Hindi Speech Recognition System with Robust Front End-Back End Features

Atul Gairola (IEEE Member)
Dehradun Institute of Technology,
Dehradun-248009, Uttarakhand, India

Swapna Baadkar
Dehradun Institute of Technology,
Dehradun-248009, Uttarakhand, India

ABSTRACT

The ideal aim of a speech recognition system is efficient and accurate conversion of speech signal into text message without any dependence on device, environment, and speaker. In this paper a system for Hindi speech recognition is discussed employing robust front end- back end techniques. At front end MF-PLP is used for feature extraction while continuous density HMM is used at the back end for evaluation. A comparison of MFCC, PLP & MF-PLP is also presented to show the robust characteristics of MF-PLP.

General Terms

Hindi conversational speech, Extraction & Evaluation

Keywords

Feature Extraction, Front End, Back End, MFCC, PLP, MF-PLP, CDHMM.

1. INTRODUCTION:

When it comes to speech recognition, we see that there has been a lot of research for languages like Chinese, French, and Arabic after dominating English. In this context, we also found out that very few researches have been done in the field of Hindi speech recognition system. Though the technology has advanced to many realms still a very low percentage of computer literates in India are able to benefit from it. The variable nature of the Hindi Dialect has made it difficult to perform connected word recognition for Hindi language. After English and mandarin, Hindi is the third most widely spoken language in the world, therefore a speech recognition system for Hindi is expected to be used with great diversity. The speech recognition technology has developed remarkably over the last 50 years. The technology trends now demands a conversational speech based interface with immense processing power, robust accuracies and independence from device, speaker & environment problems[2].

The two major components on which the recognition performance of a speech recognition system depends are feature extraction unit and classification, training & testing unit[2]. Feature extraction unit is mainly responsible for the recognition performance of the system along with reasonable amount of computation.

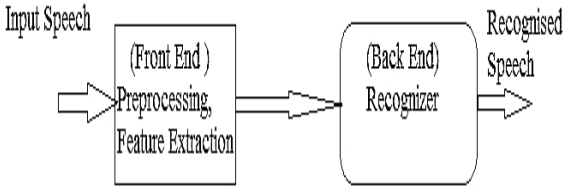

Figure 1. Basic Structure of a Speech Recogniser.

In the present work the best techniques are experimented and then applied to propose a model that can be used as a basis for Hindi conversational speech interface to pace with the current technologies in the concerned field[3]. The proposed model is a work intended to encourage researches in the field of Hindi conversational speech recognition. The dependency on keyboard for text work in Hindi is focused to compromise by developing a complete speech based interface for Hindi language.

2. DATABASE PREPARATION

Hindi is evolved from Devnagri script which is believed to be a derivative of some ancient Brahmi script. The languages that have been evolved from Brahmi or Devnagri share a common phonetic structure. In this research a small database is prepared using 30 speakers of which 15 are female speakers while 15 are male speakers. Speakers in the age group of 20 to 28 are selected from different students studying in different colleges in Dehradun city (Uttarakhand, India). This is done to produce appropriate pronunciations of Hindi digits. From every speaker 50 sets of connected digits were recorded. The 50 sets were recorded, again for creating a database for noisy environments. NOISEX-92[6] database is used for adding different noises which are down sampled at 16kHz at an SNR of 5dB-20dB. For maintaining a small database we used babble, white and pink noise levels.

3. FEATURE EXTRACTION

The main component of a feature extraction unit is the front end which reduces the complexity of the raw speech and makes it suitable for feeding to the recognition system. The techniques used in this paper are discussed below.

3.1. Mel Frequency Cepstrums features.

Mel frequency cepstrum is the most widely used feature extraction method. The speech features are extracted through a Mel spaced bank of filters which receives a windowed signal from previous stages[4]. Application of discrete cosine transform(DCT) to the filtered output converts it into 24 cepstral coefficients. After further de-correlation only 13 coefficients are used[6].

3.2. Perceptual linear predictive features.

In 1990, Hermanskey[1],[9] proposed the perceptual features method in which linear predictive approach is combined with discrete fourier transform(DFT) to obtain the power spectrum of speech utterances. The steps involved in PLP technique[9] are shown in figure and are summarized below.

Step 1: Computation of the Power spectrum.

The windowed speech is computed for power spectrum by employing short time Fourier transform with its squared magnitude on each and every frame of speech.

Step 2: Grouping of critical bands.

The power spectrum obtained in Step 1 is then warped into Bark scale and convoluted with power spectrum of band filters which are equally spaced in Bark domain. The spectral resolution is achieved.

Step 3: Loudness Monitoring.

The loudness in the perceived speech utterances are monitored by using an equal loudness function applied to filter bank values. The intensity of the spectral amplitudes are compressed using IFFT.

Step 4:All pole Autoregressive modeling.

The Auto-regression through inverse DFT produces the autoregressive coefficients to perform the all pole modeling.

Step 5: Coefficient Conversion.

Finally cepstral analysis is performed on autoregressive coefficients to convert them into required cepstral coefficients.

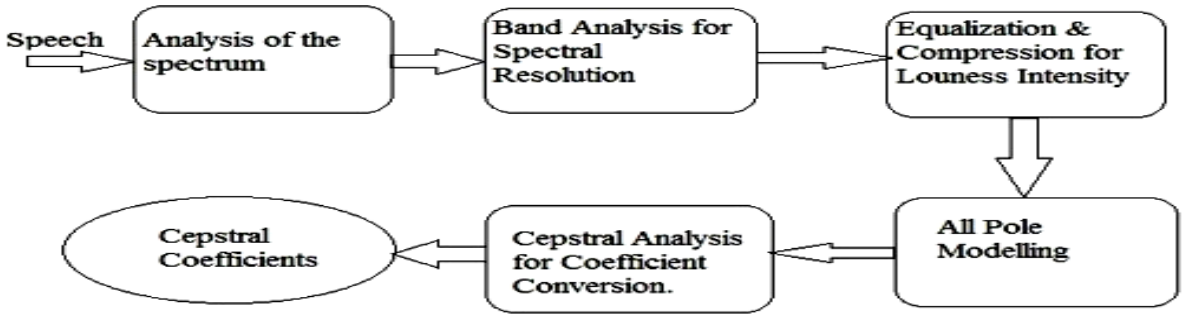

Figure 2. Steps for computing Perceptual LP Coefficients.

4.MODEL FOR RECOGNITION

The stochastic method we used to model the sequence of speech vectors is based on continuous density HMM[11],[12]. In the present work we have used a 3-state left-to-right continuous density HMM with a provision of separate duration density, associated with each phone model. The selection of phone contexts is automatic and based on their frequency levels[7]. The main advantage of continuous density HMM method is parameter adaptability due to which high precision results can be obtained without the need of smoothing techniques[8],[10]. The transition matrix is known and kept fixed to reduce the number of parameters being adapted.

5. EXPERIMENTAL RESULTS.

The feature vectors are obtained by sampling speech at 12kHz with frame rate of 10ms & windowed at 25ms,Hamming window. From the database of 30 speakers we have selected 20 speakers for creating training database while 10 speakers are selected for testing. First the clean database of 50 sets is created and then a noisy database is created using white and pink noise levels from the NOISEX-92[6] database. Earlier we recorded 30 to 35 sets during training procedure but the results were inadequate since the variance was not properly estimated. On increasing the number of training sets a slight increase in the performance was recorded. So the present work was extended to 50 sets. 30 words spoken by different speakers are randomly chosen for testing. The recognition rate is estimated as:

$$Rr = Dw / Tw$$

Rr - Rate of recognition

Dw - Number of successful detections(words).

Tw – Number of words in entire test set.

5.1. Performance comparison of MFCC,PLP and MF-PLP for clean database.

The insertion of Mel filter in perceptual features is observed to produce high recognition efficiency than MFCC and PLP[5],[9] alone. The recognition accuracy is also influenced by spectral resolution, pre-emphasis, discrete cosine transform and power law. The comparison of different feature extraction methods is shown in figure 3.

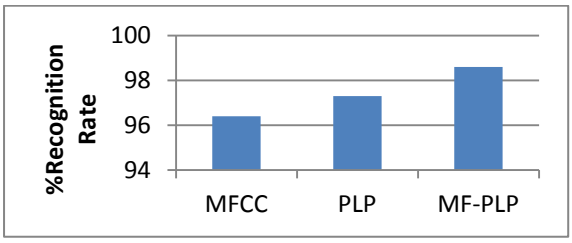

Figure 3. Performance of Feature Extraction methods with clean database.

5.2 Performance comparison of features with Noise insertion.

The experiments are performed for 10dB and 20dB SNR levels as the system performance was degraded at 5dB level. At 20dB SNR and above the system produced nearly similar results as that of a noise free clean database.

At 10dB level there is a considerable decrease in system performance but it is also observed that MF-PLP performed better than MFCC & PLP which is a measure of robustness of the system in presence of noise. Recognition rate of different features in presence of noise levels are shown in figure 4, figure 5 & Table 1 respectively.

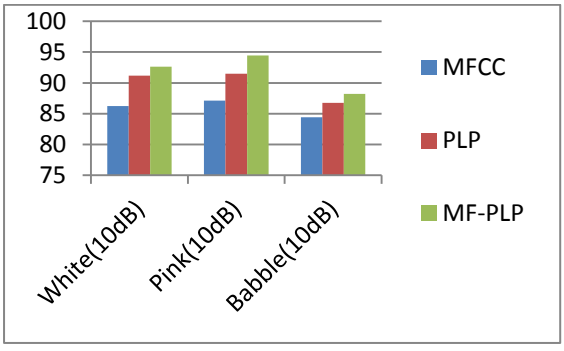

Figure 4. Performance comparison of Feature Extraction methods at 10dB Noise level.

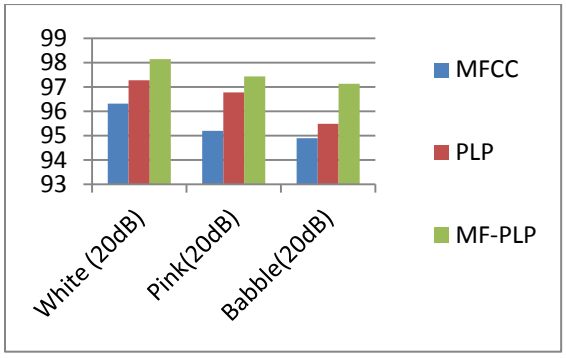

Figure 5. Performance comparison of Feature Extraction methods at 20dB Noise level.

Table 1. Performance Comparison of Feature Extraction Methods.

Different Noise Levels	Features/ SNR Levels	Percentage rate of recognition		
		MFCC	PLP	MF-PLP
White Noise	10dB	86.23	91.17	92.63
	20dB	96.32	97.27	98.14
Pink Noise	10dB	87.13	91.49	94.46
	20dB	95.19	96.78	97.43
Babble Noise	10dB	84.41	86.76	88.19
	20dB	94.90	95.48	97.13

6. CONCLUSION

The conclusion derived from all experiments favors MF-PLP as the best feature extraction method in clean as well as noisy environments. For generating acoustic models the use of continuous density HMM also proved to be equally favorable. Results show that word model gives maximum accuracy for small database of 500 words. The Model proposed here will encourage researches to develop systems for Hindi conversational speech recognition.

ACKNOWLEDGEMENTS

I would also like to thank Dr. S.C. Gupta (Emeritus Professor) and Dr. Sandip Vijay,HOD(ECE & AEI),Dehradun Institute of Technology, for providing me with valuable suggestions and guidance during my study.

7. REFERENCES

[1] H. Hermansky, "Perceptually predictive (PLP) analysis of speech," Journal of Acoustic Society of America, vol. 87, 1990, pp. 1738-1752.

[2] A.O. Afolabi, A. Williams, and O. Dotun, "Development of a text dependent speaker identification security

system", Research Journal of Applied Sciences, 2 (6), pp. 677-684, 2007.

[3] K. Samudravijaya, Barot & Maria, "A Comparison of Public-Domain Software Tools for Speech Recognition", In WSLP, pp.125-131, 2003.

[4] R.Josef and P. Pollak , "Modified Feature Extraction Methods in Robust Speech Recognition", Radioelektronika, 17th IEEE International Conference, pp.1-4, (2007).

[5] Andra´s Zolnay , Daniil Kocharov , Ralf Schlüter and Hermann Ney, "Using multiple acoustic feature sets for speech recognition", Science direct, Speech Communication 49 , pp. 514–525, 2007.

[6] study on the effect of additive noise on automatic speech recognition system", Reports of NATO Research Study Group (RSG.10), 1992.

[7] N. Goel, S.Thomas, M. Agarwal et al. "Approaches to Automatic Lexicon Learning with Limited Training Examples", Proc. of IEEE Conference on Acoustic Speech and Signal Processing, 2010.

[8] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", IEEE Transaction of Acoustic, Speech and Signal Processing, Vol.27, No. 2, 1979, pp. 113-120.

[9] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Transaction on Speech and Audio Processing, Vol.2, No. 4, 1994, pp. 578-589.

[10] S. Young, "A Review of Large Vocabulary Continuous Speech Recognition", IEEE Signal Processing Mag., Vol.13, 1996, pp. 45-57.

[11] C.H. Lee, J. L. Gauvain, R. Pieraccini, and L. R. Rabiner, "Large Vocabulary Speech Recognition using Subword Units", Speech Communication, Vol.13, 1993, pp. 263-279.

[12] X.D. Huang, H.W. Hon, M.Y. Hwang, and K.F. Lee, "A comparative study of discrete, semi continuous and continuous hidden Markov models," Computer Speech and Language, vol. 7(4), 1993, pp. 359-368.