.712 filter.

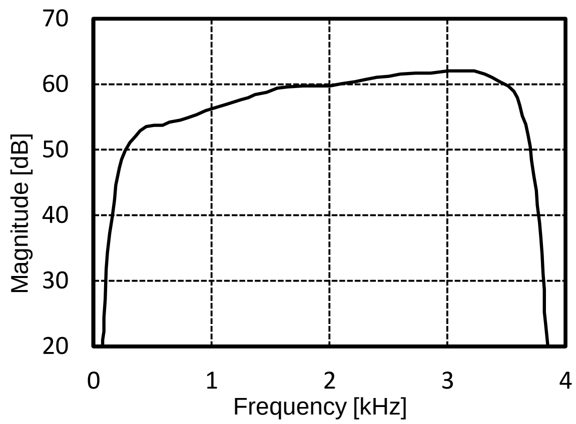

Fig. 4: Frequency response of MIRS filter.

The recognition accuracy (Acc) is evaluated as follows:

$$Acc = \frac{N - D - S - I}{N} \times 100\% \qquad (12)$$

where N is the total number of words. D, S and I are deletion, substitution and insertion errors, respectively.

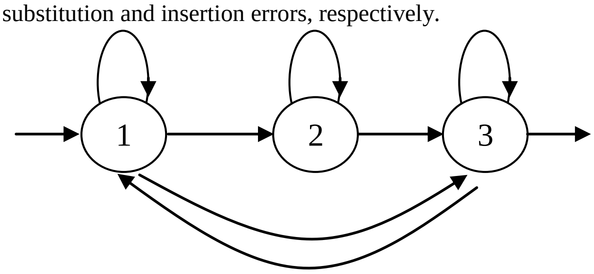

Fig. 5: Possible transition in the 3-state pause model 'sil'.

4.2 Experimental Results

The detail recognition results are presented in this section. The word accuracy for Mel-LPC without applying BEQ is listed in Table 2 for test set A, which is considered as baseline result. The average word accuracy over all noises within the set A and over SNRs 20 to -5 dB is found to be 59.05% for the baseline, while the average recognition performance of Mel-LPC with BEQ is 65.66% as listed in Table 3. It is also observed that noticeable improvements in recognition accuracy are achieved for babble and car noises with BEQ as compared to baseline performance. The average recognition accuracy does not differ significantly for subway noise. From Tables 2 and 3 it has been also observed that greater improvements are found to be at SNRs 10 to 0 dB conditions. It has also been noticeable that the recognition accuracy is not improved significantly for exhibition noise at any level of SNR conditions.

In the case of set C, the baseline performance is found to be 63.99% while the accuracy with BEQ is 64.65% on the average as shown in Tables 4 and 5, respectively. It has also been observed that the improvement is obtained only at SNR conditions of 5 to -5 dB for both subway and street noises.

As compared to our previous work [12] where CMN has been used to minimize channel distortion, it has been found that BEQ gives outstanding performance over CMN for car noise only.

5. CONCLUSION

An HMM-based automatic speech recognizer (ASR) was developed and as an enhancement technique in the cepstral domain, the performance of BEQ on Mel-LPC was evaluated on test sets A and C of Aurora 2 database. It is observed that the BEQ has significant effect for babble and car noises. It is also found that BEQ exhibits best performance for car noise. The average word accuracy does not differ significantly for subway noise and slightly degrades for exhibition noise of set A after applying BEQ. The recognition performance for babble and car noises has been improved from 48.06% to 61.08% and from 53.77% to 76.28%, respectively. The improvement is also significant for 15 to 0 dB speech signals. No significant improvement has been found for any level of exhibition noise. The overall recognition accuracy for test set A has been improved from 59.05% to 65.66%.

For test set C, the BEQ is only effective for SNR conditions 5 to -5 dB.

Table 2: Word accuracy [%] for Test set A without BEQ (baseline).

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
Subway	98.71	96.93	93.43	78.78	49.55	22.81	11.08	68.30
Babble	98.61	89.96	73.76	47.82	21.95	6.80	4.44	48.06
Car	98.54	95.26	83.03	54.25	24.04	12.23	8.77	53.77
Exhibition	98.89	96.39	92.72	76.58	44.65	19.90	11.94	66.05
Average	98.69	94.64	85.74	64.36	35.05	15.44	9.06	59.05

Table 3: Word accuracy [%] for Test set A with BEQ.

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
Subway	97.14	93.37	88.24	74.79	54.93	27.54	7.18	67.77
Babble	96.83	92.65	85.40	71.61	44.11	11.64	-8.40	61.08
Car	97.35	96.18	94.72	87.24	68.86	34.42	3.67	76.28
Exhibition	97.56	89.63	82.57	65.32	38.63	11.23	-1.88	57.48
Average	97.22	92.96	87.74	74.74	51.64	21.21	0.15	65.66

Table 4: Word accuracy [%] for Test set C without BEQ (baseline).

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
Subway	99.08	93.89	86.95	72.15	43.75	18.18	8.87	62.99
Street	98.46	93.95	87.7	71.98	47.67	23.67	11.97	65.00
Average	98.77	93.92	87.33	72.07	45.71	20.93	10.42	63.99

Table 5: Word accuracy [%] for Test set C with BEQ.

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
Subway	97.08	79.86	74.46	67.24	57.32	42.83	22.54	64.35
Street	96.52	83.74	77.66	67.41	56.74	39.21	19.38	64.96
Average	96.80	81.80	76.06	67.33	57.03	41.02	20.96	64.65

6. REFERENCES

[1] Bateman, D. C., *et al*., 1992. Spectral contrast normalization and other techniques for speech recognition in noise. Proc. of ICASSP ’92, I: 241-244.

[2] Vaseghi, S. V. and B.P. Milner, 1993. Noise-adaptive hidden Markov models based on Wiener filters. Proc. of Eurospeech ’93, II: 1023-1026.

[3] Islam, M. B., K. Yamamoto, H. Matsumoto, 2007. Mel-Wiener filter for Mel-LPC based speech recognition. IEICE Transactions on Information and Systems, E90-D (6): 935-942.

[4] Boll, S. F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech and Signal Processing, 27(2): 113-120.

[5] Lim, J. S. and A.V. Oppenheim, 1979. Enhancement and bandwidth compression of noisy speech. Proc. of the IEEE, 67(2): 1586-1604.

[6] Lockwood, P. and J. Boudy, 1992. Experiments with a nonlinear spectral subtractor (nss), hidden Markov models and the projection or robust speech recognition in cars. Speech Commun., 11(2-3): 215-228.

[7] Agarwal, A. and Y. M. Cheng, 1999. Two-stage Mel-warped Wiener filter for robust speech recognition. Proc. of ASRU ’99: 67-70.

[8] Zhu, Q. and A. Alwan, 2002. The effect of additive noise on speech amplitude spectra: A Quantitative analysis. IEEE Signal Processing Letters, 9(9): 275-277.

[9] Atal, B., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55(6): 1304-1312.

[10] Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust., Speech and Signal Processing, ASSP-29: 254-272.

[11] Mokbel, C., *et al*., 1984. Compensation of telephone line effects for robust speech recognition. Proc. of ICSLP ’94: 987-990.

[12] Rahman, M. M., *et al*., 2012. Performance evaluation of CMN for Mel-LPC based speech recognition in different noisy environments. International Journal of Computer Applications, 58(10): 6-10.

[13] Gales, M. J. F. and S. J. Young, 1993a. HMM recognition in noise using parallel model combination. Proc. of Eurospeech ’93, II: 837-840.

[14] Gales, M. J. F. and S. J. Young, 1993b. Cepstral parameter compensation for HMM recognition in noise. Speech Communication, 12(3): 231-239.

[15] Varga, A. P. and R. K. Moore, 1990. Hidden Markov model decomposition of speech and noise. Proc. of ICASSP ’90, 2: 845-848.

[16] Davis, S. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-28(4): 357-366.

[17] Hermansky, H., 1987. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87(4): 17-29.

[18] Virag, N., 1995. Speech enhancement based on masking properties of the auditory system. Proc. ICASSP ’95: 796-799.

[19] Itakura, F. and S. Saito, 1968. Analysis synthesis telephony based upon the maximum likelihood method. Proc. of 6th International Congress on Acoustics, Tokyo: C-5-5, C17-20.

[20] Atal, B. and M. Schroeder, 1968. Predictive coding of speech signals. Proc. of 6th International Congress on Acoustics, Tokyo: 21-28.

[21] Makhoul, J. and L. Cosell, 1976. LPCW: An LPC vocoder with linear predictive warping. Proc. of ICASSP ’76: 446-469.

[22] Itahashi, S. and S. Yokoyama, 1987. A formant extraction method utilizing mel scale and equal loudness contour. Speech Transmission Lab.-Quarterly Progress and Status Report (Stockholm) (4): 17-29.

[23] Rahim, M. G. and B. H. Juang, 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. IEEE Trans. on Speech and Audio Processing, 4(1): 19-30.

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

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

[26] Matsumoto, H., *et al*., 1998. An efficient Mel-LPC analysis method for speech recognition”, Proc. ICSLP ’98: 1051-1054.

[27] Zwicker, E. and E. Terhardt, 1980. Analytical expressions for critical band rate and critical bandwidth as a function. J. Acoust. Soc. Am., 68: 1523-1525.

[28] Lindsay, P. H. and D. A. Norman, 1977. Human information processing: An introduction to psychology. 2nd Edn., Academic Press.

[29] Nakagawa, S., *et al*., ed., 2005. Spoken language systems. Ohmsha, Ltd., Japan, ch.7.

[30] Markel, J. and A. Gray, 1976. Linear prediction of speech. Springer-Verlag.

[31] Mauuary, L., 1996. Blind equalization for robust telephone based speech recognition”, Proc. EUSPICO ’96: 125-128.

[32] Mauuary, L., 1998. Blind equalization in the cepstral domain for robust telephone speech recognition. Proc. EUSPICO ’98, 1: 359-363.

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

[34] Leonard, R. G., 1984. A database for speaker independent digit recognition. ICASSP84, 3: 42.11.

[35] ITU recommendation G.712, 1996. Transmission performance characteristics of pulse code modulation channels.