IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Comparative Study of Feature Extraction Techniques for Hindi Speech Recognition System on HTK-Toolkit

Malay Kumar

PhD Scholar, Department of Computer Science and Engineering, NIT Raipur, Chhattisgarh, India

Abstract: Hindi is a language of masses. It is the national language of India and widely popular in Indian sub-continent. It is important to develop interactive system which comprehend Hindi language. This article presents the implementation of Hindi speech recognition system. The systems have been developed on HTK-Toolkit V 3.4 in Linux environment (Ubuntu 10.04.3 LTS). The Hindi speech recognition system has been developed for 101-word vocabulary size. Each word is uttered for a number of times to capture all the acoustic variability's. The system has been developed in two parts namely front-end and back-end. Front-end part covers preprocessing and feature extraction while back-end covers acoustic modeling, language modeling and recognition.The comparative analysis shows that MFCC perform better in same training and testing conditions while PLP perform better in mismatch conditions while both the feature extraction techniques outperform LPCC.

Keywords: MFCC, LPCC, PLP, HTK, Hindi Speech Recognition Engine.

I. INTRODUCTION

In last fifty years, many speech recognition strategies have been given proposed and implemented. These strategies span many sciences, which includes signal processing, pattern recognition, artificial intelligence, statistics, information theory, probability theory, computer algorithms, psychology, linguistics, and even biology.

Automatic speech recognition has been matured markedly during this time. This is due in part to the increase in available computing power, and in part to more sophisticated modelling techniques. The introduction of the HMM in the 1970s [1-2], and a statistical framework for ASR, has proven the most successful approach to date, and is the basis for current state-of-the-art speech recognizers. Speech recognition system for Hindi language is developed using HTK-Toolkit. The system is developed for limited vocabulary size of 101 words. The developed system recognizes isolated as well as connected words and gives the output transcriptions in Hindi. The system is developed using two approaches, the firstly the system is model for original Hindi words and secondly the system is model for English transcribed words where the English transcribed system gives the output in Hindi text by using lookup table. In the comparative study of feature extraction module, three feature extraction techniques are compared namely, MFCC, PLP and LPCC.

II. SYSTEM ARCHITECTURE

The developed speech system mainly consists of two modules: front-end module and back-end module. Firstly, data preparation is carried out. All the words of the vocabulary are uttered a number of times. Since speech sound cannot be directly processed by speech recognition system because an acoustic signal is an analog signal. Acoustic signal have to be represented in a more compact and efficient form which is achieved using acoustic analysis. Back-End module is used to generate the system model which is to be used during testing [9-10].

A. Data Preparation

To implement a speech recognition system, a basic requirement is speech and text corpus. In this implementation work self-developed speech and text corpus is used. A unidirectional Sony microphone of 120 VA is used for the preparation of speech corpus. Data is collected using 4 people (3 males, 1 female). Recording is done using system command brec. The properties of data are: sample is taken at sampling rate of 16000 Hz, bit rate 16-bit and the file format is PCM .wav. HTK also supports .sig file format, but it has compatibility issues. This file format is only supported by HTK, while .wav file format is supported by many other recognition tools. Data is prepared for limited vocabulary size of 101 words. Each word is uttered for ten times, so that speech corpus can capture most of the acoustic variability's. Text corpus is prepared manually using wave surfer. It takes lots of human hours but manually prepared speech and text corpus produce better results, if it is prepared with proper precautions [7-8].

System vocabulary of 101 words

भारतएकमहानदेशहै | भारतकेप्रधानमंत्रीमनमोहनसिंहहै |
बाघभारतकाराष्ट्रीयपशुहै | मोरभारतकाराष्ट्रीयपक्षीहै |
दिल्लीभारतकीराष्ट्रीयराजधानीहै |
दिल्लीमुंबईकोलकातातथाचेन्नईमहानगरहैं

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

|मुंबईभारतकीआर्थिकराजधानीहै | गंगाभारतकीराष्ट्रीयनदीहै | मैंएकछात्रहूँ | मैएनआईटीकाछात्रहूँ बाघभारतकाराष्ट्रीयपशुहै | मोरराष्ट्रीयपक्षीहै | |रामसीताकेपतितथासीतारामकीपत्नीहै | मेरेपिताअध्यापकहैं | हिन्दीभारतकीराष्ट्रीयभाषाहै | चंडीगढ़एकसुंदरशहरहै | मेराभाईरामहै | मेरीबहनसीताहै | चंडीगढ़पंजाबऔरहरियाणाकीराजधानीहै | मेरेपिताएकशिक्षकहै |

Vocabulary: The system is developed on a limited size vocabulary. The system is developed for 101 words. System performs well for any combination of these words for making sentences.

B. Feature Extraction

The acoustic signal is cannot be directly processed by the computer. Hence pre-processing is carried out to convert the input acoustic speech signal into a form that can be processed by the recognizer. During pre-processing, firstly the speech input is converted into the digital form. To convert the speech signal into the digital form, sampling can be done at the rate of 8000 Hz to 48000 Hz or any other sampling rate which is supported by the soundcard of system.

Speech recognition system cannot process digital waveforms directly. These have to be represented in a more compact and efficient way. For the purpose to reduce the dimensionality of the speech signal on the typical order of 80:1, feature extractors maintain much of the relevant characteristics of the original speech and eliminate the extraneous information for this, firstly digitized input is flattened using filters and then essential features having acoustic correlation with the speech input are extracted using feature extraction.

MFCCMel Frequency Cepstral Generation [3] The final procedure for the Mel Frequency Cepstral Coefficient (MFCC) consist of performing the Inverse DFT on the logarithm of the filter bank output .The inverse DFT reduces to a Discrete Cosine Transformation (DCT). The DCT has property to produce highly uncorrelated features. The DCT of filter bank output is given by

$$Y_t^m(k) = \sum_{m=1}^{M} log\,\{|Y_t^m|\}.\,cos\left(k\left(m-\frac{1}{2}\right)\frac{\pi}{m}\right) \qquad (1)$$
$$k = 0, 1, 2.........L$$

PLP Perceptual Linear Prediction (PLP) method proposed by [4] demonstrated a further improvement over the LPCC which takes advantage of three principal characteristics derived from the psycho-acoustic properties of the human hearing viz., spectral resolution of the critical band, equal loudness curve adjustment and application of intensity-loudness power law.

LPCC

Linear prediction [5] is a good tool for analysis of speech signals. Linear predication models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal. In speech coding, the success of LPC have been explained by the fact that an all pole model is a reasonable approximation for the transfer function of the vocal tract. All pole models is also appropriate in terms of human hearing, because the ear is more sensitive to spectral peaks than spectral valleys. Hence an all pole model is useful not only because it may be a physical model for a signal, but because it is a perceptually meaningful parametric representation for a signal.

III.COMPARATIVE FEATURE PARAMETER

The experimental parameter for MFCC, PLP, and LPCC are as follows: Since the experiment uses same corpus the input file specification is similar in all parameters table. Such as file format, sampling rate, bit rate, type of channel. In the experiment same type of window is also used. The main difference can be seen in target kind and number of coefficient. The complete list of parameters and their value are given in the table[5-6].

TABLE I PARAMETERS OF MFCC

S.No	Parameter	Value
1	Input File Format	.wav
2	Sampling Rate	16000Hz
3	Bit Rate (bits per sample)	16
4	Type of Channels	Mono
5	Window Size	250000.0 (25 msec.)
6	Frame Periodicity	100000.0 (10 msec.)
7	Window used	Hamming
8	Number of Filter-bank channels	26
9	Target Kind	MFCC_0_D_A (MFCC with energy, delta (Δ) and acceleration (ΔΔ) coefficients.
10	Number of MFCC Coefficients	12
11	Pre-emphasis Coefficient	0.97
12	Length of Cepstral Liftering	22
13	Energy Normalisation	True

TABLE IIPLP PARAMETERS

S.No	Parameter	Value
1	Input File Format	.wav
2	Sampling Rate	16000Hz
3	Bit Rate (bits per sample)	16
4	Type of Channels	Mono
5	Window Size	250000.0 (25 msec.)
6	Frame Periodicity	100000.0 (10 msec.)

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

7	Window used	Hamming
8	Number of Filter-bank channels	26
9	Target Kind	PLP_0 (PLP with Energy)
10	Number of PLP Coefficients	13
11	Pre-emphasis Coefficient	0.97
12	Length of Cepstral Liftering	22
13	Energy Normalisation	True

LPC provides good model of speech signal. It is a Production based method. Speech sample at time n can be represented as a linear combination of p previous samples. LPC represents low dimension feature vectors using spectral envelope and provides linear characteristics. LPC leads to a reasonable source-vocal tract separation.

TABLE III LPC PARAMETERS

Target Kind	LPCCPESTRA (Linear Predictive Cepeatrial Coefficient)
Number of LPCCPESTRA Coefficients	12

IV. EXPERIMENTAL ANALYSIS

The features has been compared in clean environment and in general field conditions with known and unknown speakers (known speakers means their data sample is recorded in corpus while known speakers means system does not have their samples). Fig. 1 shows the output of each feature extraction technique respectively. Table IV presents the recognition results in terms of number spoken words and number of recognize word. The recognition results shows that the MFCC is better when testing and training conditions are same but in mismatch condition PLP outperforms MFCC, while both the MFCC and PLP perform better that LPCC in clean environment and general field conditions.

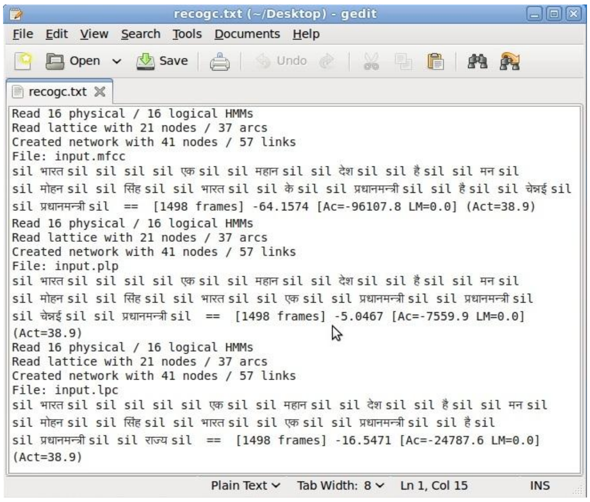

Fig. 1Recognition Results of Feature Extraction Techniques

The recognition result uses all three feature extraction techniques; the speech recognition system uses same testing data to produce the result for all the three feature extraction techniques. The experimental result shows that MFCC feature extraction technique produces more accurate results.

The system was tested using the test data prepared separately by a set of speakers. Each speaker was asked to utter some words of the vocabulary. Some test data was collected from the training data. Four speakers were selected to collect the test data. Out of these four speakers, two were those used for collecting the training data. Thus test data contains three types of sounds: sound used for training the system, sound spoken by the speaker whose other sound files were used for training the system, and sounds of a speaker that does not participate in training. Recognition results in Fig. 2 show that LPCC, MFCC, PLP produces 89.56%, 92.17%, 90.04 % respectively correct recognition. While in general field conditions the percentage of correct word recognition is respectively 85.21, 86.08 and 87.82.

TABLE IV COMPARISON OF RESULTS OF FEATURE EXTRACTION TECHNIQUES

Feature Extraction Technique	Test Condition	Training 1	Training 2	Trail 1	Trail 2
	Clear	50	55	60	65
LPCC	Field	46	53	51	56
	Clear	45	52	48	51
	Field	49	55	52	56
MFCC	Clear	46	52	49	51
PLP	Field	48	54	51	55
	Clear	48	53	50	53

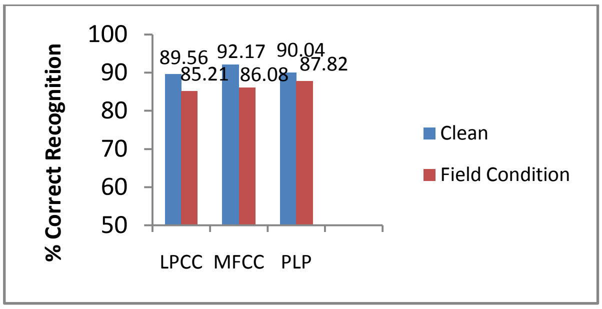

Fig. 2 Percentage of Correct Recognition

V. CONCLUSION

The experimental results of feature comparison show that MFCC is better when testing and training conditions are same but in mismatch condition PLP outperforms MFCC, while both the MFCC and PLP perform better than LPCC in clean environment and general field conditions. It has been found that the system is performing well with more

 DOI 10.17148/IJARCCE.2016.58128

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

vocabulary-size compared to the other reported similar works. This implementation opens an area for the development of system for large vocabulary size and improvement of the system's accuracy for unknown speakers.

REFERENCES

[1] B. A. Q. Al-Qatab and R. N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)", Paper presented at International Symposium in Information Technology (ITSim),Kuala Lumpur, June 2010.
[2] K. Kumar and R. K. Aggarwal, "A Hindi speech recognition system for connected words using HTK", Int. J. of Computational Systems Engineering, vol.1, no.1, pp. 25 – 32, 2012.
[3] H. Hermansky, "Perceptually predictive analysis of speech", Journal of Acoustic Society of America, vol. 87, pp. 1738-1752, 1990.
[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, pp.357-366, 1980.
[5] J. D. Markel and A. H. Gray, "Linear prediction of speech, New York: Springer-Verlag, 1976.
[6] Chao Hao and Liu Wenju, "Improved Syllable Based Acoustic Modeling by Inter-Syllable Transition Model for Continuous Chinese Speech Recognition", Pattern Recognition, pp. 1-4, 2009.
[7] K. S. Rao and B. Yegnanarayana, "Modeling durations of syllables using neural networks" Computer Speech and Language, vol. 21, no. 2, 282–295, 2007.
[8] M. Plauche, N. Udhyakumar, C. Wooters, J. Pal and D. Ramachadran, "Speech recognition for illiterate access to information and technology", In Proceedings of First International Conference On ICT and Development, 2006.
[9] M. Kumar, A. Verma and N. Rajput, "A large vocabulary speech recognition system for hindi," Journal of IBM Research, vol. 48, pp. 703-715, 2004
[10] M Kumar, R K Aggarwal, G Leekha, and Y Kumar, "Ensemble feature extraction modules for improved hindi speech recognition system" International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

BIOGRAPHY

Malay Kumar received M. Tech. degree in Computer Engineering from National Institute of Technology, Kurukshetra, Haryana, India, in 2012. He is currently working toward the PhD degree in computer science and engineering from National Institute of Technology Raipur, Chhattisgarh, India. His current research interest includes Hindi speech recognition systems, HCI systems, Devices for Disables, secure outsourcing of mathematical and scientific application, encryption schemes, scheduling, distributed systems, and Cloud Computing.

 DOI 10.17148/IJARCCE.2016.58128