On the Use of Perceptual Line Spectral Pairs Frequencies for Speaker Identification

Md. Sahidullah and Goutam Saha
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur, India, Kharagpur-721 302
Email: sahidullah@iitkgp.ac.in, gsaha@ece.iitkgp.ernet.in

***Abstract*—Line Spectral Pairs Frequencies (LSFs) provide an alternative representation of the linear prediction coefficients. In this paper an investigation is carried out for extracting feature for speaker identification task which is based on perceptual analysis of speech signal and LSF. A modified version of the standard perceptual analysis is applied to obtain better performance. We have extracted the conventional LSF from the perceptually modified speech signal. State-of-the art Gaussian Mixture Model (GMM) based classifier is employed to design the closed set speaker identification system. The proposed method shows significant performance improvement over existing techniques in three different speech corpuses.**

***Index Terms*—Speaker Identification, Line Spectral Pairs Frequencies, Perceptual Linear Prediction, Gaussian Mixture Model (GMM).**

I. Introduction

Speaker Identification (SI) [1] is the task of determining the identity of a subject by its voice. A robust acoustic feature extraction technique followed by an efficient modeling scheme are the key requirements of an SI system. Feature extraction transforms [2] the crude speech signal into a compact but effective representation that is more stable and discriminative than the original signal. The central idea behind the feature extraction techniques for speaker recognition system is to get an approximation of short term spectral characteristics of speech for characterizing the vocal tract. Most of the proposed speaker identification systems use Mel Frequency Cepstral Coefficient (MFCC) or Perceptual Linear Predictive Cepstral Coefficient (PLPCC) for parameterizing speech. These cepstral coefficients parameterize the short term frequency response of speech signal to characterize the vocal tract information.

In this paper, a new spectral feature is proposed, which is inspired by Line Spectral Pairs (LSFs) frequency representation of Linear Predictive Coefficients (LPC) and is coupled with perceptual analysis of speech. LSFs are popular to represent linear prediction coefficients in LPC based coders for filter stability and representational efficiency. It also has other robust properties like ordering related to the spectral properties of the underlying data. The vocal tract resonance frequencies fall between the two pairs of LSF frequencies [3], [4]. These properties make LSFs popular for analysis, classification, and transmission of speech signal. Earlier, LSP was successfully introduced in speaker recognition task [5]. In this paper we have modified the conventional LSF using perceptual analysis. The conventional perceptual analysis [6] is modified for improving the performance of speaker recognition. The approach may contrasted with the method described in [7]. The LSF coefficients extracted is referred as *Perceptual LSF (PLSF)* throughout this paper.

In brief, the emphasis of this work is to efficiently extract LSF coefficients from perceptually modified speech signal; and finally use those coefficients for training the individual speaker models. Speaker Identification experiment is performed using this newly proposed feature using Gaussian Mixture Model (GMM) [8], [9] as a classifier. Three popular speech corpuses: POLYCOST, YOHO, and TIMIT are used for conducting experiments and evaluating the performance of PLSF feature based SI system.

II. Theoretical Background

A. Linear Prediction Analysis

In the LP model, $(n-1)$-th to $(n-p)$-th samples of the speech signal are used to predict the n-th sample. The predicted value of the n-th speech sample [10] is given by

$$\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k) \tag{1}$$

where $\{a(k)\}_{k=1}^{p}$ are the Predictor Coefficients (PC) and $s(n)$ is the n-th speech sample.The value of p is chosen such that it could effectively capture the real and complex poles of the vocal tract in a frequency range equal to half the sampling frequency.

Using the $\{a(k)\}_{k=1}^{p}$ as model parameters, equation (2) represents the fundamental basis of LP representation. It implies that any signal can be defined by a linear predictor and its prediction error.

$$s(n) = -\sum_{k=1}^{p} a(k)s(n-k) + e(n) \tag{2}$$

The LP transfer function can be defined as,

$$H(z) = \frac{G}{1+\sum_{k=1}^{p} a(k)z^{-k}} = \frac{G}{A(z)} \tag{3}$$

where G is the gain scaling factor for the present input and $A(z)$ is the p-th order inverse filter. These LP coefficients itself can be used for speaker recognition as it contains some speaker

978−1−4244−6385−5/10/$26.00 ©2010 IEEE

specific information like vocal tract resonance frequencies, their bandwidths etc. Various derivatives of LP coefficients are formulated to make them robust against different kinds of additive noise. Reflection Coefficient (RC), Log Area Ratio (LAR), Linear Prediction Cepstral Coefficients (RC), Inverse Sine Coefficients (IS), Line Spectral Pairs Frequencies (LSF) are such representations [1], [2].

B. Line Spectral Pairs Frequencies (LSF)

The LSFs are representation of the predictor coefficients of the inverse filter $A(z)$. At first $A(z)$ is decomposed into a pair of two auxiliary $(p+1)$ order polynomials as follows:

$$\begin{aligned} A(z) &= \frac{1}{2}(P(z)+Q(z)) \\ P(z) &= A(z) - z^{-(p+1)}A(z^{-1}) \\ Q(z) &= A(z) + z^{-(p+1)}A(z^{-1}) \end{aligned} \tag{4}$$

The LSF are the frequencies of the zeros of $P(z)$ and $Q(z)$. It is determined by computing the complex roots of the polynomials and consequently the angles. It can be done in different ways like complex root method, real root method and ratio filter method. The root of $P(z)$ and $Q(z)$ occur in symmetrical pairs, hence the name Line Spectrum Pairs (LSF). $P(z)$ corresponds to the vocal tract with the glottis closed and $Q(z)$ with the glottis open [3]. However, speech production in general corresponds to neither of these extreme cases but something in between where glottis is not fully open or fully closed. For analysis purpose, thus, a linear combination of these two extreme cases are considered.

On the other hand, the inverse filter $A(z)$ is a minimum phase filter as all of its poles lie inside the unit circle in the z-plane. Any minimum phase polynomial can be mapped by this transform to represent each of its roots by a pair of frequencies with unit amplitude. Another benefit of LSF frequency is that power spectral density (PSD) at a particular frequency tends to depend only on the close by LSF and vice-versa. In other words, an LSF of a certain frequency value affects mainly the PSD at the same frequency value. It is known as localization property, where the modification to PSD have a local effect on the LSF. This is its advantage over other representation like LPCC, Log Area Ratio (LAR) where changes in particular parameter affect the whole spectrum. The LSF parameters are themselves frequency values directly linked to the signal's frequency description.

In [11], it is stated that LSF coefficients are sufficiently sensitive to the speaker characteristics. Though popularity of LSF remains in low bit rate speech coding [12], [13], it is also successfully employed in speaker recognition [1], [5].

C. Perceptual Linear Prediction (PLP) Analysis

The PLP technique converts speech signal in meaningful perceptual way through some psychoacoustic process [6]. It improves the performance of speech recognition over conventional LP analysis technique. The various stages of this method are based on our perceptual auditory characteristics. The significant blocks of PLP analysis are as follows:

1) Critical Band Integration: In this step the power spectrum is wrapped along its frequency axis into Bark frequency. In brief, the speech signal passed through some trapezoidal filters equally spaced in Bark scale.

2) Equal Loudness Pre-emphasis: Different frequency components of speech spectrum are weighted by a simulated equal-loudness curve.

3) Intensity-loudness Power law: Cube-root compression of the modified speech spectrum is carried out according to the power law of hearing [14].

In addition, RASTA processing [15] is done with PLP analysis as an initial spectral operation to enhance the speech signal against diverse communication channel and environmental variability. The integrated method is often referred as RASTA-PLP.

III. Proposed Frame Work: PLSF

The contribution of the present work is in combining strength of Perceptual Linear Prediction (PLP) with LSF for automatic speaker identification. Towards this, a modification in standard PLP scheme is investigated and a strategy is formulated to use modified PLP coefficient for generation of LSFs. A drawback of PLP analysis technique is that the nonlinear frequency wrapping stage or critical band integration stage introduces undesired spectral smoothing. We have analyzed the scatter plot of training data of two first features of two male (fig. 1) and two female (fig. 2) speakers including and excluding the critical band integration step. It is very clear from both the figures is that the speaker's data are more separable if critical band integration step is ignored.

Contrasted with the work [7], we include pre-emphasis in this part of the scheme and LSF. Perceptual weighting of different frequency component enhances the speech signal according to the listening style of human beings. Pre-emphasis stage, is however to emphasize the high frequency component of speech to overcome the roll-of factor of -6dB/octave due to speaking characteristics of human being. Hermansky also

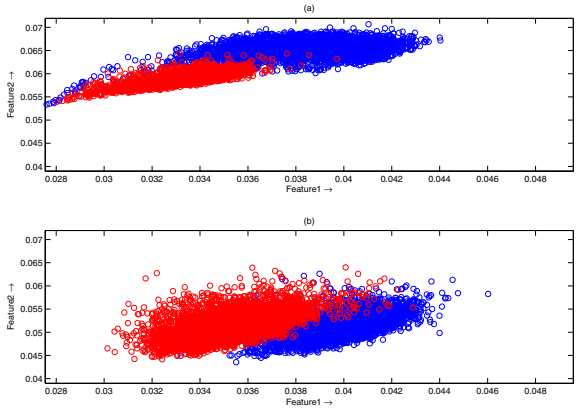

Fig. 1. Scatter plot of first two features of training data for two male speakers (shown using red and blue color) (from POLYCOST database) for (a) With Critical Band Integration step (b) Without Critical Band step.

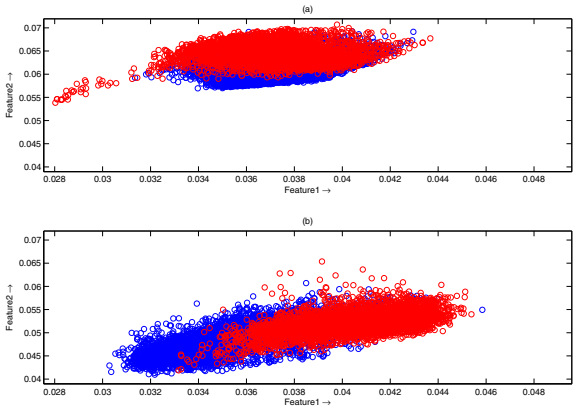

Fig. 2. Scatter plot of first two features of training data for two female speakers (shown using red and blue color) (from POLYCOST database) for (a) With Critical Band Integration step (b) Without Critical Band step.

has included this step in his work. We have also experimentally observed that it also has contribution in improving the performance.

The overall schematic diagram of the proposed Perceptual Line Spectral Pairs feature extraction technique which is based on modified perceptual linear prediction analysis, is shown in Fig. 3.

The proposed perceptual operation represents the lower frequency region more accurately than the higher frequency zone. In fig. 4 comparative plots of speech spectrum, LP- spectrum, LSF of a speech speech frame (a) and its perceptual version (b) are shown. The spectral peaks which are sharply approximated by conventional LP are smoothed by modified PLP. This property enables this technique to carry the information regarding the variability of a formant frequency with in the particular speaker. The spectral tilt carries speaker related information [16]. The perceptual modification of spectral information may carry speaker dependant information which was removed by conventional PLP [Sec. II-D in [6]].

LSFs reveal vocal tract spectral information including mouth shape, tongue position and contribution of the nasal cavity. Its perceptually motivated version represents those characteristics more effectively and hence is expected to improve speaker recognition performance.

IV. Speaker Identification Experiment

A. Experimental Setup

1) Pre-processing stage: In pre-processing step silence portions are removed from the speech signals. Then, each utterance is pre-emphasized with a pre-emphasis factor of 0.97. Consequently, the signal is framed into segments of 20ms keeping 50% overlap with adjacent frames and they are windowed using hamming window function.

2) Classification & Identification stage: Gaussian Mixture Modeling (GMM) technique is used to get probabilistic model for the feature vectors of a speaker. The idea of GMM is to use weighted summation of multivariate gaussian functions to represent the probability density of feature vectors and it is given by

$$p(\mathbf{x}) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}) \tag{5}$$

where $\mathbf{x}$ is a d-dimensional feature vector, $b_i(\mathbf{x})$, $i = 1, ..., M$ are the component densities and p_i, $i = 1, ..., M$ are the mixture weights or *prior* of individual gaussian. Each component density is given by

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma_i}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_i})^t \mathbf{\Sigma_i}^{-1} (\mathbf{x} - \boldsymbol{\mu_i}) \right\} \tag{6}$$

with mean vector $\boldsymbol{\mu_i}$ and covariance matrix $\mathbf{\Sigma_i}$. The mixture weights must satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$ and $p_i \geq 0$. The Gaussian Mixture Model is parameterized by the mean, covariance and mixture weights from all component densities and is denoted by

$$\lambda = \{p_i, \boldsymbol{\mu_i}, \mathbf{\Sigma_i}\}_{i=1}^{M} \tag{7}$$

In these experiments, the GMMs are trained with 10 iterations of Expectation Maximization(EM) algorithm where clusters are initialized by vector quantization algorithm.

In closed set SI task, an unknown utterance is identified as an utterance of a particular speaker whose model gives maximum log-likelihood. It can be written as

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} p(\boldsymbol{x}_t | \lambda_k) \tag{8}$$

where $\hat{S}$ is the identified speaker from speaker's model set $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_S\}$ and S is the total number of speakers.

3) Databases for experiments: *POLYCOST Database:* Mother tongue files (MOT) of POLYCOST database were used for evaluating performance. All speakers (131 after deletion of three speakers due to insufficient data) in the database were registered as clients.

YOHO Database: All the 138 speakers were used in evaluation purpose. Speech data of enrollment section were used for creating speaker models, where as all the test utterances i.e. $138 \times 40 = 5520$ speech files were used to evaluate the performance of the system.

TIMIT Database: TIMIT is a noise-free speech corpus recorded with a high-quality microphone sampled at 16 kHz [7]. This database consists of total 630 speakers. In this paper, we have utilized all 168 speakers in the testing folder of TIMIT for conducting the experiments. Each speaker has 10 utterances; the first five of them are used for training the speaker model and the remaining fives are used for testing purpose. The total number of utterances under test becomes $168 \times 5 = 840$.

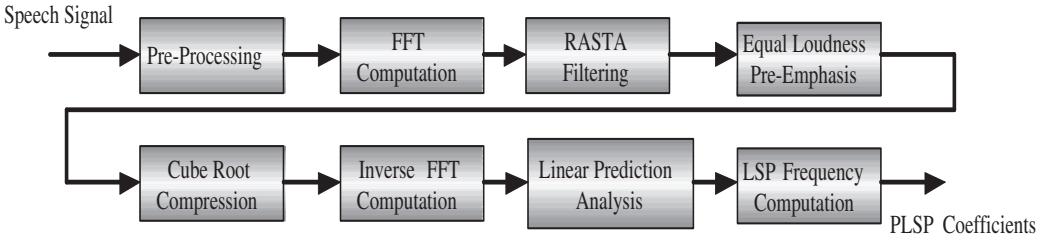

Fig. 3. Block diagram showing different stages Perceptual Line Spectral Pairs (PLSF) frequency based feature extraction technique.

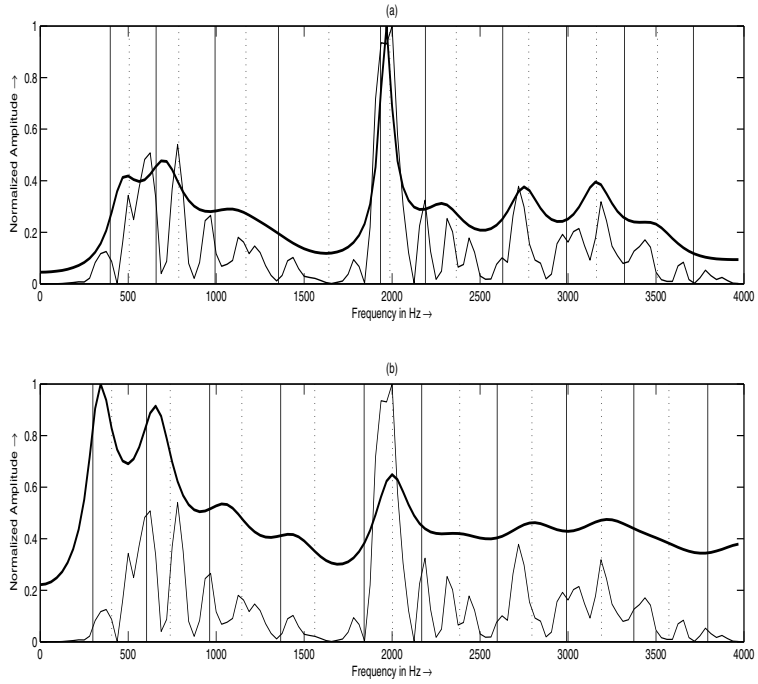

Fig. 4. Plot showing (a) Speech spectrum (light line), LP- spectrum (dark line) and LSF (Vertical Lines) & (b) Speech Spectrum (light line), PLP- spectrum (dark line) and PLSF (Vertical Lines). The odd LSFs are denoted using continuous lines and the even LSFs are denoted using dotted lines.

4) Score Calculation: In closed-set speaker identification problem, identification accuracy as defined in [8], is given by the equation (9) is followed.

$$\text{Percentage of identification accuracy } (\textbf{PIA}) = \frac{\text{No. of utterance correctly identified}}{\text{Total no. of utterance under test}} \times 100 \quad (9)$$

B. Speaker Identification Experiments and Results

We have evaluated the performance of SI system based on each databases using PLSF feature as frontend processing block. The experiments are conducted using GMM based classifier of different model orders depending on the amount of training data. To show a comparative study experiments are also carried out using the different baseline features which are widely used for speaker identification task. Other than PLSF, the performance of SI system is evaluated for MFCC, PLPCC, and PLAR. Identification accuracy is also shown for LSF to infer the significant of perceptual analysis. The feature dimension is fixed at 19 for techniques for better comparison. In LP based systems 19 filters are used for all-pole modeling of speech signals. On the other hand 20 filters are used for MFCC, and 19 coefficients are taken after discarding the first co-efficient which represents dc component. The detail description are available in [17]. The process of extraction of

TABLE I
RESULTS (**PIA**) OF GMM FOR POLYCOST

Model Order	LSF	MFCC	PLPCC	PLAR	**PLSF**
2	60.7427	63.9257	62.9973	65.5172	**65.6499**
4	66.8435	72.9443	72.2812	74.1379	**74.5358**
8	75.7294	77.8515	75.0663	78.3820	**80.6366**
16	78.1167	77.8515	78.3820	78.7798	**82.7586**

TABLE II
RESULTS (**PIA**) OF GMM FOR YOHO

Model Order	LSF	MFCC	PLPCC	PLAR	**PLSF**
2	70.7428	74.3116	66.5761	83.4420	**77.7355**
4	81.3768	84.8551	76.9203	90.0000	**88.8043**
8	90.4529	90.6703	85.3080	94.0580	**94.0942**
16	93.2246	94.1667	90.6341	95.6703	**96.0326**
32	95.5978	95.6522	93.5326	96.5036	**97.0833**
64	96.5761	96.7935	94.6920	96.9746	**97.4094**

TABLE III
RESULTS (**PIA**) OF GMM FOR TIMIT

Model Order	LSF	MFCC	PLPCC	PLAR	**PLSF**
2	91.3095	95.3571	82.2619	88.2143	**95.8333**
4	92.1429	97.1429	93.9286	95.8333	**98.5714**
8	97.8571	98.3333	96.7857	99.1667	**99.0476**
16	99.2857	99.5238	98.0952	98.5714	**99.6429**
32	98.9286	99.0476	98.4524	98.8095	**99.4048**

other features are also available in [1], [7], [18].
The results are shown in Table I, II & III for POLYCOST, YOHO, and TIMIT database respectively. The last columns of each table correspond to results on proposed PLSF based SI system while the rest are based on other baseline features. The proposed feature based system outperforms the other existing conventional techniques as well as recently proposed perceptual feature, PLAR [7]. The POLYCOST database consists of speech signals collected over telephone channel. The improvement for this database is significant compared the other two i.e YOHO and TIMIT which are micro-phonic.

LSF frequency coefficients are extracted from the LP polynomial. PLSF are formulated by adding perceptual flavor to it. Therefore, it has both the advantages of LSF and PLP. On the other hand the conventional RASTA-PLP is modified by removing the critical band integration stage. It also helps in improving the identification accuracy.

V. CONCLUSION

The objective of this paper is to propose a feature extraction technique for improving the performance of SI systems. The proposed technique which exploits the advantages of line spectral pairs frequency parameters and perceptual analysis, gives improved identification accuracy for three large population speech corpuses for different numbers of Gaussians. PLSFs are well suited for quantization process just like LSFs. It is expected that VQ modeling based SI system using PLSF front-end can give significant performance compared to other features based ones. This can be incorporated where recognition time provided is less as well as the duration of the test utterance. It is also anticipated that the proposed feature can also be employed in speaker recognition task and will decrease equal error rate (EER) to improve the performance of ASR systems in latest NIST databases.

REFERENCES

[1] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep 1997.

[2] T. Kinnunen, "Spectral features for automatic textindependent speaker recognition," Ph.D. dissertation, University of Joensuu, 2004.

[3] T. Bäckström and C. Magi, "Properties of line spectrum pair polynomials: a review," *Signal Process.*, vol. 86, no. 11, pp. 3286–3298, 2006.

[4] I. V. McLoughlin, "Review: Line spectral pairs," *Signal Process.*, vol. 88, no. 3, pp. 448–467, 2008.

[5] C.-S. Liu, W.-J. Wang, M.-T. Lin, and H.-C. Wang, "Study of line spectrum pair frequencies for speaker recognition," *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 277–280 vol.1, Apr 1990.

[6] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990. [Online]. Available: http://link.aip.org/link/?JAS/87/1738/1

[7] W. H. Abdulla, "Robust speaker modeling using perceptually motivated feature," *Pattern Recogn. Lett.*, vol. 28, no. 11, pp. 1333–1342, 2007.

[8] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[9] D. A. Reynolds, "A gaussian mixture modeling approach to text-independent speaker identification," Ph.D. dissertation, Georgia Institute of Technology, Sept 1992.

[10] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[11] F. Soong and B. Juang, "Line spectrum pair (lsp) and speech data compression," vol. 9, Mar 1984, pp. 37–40.

[12] A. Lepschy, G. Mian, and U. Viaro, "A note on line spectral frequencies [speech coding]," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 8, pp. 1355–1357, Aug 1988.

[13] A. G. Bishnu S. Atal, Vladimir Cuperman, *Advances in Speech Coding*. Springer, 2003.

[14] S. S. Stevens, "On the psychophysical law," *Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.

[15] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[16] N. B. Yoma and T. F. Pegoraro, "Robust speaker verification with state duration modeling," *Speech Communication*, vol. 38, no. 1-2, pp. 77 – 88, 2002.

[17] S. Chakroborty, "Some studies on acoustic feature extraction, feature selection and multi-level fusion strategies for robust text-independent speaker identification," Ph.D. dissertation, Indian Institute of Technology, 2008.

[18] L. Rabiner and H. Juang B, *Fundamental of speech recognition*. First Indian Reprint: Pearson Education, 2003.