Comparative study between Mel-LP and LP-Mel based front-ends for noisy speech recognition using HMM

[1]M. Shohidul Islam, [2]M. Babul Islam, [1]M. Mojahidul Islam, [1]M. Shamim Hossain, [1]M. Muntasir Rahman
[1]Dept. of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh.
[2]Dept. of Applied Physics and Electronic Engineering, University of Rajshahi, Rajshahi, Bangladesh.
shohid7@gmail.com, babul.apee@ru.ac.bd, m_mujahidul@yahoo.com, shamimmalitha@yahoo.com, muntasir_rahman6@yahoo.com

***Abstract*—Since the parameterization in the perceptually relevant aspects of short-term speech spectra in ASR front-end is advantageous for speech recognition, such as Mel-LPC, LPC-Mel, MFCC etc., in this paper, Mel-LP and LP-Mel based front-ends have been designed for automatic speech recognition (ASR). The speech classifier of the developed ASR is based on Hidden Markov Model (HMM) as it can successfully cope with acoustic variation and lack of word boundaries of speech signal. The performance of the developed system has been evaluated on test set A of Aurora-2 database both for Mel-LP and LP-Mel based front-ends. It has been found that the Mel-LP based front-end is more effective for noise type subway, babble and exhibition; on the other hand, LP-Mel based front-end is suitable for car noise. The average word accuracy for Mel-LPC has been found to be 59.05%, while for LPC-Mel, it has been 54.45%.**

Keywords*—*Mel-LP, LP-Mel, HMM, Bilinear transformation, Noisy speech recognition.

I. INTRODUCTION

Research in speech recognition has produced numerous algorithms and commercially available speech recognizers that all work to some extent. Among these, statistical approach, in particular, the Hidden Markov Model (HMM) is the most prevailing approach that has proved its practical and theoretical soundness. In speech recognition, there are two main problems – one is acoustic variation due to speaker variability, mood, environment, especially additive noise and the other one is lack of word boundaries. The most successful solution is to use a stochastic model of speech, in particular the HMM, since it can cope with the above problems [1].

Speech recognition systems include an initial processing stage that converts speech signals into sequences of observation vectors, which represent the short-term spectrum of the speech signal useful for further processing. Most of these front-ends are based on standard processing techniques such as filter-bank or linear prediction (LP).

Designing a front-end incorporating auditory-like frequency resolution improves recognition accuracy [2, 3, 4]. Therefore, we need to parameterize the perceptually relevant aspects of short-term speech spectra and their dynamics in ASR front-end, in order to enhance the performance of Automatic Speech Recognition (ASR).

In nonparametric spectral analysis, Mel-frequency Cepstral Coefficient (MFCC) [2] is one of the most popular spectral features in ASR. This parameter takes account of the nonlinear frequency resolution like the human ear.

In parametric spectral analysis, the linear prediction coding (LPC) analysis [5, 6] 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 [3, 7, 8, 9].

In contrast to the different spectral modification, Strube [10] proposed an all-pole modeling to a frequency warped signal which is mapped onto a warped frequency scale by means of the bilinear transformation [11], and investigate several computational procedures. However, the methods proposed in [11] to estimate warped all-pole model have been rarely used in automatic speech recognition. Recently, as an LP-based method, a simple and efficient time-domain technique to estimate all-pole model on the mel-frequency scale is proposed in [12], which is referred to as a "Mel-LPC" analysis. The prediction coefficients are estimated without any approximation by minimizing the prediction error power at a two-fold computational cost over the standard LPC analysis.

In this paper an HMM based automatic speech recognition (ASR) system is developed. As front-end features both the Mel-LP and LP-Mel cepstral coefficients, that is, Mel-LPC and LPC-Mel are used and the effectiveness of these features on noise category is evaluated for HMM based noisy speech recognition.

The rest of the paper is organized as follows. The Mel-LP and LP-Mel analyses are introduced in Section II and III, successively. Section IV deals with experimental setup and recognition results. Finally, conclusion is presented in Section V.

II. MEL-LP ANALYSIS

The frequency-warped signal $\tilde{x}[n]$ $(n = 0, \ldots, \infty)$ obtained by the bilinear transformation [11] of a finite length windowed signal $x[n]$ $(n = 0, 1, \ldots, N-1)$ is defined by

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

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 [13, 14] at the sampling frequency of 8 kHz, respectively.

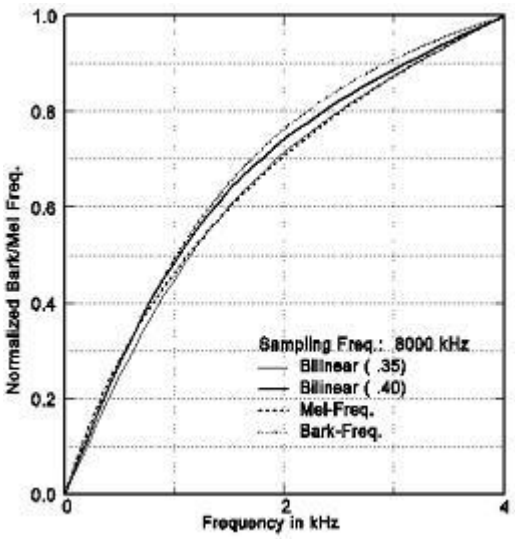

Figure 1. The frequency mapping function by bilinear transformation.

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

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

where $\tilde{a}_k$ is the *k*-th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy [10].

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 [12, 15]. 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. 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 $\left\{\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})\right\}^{-1}$.

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

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

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

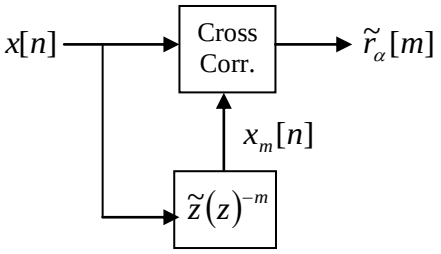

Figure 2. Generalized autocorrelation function.

The mel-cepstral coefficients can also be calculated directly from mel-prediction coefficients $\{\tilde{a}_k\}$ [16] 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.

III. LP-MEL ANALYSIS

In linear prediction analysis, the vocal tract transfer function is modeled by an all-pole filter given by

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{11}$$

where a_k is the *k*-th mel-prediction coefficient.

On the basis of minimum mean square prediction error for a finite length windowed signal $x[n]\,(n = 0, 1, \ldots, N-1)$, $\{a_k\}$ are obtained by Durbin's algorithm from the autocorrelation coefficients $r[m]$ of $x[n]$ defined by

$$r[m] = \sum_{n=0}^{N-1-m} x[n]x[n+m] \tag{12}$$

Finally, the Mel-prediction coefficients have been obtained by Openheim and Jhonson recursion [10]. From the Mel-prediction coefficients, the LP-Mel cepstral coefficients are obtained using (10).

IV. EVALUATION ON AURORA-2 DATABASE

A. Experimental Setup

The proposed system was evaluated on Aurora-2 database [17], which is a subset of TIDigits database 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.

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 I.

TABLE I. DEFINITION OF TRAINING AND TEST DATA.

	Data set	**Noise Type**	**SNR [dB]**
Training	Clean	–	∞
Test	Test set A	Subway, Babble, Car, Exhibition	clean, 20, 15, 10, 5, 0, -5

The reference recognizer was based on HTK (Hidden Markov Model Toolkit). 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. 2. 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.

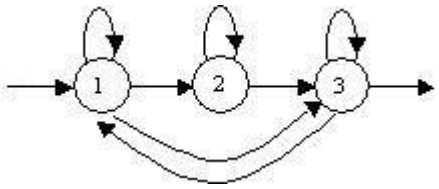

Figure 3. Possible transition in the 3-state pause model 'sil'.

The recognition experiments were conducted with a 12th order prediction model both for Mel-LP and LP-Mel analyses. 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 both for Mel-LP and LP-Mel based front-ends.

B. Recognition Results

The detail recognition results have been presented in this section both for Mel-LP and LP-Mel based front-ends. The recognition accuracy for Mel-LPC and LPC-Mel are listed in Table II and Table III, successively. The average recognition accuracy for Mel-LPC and LPC-Mel are found to be 59.05% and 54.45%, respectively.

From Table II, we have found that the average word accuracy obtained for Mel-LPC are 68.30%, 48.06% and 66.05% for noise type subway, babble and exhibition, consecutively. On the other hand, in the case of LPC-Mel front-end the average recognition accuracy for noise category subway, babble and exhibition are found to be 63.93%, 44.11% and 55.56%, respectively which are presented in Table III. The comparative word accuracy between Mel-LPC and LPC-Mel is also presented graphically in Fig. 4 for different noise groups. It has been shown that in the case of car noise, LPC-Mel gives slightly better accuracy than that of Mel-LPC which is also presented graphically in Fig. 5.

V. CONCLUSION

An HMM based automatic speech recognition (ASR) system has been developed and a comparative study has been made between Mel-LP and LP-Mel based front-ends. It has been found that the Mel-LPC outperforms the LPC-Mel for noise category subway, babble and exhibition. On the contrary, the LPC-Mel gives slightly better word accuracy than that of the Mel-LPC. On the average, the word accuracy for the Mel-LPC is found to be 59.05% while the accuracy for the LPC-Mel is found to be 54.45%.

From the above discussion we can conclude that the preprocessing of auditory like frequency resolution analysis is

more effective than that of postprocessing for designing front-end.

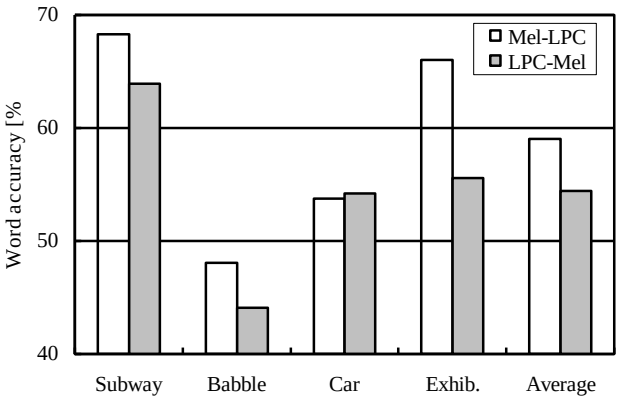

Figure 4. Comparative average word accuracy between Mel-LPC and LPC-Mel for different noise category.

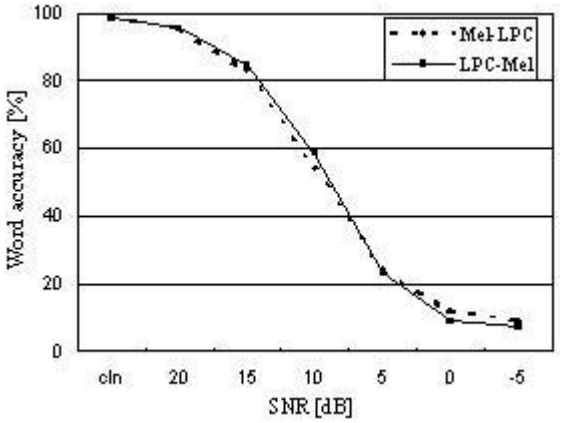

Figure 5. Comparative word accuracy between Mel-LPC and LPC-Mel for car noise.

TABLE II. WORD ACCURACY (%) FOR MEL-LP FRONT-END (MEL-LPC).

Noise	SNR [dB]							**Average**
	Clean	20	15	10	5	0	-5	**(20 to 0 dB)**
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 III. WORD ACCURACY (%) FOR LP-MEL FRONT-END (LPC-MEL).

Noise	SNR [dB]							**Average**
	Clean	20	15	10	5	0	-5	**(20 to 0 dB)**
Subway	98.83	96.32	91.19	72.09	40.65	19.40	9.43	**63.93**
Babble	98.91	88.33	70.86	43.20	17.14	1.03	-0.91	**44.11**
Car	98.69	95.47	84.46	58.28	23.53	9.25	7.43	**54.20**
Exhibition	98.73	94.17	84.76	58.99	28.42	11.48	7.81	**55.56**
Average	**98.79**	**93.57**	**82.82**	**58.14**	**27.44**	**10.29**	**5.94**	**54.45**

REFERENCES

[1] M. Babul Islam, "Wiener filter for Mel-scaled LP based noisy speech recognition," Doctoral thesis, Shinshu University, Japan, 2007.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, No. 4, pp. 357-366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 17-29, 1987.

[4] N. Virag, "Speech enhancement based on masking properties of the auditory system", Proc. ICASSP'95, pp.796-799, 1995.

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

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

[7] Makhoul and L. Cosell, "LPCW: An LPC vocoder with linear predictive warping", Proc. of ICASSP '76, pp. 446-469, 1976.

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

[9] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition ", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, pp. 19-30, 1996.

[10] H. W. Strube, "Linear prediction on a warped frequency scalle", J. Acoust. Soc. Am., vol. 68, no. 4, pp. 1071-1076, 1980.

[11] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," IEEE Proc., vol. 60, no. 6, pp. 681-691, 1972.

[12] H. Matsumoto, Y. Nakatoh and Y. Furuhata, "An efficient Mel-LPC analysis method for speech recognition", Proc. ICSLP '98, pp. 1051-1054, 1998.

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

[14] P. H. Lindsay and D. A. Norman, “Human information processing: An introduction to psychology”, 2nd Ed., pp. 163, Academic Press, 1977.

[15] S. Nakagawa, et al., ed., “Spoken language systems,”Ohmsha, Ltd., Japan, ch.7, 2005.

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

[17] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” ISCA ITRW ASR 2000, September 2000.