MEL-LP BASED GENERALIZED CEPSTRAL ANALYSIS FOR NOISY SPEECH RECOGNITION USING HMM

Md. Rashedul Islam[1], Dr. Md. Babul Islam[2], Md. Firoz Ahmed[3], Md. Najmul Hossain[4], Md. Abdur Rahim[5]

[1,4]Department of Electronic and Telecommunication Engineering,
Pabna University of Science and Technology, Pabna-6600, Bangladesh
rashed.ru11@gmail.com, rony.ru85@gmail.com

[2]Department of Applied Physics and Electronic Engineering,
Rajshahi University, Rajshahi, 6205, Bangladesh.

[3]Department of Information and Communication Engineering,
Rajshahi University, Rajshahi, 6205, Bangladesh.

[5]Department of Computer Science and Engineering,
Pabna University of Science and Technology, Pabna-6600, Bangladesh
rahim_bds@yahoo.com

Abstract: *This paper deals with LP based Mel-Generalized cepstrum which has been used as front-end for Hidden Markov Model (HMM) based speech recognition and it incorporates equal-loudness power law as well as auditory-like frequency resolution. To utilize the generalized cepstral representation, the model spectrum can be varied continuously from the all-pole spectrum to that represented by the cepstrum according to the value of γ. The performance of Mel-LP based generalized cepstral analysis has been evaluated on Aurora-2 database for HMM based speech recognition. The word accuracy for Mel-Generalized cepstral analysis is found to be 63.63% for test set A. On the contrary, the conventional Mel-LPC gives 59.05% word accuracy.*

Keywords: Aurora-2 database, Mel-Generalized cepstrum, Bilinear transformation, Mel-LPC

1. Introduction

Linear prediction [1], [2] is a widely accepted method for obtaining all-pole representation of speech. However, in some cases, for instance, nasal sounds spectral zeros are important and a more general modeling procedure is required. However, cepstral modeling based on linear prediction can represent poles and zeros with equal weights; the cepstral method [3] with a small number of cepstral coefficients overestimates the bandwidths of the formants. To overcome this problem, the generalized cepstral analysis method [4], [5] can be used. The generalized cepstral coefficients [6] are identical with the cepstral and AR coefficients when a parameter γ equals 0 and -1, respectively. Thus, utilizing the generalized cepstral representation, the model spectrum can be varied continuously from the all-pole spectrum to that represented by the cepstrum according to the value of γ.

Since the human ear has high resolution at low frequencies, introducing the similar characteristics to the model spectrum will be more effective for encoding speech signal. Therefore, mel-generalized spectral model is one of the appropriate methods to estimate the auditory-like feature parameters as it includes the intensity-loudness power law by the generalized logarithmic function [7] as well as auditory frequency resolution.

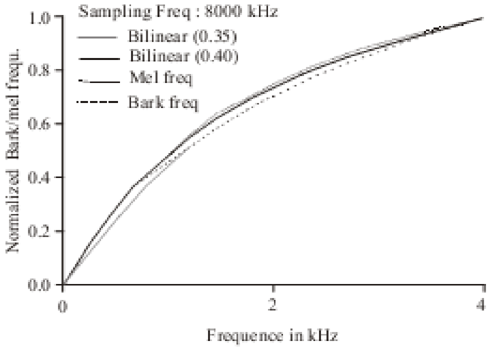


Fig.1: The frequency mappings function by bilinear transformation.

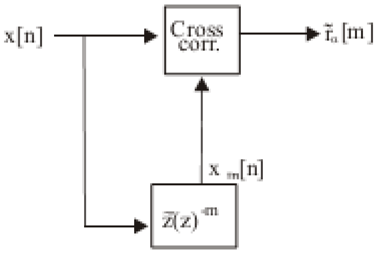


Fig.2: Generalized autocorrelation function

2. Mel-LP Analysis

The frequency warped signal $\tilde{x}[n]$ (n=0… α) obtained by the bilinear transformation [8] of a finite length windowed signal $x[n]$ (n = 0, 1, 2… N –1) is defined by:

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

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

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

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

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

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

This phase function determines a frequency mapping. In the frequency domain, the spectrum $X(e^{j\lambda})$ on the linear frequency axis λ is converted to the frequency warped spectrum $\tilde{X}(e^{j\tilde{\lambda}})$ on the mel-frequency axis $\tilde{\lambda}$ by the above frequency mapping function given by Eq. 2.3. Fig. 2.1 shows the approximated frequency mapping functions of the bark and the Mel scales (solid lines) at the sampling frequency of 8 kHz. This figure also shows the following analytical expressions of the "Mel" and Bark scales (dotted lines) based on psychoacoustic works [9], [10]:

$$Bark = 13 \tan^{-1}(0.00076\ f) + 3.5 \tan^{-1}(f / 7500) \quad \text{…………(4)}$$

and

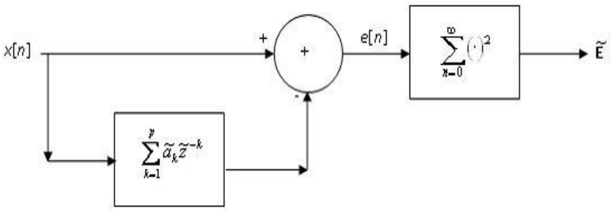
$$Mel = 2595 \log_{10}(1 + f / 700) \quad \text{…………(5)}$$

where f is the frequency in Hz. As shown in Fig 1, α = 0.35, α = 0.40 can approximate the mel-scale and bark-scale at the sampling frequency of 8 kHz, respectively.

In Mel-LP analysis, the spectral envelope of $\tilde{X}(\tilde{z})\tilde{W}(\tilde{z})$ is approximated by the following all-pole model on the linear frequency domain,

$$\tilde{H}_a(\tilde{z}) = \frac{\tilde{\sigma}_e}{1 + \sum_{k=1}^{p} \tilde{a}_k \tilde{z}^{-k}} \quad \text{…………(6)}$$

where $\tilde{a}_k$ is the k-th mel-prediction coefficient and $\tilde{\sigma}_e^{\ 2}$ is the residual energy [11].



x[n]
+
+
-
e[n]
$\sum_{n=0}^{\infty}(\cdot)^2$
$\tilde{E}$
$\sum_{k=1}^{p}\tilde{a}_k\tilde{z}^{-k}$

Fig.3: Mel-LP analysis on the linear frequency scale.

The model $\tilde{\mathrm{H}}_a(\tilde{z})$ is estimated on the basis of minimization of mean square error (MMSE) as shown in Fig. 3. Since $\tilde{x}[n]$ is an infinite sequence, the prediction error signal is also an infinite sequence. Thus, the total error energy $\tilde{E}$ over an infinite sequence is given by:

$$\tilde{E} = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{p} \tilde{a}_k x_k[n] \right)^2 \quad \text{…………(7)}$$

Where $x_k[n]$ is the output signal of a k-th order all-pass filter $\tilde{z}(z)^{-k}$ excited by $x_0[n] = x[n]$. As a result of minimizing $\tilde{E}$, the mel-prediction coefficients $\{\tilde{a}_k\}$ are obtained by solving for the following normal equations:

$$\sum_{k=1}^{p} \tilde{\phi}(m,k)\tilde{a}_k = -\tilde{\phi}(0,m), \quad (m = 1,..., p) \quad \text{…………(8)}$$

$$\text{where} \quad \tilde{\phi}(m,k) = \sum_{n=0}^{\infty} x_m[n]x_k[n] \quad \text{…………(9)}$$

In the warped frequency domain, Eq.9 can be rewritten as:

$$\tilde{\phi}(m,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} | \tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}}) |^2 . e^{j(m-k)\tilde{\lambda}} d\lambda \quad \text{…………(10)}$$

where the frequency weighting function $\tilde{W}(e^{j\tilde{\lambda}})$ is defined by:

$$\tilde{W}(\tilde{z}) = \frac{\sqrt{1-\alpha^2}}{1 + \alpha\tilde{z}^{-1}} \quad \text{…………(11)}$$

which is derived from

$$\frac{d\lambda}{d\tilde{\lambda}} = | \tilde{W}(e^{j\tilde{\lambda}}) |^2 \quad \text{…………(12)}$$

Eq.8 indicates that $\tilde{\phi}(m,k)$ reduces to the autocorrelation function of the signal whose Fourier transform is equal to the frequency warped and frequency weighted spectrum $\tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})$ .This autocorrelation function is called as "generalized autocorrelation function". Fig. 2 illustrates the calculation procedure of generalized autocorrelation function. From Eq.8, it should be noted that $\tilde{\phi}(m,k)$ is a function of the difference $(k-m)$. Thus, $\tilde{\phi}(m,k)$ can be calculated from the sum of finite terms without any approximation,

$$\tilde{\phi}(m,k) = \tilde{r}[k-m] = \sum_{n=0}^{N-1} x[n].x_{|k-m\ |}[n] \text{…………(13)}$$

Therefore, to solve for $\tilde{a}_k$ and $\tilde{\sigma}_e$, the generalized autocorrelation coefficients of the input signal $x[n]$ is required instead of autocorrelation coefficients in the traditional LP analysis [12], [13]. Since the mel-prediction coefficients $\{\tilde{a}_k\}$ are obtained from the generalized autocorrelation function of the input signal $x[n]$, the proposed system enhances the speech signal in the generalized autocorrelation domain. Although the estimated model given by Eq. 6 includes the frequency weighting $\tilde{W}(e^{j\tilde{\lambda}})$, this can be easily removed by inverse filtering in the generalized autocorrelation domain using $\{\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})\}^{-1}$, which leads to the mel-autocorrelation function $\tilde{r}_{\alpha}[m]$:

$$\tilde{r}_{\alpha}[m] = \beta_0 \tilde{r}[m] + \beta_1 \{\tilde{r}[m-1] + \tilde{r}[m+1]\} \quad \text{…….(14)}$$

$$\text{and} \quad \beta_1 = \alpha(1-\alpha^2)^{-1/2} \quad \text{…………(15)}$$

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

$$\log \tilde{\mathrm{H}}_a(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{c}_k \tilde{z}^{-n} \quad \text{…………(16)}$$

where $\{\tilde{c}_k\}$ are the mel-cepstral coefficients.

The mel-cepstral coefficients can also be calculated directly from mel-prediction coefficients $\{\tilde{a}_k\}$[14] as shown in Fig.4, using the following recursion:

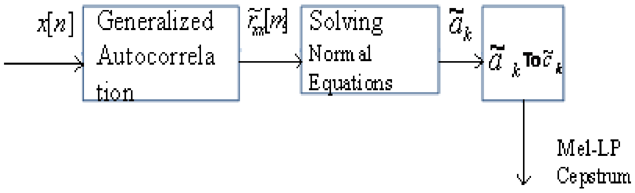


Fig.4: Block diagram for the calculation of Mel-LP cepstrum

$$\tilde{c}_k = -\tilde{a}_k - \frac{1}{k}\sum_{j=1}^{k-1}(k-j)\tilde{a}_k \tilde{c}_{k-j} \quad \text{…………(18)}$$

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

3. Generalized LP Cepstrum Analysis

In auditory perception, intensity-loudness compression is not logarithmic, but is rather cubic-root characteristic as used in the PLP analysis. In order to incorporate this auditory characteristic into the spectral representation, the following generalized logarithmic function [6] has been introduced instead of the logarithmic function:

$$s_{\gamma}(w) = \begin{cases} (w^{\gamma}-1)/\gamma, & 0 < |\gamma| \le 1 \\ \ln w, & \gamma = 0 \end{cases} \quad \text{…………(19)}$$

where $s_{\gamma}(w)$ approaches the logarithmic function as $\gamma \to 0$.

The generalized cepstrum [6] for a minimum phase sequence is defined by substituting this function to the logarithmic function in the definition of the cepstrum. The generalized cepstrum for $H(z)$ is computed by the recursion similar to the conventional cepstrum.

.

4. Mel-Generalized Cepstral Analysis

Although all-pole modeling is effective and simple to estimate, it is not appropriate to represent the spectra with zeroes as in the case of nasal sounds. In order to take both the pole and zero into account and also to incorporate the auditory characteristics of frequency resolution and the intensity-loudness power law, the mel-generalized cepstral analysis has been shown to be effective [7]. In this analysis, using the mel-generalized cepstral coefficients, a spectrum is modeled by

$$H(z) = s_{\gamma}^{-1}\left(\sum_{m=0}^{p}\tilde{c}_{\gamma}(m)\tilde{z}^{-m}\right)$$

$$= \begin{cases} \left(1+\gamma\sum_{m=0}^{p}\tilde{c}_{\gamma}(m)\tilde{z}^{-m}\right)^{1/\gamma}, & 0 < |\gamma| \le 1 \\ \exp\sum_{m=0}^{p}\tilde{c}_{\gamma}(m)\tilde{z}^{-m}, & \gamma = 0 \end{cases}$$

…………(20)

This spectral representation includes various types of models depending on the value of α and γ. The models when γ equals -1 and 1 are identical to the all-pole and all-zero models, respectively, and γ = 0 leads the conventional cepstral model. The parameter α controls the degree of frequency warping by Eq. 3.

The model parameters are estimated so as to minimize a criterion used in the unbiased estimation of the log spectrum [15]:

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi}\{\exp R(\lambda) - R(\lambda) - 1\}d\lambda, \quad \text{…………(21)}$$

where

$$R(\lambda) = \ln\left(\left|X\left(e^{j\lambda}\right)\right|^2\right) - \ln\left(\left|H\left(e^{j\lambda}\right)\right|^2\right).$$

….(22)

For $\gamma = -1$ this minimization is solved directly by a set of linear equations, while for $-1 < \gamma < 0$ the optimal estimates are obtained by an iterative algorithm whose convergence is quadratic.

5. Evalution on Aurora-2 Database

5.1 Experimental Setup

Place The proposed system was evaluated on Aurora 2 database [16] 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.

To consider realistic the 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 the ITU (1996) [17].Their frequency responses have been shown in Fig. 5

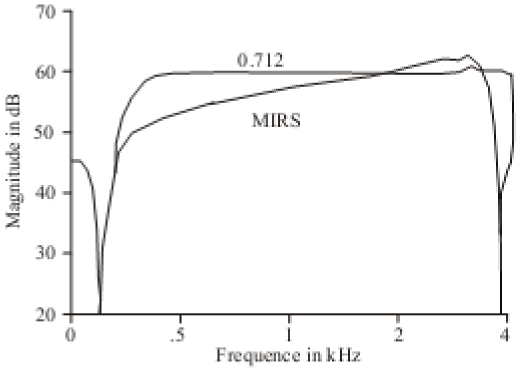


Fig.5: Frequency responses of G712 and MIRS filters

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.

Training Model	Filter	Noise	SNR [dB]
Clean	G.712	**…**	∞
Multi	G.712	Subway, car ,babble, exhibition	20, 15, 10, 5, 0, -5 and clean

Table-1**:** Definition of training data.

The recognition experiments were conducted with a 12th order Mel-LP 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. For mel-generalized cepstral analysis the value of γ was set to 0.3 since it gives better performance.

As front-end, 14 cepstral coefficients and their delta coefficients including 0th terms were used. Thus, each feature vector size is 28. The acceleration coefficients were not used in the proposed system, because it was found that their incorporation could not improve the word accuracy, especially in low SNR conditions [18].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 with16 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. 6. 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:

$$Acc = \frac{N-D-S-I}{N} \times 100\% \quad \text{…………(23)}$$

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

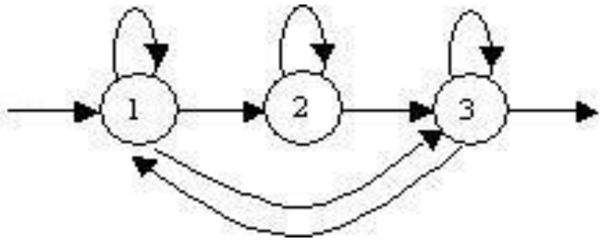


Fig. 6: Possible transition in the 3-state pause model 'sil'.

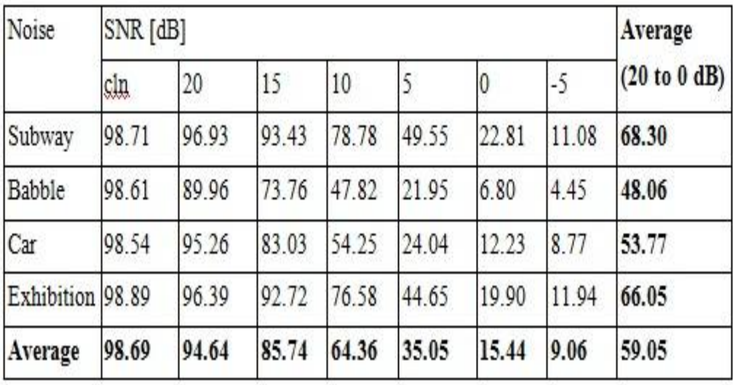
6. Experimental Results

The detail recognition results are presented in this section**.** The detail recognition results using the Mel-LPC feature parameter with the value of warping factor of 0.35 is tabulated in Table 2. For the Mel-Generalized cepstral parameter with the same value of warping factor and γ = 0.30 is listed in Table 3.

Form these tables it is observed that for Mel-LPC, the average recognition accuracy for noise categories subway, babble, car and exhibition are found to be 68.30%, 48.06%, 53.77% and 66.05%, on the other hand, for Mel-Generalized cepstral, the average recognition accuracy found to be 66.14%, 61.15%, 63.30% and 63.93%, respectively.

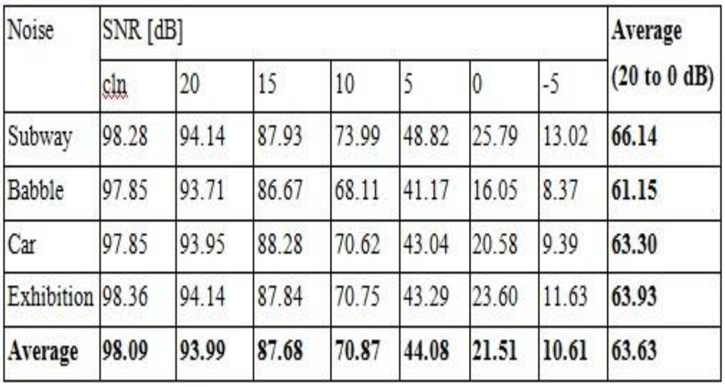
On the average, the word accuracy for the Mel-LPC is found to be 59.05% while the accuracy for the Mel-Generalized cepstral parameter is found to be 63.63%.

Table 2: Word accuracy for Mel-LPC with the value of warping factor, $\alpha = 0.35$



Noise	SNR [dB]							**Average**
	cln	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	4.45	**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 mel-generalized cepstral parameter with $\alpha = 0.35$ and $\gamma = 0.30$.



Noise	SNR [dB]							**Average**
	cln	20	15	10	5	0	-5	**(20 to 0 dB)**
Subway	98.28	94.14	87.93	73.99	48.82	25.79	13.02	**66.14**
Babble	97.85	93.71	86.67	68.11	41.17	16.05	8.37	**61.15**
Car	97.85	93.95	88.28	70.62	43.04	20.58	9.39	**63.30**
Exhibition	98.36	94.14	87.84	70.75	43.29	23.60	11.63	**63.93**
Average	**98.09**	**93.99**	**87.68**	**70.87**	**44.08**	**21.51**	**10.61**	**63.63**

7. Conclusion

From the recognition experiments, it has been found that the Mel-Generalized cepstrum outperforms the Mel-LPC. The word accuracy for Mel-Generalized cepstrum and for Mel-LPC was found to be 63.63% and 59.05%, respectively, for test set A. After evaluating the performance of Mel-Generalized cepstrum and Mel-LPC the average ward accuracy are obtained 61.15% for babble and 63.30% for car, whereas, in the case of Mel-LPC 48.06% is obtained for babble and 53.77% for car.

For Mel-LPC, the average word accuracy is obtained 68.30% for subway and 66.05% for exhibition, on the other hand, in the case of Mel-Generalized cepstrum 66.14% is obtained for subway and 63.93% is obtained for exhibition.

From the above discussion we can conclude that Mel-Generalized cepstral analysis is more effective than Mel-LPC analysis for noise type's babble and car; on the other hand Mel-LPC is more suitable for subway and exhibition noises.

References

1. F. Itakura and S. Satio, "A statistical method for estimation of speech spectral density and formant frequencies," Trans. IECE, vol. J53-A, pp.35-42, Jan. 1970 (in Japanese). Translation: R. W. Schafer and J. D. Markel, ed., Speech Analysis. New York: IEEE Press, 1979, pp.295-302.
2. B. S. Atal and S.L. Hanauer: "Speech analysis and synthesis by linear prediction of the speech wave", in J. Acoust. Soc. America, vol. 50, no. 2, pp.637-655, Mar. 1971.
3. A. V. Oppenheim and R.W. Schafer, "Homomorphic analysis of speech," IEEE Trans. Audio and Electroacoust, vol. AU-16, pp.221-226, June. 1968.
4. K. Tokuda, T. Kobayashi and S. Imai, "Generalized cepstral analysis of speech-unified approach to LPC and cepstral method," in Proc. ICSLP-90, 1990, pp.37-40.
5. K. Tokuda, T. Kobayashi, R. Yamamoto and S. Imai, "spectral estimation of speech based on generalized cepstral representation," Trans. IEICE, vol. J72-A, pp.457-465, Mar. 1989(in Japanese). Translation: Electronics and Communications in Japan (Part 3), vol. 73, no. 1, pp.72-81, Jan. 1990.
6. T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," IEEE Trans. Acoust., Speech, Signal processing, vol. ASSP-32, pp.1087-1089, Oct. 1984.
7. K. Tokuda, *et. al.*, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," Proc. of ICSLP94, pp.1043-1046, 1994.
8. A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," IEEE Proc.,
   vol.60, no.6, pp.681-691, 1972.
9. 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.
10. P. H Lindsay and D. A. Norman, " Human information processing: An introduction to Psychology," 2nd Ed., pp. 163, Academic Press, 1977.
11. H. W Strube,"Linear prediction on a warped frequency scale," J. Acoust. Soc. Am.,
    vol. 68, no. 4, pp. 1071-1076, 1980.
12. Moreno P., Raj B., Gouvea E. and Stern R. "Multivariate Gaussian-Based Cepstral Normalization for Robust Speech Recognition. *ICASSP95*.
13. Neumeyer L. and Weintraub M. "Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques".*ICASSP95*.
14. J. Markel and A. Gray, "Linear prediction of speech", Springer-Verlag, 1976.
15. S. Imai and C. Furuichi, "Unbiased estimator of log spectrum and its application to speech signal processing," in Proc. 1988 EURASIP, Sep.1988, pp.203-206.
16. H. G. Hirsch and D. Pearce, "The AURORA Experimental framework for the Performance evaluation of speech recognition system under noisy conditions", Proc. ISCA ITRW ASR., 181:188, 2000.
17. ITU recommendation G.712, "Transmission performance characteristics of pulse code modulation channels", 1996.
18. M. B. Islam, "Wiener filter for Mel-scaled LP based noisy speech recognition," Doctoral thesis, Shinshu University, Japan, March, 2007.