Speaker Independent Isolated Tamil Words for Speech Recognition using MFCC, IPS and HMM

K.Murali Krishna, M.Vanitha Lakshmi

Abstract— The process of converting an acoustic waveform into the text resembling the information, conveyed by the speaker is termed as speech recognition. Nowadays, normally Hidden Markov Model (HMM) based speech recognizer with Mel Frequency Cepstral Coefficient (MFCC) feature extraction is used. The proposed speech feature vector is generated by projecting an observed vector onto an Integrated Phoneme Subspace (IPS) based on Independent Component Analysis (ICA) or Principal Component Analysis (PCA). The performance of the new feature has to be evaluated for Isolated Tamil Word Speech Recognition. The proposed method is expected to provide higher recognition accuracy than conventional method in clean environment.

Index Terms— Hidden Markov Model Tool Kit (HTK), recognition accuracy, dimensionality reduction, linear transformation matrix.

1 INTRODUCTION

SPEECH is the most natural means of information exchange among human beings. Communication is done by speech production and perception. To communicate with a machine one requires interfaces like keyboard, mouse and screen etc., operated with the help of software. A simple alternative to these hardware interfaces is an Automatic Speech Recognition (ASR) system [1]. ASR system consists of an acoustic front-end and a statistical pattern classifier. The front-end enables the appropriate feature extraction. The chosen feature vectors should contain only the relevant information enabling the unambiguous classification into individual phonetic classes considering the vector dimension. Speech recognition systems for regional languages spoken in different countries with rural background and low literacy rates appear to be still evolving. Speech recognition systems can be characterized by speaking style, speaking mode, environment, vocabulary, acoustic model and language model. Feature extraction for speech recognition points at an efficient representation of spectral and temporal information of nonstationary speech signals.

This feature extraction stage transforms the input speech waveform in a series of low-dimensional vectors, each comprising a short segment of the acoustical speech input to reduce the computational demands of the HMM classifier. Some examples of common speech features are: LPC, PLP and MFCC. In the present work, MFCC features are more commonly used for speech recognition. The MFCC features are obtained by applying Discrete Cosine Transform (DCT) to Logarithm of Mel Filter bank Energies (LMFE).

- *K.Murali Krishna is currently pursuing Master's Degree programme in Communication Systems in S.A. Engineering College affiliated to Anna University, Chennai, India. E-mail: muralik87@gmail.com*
- *M.Vanitha Lakshmi is currently working as Assistant Professor in the ECE Department of S.A. Engineering College affiliated to Anna University, Chennai, India. E-mail: vanithahitesh08@yahoo.co.in*

There are many techniques available that improve MFCC features from different aspects. Some works try to make MFCC more robust to channel and additive noises using weighting or compression of Mel sub-band energies [2] [3][4]. In other group of methods, we try to overcome the disadvantages of DCT in clean or noisy conditions [5][6]. DCT is a non-adaptive procedure that projects LMFE in the direction of global variance which achieves only partial decorrelation of features.

In order to get over the partial decorrelation, several methods have been proposed to take the place of DCT and decorrelate LMFE. Some examples of such methods are: Principal Component Analysis (PCA) [5][6][7] and Independent Component Analysis (ICA) [6]. PCA, based on the principle of minimum reconstruction error, projects the data (LMFE) in the direction of maximum variability [5][6]. But, there is no guarantee that variability given by PCA is useful for speech recognition. While PCA removes the second order dependencies of the feature vector components, ICA removes also higher order dependencies and minimizes the mutual information between the feature vector components. ICA is the linear and a supervised dimensional reduction algorithm. Linear Discrimination Analysis (LDA) tries to separate classes using linear hyper planes. Such basis vectors are sought which try to maximize the linear class separation. Integrated Phoneme Subspace (IPS) is a transformation based on combining PCA and ICA which provides higher accuracy than existing methods and tries to incorporate phonemic information into the feature space.

2 SPEECH FEATURE EXTRACTION

2.1 Mel Frequency Cepstral Coefficients (MFCC)

Technique of evaluating MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is obtained. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to reduce the discontinuities of a signal. Then Discrete Fourier Transform will be used to generate the Mel filter bank. According to Mel

frequency warping, the width of the triangular filters gets modified and so the log total energy in a critical band around the center frequency is included. The warping process gives rise to the numbers of coefficients at the end. Finally the Discrete Fourier Transform in the inverse mode is used for the calculation of the cepstral coefficients. It converts the log of the frequency domain coefficients to the frequency domain where N is the length of the DFT. The most remarkable downside of using MFCC is its sensitivity to noise due to its dependence on the spectral form. Methods that utilize information in the periodicity of speech signals could be used to get over this problem. The steps required for MFCC feature extraction are explained clearly using Fig. 1. MFCC can be computed using (1).

$$\text{Mel (f)} = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

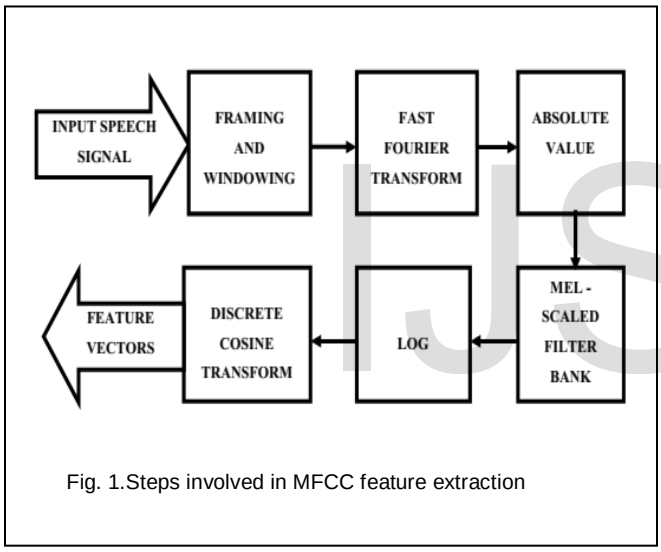

Fig. 1.Steps involved in MFCC feature extraction

2.2 Principal Component Analysis (PCA)

Principal Component Analysis is a method to display the data in the low-dimensional subspace. The corresponding projection matrix is called Karhuneu-Loeve Transform (KLT). The reconstruction error will be the smallest possible among linear transformations, when the original feature vectors are subjected to a lower-dimensional linear subspace using KLT. In the original feature space and in the projection space, the reconstruction error is measured as the mean-square error between the data vectors. The rows of the D′ X D KLT transformation matrix consist of the D′ eigenvectors corresponding to the D′ largest eigenvalues of the covariance matrix of the training data. These eigenvectors are the principal axes of the data set. KLT decorrelate the feature vectors, which strengthen modelling the data with diagonal Gaussians.

Suppose that $x_1, x_2, \ldots, x_M$ are N x 1 vectors. Then, according to [8] and [9], PCA can be performed in the following steps.

1. Firstly, the sample mean vector $\bar{x}$ as the average of the elements of the input column vectors is evaluated.

$$\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i \quad (2)$$

2. For PCA to work properly, the computed mean has to be subtracted from each input vector. The new subtracted vectors then are:

$$\Phi_i = x_i - \bar{x} \quad (3)$$

This step produces a data set whose global mean is zero.

3. Then, the N x M centered matrix A is created from the subtracted vectors:

$$A = [\Phi_1\ \Phi_2 \ldots\ldots \Phi_M] \quad (4)$$

4. Next, the covariance matrix C of the matrix A is obtained as follows:

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi^T{}_n = \frac{1}{M}\sum_{n=1}^{M} (x_i - \bar{x})\,(x_i - \bar{x})^T = AA^T \quad (5)$$

The covariance matrix is a symmetric N x N square matrix with all the elements are real.

5. In the subsequent step, the eigenvalues $\lambda_i, i \in \langle 1; N \rangle$ of the covariance matrix are computed. It is necessary to rank them in decreasing order as given below:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_N \quad (6)$$

6. Then, the eigenvectors $u_1, u_2, \ldots, u_N$ corresponding to their eigenvalues are determined. Since the matrix C is symmetric, the eigenvectors form a linear basis. This means that any vector x or actually $(x - \bar{x})$ can be expressed as a linear combination of the eigenvectors):

$$(\mathrm{x} - \bar{\mathrm{x}}) = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i \quad . \tag{7}$$

7. The dimensionality reduction step is executed by keeping only the eigenvectors corresponding to the K largest eigenvalues:

$$(\hat{\mathrm{x}} - \bar{\mathrm{x}}) = \sum_{i=1}^{K} b_i u_i \ , where\ K \ll N. \quad . \tag{8}$$

The representation of $\hat{\mathrm{x}} - \bar{\mathrm{x}}$ into the basis $\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_K$ is thus $\mathrm{Y} = [\,\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_K]^{\mathrm{T}}$.

8. Finally, the linear transformation $\mathrm{R}_N \rightarrow \mathrm{R}_K$ is computed as multiplication of the following matrices:

$$\mathrm{Y} = U^T (\mathrm{x} - \bar{\mathrm{x}}) \quad . \tag{9}$$

Where Y represents the transformed data and U^{T} is the PCA transformation matrix.

2.3 Independent Component Analysis (ICA)

The idea behind the Independent Component Analysis in the feature extraction is to minimize the redundancy of the original feature vector components. While PCA eliminates the second order dependencies of the features vector components, ICA removes also higher order dependencies (reduces the mutual information between the feature vector components).

The data model of the linear ICA is x = As, where x is the original feature vector is, s is the vector of the underlying (independent) sources, and A is a mixing matrix. Only x is observed, and the goal is to estimate both A and s trying to find the sources s which are statistically independent.

The column vectors of the mixing matrix correspond to the building blocks of the data in the generative model. When the mixing matrix A has been estimated from the training data, the transformation matrix for obtaining a new feature representation is its inverse. $\mathrm{W} = \mathrm{A}^{-1}$. When data vector x is projected to the row vectors of W, the components of the new feature represent the activations of the source s. ICA representation is usually sparse, i.e., only few sources are active at the same time. Fast fixed-point algorithm was used for computing the ICA basis and maximizes negentropy [10].

The FastICA algorithm for finding one w that derives one independent component is as follows.

1. Center the data to make its mean zero.
2. Whiten the data to give z.
3. Choose an initial (e.g., random) vector W of unit norm.
4. $$\text{Let } \mathrm{W} \longleftarrow \mathrm{E}\{\mathrm{z}\, g(\mathrm{W}^{\mathrm{T}}\mathrm{z})\} - \mathrm{E}\{g'(\mathrm{W}^{\mathrm{T}}\mathrm{z})\}\,\mathrm{W},$$

where g is the function that gives approximation of negentropy.

5. $$\text{Let } \mathrm{W} \longleftarrow \frac{\mathrm{W}}{\|\mathrm{W}\|}$$

6. If it is not converged, go back to step 4.

Once W is estimated, the final step is to project the signal into the space created by ICA.

New dataset = $\mathrm{W}_{\mathrm{ica}}$ * Mean Adjusted original data,

Where, $\mathrm{W}_{\mathrm{ica}}$ is the transformation matrix obtained from FastICA algorithm.

3 ACOUSTIC MODELLING

In the Markov chain, each state corresponds to a deterministically observable event; i.e., the output of such sources in any given state is not random. A natural extension to the Markov chain introduces a non-deterministic process that generates output observation symbols in any given state. Thus, the observation is a probabilistic function of the state. This new model is known as a Hidden Markov Model (HMM), which can be seen as a double-embedded stochastic process with an underlying stochastic process (the state sequence) not directly observable. This underlying process can only be probabilistically connected with another observable stochastic process generating the sequence of features one can observe.

Given the definition of HMMs above, three basic limitations of interest must be highlighted before they can be applied to real-world applications.

1. **The Evaluation Problem –** Given a model Φ and a sequence of observations$\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_T)$, what is the probability $\mathrm{P}(\mathrm{X} \mid \Phi)$; i.e., the probability of the model that generates the observations?
2) **The Decoding Problem –** Given a model Φ and a sequence of observations $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_T)$,, what is the most likely state sequence $\mathrm{S} = (\mathrm{s}_0, \mathrm{s}_1, \ldots, \mathrm{s}_T)$ in the model that produces the observations?
3) **The Learning Problem –** Given a model Φ and a set of observations, how to change the model parameter $\hat{\Phi}$ to maximize the joint probability (likelihood) ? $\prod_x \mathrm{P}(\mathrm{X} \mid \Phi)$?

The Acoustic modelling phase handles, the creation and updating of the statistical model used to model the sound

wave observations. The statistical model employed is the HMM. The final goal is to create a single HMM for each word in the recognition vocabulary. The type of HMM used would be simple left to right with a fixed number of states. The first and the last state would be assumed to be non- emitting, which means while middle states are emitting these states could either jump to another state but not both.

After the creation of the HMM skeleton, its parameters would be estimated from the extracted features of the waveform (feature file). Now the feature vectors set X contains the MFCC, delta, and acceleration coefficients with the energy. These vectors will be used to estimate the parameters of the HMM.

4 EXPERIMENTS AND RESULTS

4.1 Speech Corpus

The data collected for training are 5 basic elements of nature as shown in Table 1. uttered by 3 speakers in a normal room with minimal external noise. Each element's name is uttered 9 times by each speaker on the basis of 3 utterances per file using wavesurfer tool. Then feature extraction and speech recognizer carried out using MatLab and HTK tool [11].

4.2 Processing

The following steps are carried out to build simple isolated word recognition with whole word models using HTK [12]:

TABLE 1
SAMPLE ELEMENTS CONSIDERED FOR THE RECOGNITION

Elements Name in Tamil	Elements Name in English
ஆகாயம்	SKY
காற்று	AIR
நீர்	WATER
நெருப்பு	FIRE
நிலம்	LAND

1. Constructing the grammar and word network (HParse)
2. Constructing a dictionary for the models (HDMan)
3. Extracting the features (MATLAB or HCopy)
4. Training the Acoustic Model (HERest)
5. Evaluating the recognizer result from the test data (HVite)
6. Reporting recognition result (HResults)

4.3 Recognition Results

The performance of the system is tested against speaker independent parameter by using two types of speakers: one who is involved in training and the other involved in testing. The experiment set details used for the recognition using MFCC feature extraction is shown in the Table 2.

The average performance of the system using MFCC feature extraction lies in the range of 87% to 88% with word error rate 12% to 13%. The Isolated Tamil Word Speech Recognition was done using HTK for the sample data of Elements. Figure 2 shows the graphical representation of the performance of the speech recognizer for the given Set A using MFCC. The word nilam (Land) is substituted by neer (Water) and the word kaatru (Air) is substituted by aagayam (Sky).

TABLE 2
RECOGNITION IN NORMAL ROOM ENVIRONMENT BY SPEAKERS 1, 2 AND 3

SET	Speakers used in Training	Speakers used in Testing	Recognition Accuracy (%)	Word Error Rate (%)
A	1,2	3	86.67	13.33
B	2,3	1	88.89	11.11
C	1,3	2	86.67	13.33

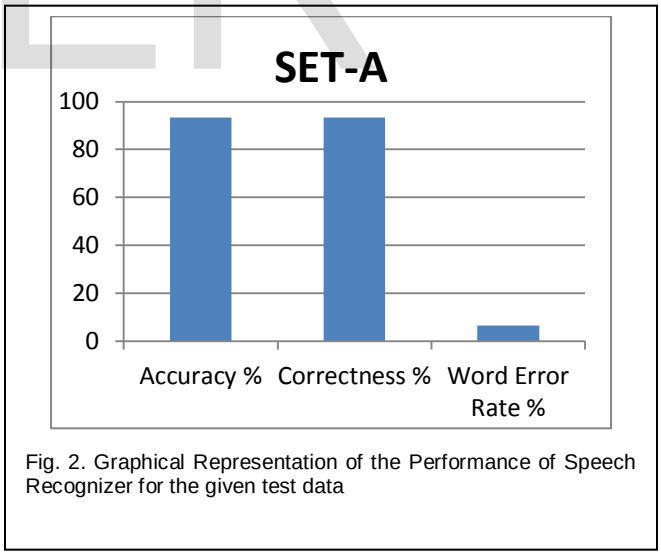

Fig. 2. Graphical Representation of the Performance of Speech Recognizer for the given test data

Similarly, the following Table 3 details about the performance of recognizer using PCA.

TABLE 3
RECOGNITION IN NORMAL ROOM ENVIRONMENT BY SPEAKERS 1, 2 AND 3

SET	Speakers used in Training	Speakers used in Testing	Recognition Accuracy (%)	Word Error Rate (%)
A	1,2	3	93.33	6.67
B	2,3	1	88.89	11.11
C	1,3	2	91.11	8.89

5 CONCLUSION

In this paper, feature extraction is carried out using MFCC and IPS. HMM is used in speech modelling. The process is carried out in clean environment with isolated words of limited vocabulary. Errors are introduced by the recognizer and corresponding accuracy have been measured. The performance of PCA speech recognizer seems to give more recognition accuracy than MFCC.

The above work can be carried out in noisy environment and the performance of the feature extraction process can be tested.

REFERENCES

[1] R. Rabiner, and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall International, New Jersey, 1993.

[2] B Nasersharif,., A Akbari,.,"A Framework for Robust MFCC Feature Extraction Using SNR-Dependent Compression of Enhanced Mel Filter Bank Energies", International Conference on Spoken Language Processing(ICSLP), pp. 33-36, 2006.

[3] D. Zhu, S. Nakamura, K.K. Paliwal, R. Wang, "Maximum likelihood sub-band adaptation for robust speech recognition", Speech Communication, Vol. 47, Iss. 3, pp. 243-264, 2005.

[4] H.Y. Cho, Y.H. Oh, "On the use of channel-attentive MFCC for robust recognition of partially corrupted speech", IEEE signal processing letters, Vol.11, No. 6, pp. 581-584, 2004.

[5] P. Somervuo, "Experiments With Linear And Nonlinear Feature Transformations In HMM Based Phone Recognition" IEEE Int. Conf. on Acoustics, Speech and Signal processing, vol. I, pp. 52-55, 2003.

[6] I. Potamitis, N. Fanotakis, G. Kokkinakis, "Spectral and cepstral projection bases constructed by independent component analysis", International Conference on Spoken Language Processing(ICSLP), vol.3, pp. 63-66, 2000.

[7] I. T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.

[8] G. Bebis, Principal Components Analysis. University of Nevada

[9] L. I. Smith, A tutorial on Principal Components Analysis. 2002

[10] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, JohnWiley & Sons, 2001, http:/www.cis.hut.fi/projects/ica/fastica/index.shtml

[11] Steve Young et al, 2006, The HTK Book. Version 3.4

[12] Akila, A & Chandra, E 2013, 'Isolated Tamil Word Speech Recognition System Using HTK', International Journal of Computer Science Research and Application (IJCSRA), vol. 3, no. 2, pp. 30-38.