SPECTRAL FEATURES ANALYSIS FOR HINDI SPEECH RECOGNITION SYSTEM

Kanika Garg[1], Bharat Gupta[2], Sakshi[3]

[1]*PhD Research Scholar, SC&SS, Jawaharlal Nehru University, Delhi, India*
[2]*Assisstant Professor, AESPL, Jodhpur, India*
[3]*M.Phil Research Scholar, Department of Educational Studies, JMI, Delhi, India*

Abstract

Automatic speech recognition refers to recognizing the speech utterances and converting them to text through machines. For this purpose, the features forms an extremely important part. The richness of features will predict the performance of the overall system. So, this paper deals with the various speech features that can used for Hindi speech that has been tested for many other languages. In this work, MFCC, PLP, EFCC and LPC have been tested against Hindi Speech Corpus using HMM toolkit HTK 3.4.1. These features have been evaluated using common environment. The main objective of this paper is to summarize and compare the traditional and newer feature extraction methodology in automatic speech recognition system. This work favours EFCC features over other features. EFCC have shown a significant improvement in noisy environment in automatic speech recognition system.

Keywords: *ASR, MFCC, EFCC, PLP, LPC*

----------***----------

1. INTRODUCTION

Automatic speech recognition systems are the systems that are able to understand the human utterances and convert them into valid readable format automatically without human intervention. In simple words, ASR is the system that converts vocal sounds into the legitimate text. The utterances are converted into text by the machine[1]. This method provides man-machine communication possible. In present time, generally all the commands and any transfer of information from man to machine or vis-à-vis are provided using input devices like keyboard, touchpad etc. or using output devices like monitor, printer etc. However, this is not particularly useful for the people who are computer illiterate as this needs a special training. Also, these methods are time consuming and has a higher dependency on the external devices used with the computer to work properly. In this way, humans can provide commands to the systems using their own native languages which needs no extra training.

Hindi is a resource deficient language so it is difficult for the researchers to achieve higher accuracy in its speech recognition. Work has been done by the researchers for this retroflexed language in noisy and noise free environments but still the benchmarks are yet to achieve. A lot of work has been done in this field. Various methods have been proposed and used to make this efficient. But one problem that has not been able to meet the reality grounds is the parameterization of the speech. The feature extraction from speech is really an important task in automatic speech recognition. So, in this work, we are to address this issue and an effort has been made to provide Hindi speech recognition a better feature extraction method, so as to improve overall accuracy and efficiency of the system.

Various feature extraction methods like Mel frequency cepstral coefficient (MFCC), perceptual linear prediction (PLP)[2], ERB scale cepstral coefficient (EFCC) [3]and wavelet pattern cepstral coefficients (WPCC)[4] were tested against the speech corpus. The speech taken in this work is recorded in noisy environment, so, initially noise reduction method has been used[5]. In our previous work, we have concluded that gamma tone filters provide the best noise reduction in our suitable conditions. So, gamma tone filters[3] have been used in this work. Then the feature extraction methods have been applied to the speech. The final step is to recognise speech words that has been done using HMM toolkit HTK 3.4.1. Hcopy tool is used to extract MFCC and PLP features.

The various sections of this paper are organised as follows: Section 2discusses related work done in this field. Section 3 discusses about the basic methodology used to carry out this work. Section 4 discusses the various feature extraction technique. Section 5 discusses the implementation scheme and various results obtained. Section 6 discusses the conclusion. Section 7 discusses the future work.

2. RELATED WORK

In 2010, Ranjan et.al. [6]performed speech recognition by identifying speaker in text dependent multilingual environment. They had used artificial neural network with speech feature vectors. In their work, they had used various speech features like linear predictive coding (LPC), reflection coefficients (RC), formant frequencies, number of zero crossing etc. Then these are taken as input for ANN using clustering algorithm for training and back propagation as their learning algorithm. They have achieved 92.7% identification rate using clustering algorithm.

In 2010, K. Mehta [3] used feature extraction method that work on human auditory based systems. In their work, they had employed gamma tone filters to initially reduce the noise from the noisy speech and then the various features were extracted. They had used Mel frequency cepstral coefficients (MFCC) and ERB scale cepstral coefficients (EFCC) in noisy environment to extract features. Their work showed that ERB based features outperformed the standard MFCC features by approximately 10.5%.

In 2011, S. G. Koolagudi[7] analysed emotions out of Hindi speech. They had used spectral features as well as prosodic features to accomplish their work. For prosody, they had used energy, pitch and duration as features. These features help in recognizing emotional aspects. For spectral features they had used standard feature extraction method i.e. Mel frequency cepstral coefficient (MFCC). Using these features they were able to achieve around 81 % success rate in their work.

In 2013, S. Tripathy[8]used HMM as classifier and Mel frequency cepstral coefficient (MFCC) and Linear Predictive Coding (LPC) as feature extraction algorithms. They had tested their work on speaker dependent as well as speaker independent conditions. Their work had shown that MFCC had outperformed LPC in every condition.

In 2014, A. Kumar[9]investigated various Gaussian mixture HMM using Mel frequency cepstral coefficient (MFCC) as features. They had used 16 KHz sampling rate with hamming window size of 25 milliseconds and frame size as 10 milliseconds. Their experiments showed the best performance with four component Gaussian mixtures.

In 2015, A. Biswas et. al. [10] performed speech recognition using harmonic energy based features. They proposed a novice approach of new wavelet packet sub- band- based energy features. This technique helped in unvoiced as well as voiced phoneme recognition.

3. METHODOLOGY

The diagram shows the basic approach followed. Almost all the features discussed in this paper are finally computed using HMM toolkit. Before extracting the features various steps need to be followed. Fig-1 shows the basic methodology used in this paper.

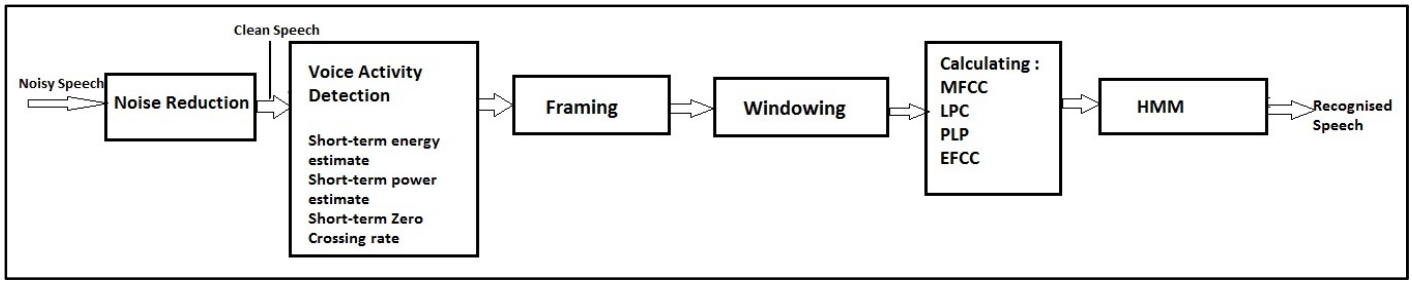

Fig-1: Basic Methodology in ASR

3.1 Noise Reduction

This step is really important in speech recognition system as all the recordings are done in the environment that are more or less susceptible to noise. Basically, noise reduction or speech enhancement methods are used before feeding the speech signals to the ASR. This helps in reducing the noise or distortion from background or the transmission medium. For higher accuracy in automatic speech recognition systems, a powerful noise reduction technique is required. In this work, we have employed spectral subtraction as they are simple to implement and provide promising results.

3.2 Voice Activity Detection

This step helps in recognising the speech utterances and unvoiced speech components. It helps in detecting the presence or absence of voice in the speech signals. It is therefore called as speech detection.There are various ways of doing this. In this paper, we have used some features as Short-term energy estimate, short-term power estimate, short-term zero crossing rate. These combined features give the outlook of spatial and temporal pattern in the speech. These patterns helps in determining the termination of speech boundaries. Energy is one of the most important feature for VAD but it loses its strength in lower SNR condition with the noisy speech. So, we have used two more features to improve the strength of VAD. Zero crossing rate in a given frame refers to the number of times the sample changes its sign. In narrowband signals it depicts the frequency content.

3.3 Framing & Windowing

After speech enhancement/ noise removal and voice activity detection, there is a need of framing and windowing. Speech signals are time variant and non-stationary, due to which framing is done so as to take small chunks of signals to process at once. Small chunks are taken as it is considered that spectral features are invariant when taken in taken window of signals. Generally the frame length varies from 5-10 milliseconds and window length varies from 20-25 milliseconds. Higher overlapping of around 50% - 70% on consecutive frames is taken which helps in smoother change in parametric values with higher computational power. After the framing is done, each frame is multiplied with the suitable windowing function. The various windows that can be used are hamming, rectangular, hanning, barlett etc.

3.4 Calculating Cepstral Features

After the framing and windowing, the various features are extracted from each frame. In this work we have used only the cepstral features. Various feature extraction techniques are Mel-frequency cepstral coefficients (MFCC), ERB-scale frequency cepstral coefficients (EFCC), perceptual linear prediction (PLP), wavelet, RASTA, HLDA etc. In this work, we have used MFCC, LPC, PLP and EFCC.

3.5 Speech Recognition

After the features have been extracted, knowledge models are employed on the features to recognize the unknown speech utterances. This step includes acoustic models, lexical models and language models. Acoustic models can be implemented using Hidden Markov Models (HMM), Support Vector Machines (SVM), Deep Neural Networks (DNN) etc.

4. FEATURE EXTRACTION

We have discussed above the various methods available for feature extraction. In this work we have emphasized on cepstral features like MFCC, EFCC, PLP and LPC.

4.1 Mel-Frequency Cepstral Coefficient (MFCC)

This feature extraction method is considered as one of the standard method. This method has given promising results with all the speech recognition systems in noise free environments[2]. The use of 39 Mel coefficients is generally common in automatic speech recognition [9]. The main drawback of this method is that it doesn't work in noisy speech as it is dependent on spectral features only. This method works on human auditory system so to approximate the mel-scale, non-linear frequency scale is used which provide linear values below a threshold value(generally of 1 KHz) and log values above it.These Mel-features correspond to thecepstrum of the log filter bank energies.

4.2 Perceptual Liner Prediction (PLP)

The PLP model was developed in 1990 by Hermansky. This feature extraction technique relates to the psychophysics of the human auditory system. It is based on short term spectrum and make a series of psychophysical changes to the spectrum. Unlike MFCC, PLP makes use of Bark-spaced filter bank. This covers the frequency range of 0-5000 Hz. Equal loudness pre-emphasis and intensity loudness compression is done in this method. The various methods are shown infig-2.

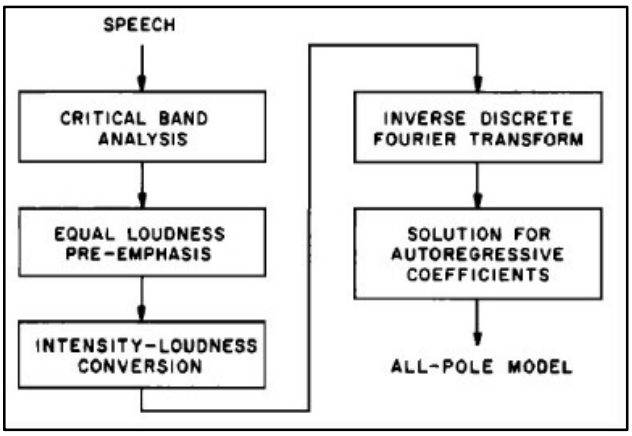

Fig-2: Perceptual Linear Prediction [11]

4.2 Linear Predictive Coding (LPC)

LPC is a feature extraction technique that is based on human speech production[12]. It helps in encoding good quality speech even at low bit rates. This works on the principle that present speech samples can be approximated using past speech samples. These features helps in minimizing the difference between the actual speech samples and the predicted speech samples.It is shown in Fig-3.

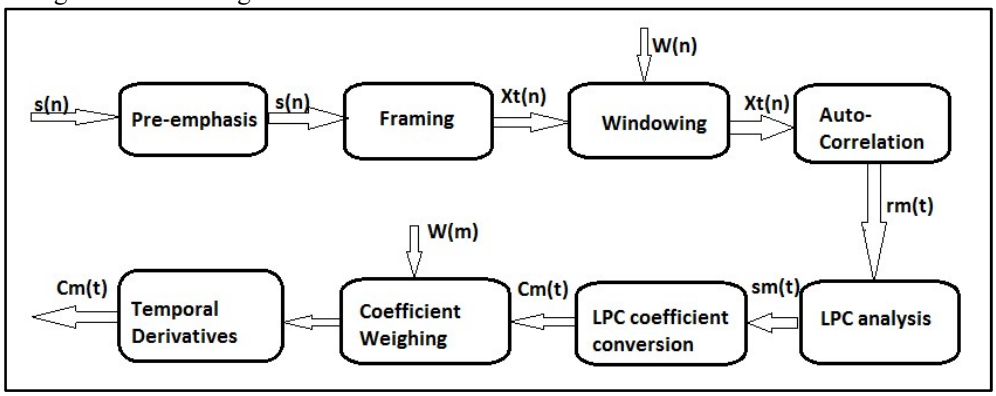

Fig-3: Linear Predictive Coding

4.3 ERB Scale Based Cepstral Features (EFCC)

This feature technique works on ERB scale cepstral coefficients. It is calculated in the same manner as MFCC's are calculated except in the filter bank and compression technique. In this, equivalent rectangular bandwidth scale is used [13]. ERB scale is more physiologically motived towards human auditory periphery. When subjected to white noise, it passes the same power as of real filter. For this, normalized filter bank based on gamma tone filters is applied to spectrum and then instead of using the log-compression as used in calculation of MFCC, power law non-linearity is used(Fig-4) . In noisy speech environment, EFCC's showed better results than MFCC.

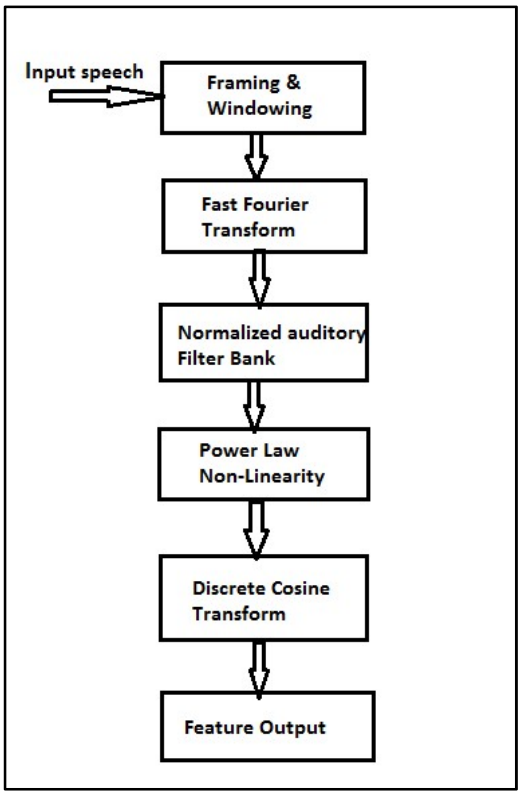

Fig-4: ERB scale cepstral coefficients

5. IMPLEMENTATION AND RESULTS

Proposed work is done in four major steps:

a) Collection of speech samples
b) Noise cancellation/speech enhancement
c) Feature Extraction
d) Speech Recognition

For collection of speech samples, recording of sentences is done using microphone in noise free environment in wav files which is connected with a computer at sampling rate of 16 KHz. This is stored in 16-bit PCM in mono-mode encoded in waveform format. There are two male speakers and one female speakers. Total 150 sentences were recorded in which there are 394 different utterances of words are present. The training and testing test have been created in the ratio 3:2. The recording has been performed several times for each sentence so as to get the best quality speech sample.

Then, as the sentences were recorded in negligible noise.So, white Gaussian noise is added to the speech samples. Noise is such added so as to get SNR 20 dB, 15 dB and 10 dB. So, in this work, spectral subtraction method is used to remove the added noise from the speech utterances. Noise cancellation technique also offers some error and are not fully reliable, so, there will be a reflection of this technique on final results. After removal of noise frames are chosen for 10 milliseconds, 15 milliseconds and 20 milliseconds. Hamming window of 20 milliseconds was used.

After, the framing and windowing, cepstral feature extraction techniques are employed. These features were extractedusing HTK toolkit:

- MFCC
- LPC
- PLP
- EFCC

Hcopy tool is used to extract MFCC features in HTK toolkit. For MFCC features log energy is computed and then log filter banks were computed as these features corresponds to them.

The PLP features are also based on Mel-filter bank. The Mel cepstral features were taken on equal-loudness curve and then they were weighted. Then the cube root was taken of these coefficients to compress them. Finally, these coefficients were converted to cepstral coefficients.

When the past speech samples and linear combination is known, then LPC is used. LPC cepstral features can also be generated with slight changes in the method of obtaining MFCC. They used power spectrum rather than its log. Auto regression model is used and helps in minimizing the difference between the actual speech samples and the predicted speech samples.

EFCC are extracted in the similar way as MFCC. The major difference lies in the filter bank and compression technique. For EFCC, we have used normalized filter bank. Instead of using Mel triangular scale, rectangular bandwidth scale is used as it suits the human auditory system. Also, power-law non-linearity is used for loudness compression rather than log compression.

After all the features were extracted, speech recognition is done using HMM model using HTK. It uses 5 states. HMM model for training is initialised using HInit. After training, the HMM parameters were re-calculated using tool HRest iteratively until the result converges.

The system was tested using the 40% spoken utterances containing both male and female voices. The word accuracy rate achieved with MFCC for 10 milliseconds frame and at 20dB SNR is 79.02%, at 15 dB is 70.25% and at 10 dB is 53.9%. The word accuracy rate achieved with PLP for 10 milliseconds frame and at 20 dB is 80.04%, at 15 dB is 78.2% and at 10 dB is 73.8%. The word accuracy rate achieved with LPC for 10 milliseconds frame at 20 dB is 68.7%, at 15 dB is 67.1% and at 10 dB is 50.3%. The word accuracy rate achieved with EFCC for 10 milliseconds frame at 20 dB is 82.4%, at 15 dB is 87.6% and at 10 dB is 88.7 %. The results are tabulated in Table-1.

Table-1: Experimental Results

	20 dB	15 dB	10 dB
MFCC	79.02%	7.25%	53.9%
PLP	80.04%	78.2%	73.8%
LPC	68.7%	67.1%	50.3%
EFCC	**82.4%**	**87.6%**	**88.7%**

6. CONCLUSION

In this work, the results show that EFCC outperformed all the other feature extraction techniques. EFCC performed well in all the conditions. Then, PLP performed better than MFCC and LPC. The power of MFCC degraded because of the white Gaussian noise added in the clean speech. Further, the performance has been affected because of noise reduction methods used. Overall, the system performed well in noisy environment with EFCC features.

FUTURE WORK

In future work, this system will be tested for speech corpus with large vocabulary and speech utterances. The new and hybrid feature extraction methods will be tested to further improve the efficiency and accuracy of the system. Further, the system can be tested on standard Hindi corpuses available.

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success. I would like to take this opportunity to express my profound gratitude to my friend Goonjan Jain who with her thought provoking views, veracity and whole hearted cooperation helped me in doing this work.

REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing, I. USA: Prentice-Hall, 2001.

[2] S. H. Choi, H. K. Kim, and H. S. Lee, "LSP weighting functions based on spectral sensitivity and mel-frequency warping for speech recognition in digital communication," vol. 1, pp. 0–3, 1999.

[3] K. Mehta and R. S. Anand, "Robust Front-end and Back-end Processing for Feature Extraction for Hindi Speech Recognition," no. 2, pp. 1–4, 2010.

[4] J. R. Karam, W. J. Phillips, W. Robertson, and M. M. Artimy, "New wavelet packet model for automatic speech recognition system," Can. Conf. Electr. Comput. Eng. 2001. Conf. Proc. (Cat. No.01TH8555), vol. 1, pp. 511–514, 2001.

[5] J. Li, L. Deng, Y. Gong, and S. Member, "An Overview of Noise-Robust Automatic Speech Recognition," vol. X, no. X, pp. 1–33, 2013.

[6] R. Ranjan, S. K. Singh, and A. Shukla, "Text-Dependent Multilingual Speaker Identification for Indian Languages using Artificial Neural Network," pp. 632–635, 2010.

[7] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "IITKGP-SEHSC : Hindi speech corpus for emotion analysis," 2011.

[8] S. Tripathy, N. Baranwal, and G. C. Nandi, "A MFCC based Hindi speech recognition technique using HTK Toolkit," 2013 IEEE 2nd Int. Conf. Image Inf. Process. IEEE ICIIP 2013, pp. 539–544, 2013.

[9] Kuamr, M. Dua, and T. Choudhary, "Continuous Hindi Speech Recognition Using Gaussian Mixture HMM," no. 1, pp. 0–4, 2014.

[10] Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Admissible wavelet packet sub-band-based harmonic energy features for Hindi phoneme recognition," pp. 511–519, 2015.

[11] H. Hermansky, "Perceptual Linear Predictive(PLP) Analysis of Speech," J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752, 1990.

[12] D. O'Shaughnessy, "Linear predictive coding," Potentials, IEEE, vol. 7, no. 1, pp. 29–32, 1988.

[13] W. W. H. Abdulla, "Auditory Based Feature Vectors for Speech Recognition Systems," Adv. Commun. Softw. Technol., pp. 231–236, 2002.