www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 3 March 2013 Page No. 821-824

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

[1]Md. Jashim Uddin*, [2]S.M.A. Rahim, [3]Md. Abdul Al Mohit, [4]M. Shohidul Islam

[1]Dept. of Information & Communication Engineering, Islamic University, Kushtia, Bangladesh.

[2]Dept. of Applied Physics Electronics & Communication Engineering, Islamic University Kushtia, Bangladesh.

[3]Dept. of Mathematics, Islamic University, Kushtia, Bangladesh.

[4]Dept. of Computer Science & Engineering, Islamic University, Kushtia, Bangladesh.

jashim.iu@gmail.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, MFCC 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 MFCC and LP-Mel based front-ends. It has been found that the MFCC based front-end is more effective for noise type subway, babble, car and exhibition. The average word accuracy for MFCC has been found to be 59.21%, while for LPC-Mel, it has been 54.45%.*

Keywords- MFCC, 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 MFCC and LP-Mel cepstral coefficients, that is, MFCC 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 MFCC 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. MFCC ANALYSIS

The signal processing front end summarizes the spectral characteristics of the speech waveform into a sequence of acoustic vectors that are suitable for processing by the acoustic model. In filter-bank based systems, MFCC [13] is widely used spectral features. This parameter takes account of the nonlinear frequency resolution as like the human ear.

Mel Filter Bank: The mel-scale filter-bank is illustrated in Figure-1. The Mel frequency scale is linear up to 1000 Hz and logarithmic thereafter. As can be seen, a set of overlapping Mel filters are made such that their center frequencies are equidistant on the Mel scale which is defined by (II.I). Usually, the triangular filters are spread over the whole frequency range from zero up to the Nyquist frequency [14].

$$Mel = 2595\log_{10}(1 + f/700) \qquad \text{(II.I)}$$

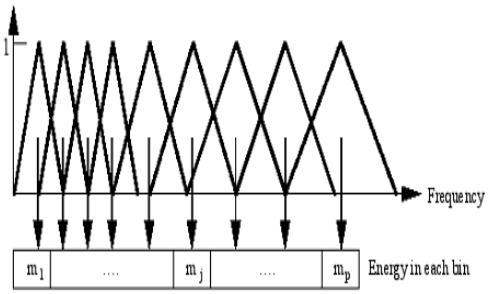


Figure- 1:Mel-scale filter bank

Figure-2 shows the stages of this transformation. To implement this filter-bank, first, Fourier transform is applied to the preemphasized and windowed speech signal and the magnitude is calculated. Each FFT magnitude coefficient is then multiplied by the corresponding filter gain and the results are accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filter-bank channel.

As an alternative, the power can be used rather than the magnitude of FFT in the binning process. The transformation is:

$$Y[n] = \sum_{i=0}^{N/2} X[i] \times MelWeight[n][i], \quad 0 < n < Number\ of\ filters \qquad \text{(II.II)}$$

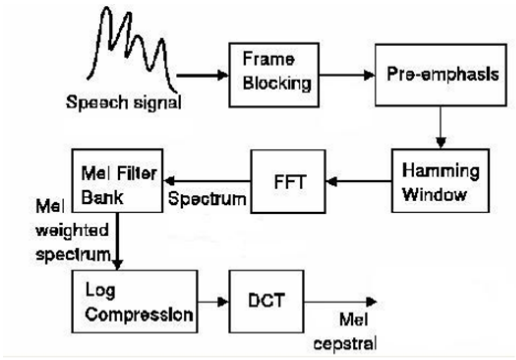


Figure- 2: MFCC feature extraction technique.

Log Compression: The range of the values generated by the Mel filter bank is reduced by replacing each value by its natural logarithm. This is done to make the statistical distribution of the spectrum approximately Gaussian - a requirement for the subsequent acoustic model. The transformation is[15]:

$$m_n = \ln(Y[n]), \qquad 0 < n \le \text{Number. of filters} \qquad \text{(II.III)}$$

DCT: The discrete cosine transform is used to compress the spectral information into a set of low order coefficients. This representation is called the Mel-cepstrum which is calculated as follows [14]:

$$C_i = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} m_n \cos\left(\frac{\pi}{N}(n + 0.5)\right) \qquad \text{(II.IV)}$$

where, N is the number of filter-bank channels.

III. LP-MEL ANALYSIS

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

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad \text{(III.I)}$$

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

$$r[m] = \sum_{n=0}^{N-1-m} x[n]x[n+m] \qquad \text{(III.II)}$$

Finally, the LP-Mel cepstral coefficients are obtained using (III.III).

$$c_k = -\tilde{a}_k - \frac{1}{k}\sum_{j=1}^{k-1}(k-j)\tilde{a}_k c_{k-j} \qquad \text{(III.III)}$$

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 Figure-3. 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.

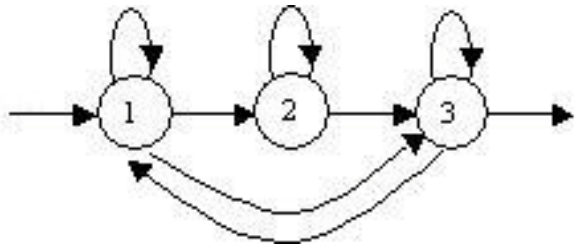


Figure 3. Possible transition in the 3-state pause model 'sil'.

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

B. Recognition Results

The detail recognition results have been presented in this section both for MFCC and LP-Mel based front-ends. The recognition accuracy for MFCC and LPC-Mel are listed in Table II and Table III, successively. The average recognition accuracy for MFCC and LPC-Mel are found to be 59.21% and 54.45%, respectively.

From Table II, we have found that the average word accuracy obtained for MFCC are 64.28%, 51.87% , 56.59% and 64.09% for noise type subway, babble, car and exhibition, consecutively. On the other hand, in the case of LPC-Mel front-end the average recognition accuracy for noise category subway, babble, car and exhibition are found to be 63.93%, 44.11%, 54.20% and 55.56%, respectively which are presented in Table III. The comparative word accuracy between MFCC and LPC-Mel is also presented graphically in Figure-4 for different noise groups.

V. Conclusion

An HMM based automatic speech recognition (ASR) system has been developed and a comparative study has been made between MFCC and LP-Mel based front-ends. It has been found that the MFCC outperforms the LPC-Mel for noise category subway, babble, car and exhibition. On the average, the word accuracy for the MFCC is found to be 59.21% while the accuracy for the LPC-Mel is found to be 54.45%.

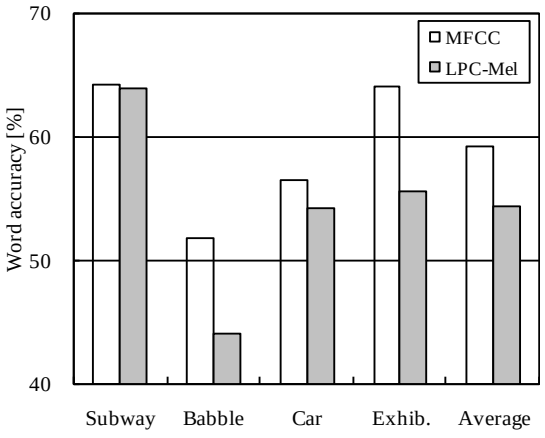
From the above discussion we can conclude that the preprocessing of auditory like frequency resolution analysis is more effective than that of postprocessing for designing front-end.

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
Subway	98.83	95.61	90.08	72.74	44.03	18.94	9.7	64.28
Babble	98.91	91.99	76.42	52.00	26.39	12.55	8.77	51.87
Car	98.78	95.71	85.12	61.59	30.51	9.99	7.01	56.59
Exhibition	98.95	95.87	90.31	74.54	42.61	17.09	8.73	64.09
Average	98.87	94.8	85.49	65.22	35.89	14.65	8.56	59.21

TABLE II. WORD ACCURACY (%) FOR MFCC FRONT-END (MFCC).

Noise	SNR [dB]							Average (20 to 0 dB)
	Clean	20	15	10	5	0	-5	
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

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



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

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] L. Rabiner and B. H. Juang, Fundamentals of speech recognition. Englewood Cliffs, NJ, 1993.

[14] Corpus-Based Methods in Language and Speech Processing (Steve Young, Cambridge University, Engineering, Department, Cambridge U.K.Gerrit Bloothooft, Research Institute for Language and Speech, Utrecht University, Utrecht Netherlands) Publisher: Kluwer Academic; Dordrecht/Boston/London)

[15] http://www.silliconintelligence.com/people/binuerciption/node3.

[16] 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.