

Parameter Compensation for Mel-LP based Noisy Speech Recognition

¹Md. Mahfuzur Rahman, ²Md. Robiul Hoque and ³M. Babul Islam

¹Department of Computer Science and Engineering, Comilla University, Comilla, Bangladesh

²Department of Computer Science and Engineering, Islamic University,
Kushtia 7003, Bangladesh

³Department of Applied Physics and Electronic Engineering, University of Rajshahi,
Rajshahi, 6205, Bangladesh

Abstract: This study deals with a noise robust distributed speech recognizer for real-world applications by deploying feature parameter compensation technique. 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. To minimize the mismatch between training and test phases, Cepstral Mean Normalization (CMN) and Blind Equalization (BEQ) have been applied to enhance Mel-LP cepstral coefficients as an effort to reduce the effect of additive noise and channel distortion. The performance of the proposed system has been evaluated on Aurora-2 database which is a subset of TIDigits database contaminated by additive noises and channel effects. The baseline performance, that is, for Mel-LPC the average word accuracy for test set A has found to be 59.05%. By applying the CMN and BEQ with the Mel-LP cepstral coefficients, the performance has been improved to 68.02 and 65.65%, respectively.

Key words: Aurora-2 database, BEQ, bilinear transformation, CMN, Mel-LPC

INTRODUCTION

Speech recognition systems have evolved from laboratory demonstrations to a wide variety of real-life applications, for instance, in telecommunication systems, question and answering systems, robotics, etc., Distributed Speech Recognition (DSR) system is being developed for portable terminals. These applications require such Automatic Speech Recognizers (ASRs) which can be able to maintain the performance at an acceptable level in a wide variety of environmental situations has emerged. The performance of ASRs has reached to a satisfactory level under controlled and matched training and recognition conditions. However, their performance 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 (Bateman *et al.*, 1992; Vaseghi, 1993). Consequently, a serious mismatch is occurred between training and recognition conditions, resulting in degradation in recognition accuracy.

Noise robustness can be achieved in many ways, such as, enhancement of input signal either in time domain (Islam *et al.*, 2007) or in frequency domain (Boll, 1979;

Lim, 1979; Lockwood, 1992; Agarwal and Cheng, 1999; Zhu, 2002), enhancement in cepstral domain (Atal, 1974; Furui, 1981; Mokbel *et al.*, 1984), that is, feature parameter compensation, and acoustic model compensation or adaptation (Gales, 1993a, b; Varga, 1990).

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 CMN and 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 (Davis and Mermelstein, 1980; Hermansky, 1987; Virag, 1995). In nonparametric spectral analysis, Mel-frequency Cepstral Coefficient (MFCC) (Davis and Mermelstein, 1980) is one of the most popular spectral features in ASR. This parameter takes account of the nonlinear frequency resolution like the human ear.

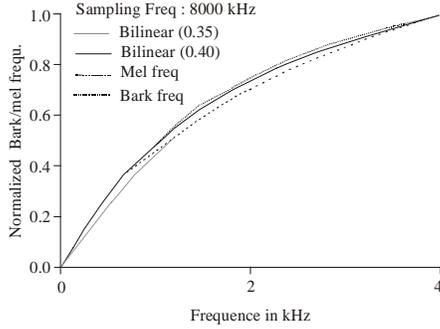


Fig. 1: The frequency mapping function by bilinear transformation

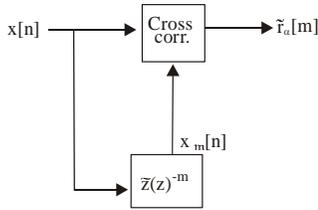


Fig. 2: Generalized autocorrelation function

In parametric spectral analysis, the Linear Prediction Coding (LPC) analysis (Itakura and Saito, 1968; Atal and Schroeder, 1968) 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 (Hermansky, 1987; Makhoul and Cosell, 1976; Itahashi and Yokoyama, 1987; Rahim and Juang, 1996).

In contrast to the different spectral modification, Strube (1980) proposed an all-pole modeling to a frequency warped signal which is mapped onto a warped frequency scale by means of the bilinear transformation (Oppenheim and Johnson, 1972), and investigated several computational procedures. However, the methods proposed by Oppenheim and Johnson (1972) 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.* (1998), 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 this study, the Mel-LP analysis along with CMN and BEQ has been used to reduce the mismatch between training and test sessions for designing a noise-robust DSR front-end.

MEL-LP ANALYSIS

The frequency-warped signal $\tilde{x}[n]$ ($n = 0, \dots, \infty$) obtained by the bilinear transformation (Oppenheim and Johnson, 1972) of a finite length windowed signal $x[n]$ ($n = 0, 1, \dots, N-1$) is defined by:

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

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

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad (2)$$

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

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

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

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 (Zwicker and Terhardt, 1980; Lindsay and Norman, 1977) at the sampling frequency of 8 kHz, respectively. Now, the all-pole model on the warped frequency scale is defined as.

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

where \tilde{a}_k is the k -th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy (Strube, 1980).

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.

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

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 (Matsumoto *et al.*, 1998, Nakagawa, 2005). First, the generalized autocorrelation coefficients are calculated as:

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

where $x_m[n]$ is the output signal of an m -th order all pass filter 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. Due to the frequency warping, $\tilde{r}_\alpha[m]$ includes the frequency weighting $\tilde{W}(e^{j\tilde{\lambda}})$ defined by:

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

which is derived from:

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

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:

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

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

The mel-cepstral coefficients can also be calculated directly from mel-prediction coefficients $\{\tilde{a}_k\}$ (Markel and Gray, 1976) using the following recursion:

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

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

MEL-LP PARAMETER COMPENSATION

Cepstral mean normalization: A robust speech recognition system must adapt with its acoustical environment or channel. To bring this concept in effect, a number of normalization methods have been developed in the cepstral domain so far. The simplest but effective cepstral normalization method is the Cepstral Mean Normalization (CMN) technique (Atal, 1974). In CMN the mean of the cepstral vectors over an utterance is subtracted from the cepstral coefficients in each frame as given below:

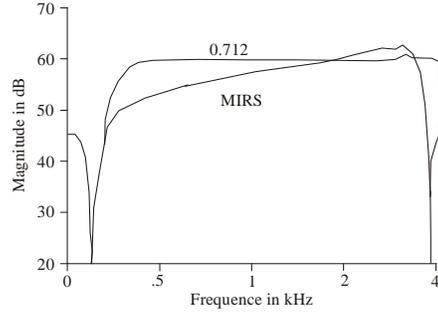


Fig. 3: Frequency response of G.712 and MIRS filters

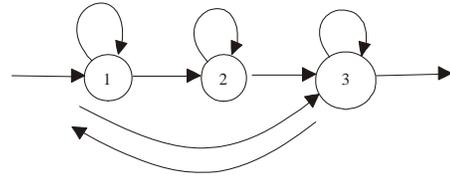


Fig. 4: Possible transition in the 3-state pause model 'sil'

Table 1: Definition of training and test data

	Filter	Data set	Noise type	SNR [dB]
Training	G.712	Clean	-	∞
Test	G.712	TestsetA	Subway, babble, car, exhibition	clean, 20, 15, 10, 5, 0, -5

$$c_m[n] = c[n] - \frac{1}{N} \sum_{n=0}^N c[n] \quad (11)$$

where $c[n]$ and $c_m[n]$ are the time-varying cepstral vectors of the utterance before and after CMN, respectively, and N is the total number of frames in the utterance.

The average of the cepstrum vectors over the speech interval represents the channel distortion, which does not use any knowledge of the environment (Furui, 1981). As the channel distortion is suppressed by CMN, it can be viewed as parameter filtering operation. Consequently, CMN has been treated as high-pass and band-pass filters (Mokbel *et al.*, 1984). The effectiveness of CMN for the combined effect of additive noise and channel distortion is limited. Acero and Stern, (1990) have developed more complex cepstral normalization techniques to compensate the joint effect of additive noise and channel distortion.

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 (Mauuary, 1996, 1998). 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

Table 2: Word accuracy [%] for MLPC without CMN and 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	4.444	8.06
Car	98.54	95.26	83.03	54.25	24.04	12.23	8.775	3.77
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 MLPC with CMN

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
Subway	99.02	96.41	92.05	78.66	50.23	24.78	16.09	68.43
Babble	98.82	97.37	93.80	82.22	55.32	25.76	13.30	70.90
Car	98.87	96.96	92.72	77.42	42.77	22.55	13.18	66.49
Exhibition	99.07	96.08	91.67	76.70	45.97	20.98	11.60	66.28
Average	98.95	96.71	92.56	78.75	48.58	23.52	13.55	68.03

Table 4: Word accuracy [%] for MLPC 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

square error computed as a difference between the current and reference cepstrum.

In this study, the same algorithm is used as that implemented in Islam *et al.* (2007) with same values of different parameters.

EVALUATION ON AURORA-2 DATABASE

Experimental setup: The proposed system was evaluated on Aurora-2 database (Hirsch and Pearce, 2000), which is a subset of TIDigits database (Leonard, 1984) 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, (1996). Their frequency characteristics have been shown in Fig. 3.

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.

The recognition experiments were conducted with a 12th order prediction model of Mel-LPC analysis. The preemphasized speech signal with a preemphasis 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 illustrates in Fig. 4. 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.

The recognition accuracy (*Acc*) is evaluated as follows:

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

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

Recognition results: The detail recognition results are presented in this section. The word accuracy for Mel-LPC without applying CMN and BEQ is listed in Table 2 which is considered as baseline result. The average word accuracy over all noises within the set A and over SNRs 20 to 0 dB is found to be 59.05% for the baseline.

The word accuracy with CMN and BEQ are given in Table 3 and 4, respectively. The average performance of Mel-LPC with CMN is found to be 68.03%, whereas, 65.66% accuracy is obtained for BEQ. It is also observed that the larger improvements are achieved for babble and car noises both for CMN and BEQ as compared to baseline performance. The average recognition accuracy does not differ significantly for subway and exhibition noise conditions.

CONCLUSION

An HMM-based Automatic Speech Recognizer (ASR) was developed and the performance of Mel-LPC with CMN and BEQ was evaluated on test set A of Aurora-2 database. It is observed that the performance of CMN is slightly better than that of BEQ on the average. It has also been found that for noise type babble CMN is more effective than BEQ. On the other hand, BEQ outperforms CMN for noise condition car. The average word accuracy does not differ significantly for noises subway and exhibition after applying CMN and BEQ. The overall recognition performance has been improved from 59.05% to 68.03% and 65.66% for CMN and BEQ, respectively.

REFERENCES

Acero, A. and R. Stern, 1990. Environmental robustness in automatic speech recognition. Proc. ICASSP., 90: 849-852.

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

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

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.

Bateman, D.C., D.K. Bye, M.J. Hunt, 1992. Spectral contrast normalization and other techniques for speech recognition in noise. Proc. ICASSP., 92(1): 241-244.

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

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

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics Speech Signal Proces., ASSP, 29: 254-272.

Gales, M.J.F. and S.J. Young, 1993a. HMM recognition in noise using parallel model combination. Proc. Euro. Speech, 93(2): 837-840.

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

Hermansky, H., 1987. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87(4): 17-29.

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., 181: 188.

Islam, M.B., K. Yamamoto and H. Matsumoto, 2007. Mel-Wiener filter for Mel-LPC based speech recognition. IEICE Trans. Inform. Sys., E90-D (6): 935-942.

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.

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.

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

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

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

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

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.

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

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

Matsumoto, H., T. Nakatoh and Y. Furuhashi, 1998. An efficient Mel-LPC analysis method for speech recognition, Proc. ICSLP., 98: 1051-1054.

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

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

Mokbel, C., D. Jouviet, J. Monne and R. De Mori, 1984. Compensation of telephone line effects for robust speech recognition. Proc. ICSLP., 94: 987-990.

- Nakagawa, S., M. Okada and T. Kawahara, 2005. Spoken Language Systems. Ohmsha Ltd., Japan, Ch.7.
- Oppenheim, A.V. and D.H. Johnson, 1972. Discrete representation of signals. *IEEE Proc.*, 60(6): 681-691.
- Rahim, M.G. and B.H. Juang, 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Trans. Speech Audio Proces.*, 4(1): 19-30.
- Strube, H.W., 1980. Linear prediction on a warped frequency scale. *J. Acoust. Soc. Am.*, 68(4): 1071-1076.
- Varga, A.P. and R.K. Moore, 1990. Hidden Markov model decomposition of speech and noise. *Proc. ICASSP.*, 90(2): 845-848.
- Vaseghi, S.V. and B.P. Milner, 1993. Noise-adaptive hidden Markov models based on Wiener filters. *Proc. Euro. Speech*, 93(2): 1023-1026.
- Virag, N., 1995. Speech enhancement based on masking properties of the auditory system. *Proc. ICASSP*, 95: 796-799.
- Zhu, Q. and A. Alwan, 2002. The effect of additive noise on speech amplitude spectra: A Quantitative analysis. *IEEE Signal Proces. Lett.*, 9(9): 275-277.
- 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.