ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

Feature Extraction and Dimensionality Reduction using IPS for Isolated Tamil Words Speech Recognizer

K.MURALI KRISHNA[1], M.VANITHA LAKSHMI[2], S.SATHIYA LAKSHMI[3]

PG Scholar, Department of PG Studies in Engineering, S.A.Engineering College, Chennai, India[1,3]

Assistant Professor, Department of PG Studies in Engineering, S.A.Engineering College, Chennai, India[2]

Abstract: Automatic Speech Recognition (ASR), is the process of converting a speech waveform into the text quite similar to the information being communicated by the speaker. This paper aims to construct a speech recognition system for Tamil language. Mel Frequency Cepstral Coefficients (MFCC) is a commonly used feature extraction technique for speech recognition which is computed by applying DCT to the mel-scale filter bank output. Instead of DCT, Integrated Phoneme Subspace (IPS) method is used to improve speech recognition. The experimental results show that the recognition accuracy of ASR using IPS in various forms yields higher or similar output comparative to MFCC and the word accuracy of one such form of IPS (IPS-2) is 84.00%.

Index Terms: Linear transformation method, Hidden Markov Tool Kit (HTK), cepstrum, Hidden Markov Model (HMM).

I. INTRODUCTION

Speech features used for speech recognition are spectral envelope of speech, speech formants, pitch, Linear Predictive Coefficients (LPC), Mel- Frequency Cepstral Coefficients (MFCC) etc. Speech recognition is broadly classified into Speaker-dependent and speaker-independent systems. Out of which, speaker-dependent system gives more accuracy compared to speaker-independent. The reason behind is higher variance and broader probability distributions that involves different speakers. Hence speaker adaption methods are required to improve speaker-independent system.

The two prominent speech features namely spectrum and cepstrum are discussed below. Spectrum is obtained by applying Fourier Transform to the measured signal. When DCT is performed over log spectrum, cepstrum is obtained. Spectrum captures local aspects of speech. Cepstrum identifies excitation frequency of glottis. This measured signal is the product of original signal excited from glottis and impulse response of channel namely vocal tract. MFCC's base feature is Cepstrum and IPS's base feature is log spectrum or Logarithm of Mel Filter bank Energies (LMFE). MFCCs are not very noise robust which is the main deficiency. Hence, researchers go for hybrid speech extraction methods to improve accuracy in noisy environment.

II. ACOUSTIC FEATURE EXTRACTION

For a given input signal, feature extraction phase contributes feature vectors that are inputted to the next stage, recognizer. The three steps involved in feature extraction are speech analysis, compiling extended feature vector and transformation. Speech analysis produces raw features with information about the envelope of the power spectrum. Extended features are static and dynamic features. Transformation is carried out from extended feature vectors to robust, compact vectors. Sampling and quantization are done which converts analog to digital and eliminates noise from acoustic samples. The following are the two feature extraction techniques taken for the present work.

A. Mel Frequency Cepstral Coefficients (MFCC)

MFCCs [1] are essential for speech recognition on the basis of sensitivity towards human auditory system. The lower order MFCCs refer to the slowly varying spectral envelope and higher order MFCCs details the fast changes of the spectrum. The Mel frequency scale consists of two distinct ranges namely, linear spacing less than 1 KHz and logarithmic spacing greater than 1 KHz. The transformation of linear frequency to Mel frequency is done using (1)

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

MFCCs are computed through the following steps namely, framing, windowing, DFT, Mel filter bank and IDFT. Framing performed with frame width equals 25ms and frame periodicity equals 10ms. For windowing step, Hamming window is used. DFT is done via FFT for fast computation yields N samples whose value equals to be power of 2 say , N = 256. Filter bank functions are smoothing the spectrum and reducing 256 Fourier coefficients to 26 Filter bank outputs.

To overcome the issues created by noise and reverberation and in need of better recognition accuracy, the next method called Integrated Phoneme Subspace (IPS) technique is carried out for this present work.

B. Integrated Phoneme Subspace (IPS)

The methodologies used in IPS are two feature extraction methods namely Principal Component Analysis (PCA) and Independent Component Analysis (ICA) in different forms. This IPS method gathers statistically, mutually

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

independent information among phonemes by projecting input vector onto integrated space using PCA or ICA or both. This paper's experimental result is based on whole word model and not phoneme based model. The detailed working of IPS is explained in [2].

1) Principal Component Analysis (PCA)

PCA is well known for feature extraction in image processing and dimensionality reduction [3, 4]. The main objective of PCA is to direct the Principal Component in the direction of largest variance in the information. PCA algorithm centres the data by subtracting the mean and fixes direction of maximum variance as principal component.

It then steps in placing the axis in the direction which is orthogonal to the first axis and has maximum variance. After much iteration, orthogonal basis vectors are created for the given dataset. If there are little variance or noise, these axis are ignored.

PCA through Singular Value Decomposition (SVD) calculates eigen vectors and transforms high dimensional data into lower dimensional space that are uncorrelated and orthogonal. Let 'p' denotes the dimension for a given data set 'x' and 'm' be number of principal axes where, $1 \leq m \leq p$. Let 'T_m' be m th orthonormal axes to which maximum variance is preserved in projected space. The principal axes 'y' are obtained by leading eigen vectors of the global covariance matrix S.

The following relation is the linear transformation matrix performed by PCA in (2)

$$y = [y_1, y_2 \ldots\ldots, y_m] = [\, T_1^T x, \ldots\ldots, T_m^T x] \quad (2)$$
$$y = T^T x$$

2) Independent Component Analysis (ICA)

ICA is a statistical method evolved after ICA which represents a multidimensional random vector as a linear combination of non-gaussian random variables that are as independent as possible. ICA involves whitening and rotation. Whitening means PCA plus scaling, the covariance matrix of the whiten data is the identity. And PCA, uncorrelated data, the covariance matrix of the PCA transformed data has the eigenvalues in its diagonal.

Rotation concept here maximizes the non-gaussianity of the projections. FastICA is a free MATLAB program that implements the fast- fixed point algorithm. The two standard measure of non-gaussianity are kurtosis and negentropy. Negentropy is based on the information theoretic quantity of differential entropy. Kurtosis is the fourth-order cumulant. Kurtosis for gaussian random variable is zero.

Thus PCA utilizes the first and second moments of the measured data, hence relying heavily on Gaussian features. While ICA exploits inherently non- gaussian features of the data and employs higher moments.

The steps required for IPS feature extraction are explained clearly using Fig. 1.

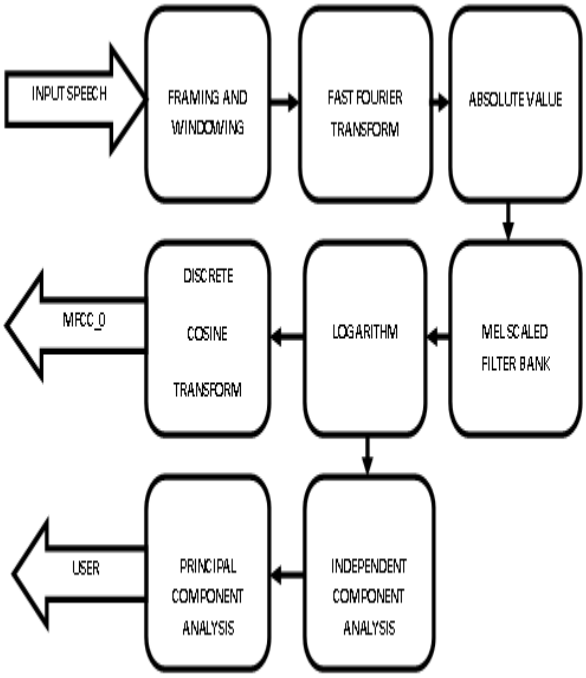


Fig.1. Steps involved in IPS feature extraction

The feature extracted are just static, hence delta coefficients can be concatenated along with the projected feature vectors. Hence, USER_D_A target feature extracted and send to modelling stage.

III. HMM BASED SPEECH MODELLING

HMMs are used for speech modelling [5]. When an observer able to view only output symbol sequence where the sequence of states is hidden, such a model called Hidden Markov Model (HMM).

The probabilistic parameters of a HMM are hidden states, observations, probability distribution of the starting state, transition probabilities and emission probabilities. In speech recognition, words are the hidden state sequence and acoustic signal is the observation sequence.

To have tractable computation, few postulates are made in HMM. Firstly, Markovian means present state is dependent only on the past state. Secondly, independency means current observation is statistically independent of the previous observations. Thirdly, stationarity means state transition probabilities are independent of the actual time at which the transitions take place.

HMM training aims to assign extracted feature vectors to HMM state. The two well known algorithms used for speech modelling are forward-backward training and Viterbi training.

The former one assigns a probability to each feature vector emitted from each hidden state and latter one assigns a feature vector to a particular state. The basic HMM definition for user defined feature extracted through IPS is shown in Fig. 2.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

```
~o <VecSize> 39 <USER_D_A>
~h "proto"
 <BeginHMM>
                <NumStates> 5
      <State> 2
                <Mean> 39
                 0.0 0.0 0.0 ..... 0.0 0.0 0.0 0.0
                <Variance> 39
                 1.0 1.0 1.0 .....1.0 1.0 1.0 1.0
      <State> 3
           <Mean> 39
                 0.0 0.0 0.0 ..... 0.0 0.0 0.0 0.0
                <Variance> 39
                 1.0 1.0 1.0 .....1.0 1.0 1.0 1.0
       <State> 4
            <Mean> 39
                 0.0 0.0 0.0 ..... 0.0 0.0 0.0 0.0
                <Variance> 39
                 1.0 1.0 1.0 .....1.0 1.0 1.0 1.0
        <TransP> 5
          0.0 1.0 0.0 0.0 0.0
          0.0 0.6 0.4 0.0 0.0
          0.0 0.0 0.6 0.4 0.0
          0.0 0.0 0.0 0.7 0.3
          0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Fig.2. A simple left-right HMM

IV. SIMULATION AND RESULTS

In this section, Tamil isolated words speech recognition were build and the performance of the extracted features are tested via recognizer.

A. System Description

The speech recognizer is developed using Hidden Markov Model Tool Kit (HTK) on the Windows platform. HTKV3.4 and Cygwin64 terminal were used. In the first step, HTK training tools are used and HMM parameters are estimated using training utterances and their related transcriptions. Second step is to transcribe unknown utterances using HTK recognition tools [6]. The system is trained with five Tamil words. To recognize the speech, Word level HMM is used.

B. Data Preparation

Training and testing phases require a set of utterances. Recording is done at normal living room environment. Sampling rate used at recording of 16000 Hz. Totally three speakers involved of which two purely used only in training and third one for testing. Speech files are recorded in .wav format using HTK command HSLab. Each speaker is asked to utter each word five times. Thus totally 75 ((3*5)*5) speech files forms the speech corpus. Of which, 50 files used for training and 25 used for testing in speaker independent mode.

C. Feature Extraction

Speech files are parameterized into a sequence of features. For MFCC, HCopy, a HTK command is used along with configuration file. For IPS, MATLAB program was written and feature files are generated. The acoustic parameters of 39 coefficients are extracted that, 13 static features and 26 derivatives as in case of IPS. While 39 MFCCs with 12 Mel Cepstrum plus log energy and delta and double delta derivates.

D. HMM Training

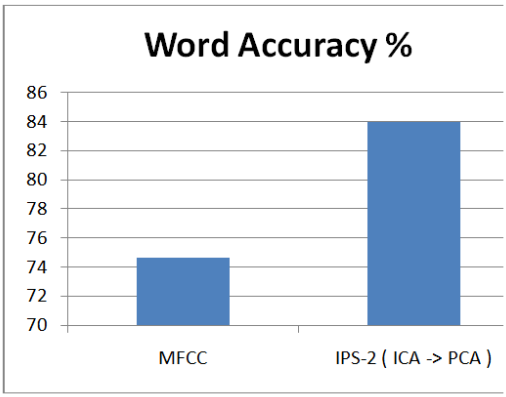
A prototype HMM model is used for initializing the word HMM models. Apart from vocabulary words, silent models namely sil and short pause are also included. In this simulation, five states HMM are used of which first and last are non-emitting states. HCompV and HERest HTK commands are used to globalise man and variance and also do embedded re-estimation for word model training.

E. Performance Evaluation

At this stage, recognizer generates transcription for testing utterance. HVite and HResults HTK commands are used for decoding and displaying recognition results. The experimental results for various forms of IPS along with MFCC are listed in the below Table 1.The recognition accuracies for MFCC and IPS-2 methods are shown in Fig. 3. For MFCC, word accuracy and word error rate of the recognizer are 74.67% and 25.33% respectively. For IPS-2, word accuracy and word error rate of the system are 84.00% and 16.00%.

TABLE I
RECOGNITION ACCURACY FOR VARIOUS FORMS OF IPS AND MFCC

Feature Extraction Name	Recognition Accuracy (%)
MFCC_0_D_A	74.67
IPS-1 [ICA (13 USER) → PCA (13+13+13 USER_D_A)]	76
IPS-2 [ICA (13+13+13 USER_D_A) → PCA (13+13+13 USER_D_A)]	84
IPS-3 [ICA (13 USER) → ICA (13+13+13 USER_D_A)]	76
IPS-4 [PCA (13+13+13 USER_D_A) → PCA (39 USER)]	80
IPS-5 [PCA (13 USER) → ICA (13+13+13 USER_D_A)]	74.67
IPS-6 [ICA (13+13+13 USER_D_A) → ICA (39 USER)]	74.67
IPS-7 [PCA (13+13+13 USER_D_A) → PCA (13+13+13 USER_D_A)]	74.67



Graphical Representation of the Performance of Speech Recognizer for given test data of MFCC and IPS Features Extraction

V. CONCLUSION

In this paper, a linear transformation method namely IPS used for feature extraction other than DCT, IPS is applied to the LMFE output. This method will project the prominent speech element onto lower order features, while noise element onto higher order ones. From the above recognition results, IPS works much better than MFCC in normal room environment.

To improve the recognition accuracy further, other than word model, sub word models can be used with large vocabulary both in normal and reverberant environments.

REFERENCES

[1] Mohit Dua, R. K. Aggarwal,Virender Kadyan and Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK", IJCSI International Journal of Computer Sciences Issues, vol. 9, issue 4, no. 1, July 2012.

[2] Hyunsin Park, Tetsuya Takiguchi and Yasuo Ariki," Integrated Phoneme Subspace Method for Speech Feature Extraction", EURASIP Journal on Audio, Speech, and Music Processing 2009, 2009:690451, 16, June 2009.

[3] G. W. Cottrell, Principal components analysis of images via back propagation, SPIE Proceedings in Visual Communication and Image Processing, Vol. 1001, 1988, pp. 1070-1077.

[4] I. T. Jolliff, Principal Component Analysis, Springer-Verlag, New York, 1986.

[5] Rajesh Kumar Aggarwal and M. Dave, " Acoustic modeling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309- 320, 2011.

[6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book,Microsoft Corporation and Cambridge University Engineering Department, 2009.