ISSN No:-2456-2165

Comparative Analysis of Speech Recognition

[1]Anusha U, [2]Avantika Holla S, [3]Eshwar Thilak R, [4]Gowthami C P, [5]Hamsaveni M
Assistant Professor
Vidyavardhaka College of Engineering, Mysuru, India

Abstract:- communication is the most important form of exchanging information. Speech is the major mode for the transfer of information. The speech recognition system which is also called as automatic speech recognition (ASR) is a way in which the computer/system understands the speech, automates it and converts it into text. There are different techniques to achieve this result which involves speech analysis, feature extraction, modelling techniques like the hidden markov model (HMM), neural networks, and DTW algorithm.

Keywords:- *Feature Extraction, Speech Analysis, Hidden Markov Model (HMM), Neural Networks, DTW Algorithm.*

I. INTRODUCTION

Speech recognition is a system that develops methodologies that helps in recognizing and translating spoken languages into texts. The spoken words and phrases are converted into machine readable format. It also recognizes vocabularies (dictionaries) which are a list of words or utterances. Speech recognition is of different types mainly isolated words, connected words, continuous speech, spontaneous speech and voice verification.

Speech recognition mainly deals with 2 algorithms namely: acoustic modelling and language modelling. Hidden Markov Model, Dynamic Time Warping, Neural Networks, Recurrent Neural Networks and End to End speech recognition are the main approaches for speech recognition.

II. LITERATURE SURVEY

An approach was developed to find the optimal forgetting factors in sequential estimation algorithms. By making use of gradient based algorithms forgetting factor is tracked and by differentiating the term-wise recursion the derivatives are calculated. By doing this it was possible to simultaneously optimize both the forgetting factor and the required parameter. When tested on different speech recognition tasks corrupted by artificially added noise and by field noise, it was observed that the proposed system performed better than the batch estimation techniques. A dynamic one pass decoder was used for decoding which resulted in the WER of 4.7% for speech without any added noise. The system made use of two types of noises. The first was white Gaussian noise at 10dB SNR and by mixing two white Gaussian noises at 10 and 5 dB SNR respectively the second noise was obtained. These noises led to an increase in word error rate to about 46.6% and 51.6% respectively. [1]

The fast match is a system where accept/reject decision is made at the phoneme level. Likelihood ratio test was developed which was performed at each and every frame by making use of hypothesis testing framework with an intention to develop an optimal solution. An acoustic fast-match consisted of a 2-pass algorithm in which the fast-match applied in the first-pass made use of biphone acoustic model and a bigram language model. By doing this it was possible to limit the search space. Further the selected candidates were rescored using triphone acoustic models and a backward trigram models in the second pass. Based on the result of the test, a phoneme was either accepted or rejected. If the phoneme passes, an arc was expanded. This method led to savings in search space because if a phoneme is rejected then automatically all arcs carrying that phoneme identity will be rejected. The proposed model led to 20-30% speed up in overall search without any loss in accuracy. [2]

Template matching was used to overcome the problems of Hidden Markov Models (HMMs). The key problems of HMMs was that it discards time dependencies information and is susceptible to overgeneralization. The recognizer here used the Dynamic Time Warping (DTW) algorithm. However it led to increase in the overall search space. Hence top-down search was replaced with the bottom-up approach and also DTW algorithm was extended with a sub word unit mechanism and a class sensitive distance measure suggested by the Hidden Markov Model systems. Using 30ms overlapping frames the 16 kHz audio input was transformed into a feature vector of 25 dimensions during preprocessing. The database consisted of only single phoneme templates. The template based matching systems led to performance worse than the HMM systems but when combined with the HMM, the combination of both the systems led to a decrease in word error rate to about 17%. [3]

A system mainly designated for Hindi language was developed where in, to train and identify the speech Hidden Markov Model was used and for feature extraction Mel Frequency Cepstral Coefficients (MFCC) was used. Hidden Markov Model toolkit (HTK) was used to accomplish this task. By making use of acoustic word models it recognizes the isolated words. The architecture consisted of mainly two modules, training and testing modules. During the testing phase, the model produced by the training model was used. The system was trained for a total of 30 Hindi words

collected from eight different native Hindi speakers. At 16 kHz the input speech is sampled and thereafter processed at a rate of 10ms with a 25ms Hamming window. Five speakers were used for testing the system and the results indicated that the accuracy of the generated system was about 94.63%. [4]

The end-end speech recognition with recurrent neural networks is a system that directly converts the audio sample to its corresponding text without the need of an intermediate phonetic representation. Coming to the architecture, it mainly consist of a deep bidirectional long short term memory which is a part of a recurrent neural network along with an objective function called the CTC(Connectionist Temporal Classification).There is a direct optimization of WER due to presence of an objective function which is introduced to train the network which eventually minimizes the expectation of an arbitrary transaction loss function .this architecture compresses the earlier speech system where most of them are replaced by a RNN architecture. The input data is being presented as spectrograms obtained from raw audio files using "spec gram function" .With minimal pre-processing and no explicit phonetic representation a character-level speech transcription is performed. The WER obtained is 27.3%. [5].

Neural network can be used for converting audio to text mainly for a large vocabulary conversational speech that maps to acoustic input characters, the character level language model and also beam search decoding. Here, the errors are used to train the speech recognizer which eliminates the complex infrastructure .the system is being trained and decoded by reasoning at character level. The system has no clue of when words occur within an utterance. The 2 neural network models are used within the deep bidirectional recurrent neural network, which at each time stamp maps to the acoustic input features to a probability distribution over characters and the other neural network i.e., character language model which enables to leverage high order n-grams context without the increase in number of free parameters on the model. CTC loss function is used which assumes the character output at each time stamp are independent given the inputs. After this, the decoding is being performed using beam search decoding that uses a beam search to combine both CLM and outputs of DBRNN.[6]

The acoustic can be converted to characters using Listen Attend and Spell (LAS) which is a speech recognizer. Chain rule decomposition is used to train the entire model. The speech utterances are being translated to text without the use of pronunciation model. The LAS consists of two models the listener which performs an operation called listen and consists of RNN which is the encoder. The Speller is another model which acts as the decoder RNN .This system uses an end-to-end learning framework which does not focus on making assumptions on the nature of probability distribution of output character sequence. The speech is converted to high level features and in turn, the speller coverts these high level features into output as text by notifying the probability distribution over the next character. The BLSTM RNN is being used by the listener operations. This achieved a WER of 14.1%. [7].

The character-based and dialog session can be modelled using language models like LSTM and also adds a CNN-BLSTM acoustic model. This model is developed for switchboards and callhome domains. The architecture mainly consists of two stage. The first stage includes the combination of frame level with the acoustic model. This stage is then followed by the world level through confusion networks. The CNN model architecture are of two types they are ResNet and LACE. This yielded a WER of 5.1%. [8]

III. COMPARATIVE ANALYSIS

Speech recognition when performed with different machine learning algorithms and feature extraction techniques produced different results. When HMM was used it produced a speedup on overall search of 3040%. When DTW was used along with HMM, the word error rate was reduced by 14%. In an end-to-end system bidirectional LSTM was used which is a sub unit of RNN which obtained the WER of about 27.3%. Later they upgraded the system and used BLSTM RNN for LAS which gave 14.1%. Therefore, according to analysis HMM along with DTW produced better results.

IV. CONCLUSION

A speaker independent speech recognition system has been proposed in the above information. The training of the data can be simplified using End to End approach which does not require a dictionary and it maps the audio directly to its texts without the need of determining any previous data. This uses LSTM to optimize an objective function called CTC. But the standard approach is the combination of HMM and DNN that require a language model and a phonetic dictionary beforehand. Therefore from the study LSTM with a CTC objective function is preferred.

REFERENCES

[1]. Mohamed Afify and Olivier Siohan, Sequential Estimation With Optimal Forgetting for Robust Speech Recognition,IEEE Transactions On Speech And Audio Processing, Vol. 12, No. 1, January 2004.
[2]. Mohamed Afify, Feng Liu, Hui Jiang, A New Verification-Based Fast-Match for Large Vocabulary Continuous Speech Recognition, IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4, July 2005.
[3]. Mathias De-Wachter et.al. Template based continuous speech recognition, IEEE transactions on Audio, speech and Language processing, Vol.15, No.4, May 2007.

[4]. Kuldeep Kumar R. K. Aggarwal, "Hindi speech recognition system using HTK", International Journal of Computing and Business Research, vol. 2, issue 2, May 2011.
[5]. Towards End-to-End Speech Recognition with Recurrent Neural Networks. Alex Graves Google DeepMind, London, United Kingdom Navdeep Jaitly Department of Computer Science, University of Toronto, Canada.
[6]. Lexicon-Free Conversational Speech Recognition with Neural Networks. Andrew L.Maas∗, Ziang Xie∗ , Dan Jurafsky, Andrew Y.Ng.
[7]. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. William Chan, Carnegie Mellon University, Navdeep Jaitly, Quoc Le, Oriol Vinyals Google Brain.
[8]. The Microsoft 2017 conversational speech recognition system.W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke.