Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach

A.Revathi[1], R.Ganapathy[2] and Y.Venkataramani[3]

[1]Department of ECE, Saranathan college of Engg., Trichy
revathidhanabal@rediffmail.com
[2]Department of MCA, Saranathan college of Engg., Trichy
sowgar05@yahoo.com
[3]Principal, Saranathan college of Engg., Trichy
principal@saranathan.ac.in

Abstract

This paper presents the effectiveness of perceptual features and iterative clustering approach for performing both speech and speaker recognition. Procedure used for formation of training speech is different for developing training models for speaker independent speech and text independent speaker recognition. So, this work mainly emphasizes the utilization of clustering models developed for the training data to obtain better accuracy as 91%, 91% and 99.5% for mel frequency perceptual linear predictive cepstrum with respect to three categories such as speaker identification, isolated digit recognition and continuous speech recognition. This feature also produces 9% as low equal error rate which is used as a performance measure for speaker verification. The work is experimentally evaluated on the set of isolated digits and continuous speeches from TI digits_1 and TI digits_2 database for speech recognition and on speeches of 50 speakers randomly chosen from TIMIT database for speaker recognition. The noteworthy feature of speaker recognition algorithm is to evaluate the testing procedure on identical messages of all the 50 speakers, theoretical validation of results using F-ratio and validation of results by statistical analysis using χ^2 distribution.

Keywords

Clustering methods, Speech recognition, Speaker recognition, Spectral analysis, Speech analysis, Speech processing, Vector quantization.

1. Introduction

The fundamental method of speech recognition is to decode the speech signal in a sequential manner based on the observed acoustic features of the signal and known relations between acoustic features and phonetic symbols. The pattern recognition approach to speech recognition is basically one in which the speech patterns are directly used. This method involves the training of speech patterns and recognition of patterns via pattern comparison. Training procedure adopted is able to adequately characterize the acoustic properties of the pattern. This type of characterization of speech via training is called pattern classification because the machine learns which acoustic properties of the speech class are reliable and repeatable across all training tokens of the pattern. The utility of the method is the pattern – comparison stage, which does the

direct comparison of the unknown speech with each possible pattern learned in the training phase and classifies the unknown speech according to the goodness of match of the patterns.

The performance of the speech recognition systems is given in terms of a word error rate (%) as measured for a specified technology, for a given task, with specified task syntax, in a specified mode, and for a specified word vocabulary. Robust speech recognition systems can be applied to automation of office or business, monitoring of manufacturing processes, automation of telephone or telecommunication services, editing of specialized medical reports and development of aids for the handicapped. For an example, high accuracy connected digits recognition system finds application in the recognition of personal identification numbers, credit card numbers, and telephone numbers. Continuous speech recognition systems find applications in voice repertory dialer where eyes free, hands free dialing of numbers is possible. Speech recognition is done [14] using audio visual features. Parameters of the mel-cepstrum transformation are optimized in [15] for speech recognition. HTK software tool kit is used [16] for large vocabulary speech recognition. Large margin hidden markov models are used [17] for speech recognition. Sub band correlation between feature streams is the method used in [18] for recognizing speech. Speaker independent Chinese digit speech recognition was done [22] using multi weighted neural network. Perceptual linear prediction and mel-frequency cepstral coefficients were used as features and HMM for developing training models [23] for combined speech recognition and speaker verification. This work mainly reveals the successful implementation of clustering procedure based on the formation of training speech for speech recognition also in this work.

Vocal communication between people and computers includes the synthesis of speech from text, automatic speech recognition and speaker recognition. Speaker recognition involves the speaker identification to output the identity of the person most likely to have spoken from among a given population or to verify a person's identity who he/she claims to be from a given speech input. While finger prints and retinal scans have been usually considered to be reliable ways of authenticating people, voice identification has the convenience of easy data collection over telephone. Extraction of optimum features depicting the variations in speaker characteristics also influence the accuracy.

Automatic speaker verification (ASV) has been a simpler task, since it only requires comparison between test pattern and one reference template and involves a binary decision of whether to accept or reject a speaker. The front end of the recognizer contains normalization, parameterization and feature extraction. It leads to data reduction or elimination of redundancies in the input data sequence. There are many features depicting the characteristics of the vocal tract such as LPCC, MFCC, DCTC, LSF, PLP and their use in speaker identification/speaker verification task has been discussed in [1,5-8,10]. Optimum wavelet entropy [19] parameter values were used as features and adaptive neural fuzzy inference system was used for classifying speakers. Paper [20] describes a method for speaker identification in multiple languages based on back propagation algorithm. Perceptual log area is used as feature [21] for speaker identification.

Das et.al [11] have introduced the scheme for speech processing in speaker verification. They have indicated that utterances should preferably be collected over a long period of time. Rosenberg et.al [12] have introduced new techniques for speaker verification. They have used linear prediction coefficients, pitch and intensity for evaluating the performance of the system. Guruprasad et.al [13] have used difference cepstrals obtained using low order and high order LP analysis and auto associative neural work as a pattern classifier for evaluating performance of speaker recognition and obtained the equal error rate as 19.5%.

Use of various system features in speaker/twins identification is discussed [1,5-8,10]. Use of perceptual features has been analysed for performing speaker identification task and isolated digits/continuous speech recognition in this paper. Clustering procedure is successfully implemented for speech/speaker recognition in this work. Formation of training speech is altered for implementing speaker/ speech recognition. This paper mainly emphasizes the use of clustering procedure and how the clusters are ultimately depicting the characteristics of the speech / speaker in evaluating the performance of speech / speaker recognition system.

2. Feature based on Cepstrum

The short-time speech spectrum for voiced speech sound has two components: 1) harmonic peaks due to the periodicity of voiced speech 2) glottal pulse shape. The excitation source decides the periodicity of voiced speech. It reflects the characteristics of speaker. The spectral envelope is shaped by formants which reflect the resonances of vocal tract. The variations among speakers are indicated by formant locations and bandwidth.

2.1. PLP and MF-PLP extraction

PLP (perceptual linear predictive cepstrum) speech analysis method [2-4] models the speech auditory spectrum by the spectrum of low order all pole model. The detailed procedure for PLP and MF-PLP (Mel frequency perceptual linear predictive cepstrum) extraction is given below. The block diagram for PLP and MF-PLP extraction is shown in FIG.1.

1. Compute power spectrum of windowed speech.
2. Perform grouping to 21 critical bands in bark scale or mel scale for sampling frequency of 16 kHz.
3. Perform loudness equalization and cube root compression to simulate the power law of hearing.
4. Perform IFFT
5. Perform LP analysis by Levinson -Durbin procedure.
6. Convert LP coefficients into cepstral coefficients.

The relationship between frequency in Bark and frequency in Hz is specified as in (1)

$$f(bark) = 6 * arcsinh(f(Hz)/600) \quad (1)$$

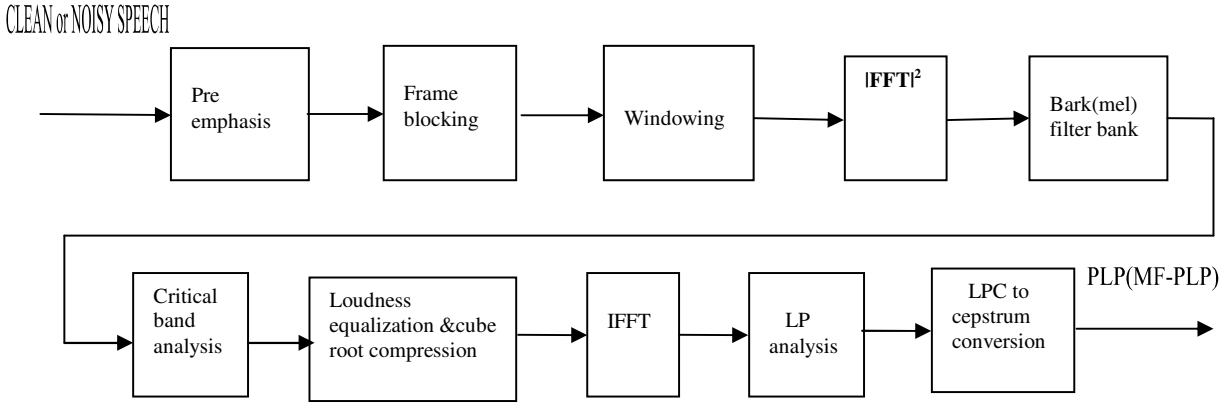

FIG.1- PLP and MF-PLP extraction model

3. Training model based on clustering technique

The way in which L training vectors can be clustered into a set of M code book vectors is by K-means clustering algorithm [9]. Block diagram for K-means clustering and classification is shown in FIG. 2.

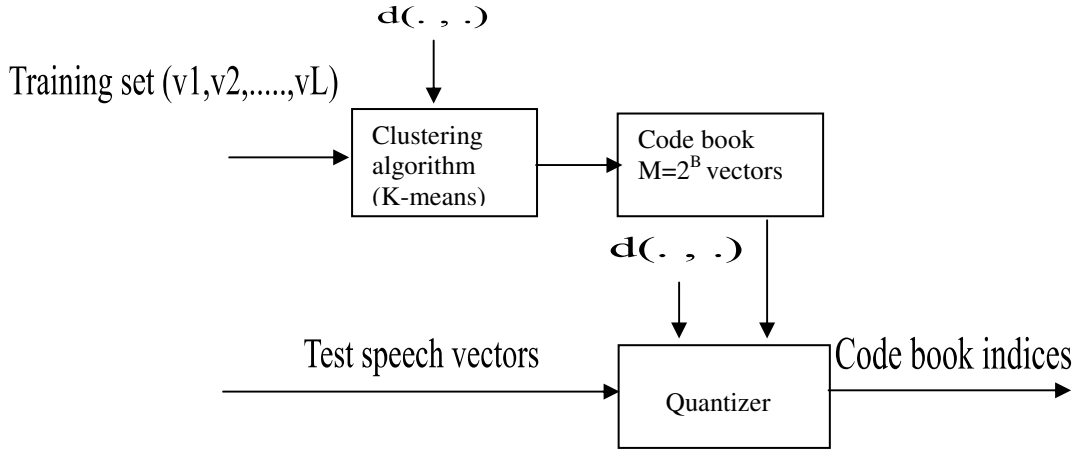

FIG.2 – Block diagram of the basic VQ training and classification structure

Classification procedure for arbitrary spectral analysis vectors that chooses the codebook vector is by computing Euclidean distance between each of the test vectors and M cluster centroids. The spectral distance measure for comparing features v_i and v_j is as in (2).

$$d(v_i, v_j) = d_{ij} = 0 \quad when \quad v_i = v_j \tag{2}$$

If codebook vectors of an M-vector codebook are taken as y_m, $1 \le m \le M$ and new spectral vector to be classified is denoted as v, then the index m^* of the best codebook entry is as in (3)

$$m^* = arg(min(d(v, y_m))) \quad for \quad 1 \le m \le M \tag{3}$$

Clusters are formed in such a way that they capture the characteristics of the training data distribution. It is observed that Euclidean distance is small for the most frequently occurring vectors and large for the least frequently occurring ones.

4. Speech recognition based on proposed features

Speech recognition system involves extraction of features from the training and testing data, building VQ codebook models [9] for all digits (0-9) and continuous speeches of speakers and testing each utterance against a certain number of speech models to detect the identity of the speech of that utterance from among the speech models. The speech database used for isolated digit recognition contains isolated digits from TI digits_1 and TI digits_2. Training data for isolated digit recognition system is formed by concatenating the speeches of the set of isolated digits pronounced by 24 speakers. Another set of isolated digits is used for evaluating speaker dependent isolated digit recognition system. Speaker independent digits recognition system is evaluated by using set

of digits pronounced by other speakers in the database. Speaker independent continuous speech recognition system is evaluated on training data formed by concatenation of dialect sentences of 24 speakers and test data from 100 speakers in the TIMIT database.

For creating a training model, speech signal is pre-emphasized using a difference operator. Hamming window is applied on differenced speech frames of 16 msecs duration with overlapping of 8 msecs. Then the features such as PLP and MF-PLP [2-4] are extracted. For each speech, VQ codebook model is developed based on K-means clustering procedure [9] for these perceptual features. In this algorithm there is a mapping from L training vectors into M clusters. Each block is normalized to unit magnitude before giving as input to the model. One model is created for each speaker.

For testing, perceptual features are extracted for test speech. Test data can be either isolated digit or continuous speech from the database. Features extracted from each test utterance are fed to the claimant models. Then the minimum distance is found between each test vector and centroid of clusters. Average of minimum distances for each speech model is determined. The test utterance best matches with a speech model which has minimum average value.

5. Speaker identification based on proposed features

The identification system involves extraction of features from the training and testing data, building VQ codebook models [9] for all enrolled speakers and testing each utterance against a certain number of claimant models to detect the identity of the speaker of that utterance from among the claimants. The speech database used for system evaluation contains 50 speakers selected randomly from 8 dialect regions in 'TIMIT' speech database.

The identification system involves extraction of features from the training data formed by combining TI random contextual variant sentences and MIT phonetically compact sentences and features from the test data formed by combining SRI dialect calibration sentences. It also involves building VQ codebook models for all enrolled speakers and testing each utterance against a certain number of claimant models to detect the identity of the speaker of that utterance from among the claimants. Present study uses the training speech of 15 seconds and test data of 4 seconds duration. Feature vectors of test speech of nearly 7 seconds duration have been considered for evaluating the performance of speaker identification system. Each speaker has been tested on an average of 75 test speech segments. All the speeches taken for analysis have been sampled at 16 kHz.

For creating a training model, speech signal is pre-emphasized using a difference operator. Hamming window is applied on differenced speech frames of 16 msecs duration with overlapping of 8 msecs. Then the features such as PLP And MF-PLP [2-4] are obtained. For each speaker VQ codebook model is developed based on K-means clustering procedure [9] for all the proposed features. In this algorithm there is a mapping from L training vectors into M clusters. Each block is normalized to unit magnitude before giving as input to the model. One model is created for each speaker.

For testing, speech signal is obtained by considering the speeches of SR1 dialect calibration sentences. The features PLP and MF-PLP are extracted for the test speech. To evaluate different test utterance lengths, the sequence of feature vectors was divided into overlapping segments of T feature vectors. The first two segments from a sequence would be

$$\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots\ldots\ldots \underset{\sim}{x}_T$$

$$\underset{\sim}{x}_{11}, \underset{\sim}{x}_{12}, \ldots\ldots\ldots \underset{\sim}{x}_{T+10}, \text{ etc.,}$$

A test segment of length T=100 feature vectors. Each segment of T vectors was treated as a separate test utterance. Features extracted from each test utterance are fed to the claimant models. Then the minimum distance is found between each test vector and centroid of clusters. Average of minimum distances for each cluster is determined. The test utterance best matches with a cluster which has minimum average value. The performance evaluation was then computed as a percent of correctly identified T-length segments over all test utterances as in (4)

$$\%\,correcti\ identification = \frac{\#\,correctly\ identified\ segmnets}{total\#\ of\ segments} x100 \quad (4)$$

6. Results and discussion

The performance of speech / speaker recognition system based on perceptual features is evaluated by finding squared Euclidean distance between test vectors and each reference value. Speech / speaker recognition rate is the number of correct choices over the total number of test speeches.

The discriminative potential of statistical parameters is commonly evaluated by F-ratio (Fisher's ratio), which is calculated as ratio of between-speech variance of parameter value means to mean within-speech variance of the same parameter values using the formula given in (4)

$$F = \frac{\sigma_1^2}{\sigma_2^2} \quad (5)$$

Where σ_1^2 is between – speech dispersion of parameter mean value, while σ_2^2 is mean within – speech dispersion of the same parameter value. For normal variables, the probability of misclassifying a speaker i to speaker j is a monotonically decreasing function of F-ratio. This F-ratio has been applied to evaluate the efficiency of feature selection in order to improve the speech recognition accuracy.

FIG.3 and FIG.4 show the comparison between the PLP and MF-PLP for speaker dependent and speaker independent isolated digit recognition in terms of their individual accuracy. Average accuracy of PLP and MF-PLP features for speaker dependent case is same and average accuracy of MF-PLP feature is better than that of PLP feature for speaker independent case.

TABLE 1 gives the details of the experimental evaluation of the features for speaker dependent and speaker independent isolated digit recognition system.

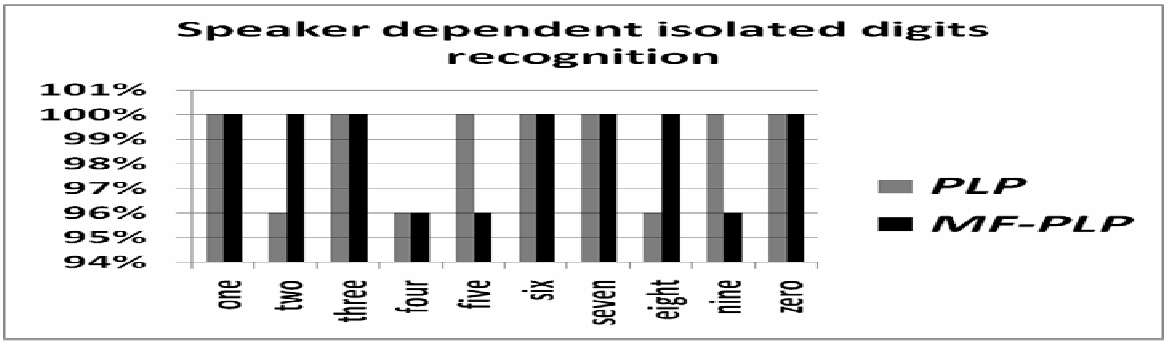

FIG.3 – Comparison chart - individual accuracy of PLP and MF–PLP (Speaker dependent)

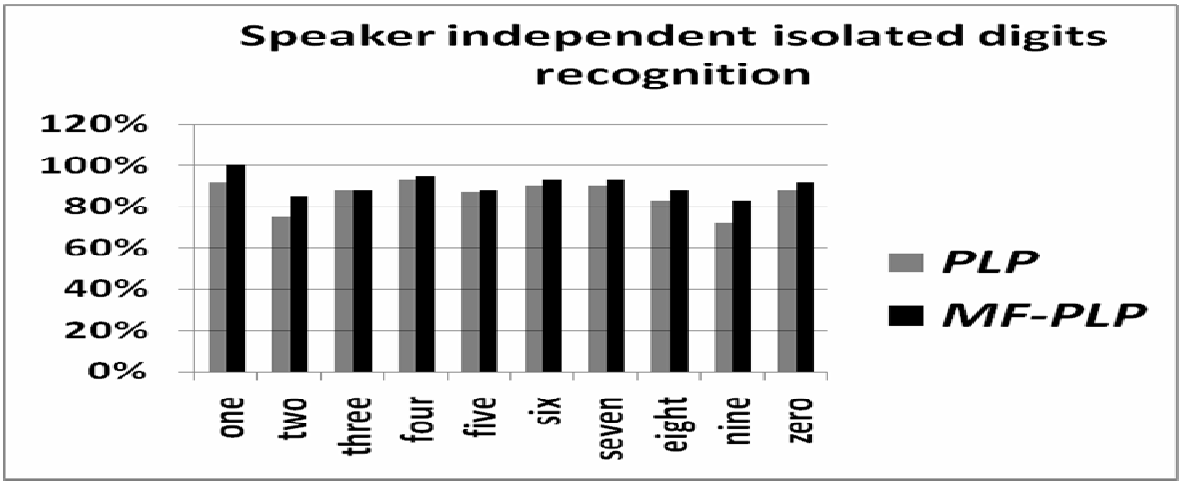

FIG.4 – Comparison chart - individual accuracy of PLP and MF–PLP (Speaker independent)

% Recognition accuracy			
Speaker dependent		Speaker independent	
PLP	MF-PLP	PLP	MF-PLP
99	99	86	91

TABLE 1 - Overall accuracy of isolated digit recognition system

From TABLE 1, it is understood that performance is same for both perceptual features for speaker dependent case, but performance is better for MF-PLP for speaker independent case. TABLE 2 indicates the performance of continuous speech recognition for clean test speech.

Feature	%Recognition accuracy	F-Ratio
PLP	99	0.0066
MF-PLP	99.5	0.0067

TABLE 2 - Overall accuracy of continuous speech recognition

From TABLE 2, it is clear that the calculation of F-ratio on the training data is monotonically increasing function of accuracy. FIG.5, FIG.6, FIG.7 and FIG.8 depict the performance of speaker recognition system in terms of overall accuracy, %FRR %FAR and %EER [5, 11-13]. FIG.5 depicts the performance of perceptual features in speaker identification for individual speakers. This clearly indicates better and consistent performance of MF-PLP in comparison with PLP. FIG.6 shows the comparative performance of perceptual features in developing speaker identification/verification system. Performance of the Speaker verification system is measured in terms of %FAR (False acceptance rate) and %FRR (False rejection rate).

Among these perceptual features, MF-PLP gives better overall accuracy of 91% for identifying speakers. It also gives low values of %FRR, %FAR and %EER. So, MF-PLP is better feature for both speech and speaker recognition.

TABLE 3 gives the details of the experimental evaluation of the features in this work and their corresponding F-ratios for composite speaker identification/verification system.

Features	% Identification accuracy	%FAR	%FRR	%ERR	F-ratio
MF-PLP	91	7.89	10.22	9	0.0206
PLP	88	10.618	12.608	11.5	0.0184

TABLE 3 – Overall accuracy of features for identification and verification with f-ratio

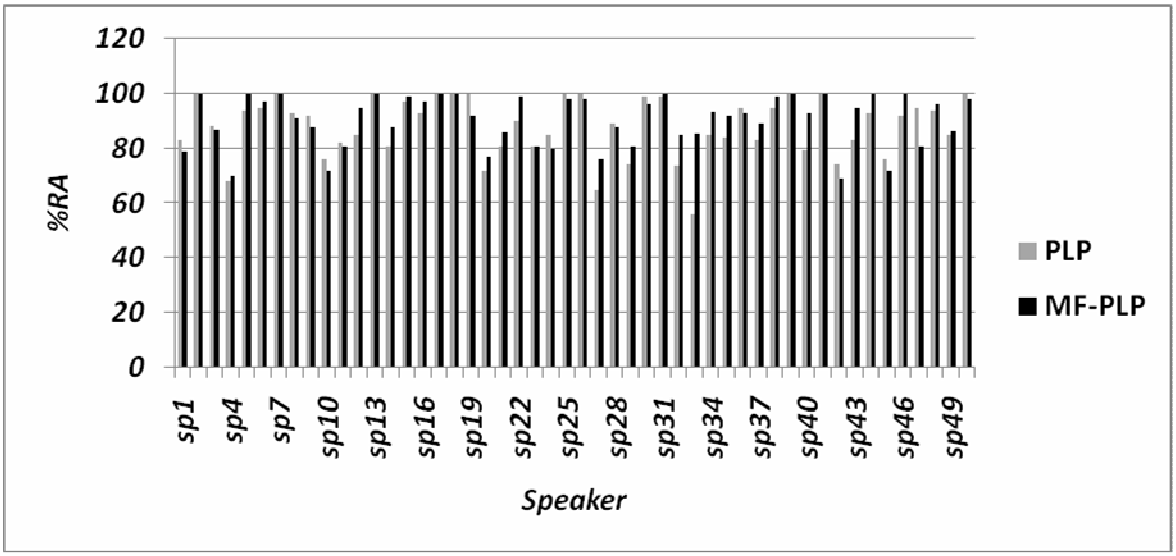

FIG.5 – Comparison chart – individual accuracy of PLP and MF-PLP for speaker identification system

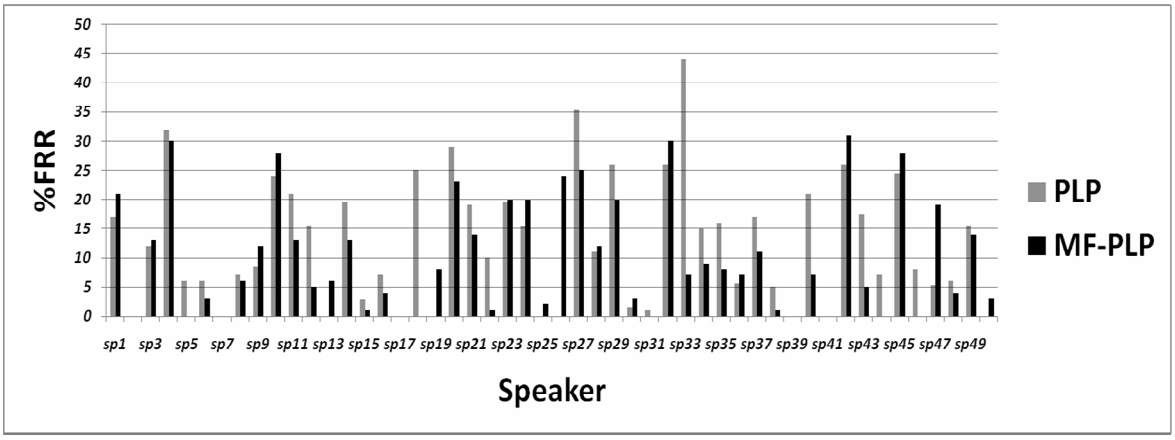

FIG.6 – Comparison chart – individual %FRR of PLP and MF-PLP for speaker verification system

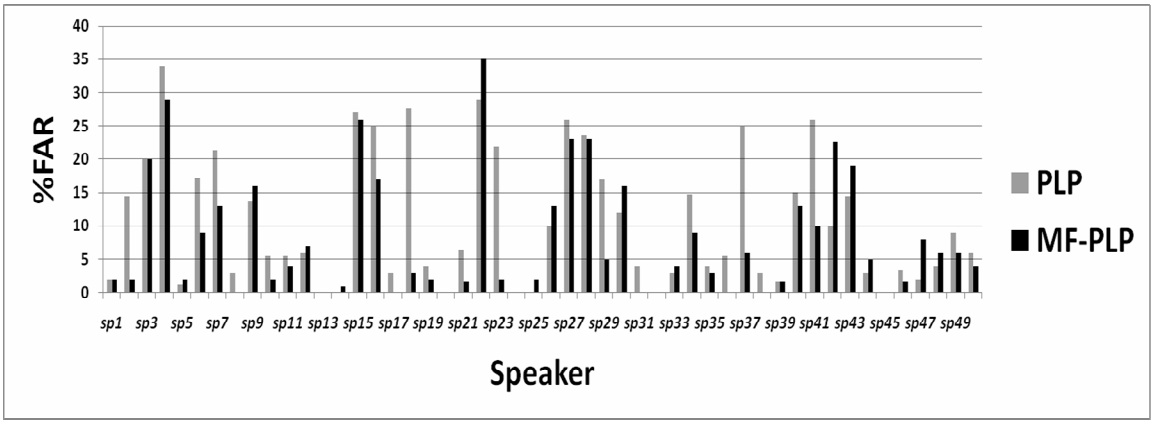

FIG.7 – Comparison chart – individual %FAR of PLP and MF-PLP for speaker verification system

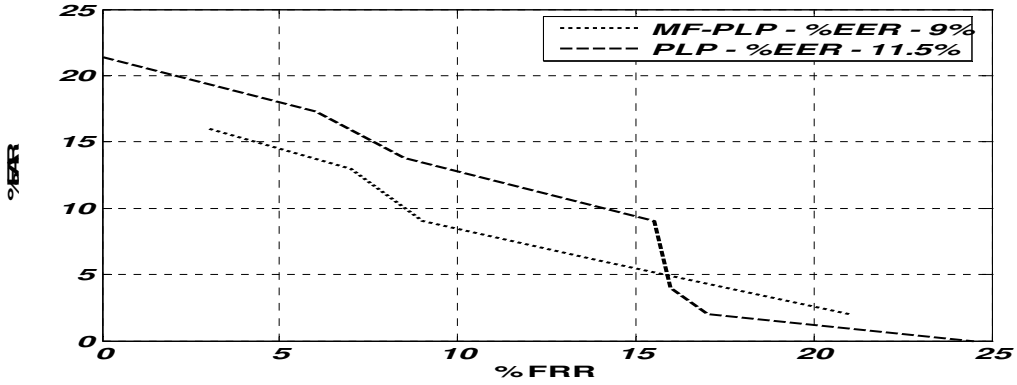

FIG.8 – Detection error trade off curves for PLP and MF-PLP

7. Statistical analysis of MF-PLP

The weighted average accuracy of MF-PLP is obtained as 91%. The better accuracy of MF-PLP is analysed using χ^2 distribution. There has been an average of more than 50 test speeches for each speaker and this is referred as expected frequency. The number of correctly identified test speeches for each speaker has been referred as observed frequency. The set of 50 enrolled speaker models has been taken as 50 attributes. Since the sample size is greater than 50, χ^2 distribution is applied to test the significance of the feature.

On the basis of the weighted average, correctly identified test speech segments for most of the speakers are more than 90%. Hence, hypothesis is set as under:

H_o : Weighted average accuracy is equal to or greater than 95%
H_1 : Weighted average accuracy is less than 95%

χ^2 test is applied at 5% level of significance. $\chi^2_{calculated}$ and $\chi^2_{0.05}$ values are 58.04 and 67.505 respectively. Since calculated value is less than the table value, null hypothesis is

accepted. Thus, weighted average accuracy obtained experimentally for MF-PLP feature is statistically justified.

8. Conclusions

This paper proposes robust perceptual features and iterative clustering approach for isolated digits and continuous speech recognition & speaker recognition and its evaluation on clean test speeches. PLP and MF-PLP are the proposed perceptual features considered for evaluation of the system performance. VQ codebook of size M = L/10 is formed to represent the L vectors of training data, thus achieving the reduction in the size of the data to be used subsequently while evaluating the test data in recognizing speech/speaker. Perceptual based features perform well in developing robust speech /speaker recognition system, because they inherently depict the perceptually important characteristics of the speech. Procedure used for speech / speaker recognition is same except the formation of training speech. It is found that MF-PLP performs better than PLP for both speaker independent isolated digits recognition and continuous speech recognition for speeches from TI Digits_1, TI Digits_2 and TIMIT databases. This feature also provides better results for speaker identification and verification in terms of better weighted average accuracy, low values of %FAR, %FRR and %EER for the test speeches considered to be identical messages for all the speakers. The noteworthy feature in this work is theoretical validation of good experimental results by using F-ratio on training data for both speech & speaker recognition and statistical validation of results for speaker recognition. Perceptual features can be in general used for both speech and speaker recognition. Another important point is that this speaker recognition system is evaluated on the identical messages for all the enrolled speakers and it reveals that perceptual features indicate the characteristics of speaker rather than the spoken content due to the formation of training speech for the purpose of text independent speaker recognition. These perceptual features also depict the characteristics of speech due to the formation of training speech in the case of speaker independent speech recognition

References

[1].A.Revathi, R.Chinnadurai & Y.Venkataramani, “T-LPCC and T-LSF in twins identification based on speaker clustering” , Proceedings of IEEE INDICON, IEEE Bangalore section, pp.25-26.September 2007.

[2].Hynek Hermansky, Kazuhiro Tsuga, Shozo Makino and Hisashi Wakita, “Perceptually based processing in automatic speech recognition”, Proceedings of IEEE International Conference on Acoustics, Speech And Signal Processing, Vol.11, pp.1971-1974, April 1986, Tokyo.

[3].Hynek Hermansky, Nelson Margon, Aruna Bayya and Phil Kohn, “The challenge of Inverse E: The RASTA PLP method”, Proceedings of Twenty Fifth IEEE Asilomar Conference on Signals, Systems And Computers, Vol.2, pp.800-804, November 1991, pacific Grove, CA, USA.

[4].Hynek Hermansky and Nelson Morgan, “RASTA processing of speech”, IEEE Transactions on Speech And Audio Processing, Vol.2, No.4, pp.578-589, October 1994.

[5]. A.Revathi and Y.Venkataramani, :”Text independent speaker identification/verification using multiple features”, Proceedings of IEEE International Conference on Computer Science And Information Engineering, April 2009, Los Angeles, USA.

[6]. A.Revathi and Y.Venkataramani, “Iterative clustering approach for text independent speaker identification using multiple features”, Proceedings of IEEE International Conference on Signal Processing And Communication Systems, December 2008, Gold coast, Australia.

[7]. A.Revathi and Y.Venkataramani, "Use of perceptual features in iterative clustering based twins identification system", Proceedings of IEEE International Conference on Computing, Communication and Networking, December 2008, India.

[8] A.Revathi, R.Chinnadurai and Y.Venkataramani, "Effectiveness of LP derived features and DCTC in twins identification-Iterative speaker clustering approach", Proceedings of IEEE ICCIMA, Vol.1, pp.535-539, December 2007.

[9]. Rabiner.L.& Juang B.H., "Fundamentals of speech recognition", Prentice Hall, NJ 1993.

[10].A.Revathi, R.Chinnadurai and Y.Venkataramani. "Use of wavelets in end point detection and denoising under low SNR constraints". International Journal of Systemic, Cybernetics And Informatics, vol.2, pp. 19-25, April 2007.

[11]. S.R.Das, W.S. Mohn, "A scheme for speech processing in automatic speaker verification", IEEE Transactions on Audio And Electroacoustics, Vol.AU-19, pp.32-43, March 1971.

[12].Aaron. E. Rosenberg, "New techniques for automatic speaker verification", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-23, No.2, pp.169-176, April 1975.

[13].Guruprasad S., Dhananjaya,. N, and B. Yegnanarayana, "AANN models for speaker recognition based on difference cepstrals", Proceedings of IEEE International Joint Conference on Neural Networks, Vol.1, pp.692-697, July 2003.

[14]. Tanveer A.Faruquie, Abhik Majmudar, Nitendra Rajput and L.V.Subramanian, "Large vocabulary audio-visual speech recognition using active shape models", Proceedings of 15th IEEE International Conference on Pattern recognition, Vol.3, pp.106-109, July 2000.

[15]. Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go and Chungyong Lee, "Optimizing feature extraction for speech recognition", IEEE Transactions on Speech and Audio Processing, Vol.11, No.1, January 2009.

[16]. P.C.Woodland, J.J.Odell, V.Vatchev and S.J.Young, "large vocabulary continuous speech recognition using HTK", Proceedings of IEEE International Conference on Acoustics, Speech and signal processing, Vol.2, pp.125-128, April 1994.

[17]. Hui Jiang, Xinwei Li and Chaojun Liu, "Large margin hidden markov models for speech recognition", IEEE Transactions on Audio, Speech and Language Processing, Vol.14, No.5, pp. 1584-1595,September 2006.

[18]. James Mc Auley, Ji Ming, Daryl Stewart and Philip Hanna, "Sub band correlation and robust speech recognition", IEEE Transactions on Speech and Audio Processing, Vol.13, No.6, pp.956-964, September 2005.

[19]. E.Avci and D.Avci, "The speaker identification by using genetic wavelet adaptive network based fuzzy inference system", International Journal on Expert Systems with Applications, Vol. 36, No.6, pp. 9928-9940, August 2009.

[20]. Prateek Agarwal, Anupam Shukla and Ritu Tiwari, "Multilingual speaker recognition using artificial neural network", Advances in Computational Intelligence, pp.1-9, 2009.

[21]. Waleed H.Abdulla, "Robust speaker modeling using perceptually motivated feature", Pattern Recognition letters, pp.1333-1342, August 2007.

[22]. Chunyan Xu, Xianbao Wang and Shoujue wang, "Research on Chinese digit speech recognition based on multi weighted neural network", Proceedings of IEEE Pacific-Asia workshop on Computational Intelligence and Industrial Applications, pp.400-403, 2008.

[23]. Anastasis Kounoudes, Antonakoudi, Vasili Ketatos and Phillippos Peleties, "Combined speech recognition and speaker verification over the fixed and mobile telephone networks", Proceedings of 24th

IASTED International Conference on Signal Processing, Pattern Recognition and Applications, pp.228-233, 2006.