A Semantic-Aware Strategy for Automatic Speech Recognition incorporating Deep Learning Models

A Santhanavijayan[1], Naresh Kumar D[2*], Gerard Deepak[1]
[1] Department of Computer Science and Engineering,
[2*]Department of Mechanical Engineering
[1,2]National Institute of Technology, Tiruchirappalli, India
[2*]*d.nareshkumar1040@gmail.com*

Abstract: Automatic Speech Recognition (ASR) is trending in the age of the Internet of Things and Machine Intelligence. It plays a pivotal role in several applications. Conventional models for Automatic Speech recognition do not yields a high accuracy rate especially in the context of native Indian Languages. This paper proposes a novel strategy to model an ASR and Speaker Recognition system for the Hindi language. A semantic aware strategy incorporating Acoustic Modeling is encompassed along with Deep Learning Techniques such as Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs). MFCC is further imbibed into the model strategically for feature extraction and the RNN for pattern matching and decoding. The proposed strategy has shown a promising performance with the speaker recognition system yielding 97% accuracy and speech recognition system with a very low word error rate of 2.9 when compared to other existing solutions.

Keywords : Automatic Speech Recognition; Deep learning; Feature Extraction; LSTM

1. Introduction

Automatic speech recognition is a method where a given audio signal is converted into text. An Automatic Speech Recognizer behaves like an interface for making the use of the machines much easier. For the communications in Indian language such as Hindi, machine models as a huge help to the public in the native speaking countries. The Majority of the population of India is unaware of other languages like English. So, Automatic Hindi speech recognition is the necessity of the citizens of our country. There are several applications in which automatic Hindi speech recognition can be put to great use, such as government offices, railway stations, information retrieval system at railway stations, bus station, aviation stations, and etc. by providing the people with the answer to their problems faced by them. Healthcare sectors also find a good variety of use for ASR like documentation processes of patients' records. It has shown significant importance for helping people with short-term memory loss by treating them with prolonged speech to help them remember the required details thus reinforcing their brain to retain memory. It also finds use in military technologies like training air traffic controllers and helicopters. Education and daily life is another domain which can make use of the ASR systems. It helps in the education of students who are physically disabled or blind.

Organization: The remaining paper organization is as follows. Section 2 provides a brief overview of the related work of research that has been conducted. Section 3 illustrates the proposed architecture. Section 4 describes the Results and Performance Evaluation. and finally, the paper is concluded in section 6.

2 Related Work

K. Kuldeep et al, [1] have formulated a system for speech recognition employing the strategy of Hidden Markov model. The methodology was adopted for the processing of continuous speech signals. The feature extraction method used was relative cepstral transform and the word level segmentation was used for modeling the system. The model was for speech in Hindi language only and performance was found to be good for audio files which were in its database used for training. Mohit D. et al., [2] have implemented an model for recognition of speech modeled by the usage of Hidden Markov model. The model uses a method called discriminative training and feature extraction is done using perpetual linear prediction along with the Mel frequency cepstral coefficients. It mainly is using heterogeneous feature vectors. N. Rajput Kumar et al. have proposed a continuous system for recognition of speech using the Gaussian mixture models. Trigram model was used for the language model. The speech segmentation was done phoneme based. The feature extraction technique used was the Mel frequency cepstrum coefficient. Wu et al., [3] have proposed a scheme for ASR incorporating joint learning in front end speech processing. Tang et al., [4] have proposed a strategic model ASR by employing a multi-task model comprising of neural networks. Z Tang et al., [5] have proposed a joint collaborative model for speech and speaker recognition that incorporates multitask training in it where the outputs of each task are backpropagated. Tarun et al., [6] have proposed a scheme that amalgamates Vector Quantization and HMM model for isolated word recognition for the Hindi language. This method also imbibes peak picking and is suitable for the Hindi Language. Mishra et al., [7] have proposed a speaker independent strategy hybridizing Revised perceptual linear prediction (RPLP), Bark frequency cepstral coefficients (BFCC) and Mel frequency perceptual linear prediction (MF-PLP) for recognizing Hindi digits. Sinha et al., [8] have proposed a strategy for context-dependent speech recognition for Hindi Language using Hidden Markov Model with continuous density. Also, HLDA is incorporated for feature reduction.. Certain Semantic approaches as depicted in [9-21] can be imbibed along with Deep Learning Models especially for the rearrangement of phonemes in order to achieve a much efficient technique, as semantic aware systems are more intelligent and efficient.

3 Proposed System Architecture

The proposed system architecture is depicted in Figure 1. The process of Feature extraction is a method of extracting a set of properties of an utterance having certain acoustic relations and matching with the speech signal. A feature extractor removes irrelevant properties from the input and retains the required properties. For doing this a part of the speech signal is taken into consideration for the purpose of processing. This portion is called window size. A frame is the required data received from a window.The range of frame is usually between 10-26 milliseconds. This range has an overlap of nearly 50%–70% between two sequential frames. The data obtained from this analysis interval is then required to be multiplied with a windowing function. The proposed model is using MFCC for extracting features. For MFCC, an audio signal is constantly changing, and the short time scales the audio signals do not change much. This is the reason the frame of the signal is presumed into 20-40 millisecond frames. Power spectrum calculation is the next step of every frame. The periodogram in the proposed approach performs a similar job by identifying which frequencies are present in the frame.

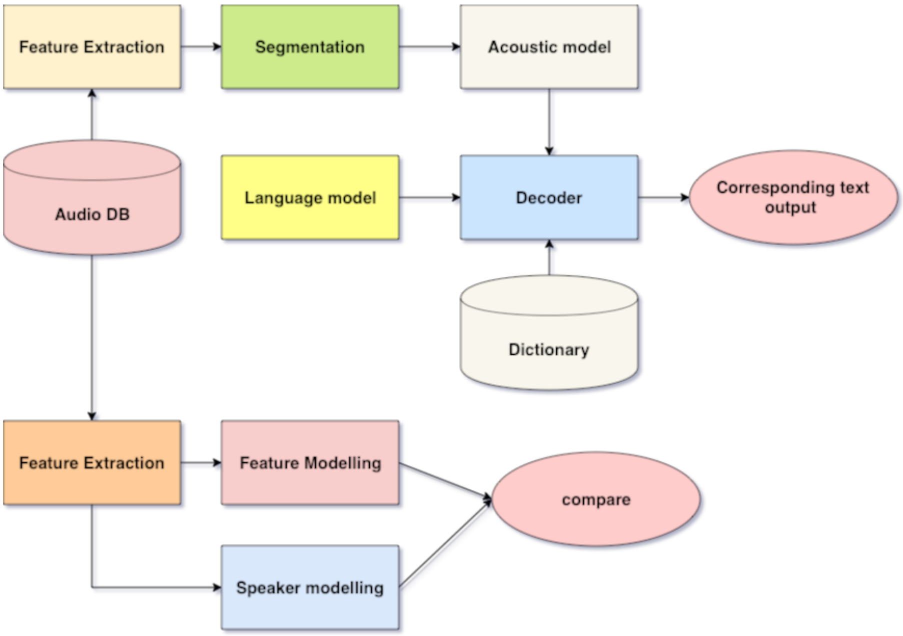

Figure 1: Proposed system architecture

There is a lot of information in periodogram spectral that is actually not required for the computation of Automatic Speech Recognition (ASR). Due to this, periodogram bin clumps are taken and are summed up inorder to obtain an awareness of the existing energy in various regions of frequency. Mel filterbank performs this process: the initial filter gives an indication of the energy existing near 0 Hertz and is very narrow. These filters get wider following with the trend of frequencies getting higher. The width that is present is directly proportional to the frequency available. At these high frequencies, these filters become less concerned about the variations. The main concern is the spot at which the maximum energy occurs. The method of spacing of the filterbanks is clearly depicted by the Mel scale. As soon as the filterbank energies are obtained, logarithm is applied and this application facilitates the use of cepstral mean subtraction. This is a technique that involves channel normalization. Computation of DCT of log filterbank is the final process. The main motives behind this processes can be stated as two. One being the fact that the filterbanks are overlapping and the other reason being the energies of these filterbanks are correlated with one another. The decorrelation of these energies is done by DCT. It should be noticed that only 12 of the present 26 DCT coefficients. The reason for this is that fast changes are represented in higher values of

DCT co-efficients and these higher values are responsible for the degradation of ASR performance. Hence improvement in performance can be achieved by dropping these values.

The process of identifying the boundaries between spoken words is called speech segmentation. The boundary detection will enable us to form the phonemes and which can be used form the dictionary of phonemes to be used by the acoustic and language models. It is necessary to take into account semantics as well as the grammar while processing the natural languages. The Speech Segmentation is achieved using HMM-based Phonetic. Segmentation scheme. Also, it has been modeled as a five-state HMM under the consideration that it is another phoneme. The phoneme HMM is initialized using a flat start training where the models were equally initialized. Therefore, there was no necessity of manually segmented database for training for the purpose of initializing the HMM. The segmentation result is probabilistic and is still a major challenge for many natural languages. Text Transcriptions are carried out using Unicode Devanagari characters is to be formed by converting the raw input given in the form of text which is further converted into Roman characters by further processing since the keywords available is having a corresponding ASCII code generated. Translation of all the input text in our database using this method is done. The input to the phoneme parser is a text which then extracts phonemes and arranges them to form a sequence of words from the given sentence and produce the list of phonemes as output.The Acoustic models are put to use to match the features observed of the given speech signal with the hypothesis for the expected phonemes. The acoustic model is one of the main components of ASR and accounts for the computational load and system. One of the common implementations is using a probabilistic method which uses the hidden market models. A training process is used for the purpose of generating a mapping of the speech units like phonemes and observations found from the acoustics. For the purpose of training, a pattern representing the features is created for each class that should correspond to the speech utterances of the corresponding class. A database rich in phonetics and database which is balanced is required for the purpose of training the acoustic models. For the purpose of transcriptions into linguistic units from acoustic features, a large variety of representations are used as whole words and syllables.

The estimation of convergence of class posteriors due to the application of cross-entropy is done by using recurrent neural networks which has SoftMax output layer. The training method generally followed is by using targets along with some alignments. Bootstrapping or flat start can be used for the purpose of aligning the acoustic feature sequences with the help of an existing model. The detection of phonetic units can be predicted using the phonemes that occur preceding it. Different models having different contexts can be modeled to enhance the modeling power as well. A three-layer long short-term memory cell recurrent neural networks have been used in the proposed system. There are eight hundred memory cells in each of long short-term memory layer. The recurrent projection layer is five hundred and twelve units. The SoftMax activation function used for the output layer. A tangent activation function is also present in each layer along with a sigmoid function for calculation. The long short-term memory is given an at an interval of twenty-five-millisecond frame of 40-dimensional logmel filterbank features. Better decisions were formed from the information got from previous frames and future frames and hence were able to make better decisions by the neural network. Computation of DCT for the energies of log filterbanks is the final step.

4 Results and Performance Evaluation

The experimentation was realized in a Python 3.6 environment. The system imbibed a RAM specification of 4GB under the Ubuntu operating system. The dataset used is from Hindi corpus-Technology for Development of Indian Languages (TDIL). The dataset had 150 speakers and each speaker has roughly around 95 audio files. The performance is evaluated using the Word Error Rate (WER) and accuracy as potential metrics. Figure 2 shows a graph depicting the average word error rate of the proposed model during training through different epochs. The system was trained for 350 epochs. The graph shows there is a decrease in the WER with each epoch passed. The first epoch of the system showed very high word error rate as the neural network was not having enough knowledge and loss from which it can predict the speech. As the model passes through the iteration in each epoch the knowledge of the neural network is becoming more and better. The neural network slowly will learn to use the parameter inputs got from the language and acoustic model and predict better. Thus, as the epochs increases, there is a decrease in the word error rate. The proposed model is outperforming because of the incorporation of attention mechanism in the sequence to sequence model of the recurrent neural network.

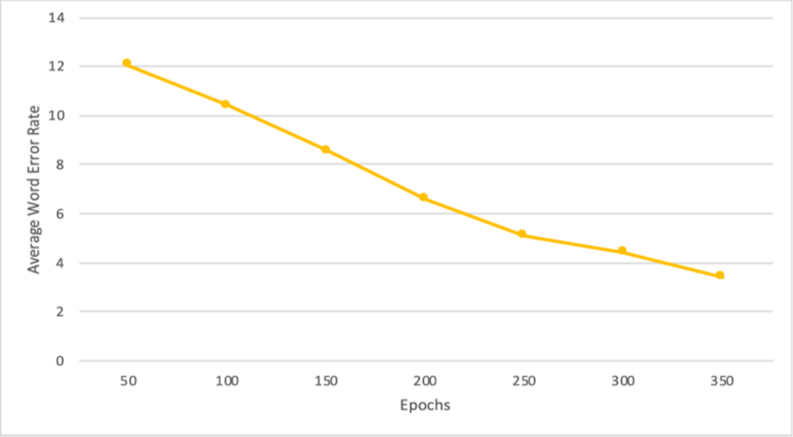

Figure 2: Graph of Average WER of the proposed model

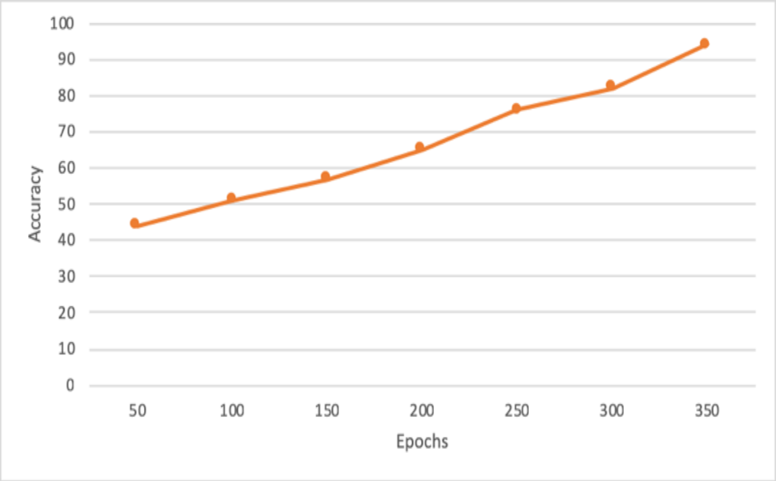

Figure 3: Graph of the accuracy of the proposed Speaker Recognition System

Figure 3 shows the graph plotted during the training period for the accuracy of the system. speaker recognition system was trained for 350 epochs using recurrent neural networks. A total of 150 speakers and each speaker has 95 audio files spoken. The accuracy of the system was observed to increase with each subsequent epoch. The final accuracy was found to be 97%. Figure 4 depicts a comparison of accuracies of different models trained in the same environment as the proposed system. The proposed system is making use of recurrent neural networks. The proposed model is better because of the incorporation of attention mechanism which helps in training the neural network in such a way that it learns from previous iterations and thus learning from the errors made in prediction in the previous iteration.

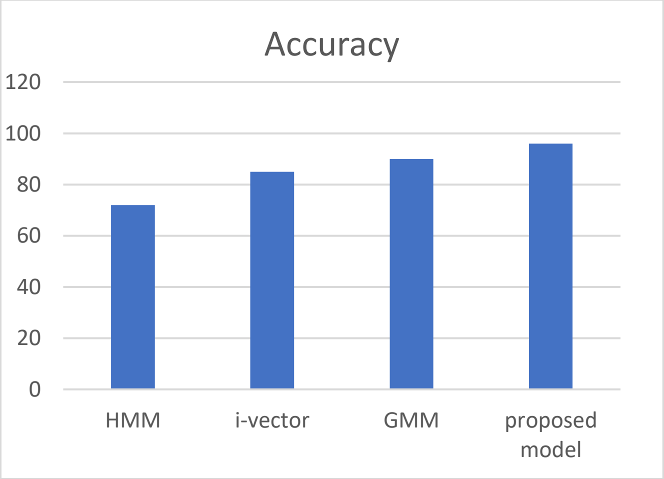

Figure 4: Comparison of existing Speaker Recognition Systems

5. Conclusions

A novel strategic approach for automatic speech recognition as well as speaker recognition encompassing a semantic-aware deep learning model. This paper devises a mechanism for continuous speech and speaker recognition of Hindi language. A triphone model is incorporated as Acoustic model and bigram model is encompassed as the Language model. The formulated strategy yielded promising performance with an accuracy of 97%, which is much more efficient when compared to the existing models. The model can be implemented on cross-platform operating systems and the results can be successfully obtained. In addition, the proposed model also yielded a very low word error rate of 2.9 which is vastly better than the other existing models. LSTM techniques incorporating Deep Learning has proven to be much more robust in speech recognition and have played a major role in obtaining the best in class accuracy in the context of Indian Languages.

References

[1] Kumar, Kuldeep, R. K. Aggarwal, and Ankita Jain. "A Hindi speech recognition system for connected words using HTK." International Journal of Computational Systems Engineering1, no. 1 (2012): 25-32.

[2] K.Mohit, Nittendra Rajput, and Asish A.Verma."A large vocabulary continuous speech recognition system for Hindi." IBM Journal of research and development 48.5.6: pp. 723-755, Nov. 2016

[3] Wu, Bo, Kehuang Li, Fengpei Ge, Zhen Huang, Minglei Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee. "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition." IEEE Journal of Selected Topics in Signal Processing 11, no. 8 (2017): 1289-1300.

[4] Z. Tang, L.Li, and D.Wang “Multitask recurrent model for speech and speaker recognition,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 20, no. 2, pp. 493–504, Mar. 2016.

[5] Tang, Zhiyuan, Lantian Li, Dong Wang, Ravichander Vipperla, Zhiyuan Tang, Lantian Li, Dong Wang, and Ravichander Vipperla. "Collaborative joint training with multitask recurrent model for speech and speaker recognition." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 25, no. 3 (2017): 493-504.

[6] Pruthi, Tarun, Samer Sakssena, and Prathip K. Das. "Swaranjali: Isolated word recognition for Hindi language using VQ and HMM." International Conference on Multimedia Processing and Systems (ICMPS), IIT Madras, Dec.2016.

[7] Mishra, A. N., Mahesh Chandra, Astik Biswas, and S. N. Sharan. "Robust features for connected Hindi digits recognition." International Journal of Signal Processing, Image Processing and Pattern Recognition 4, no. 2 (2011): 79-90.

[8] Sinha, S., Agrawal, S.S. and Jain, A., 2013, August. Continuous density Hidden Markov Model for context-dependent Hindi speech recognition. In 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1953-1958). IEEE.

[9] Gerard D and Z. Gulzar. "Ontoepds: Enhanced and personalized differential semantic algorithm incorporating ontology-driven query enrichment." Journal of Advanced Research in Dynamical and Control Systems 9, no. Special (2017): 567-582.

[10] Giri, G.L., Deepak, G., Manjula, S.H. and Venugopal, K.R., 2018. OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 265-275). Springer, Singapore.

[11] Deepak, G . and Priyadarshini, J.S., 2018. Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. Computers & Electrical Engineering, 72, pp.14-25.

[12] Deepak, G., Priyadarshini, J.S. and Babu, M.H., 2016, October. A differential semantic algorithm for query relevant web page recommendation. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA)(pp. 44-49). IEEE.

[13] Pushpa, C.N., Deepak, G., Thriveni, J. and Venugopal, K.R., 2015, December. Onto Collab: Strategic review oriented collaborative knowledge modeling using ontologies. In 2015 Seventh International Conference on Advanced Computing (ICoAC) (pp. 1-7). IEEE.

[14] Deepak, Gerard, Ansaf Ahmed, and B. Skanda. "An intelligent inventive system for personalized webpage recommendation based on ontology semantics." Int. Journal of Intelligent Systems Technologies and Applications 18, no. 1/2 (2019): 115-132.

[15] Deepak, Gerard, and Sheeba Priyadarshini. "A hybrid framework for social tag recommendation using context driven social information." International Journal of Social Computing and Cyber-Physical Systems 1, no. 4 (2016): 312-325.

[16] Deepak, G., Shwetha, B.N., Pushpa, C.N., Thriveni, J. and Venugopal, K.R., 2018. A hybridized semantic trust-based framework for personalized web page recommendation. International Journal of Computers and Applications, pp.1-11.

[17] Pushpa, C.N., Deepak, G., Thriveni, J. and Venugopal, K.R., 2016. A Hybridized Framework for Ontology Modeling incorporating Latent Semantic Analysis and Content based Filtering. International Journal of Computer Applications, 150(11).

[18] Deepak, G. and Priyadarshini, J.S., 2018. A Hybrid Semantic Algorithm for Web Image Retrieval Incorporating Ontology Classification and User-Driven Query Expansion. Advances in Big Data and Cloud Computing, p.41.

[19] Gulzar, Z., Leema, A.A. and Deepak, G., 2018. PCRS: Personalized course recommender system based on hybrid approach. Procedia Computer Science, 125, pp.518-524.

[20] Deepak, Gerard, and S. J. Priyadarshini. "Onto tagger: ontology focused image tagging system incorporating semantic deviation computing and strategic set expansion." Int. J. Comput. Sci. Bus. Inform 16, no. 1 (2016).

[21] CN, Pushpa, Gerard Deepak, Mohammed Zakir, and Venugopal KR. "Enhanced Neighborhood Normalized Pointwise Mutual Information Algorithm for Constraint Aware Data Clustering." ICTACT Journal on Soft Computing 6, no. 4 (2016).