Science and Technology Journal, Vol. 4 Issue: II ISSN: 2321-3388

Automatic Speech Recognition System: A Survey Report

Moirangthem Tiken Singh*
Dibrugarh University Institute of Engineering and Technology, Assam, India
E-mail: *tiken.m@dibru.ac.in

Abstract—This paper presents a report on an Automatic Speech Recognition System (ASR) for different Indian language under different accent. The paper is a comparative study of the performance of system developed which uses Hidden Markov Model (HMM) as the classifier and Mel-Frequency Cepstral Coefficients (MFCC) as speech features.

Keywords: Indian Language, Hidden Markov Model, Mel Frequency Cepstral Coefficients

INTRODUCTION

Automatic Speech Recognition (ASR) is a field of Computer Science with aims to design the computer system to recognize the human voice. It takes an utterance of the speech signal as input, which is captured through a microphone, or a telephone etc. and converts into a text sequence which is close as possible to the spoken data [1]. ASR was first introduced to 1950. The first attempt to develop techniques for speech recognition was based on the direct conversion to the speech signal into a sequence of phoneme-like units. But unfortunately, it was failed. The first positive results of spoken word recognition came into existence of the 1970s, when general pattern matching techniques [2] was introduced. ASR has attracted much attention to/on the last three decades and has witnessed the dramatic improvement in the last decade. Today it has different areas of application like dictation, the program controlling, automatic telephone calls, weather report information system, travel information systems etc. But its implementation is difficult due to the different accents on human beings. Therefore, the main aim of ASR today is to transform an input voice signal to its corresponding text output independent of the speakers. This paper aims to present a review of methodology and results obtained during speech recognition by various researchers. A comparison is done based on their recognition level and accuracy.

LITERATURE REVIEW

Mishra [3] and his team worked on automatic speech recognition of speaker independent connected digits with Revised Perceptual Linear Prediction (RPLP), Bark Frequency Cepstral Coefficients (BFCC) and Mel-frequency Cepstral Coefficients (MFCC) [15] with a clean dataset. Hidden Markov Model (HMM) [11] is implemented using Hidden Markov Toolkit (HTK) [12]. A N Mishra, Biswas and Chandra [4] designed a system for isolated digit recognition in Hindi. HMM is chosen as the classifier and Mel-Frequency Cepstral Coefficients (MFCC) [13] algorithm for the features extraction. They performed experiments using both HTK and Matlab [10], with both clean and noisy data. Pawar and Morade [5] designed a digit recognition system for isolated English digits with a huge database of 50 speakers using HMM and MFCC algorithm. HTK is used for training and testing purposes. Maruti Limkar [6] works on a system for speech recognition of a proposed approach to speech recognition of isolated English digit using MFCC and Dynamic time wrapping (DTW) [14] algorithm. Elitza Ivanova [7] worked on American and Chinese spoken English using HMM and HTK. Saxsena and Wahi [8] worked on Hindi digits recognition. They collected their data onto natural noise environments. Mohit Dua [1] also worked on digit recognition of the Punjabi language.

METHODOLOGY

Mishra [3] uses HTK to implement HMM for training and testing purposes for connecting Hindi digits. The database is prepared by 40 speakers, 23 female speakers and 17 male speakers using the Cool Edit software. For features extraction, MFCC and ΔMFCC along with the Revised Perceptual Linear Prediction (RPLP), Perceptual Linear Prediction (PLP) and Bark Frequency Cepstral Coefficients (BFCC) are used. MFCC is done through HTK and all other

http://doi.org/10.22232/stj.2016.04.02.10

features are extracted through Matlab and saved in HTK format. The analysis is done in both clean and noisy data. Mishra [4] performs experiments using HTK for isolated Hindi digits. Using 35 speakers for training and 5 speakers for testing, 3500 features are extracted. Out of which 350 features are chosen, 12 MFCC coefficients are obtained for each frame from where only 13 MFCC coefficients are chosen for vector quantization. 10 HMM are created for each digit. Pawar and Morade [2] use HTK for implementing HMM as a classifier. They prepared the database with 50 speakers, i.e., of 500 samples. 400 samples are used for training and 100 samples are used for testing. Clemson University Audio Visual Experiments (CUAVE) database is used for speaker independent environment. Maruti Limkar [6] uses MFCC algorithm to extract features, and implemented feature vector matching for training purposes. Dynamic Time Warping is used as a classifier rejecting the HMM. 100 samples are analyzed and results are obtained accordingly. Elitza Ivanova [7] works on a database of MPEG1 Audio Layer 3 (.mp3) samples of spoken English. 375-words passages are chosen for recording data. All .mp3 files are converted to corresponding wave format. HMM is used as the classifier for the study. Babita Saxena [8] uses MFCC for feature extraction of data collected from 10 speakers-8 for training and 2 for testing. The database is prepared in a noisy environment. HMM is used as the acoustic model here, with more than 61 context independent phonemes. Mohit Dua [1] and his team worked on automatic speech recognition on Punjabi language using classroom and open space environment. 115 distinct words are used and trained the system using HTK. MFCC and HMM are used in feature extraction and acoustic models.

RESULTS AND DISCUSSION

Mishra [3] discusses the efficiency of different feature extraction algorithm mainly PLP, RPLP, BFCC, and MF-PLP for both clean and noisy data using connected Hindi digit recognition system. The efficiency of algorithms for clean data, obtained during the procedure is shown in Fig. 1. The performance is based on the percentage of recognition. On the noisy environment, the system is tested with Babble noise, White noise, Pink noise and F16 noise. Each test is performed for three times with the SNR values of 5 dB, 10 dB, and 20 dB. The system is then evaluated with constant characteristics of 5-HMM and 9-Gaussian Mixtures. During the evaluation of 5 dB SNR ratio, the system does not work properly and give a result equal to the clean data result. For the noisy data, with all the respective noises, the system performed at its best with MF-PLP having 98% recognition rate. While MFCC produces an output of 96–97% accuracy.

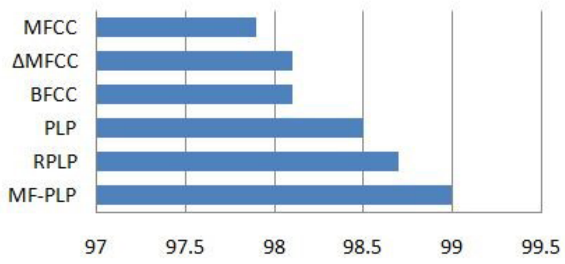

Fig. 1: Algorithm Performance on Clean Data.

MF–PLP gives us the best performance as compared to the other features extraction techniques. It is due to the fact that it incorporates Mel-filter into a perceptual linear features extraction method. PLP gives a better recognition than MFCC. It is because the signal is pre-emphasized by a simulated equal loudness curve to match the frequency magnitude. RPLP features are also shown good results for clean as well as noisy data. This is due to the fact that it takes advantage of pre-emphasis filter, Mel scale filter bank along with linear prediction and cepstral analysis [9]. Mishra [4] again performs the experiments on a system with isolated Hindi digit in the clean data environment at two different environments. In the first case, HMM is implemented using HTK. It also uses MFCC as a features extraction algorithm. Then in the second case, the same classifier and algorithm are implemented in the Matlab environment. The performance of the system in two separate environments is given in Table 1.

Table 1: Recognition Comparison.

Speaker No.	% Recognition (MATLAB)	% Recognition (HTK)
1	94	100
2	88	97
3	90	99
4	92	100
5	91	100

All the above results are obtained strictly for a clean environment. Thus, MATLAB gives an average recognition of 91% and HTK gives an average recognition of 99.2%. Pawar and Morade [2] obtains a result with 95% recognition rate. For the self-recorded database, they obtained recognition rate of 80% in average. Figure 2 represents the results obtained by Pawar and Morade [2]. Maruti Limkar [6] worked on automatic speech recognition by MFCC vectors to provide an estimate of the vocal tract filter. DTW is used to detect the nearest recorded voice with appropriate constraint. Accuracy was emphasized rather than recognition. 95% accuracy was obtained using the method. Using 200 datasets for training and 70 for testing. Elitza Ivanova [7] obtains a result with accuracy rate 70–75%. The work is done using HTK and implementing HMM. Table 2 gives the result for each digit with accuracy.

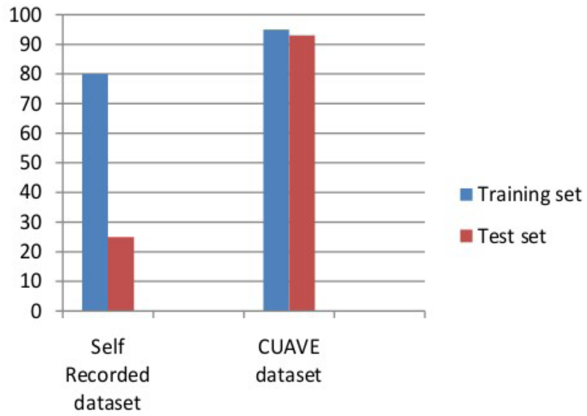

Fig. 2: Results Obtained by Ganesh S. Pawar[2].

Table 2: Results Obtained by Maruti Limkar[6].

Word	Accuracy	Word	Accuracy
Zero	80	Five	100
One	95	Six	80
Two	80	Seven	100
Three	100	Eight	100
Four	90	Nine	80

Saxsena and Wahi [8] work on digit speech recognition with 2 seen and 2 unseen speakers. Using HTK, implementing HMM and MFCC for features extraction; they obtained a result of word recognition equal to 85% during the testing with the unseen speaker. In speech recognition, the recognition rate of 95.63% and 94.08% are obtained by [1] in a classroom and open space environment respectively.

CONCLUSION

By comparison of all the work done by the respective researchers in speech recognition field, the following conclusion can be drawn as shown in Table 3. Here C, M, CL, and OS are used for the word Clean, Mixed, Classroom, and Open Space. Most of the people use Hidden Markov Model as the acoustic model. It is due to the fact that it provides better recognition and its efficiency is accepted universally.

Table 3: Comparison Statistic

Mishra	HMM	HTK	C	99
Mishra	HMM	Matlab	C	91
Pawar	HMM	HTK	CUAVE	95
Pawar	HMM	HTK	R	80
Limkar	DTW	HTK	C	95
Elitza	HMM	HTK	M	75
Babita	HMM	HTK	Noisy	85
Mohit	HMM	HTK	CL	95
Mohit	HMM	HTK	OS	94

Maruti Limkar [6] used Dynamic Time Warping which provides an accuracy rate of 95%. But recognition rate is emphasized as compared to accuracy. Ye-Yi Wang [9] proves a good accuracy but never indicates a good rate of recognition. In case of tools, HTK is chosen over Matlab, due to its efficiency in implementing HMM and being open source with better recognition rate. Most of the tools are provided with HTK for easy speech recognition. Mel-frequency cepstral coefficients are used to extract features where recognition level reduces by 1–2% as compared to MF-PLP. But simplicity makes researchers to compromises 1–2% and uses MFCC. Using HTK, MFCC algorithms is implemented directly which generates the Mel-coefficients. In noisy environments, the recognition level falls as compared to the clean database.

REFERENCES

1. Dua, M., Aggarwal, R. K., Kadyan, V. and Dua, S. 2012. Punjabi Automatic Speech Recognition Using HTK. IJCSI Int. J. Comp. Sci. 1: 359-364.
2. Pawar, G. S. and Morade, S. S. 2014. Isolated English Language Digit Recognition Using Hidden Markov Model Tool kit. Int. J. Adv. Res. Comp. Sci. Software Eng. 4: 1-5.
3. Mishra, A. N., Chandra, M., Biswas, A. and Sharan, S. N. 2011. Robust Features for Connected Hindi Digits Recognition. Int. J. Signal Processing, Image Processing and Pattern Recognition. 4: 1-12.
4. Mishra, A. N., Astik B. and Mahesh C. 2010. Isolated Hindi Digits Recognition: A Comparative Study. Int. J. Electronics Comm. Eng. 3: 229-238.
5. Ganesh S. Pawar, G. S., Sunil S. Morade, S. S. 2014. Isolated English Language Digit Recognition Using Hidden Markov Model Tool kit. Int. J. Adv. Res. Comp. Sci. Software Eng. 140: 1-7.
6. Limkar, M., Rama R. and Vidya S. 2012. Isolated Digit Recognition Using MFCC and DTW. Int. J. Adv. Electrical and Electronics Eng. 2: 11-20.
7. Elitza I., Sara K., Chris M., Marissa M., Ivo N. and Brandon, W. 2010. Recognizing American and Chinese Spoken English Using Supervised Learning. Retrieved from http://www.academia.edu/7283099.

8. Saxena, B. and Wahi, C. 2015. Hindi Digits Recognition System On Speech Data Collected in Natural Noise Environments. David C. Wyld et al. (Eds): CSITY, SIGPRO, DTMN.

9. Ye-Yi, W., Acero, A. and Ciprian, C. 2003. Is word Error rate a good indicator for spoken language understanding accuracy. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands, 23–30.

10. MATLAB 8.0 and Statistics Toolbox 8.1, The MathWorks, Inc., Natick, Massachusetts, United States.

11. Baum, L. E. and Petrie, T. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics. 37: 1554–1563.

12. Young, S. J. and Young, S. J. 1994. The HTK Hidden Markov Model Toolkit: Design and Philosophy, Entropic Cambridge Research Laboratory, Ltd, 2: 2-44.

13. Sahidullah, Md. and Saha, G. 2012. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication. 54: 543–565.

14. Sakoe, H. and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing. 26: 43–49.

15. Kamińska, D., Sapiński, T. and Pelikant , A. 2013. Comparison of perceptual features efficiency for automatic identification of emotional states from speech. 6th International Conference on Human System Interactions (HSI), Sopot, 210-213.