http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (11): 1644-1649, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.1644.1649

Robust Speech Feature Prediction Using Mel-LPC to Improve Recognition Accuracy

[1]S. Lokesh and [2]G. Balakrishnan
[1]Anna University of Technology, Tiruchirappalli, Tamil Nadu, India
[2]Indra Ganesan College of Engineering, Tiruchirappalli, Tamil Nadu, India

Abstract: The goal of the proposed study is robust speech feature prediction using mel-LPC to improve the performance of speech recognition in adverse conditions and compares the performance with those standard LPC and MFCC through English dictation system with 14,000 isolated words and 9,000 connected words. The mel-LPC feature prediction is estimated by an optimal value of frequency warping factor that can be estimated from the auto-correlation coefficients and it is computed as the inverse Fourier transform of the power spectrum to generate feature extract vector. Results of the feature extraction are a sequence of 18 mel-LPC coefficients which characteristic of the time-varying spectral properties of the speech signal and these are continuous that can map to discrete vectors in vector quantization codebook. This system is trained by 10 male and 10 female speakers and tested with 200 speakers in noisy and clean environments. Experiments results for various tasks show that with new mel-LPC feature vector system attains isolated and connected word accuracy of 97.5 and 93.2% for male speakers and 96.6 and 92.3% for female speakers with large vocabularies. The result shows that recognition accuracy is relatively higher than LPC and MFCC, respectively.

Key words: Speech recognition, mel-linear predictive coefficients, linear predictive coefficients, vector quantization, mel-frequency cepstral coefficients

INTRODUCTION

Speech is the most natural form of human communication. Speech processing has been one of the exciting areas in signal processing. Automatic Speech Recognition (ASR) system has made it possible for computer to understand human speech commands. To recognize the words, extraction of feature vector is an initial stage. There are various methods which are implemented based on the frequency scale combined into a number of spectral analysis. First, mel-Frequency Cepstral Coefficient (MFCC) is one of the most popular spectral features used in ASR (Matsumoto and Moroto, 2001; Homberg and Gelbart, 2006). In MFCC, the frequency axis is first warped to the mel-acoustic scale which is roughly linear below 1 kHz and roughly logarithmic above this point. Triangular filters which are equally spaced in the mel-scale are applied on the warped spectrum. The output of the filters is compressed using the logarithm function and cepstral coefficients are computed by applying the Discrete Cosine Transformation (DCT). Further smoothing is achieved by dropping the higher order cepstral coefficients which are known to contain mainly speaker specific information. Second, The LPC mel-cepstrum model takes into account of acoustic like frequency contribution, its frequency resolution is not improved by such a frequency warping of the LPC spectrum. To improve this inconsistency between the LPC and the acoustic analysis, several trainings have simulated the acoustic spectra before the all-pole modeling (Paul and Mustafa, 2009). Third, the Perceptual Linear Predictive (PLP) analysis is a well-known method (Hermansky, 1990), in this frequency axis is first warped to the Bark frequency scale. Trapezoidal shaped filters equally spaced on the Bark frequency scale are applied. The output of the filter bank is compressed using cubic root function which is motivated by the power law relationship between the intensity and amplitude.

In general, all ASR systems aim to automatically extract the feature vector of spoken words from input speech signals and store codebook, then performs pattern matching with reference pattern using Hidden Markov Model (HMM) training is illustrated in Fig. 1.

Existing feature extraction methods MFCC, LPC and PLP used in ASR have high computational costs and less recognition rate compared to the proposed mel-LPC analysis. The main objective of the proposed work is to extract robust speech feature coefficient using a simple and efficient time domain technique called "mel-LPC" analysis. These extracted coefficients are store in codebook which is mapped by vector quantization and trained by HMM. Future this study shows recognition rate and compares the recognition performance of

Corresponding Author: S. Lokesh, Anna University of Technology, Tiruchirappalli, Tamil Nadu, India

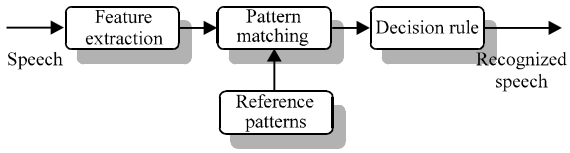

Fig. 1: General automatic speech recognition system

mel-LPC cepstrum with those of conventional cepstral parameters: the LPC mel-cepstral and the mel-frequency Cepstral Coefficients (MFCCs) through the English dictation system with 14,000 isolated words and 9000 connected words.

Mel-LPC ANALYSIS

This section combines fundamental frequency scale feature prediction based LPC analysis on mel-frequency and warped mel-frequency scale feature prediction based mel-LPC analysis.

LPC analysis on mel-frequency scale: The linear prediction method which is warped on a frequency scale is based on the standard autocorrelation method applied to the bilinear transformed speech signal. Let s[0],…,s[N-1] be the speech signal. The frequency warped signal is $\tilde{s}[n]$ where (n = 0,…,8) is obtained by:

$$\tilde{S}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{s}[\tilde{n}]\tilde{z}^{-n} \quad (1)$$

and the bilinear transformation of a finite length windowed signal s[n](n = 0,…, N-1) is defined by:

$$S(z) = \sum_{n=o}^{N-1} s[n]z^{-n} \quad (2)$$

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

$$\tilde{z}^{-1}(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad (3)$$

in the frequency domain, the spectrum $S(wj\lambda)$ on the linear $\tilde{S}(ej\tilde{\lambda})$ on the warped –frequency axis $\tilde{\lambda}$ by the frequency mapping function. The phase response of $\tilde{z}^{-1}$ s given by:

$$\tilde{\lambda} = \lambda + 2\tan^{-1}\left\{\frac{\alpha \sin\lambda}{1 - \alpha\cos\lambda}\right\} \quad (4)$$

this phase function determines a frequency mapping. The prediction error minimization in the $\tilde{z}$ domain is equivalent to minimize the output energy $\tilde{E}$ of the inverse filter:

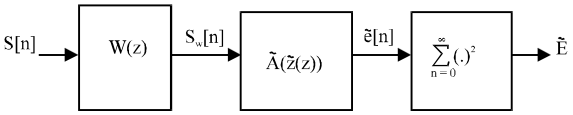

Fig. 2: Warped LPC analysis in the z domain

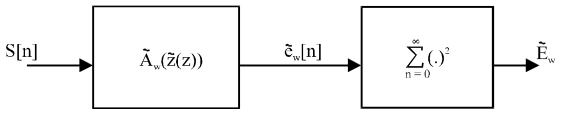

Fig. 3: Mel-LPC analysis in the z domain

$$A(\tilde{z}(z)) = 1 + \sum_{n-1}^{p} \tilde{a}_n \tilde{z}^{-n}(z) \quad (5)$$

in the z domain shown in Fig. 2 W(z) is defined by:

$$W(z) = \frac{\sqrt{1-\alpha^2}}{1 - \alpha z^{-1}} \quad (6)$$

and $|W(e^{j\lambda})|^2$ is equal to $d\tilde{\lambda}/d\lambda$.

The frequency warped signal $\tilde{s}[n]$ or the pre filtered signal $S_w[n]$ is an infinite sequence, the LPC analysis on the mel-frequency scale needs an approximation by paring $\tilde{S}[n]$ or $S_w[n]$.

Mel-LPC analysis: The idea behind the mel-LPC analysis is that a speech sample can be estimated as mel-cepstrum by applying mel-scale filter bank on linear combination of speech samples. By decreasing the sum of the squared differences between the actual speech samples and the mel-linearly predicted ones, a unique set of predictor coefficients is determined. The mel-LPC removes z domain W(z) in Fig. 2 by this proposed method directly minimize the output energy $\tilde{E}$ of the mel-inverse filter $\tilde{A}_w(\tilde{z}(z))$ in the z domain without pre-filtering S[n] as shown Fig. 3.

This alteration is equivalent to replacing s[n] in Fig. 2 by the signal whose z-transform is $S[z]W^{-1}[z]$. Thus the estimating inverse filter $\tilde{A}_w(\tilde{z}(z))$ is no longer same as $\tilde{A}(\tilde{z})$, but instead includes the effect of $W^{-1}[z]$. Then the estimated spectrum:

$$\tilde{H}_w(z) = \frac{\tilde{\sigma}_w}{1 + \sum_{n=1}^{p} \tilde{a}_{w,n} \tilde{z}(z)^{-n}} \quad (7)$$

The given Speech signal, S[0],…,S[N-1], the mel-prediction Coefficient $\{\tilde{a}_{w,i}\}$ are estimated by minimizing the prediction error energy $\tilde{E}_w$ over an infinite time interval:

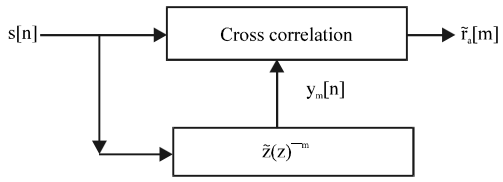

Fig. 4: The generalized autocorrelation function

$$\tilde{E}_w = \sum_{n=0}^{\infty}\left(\sum_{i=0}^{p}\tilde{a}_{w,i}\, y_i[n]\right)^2 \tag{8}$$

As a result $\{\tilde{a}_{w,i}\}$ and $\tilde{\sigma}_w$ are given by the Durbin's algorithm using the following autocorrelation function in which a unit delay is replaced by the all-pass filter:

$$\tilde{r}_\alpha[m] = \sum_{n=0}^{N-1} s[n] y_m[n] \tag{9}$$

where, $y_m[n]$ is the output signal of $\tilde{z}^{-m}(z)$ excited by x[n]. The generalized autocorrelation function is shown in Fig. 4.

The first order pre-emphasis in the z domain, cepstral coefficients (mel-LPC cepstral coefficients) derived from $\{\tilde{a}_{w,k}\}$ in the following experiments. The computational cost for the mel-LPC analysis is two times greater than that for the standard LPC analysis due to computation of $y_m[n]$ in equation. However, this computational load is much lower than existing methods and the prediction coefficients are estimated without any approximation.

VECTOR QUANTIZATION

Vector Quantization (VQ) is used for command identification in this system. It is a process of mapping feature vectors of a large vector space to a finite number of regions in that space. Each region is called a cluster and is represented by its center (centroid). A collection of all the centroids creates codebook. The quantity of data is significantly less, since the number of centroids is at least ten times smaller than the number of vectors in the original sample. This will reduce the amount of computations needed when comparing in later stages.

Codebook generation: There are many different algorithms to create a codebook. Since command recognition depends on the generated codebooks, it is important to select an algorithm that will best represent the original sample. For our system, the LBG algorithm (also known as the binary split algorithm) is used.

The algorithm is implemented by the following recursive procedure:

Step 1: Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here)

Step 2: Double the size of the codebook by splitting each current codebook y_n according to the rule: where n varies from 1 to the current size of the codebook and e is the splitting parameter. For this system, e = 0.001

$$\begin{aligned} y_n^+ &= y_n(1+\varepsilon) \\ y_n^- &= y_n(1-\varepsilon) \end{aligned} \tag{10}$$

Step 3: Nearest-neighbor search: For each training vector, find the centroid in the current codebook that is closest (in terms of similarity measurement) and assign that vector to the corresponding cell (associated with the closest centroid). This is done using the K-means iterative algorithm

Step 4: Centroid update: Update the codeword in each cell using the centroid of the training vectors assigned to that cell

Step 5: Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold

Step 6: Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed

Command matching: In the recognition phase the features of unknown command are extracted and represented by a sequence of feature vectors $\{x_1 \ldots x_n\}$. Each feature vector in the sequence X is compared with all the stored codewords in codebook and the codeword with the minimum distance from the feature vectors is selected as proposed command For each codebook a distance measure is computed and the command with the lowest distance is chosen. One way to define the distance measure is to use the Euclidean distances:

$$D = \left(\sum (x_i - y_j)^2\right)^{\frac{1}{2}} \tag{11}$$

Figure 5 describes the search of the nearest vector is done exhaustively, by finding the distance between the input vector X and each of the codewords C_1-C_M from the codebook C. The one with the smallest distance is coded as the output command.

To obtain the nearest neighbor, assume that there is a training sequence consisting of M input vector of X:

$$T = \{X_1, X_2, \ldots, X_M\} \tag{12}$$

this training sequence can be obtained from some large database. M is assumed to be sufficiently large so that all

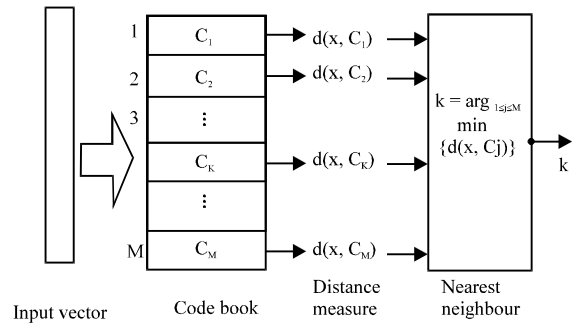

Fig. 5: Schematic of the nearest neighbor search on the VQ decoding process

the statistical properties of the source are captured by the training sequence. Let N be the number of code vectors and let:

$$C = \{C_1, C_2, ..., C_K, ..., C_M\} \quad (13)$$

represents the codebook. The distance measure of input vector and code word is measured as:

$$d(X,C) = \{d(x,C_1), d(x,C_2), .., d(x,C_K), ..., d(x,C_M)\} \quad (14)$$

Each code vector of K dimensional nearest neighbor is measured as:

$$K = \arg_{1\le j\le M} \min\{d(x,C_j)\} \quad (15)$$

TRAINING HMM

An important part of speech-to-text conversion using pattern recognition is training. Training involves creating a pattern representative of the features of a class using one or more test patterns that correspond to speech sounds of the same class. The resulting pattern (generally called a reference pattern) is an example or template, derived from some type of averaging technique. It can also be a model that characterizes the reference pattern statistics. This system uses speech samples from ten individuals during training. A model commonly used for speech recognition is the HMM which is a statistical model used for modeling an unknown system using an observed output sequence. The system trains the HMM 1500 out of 14000 isolated words 1500 out of 9000 connected words using the Baum-Welch algorithm.

All of the parameters of an HMM are arranged in a line to construct a group. For speech recognition, we use a left-right with one-order jump HMM model structure. The initial state distribution is fixed as $\pi = \{1, 0, 0, 0\}$. There are six states in this HMM. The state-transition probability distribution is A = aij, in which there are only 12 and the other aij parameters are always 0. After vector quantization, the codebooks of the speech signal feature vectors have sizes 18 coefficients for each sample. For example the observation symbol probability distribution B includes 6 words×18 parameters = 108 parameters. The HMM $\lambda = (A, B, \pi)$ has overall 6+108 =114 parameters. The training of an HMM involves searching for the best setting of all words.

The training of an HMM is based on likelihood maximization. Hence, the fitness function can be defined as the logarithmic calculation of the Viterbi algorithm for the kth observation sequences to the ith group. However, the HMM has some features. For HMM $\lambda = (A, B, \pi)$, the sum of each row vector of matrix A or B is 1.0. We have to control the gene production. When a new group is generated, the correspondence numbers of each segment of the row vector to unity.

EVALUATION AND EXPERIMENTAL SETUP

Evaluating mel-LPC warping factor: First, the optimal frequency warping factors in both the mel-LPC cepstrum and the LPC mel-cepstrum were examined in terms of the words accuracy obtained by language models. In the mel-LPC analysis, the optimal value of α for male speakers is around 0.5 which is between the mel and bark scales, whereas that for female speakers is 0.4 which is smaller than that corresponding to the mel-scale. As a result, it is clear that both mel-LPC cepstrum and the LPC mel-cepstrum outperform the LPC cepstrum, that is, for $\alpha = 0$, due to auditory-like frequency contribution. Furthermore, the mel-LPC cepstrum with the optimal frequency warping improves recognition accuracy over the LPC mel-cepstrum and the MFCC.

Experimental results with mel-LPC: Speech samples collection is mostly concerned with recording various speech samples of each distinct English word by different speakers. However, there are three main factors that must be considered when collecting speech samples which affect the training set vectors that are used to train the HMM (Richard, 2001). Those factors include who the speakers are; the speaking conditions, the transducers and the speech units. This system had 10 Male and 10 Female speakers out of whom their speech samples were collected. Those speakers belonging to different ages. Table 1 summarizes the first factor about the profiles of talkers/speakers.

Table 1: Summary of speakers profile

Age	Male speakers	Female speakers
18-21	6	4
22-35	3	4
37-40	1	2
Total	10	10

Table 2: Overall recognition rate (%) of the English dictation system

	Case 1	Case 2	Case 3
Dictation of words	10 trained male speakers (%)	100 male untrained speakers (%)	All male speakers (%)
Isolated words	100.00	94.04	97.5
Connected words	97.36	88.82	93.0

The important performance parameters are:

Recognition accuracy: The most important parameter in any recognition system is its accuracy. A recognition accuracy of 100% for all trained words, independent of the speaker, is the goal.

Recognition speed: If the system takes a long time to recognize the speech, users would become restless and the system loses its significance. A recognition time of less than 1 second is required.

The recognition accuracy is considered in three cases separately for male and female speakers:

- **Case 1:** Used samples from speakers 20 speakers (10 male and 10 female) that were also used for training
- **Case 2:** Used samples of 200 untrained speakers (1 to 200) whose voices were not used for training
- **Case 3:** Used samples from speakers 20 (10 male and 10 female) speakers and untrained speakers 100 male and 100 female to obtain the overall recognition performance

Recognition rate of the trained HMM is defined as follows:

$$RR = \frac{^{N}Correct}{^{N}Total} \times 100(\%) \quad (16)$$

where, RR is the recognition rate, $N_{Correct}$ is the number of correct recognition of testing speech samples per digit and N_{Total} is the total number of testing speech samples. The results are summarized in the following table.

Table 2 shows the overall recognition rates (%) for this system. Experimental results of the combination of mel-LPC and HMM algorithms in this system are acceptable, but could be results of the combination of improved further to obtain higher accuracy rates in noisy environment.

Table 3 shows recognition rate of isolated and connected words of male speakers which is done under clean environment. Figure 6 shows the bar chart of the recognition rate above table.

Table 3: Recognition rate of words for male speakers in three cases

	Isolated words recognition		Connected words recognition	
Environment	Male speaker	Female speaker	Male speaker	Female speaker
Clean	97.5	96.6	93.0	92.2
Noisy	91.3	90.1	89.9	88.0

Table 4: Recognition rate of words for male speakers in three cases

	Case 1	Case 2	Case 3
Dictation of words	10 trained male speakers (%)	100 male untrained speakers (%)	All male speakers (%)
Isolated words	100.00	93.21	96.6
Connected words	97.36	87.22	92.2

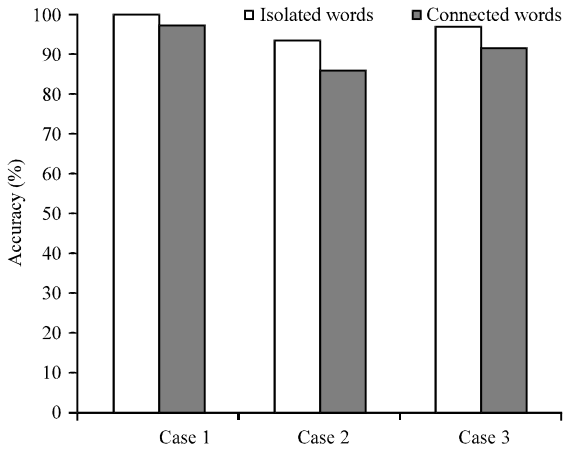

Fig. 6: Chart of recognition accuracy of words calculated for male speakers

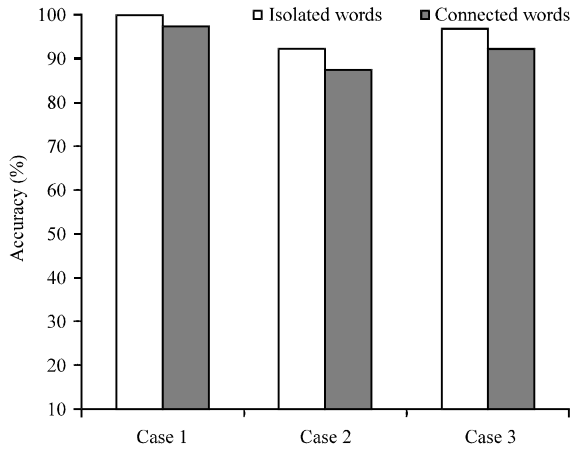

Fig. 7: Chart of recognition accuracy of words calculated for female speakers

Table 5: Comparison of recognition rates (%) for current speech recognition system with other researches system

Features extraction methods	Features classification methods	Recognition rate (%)	Reference
Mel-LPC	-	93 (Male) 93.1 (Female)	Matsumoto and Moroto (2001)
MFCC	VQ	88.88	Abu Shariah *et al.* (2007)
MFCC	-	87 to 93	Milner (2002)
PLP	-	85 to 91	
LPC	VQ and HMM	96	Podder (1997)
MFCC (Clean)	HMM (Clean)	96	Ahadi *et al.* (2003)
MFCC (Noisy)	HMM (Noisy)	78	
MFCC	VQ and HMM	57 to 98	Jackson (2005)

Table 4 shows recognition rate of isolated and connected words of Female speakers which is done under clean environment. The bar chart of the recognition rate above table is shown in Fig. 7.

From the above results, it is clearly found that the English speech recognition system performed well in both clean and noisy environment with both multi-speaker and speaker independent modes. Table 5 shows a comparison of recognition rates (%) for current speech recognition researches and systems together with feature extraction and classification techniques used.

Comparison of recognition rates (%) for current speech recognition system with other researches system is shown in Table 5, from which it is clearly shows that this system performed well in both clean and noisy environment with both male and female speakers. It is noticed that reference (Jackson, 2005) achieves higher recognition rate 98% than proposed system because that system calculates recognition rate based only on 600 vocabularies which is trained by 10 speakers and tested by 10 speakers. In proposed system, recognition rate calculated based 220 speakers with 2000 vocabularies which achieves 97.5% in clean environment.

CONCLUSION

This study has presented a simple and efficient time domain method mel-LPC and has evaluated the performance through large vocabulary. The mel-LPC cepstrum has achieved a significant improvement in recognition accuracy over the LPC mel-cepstrum and has attained slightly higher recognition accuracy than the MFCC. The results show that the system has some important advantages such as less number of feature vector co-efficient, high recognition rates, fast execution speed, robust etc. In additional, the system is trained by 20 speakers and tested by 200 untrained speakers and achieves 97.5 and 96.6% (male and female) accuracy for isolated words. In future, investigation is to improve this system by reducing number of feature vector coefficients and to improve the recognition rate by implementing new pattern recognition algorithm.

REFERENCES

Abu Shariah, M.A.M., R.N. Ainon, R. Zainuddin and O.O. Khalifa, 2007. Human computer interaction using isolated-words speech recognition technology. Proceedings of the International Conference on Intelligent and Advanced Systems, November 25-28, 2007, Kuala Lumpur, Malaysia, pp: 1173-1178.

Ahadi, S.M., H. Sheikhzadeh, R.L. Brennan and G.H. Freeman, 2003. An efficient front-end for automatic speech recognition. Electron. Circuits Syst., 1: 128-131.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87: 1738-1752.

Homberg, M. and D. Gelbart, 2006. Automatic speech recognition with an adaptation model motivated by auditory processing. Trans. Audio Speech Language Proc., 14: 43-49.

Jackson, M., 2005. Automatic speech recognition: Human computer interface for kinyarwanda language. Master Thesis, Faculty of Computing and Information Technology, Makerere University.

Matsumoto, H. and M. Moroto, 2001. Evaluation of mel-LPC cepstrum in a large vocabulary continuous speech recognition. IEEE Trans. Audio Speech Language Proc., 1: 117-120.

Milner, B., 2002. A comparison of front-end configurations for robust speech recognition. Acoust. Speech Signal Process., 1: 797-800.

Paul, A.K. and D.D. Mustafa, 2009. Bangla speech recognition system using LPC and ANN. Proceedings of the 7th International Conference on Advances in Pattern Recognition, February 4-6, 2009, Kolkata, West Bengal, pp: 171-174.

Podder, S.K., 1997. Segment-based stochastic modelings for speech recognition. PhD Thesis, Department of Electrical and Electronic Engineering, Ehime University, Matsuyama Japan.

Richard, B., 2001. Text-independent speaker recognition using source based features. Master Thesis, Wilder moth Griffith University Australia.