IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LITERATURE SURVEY ON VOICE-CONTROLLED WEB APP

[1]Raj Gandhi, [2]Romil Desai, [3]Marmik Modi, [4]Dr. Suvarna Pansambal

[123]Student, [4]Professor

Department of Computer Engineering,

Atharva College of Engineering, Mumbai, India

Abstract: Voice or speech recognition systems allow a user to make a hands-free request to the computer, which in turn processes the request and serves the user with appropriate responses. A voice-controlled system embedded in a web application can enhance user experience and can provide voice as a means to control the functionality of e-commerce websites. In this paper, we have reviewed speech recognition systems (SRS) architectures and present their advantages, disadvantages, and the approach followed by various authors to get the desired result.

I. Introduction

Speech recognition is a process of converting the sound of words or phrases spoken by humans into text that closely matches the original text. The goal of speech recognition is to enable people to communicate more naturally and effectively. This often requires deep integration with many natural language processing (NLP) components. Speech recognition can be applied to many domains and applications. By allowing users to control the functionality of the applications with their voice, Speech recognition can enhance users' browsing experience, and allow users to effectively convey their instructions and requests using natural languages.

II. Literature Survey

W. Chan et.al [1] introduced a new way of converting speech to text. The algorithm is an attention-based neural network that can directly transcribe acoustic signals to characters.

It is trained end-to-end and has two major components.

a. Listener: It consists of an RNN encoder that transforms the input into high-level features.
b. Speller: The input from the listener is then passed on to the RNN decoder that attends to high-level features and spells out the transcript one character at a time.

This paper identifies the problem of out of vocabulary words by generating characters instead of a whole word which might not be correct and can further lead to the wrong prediction of other words as well. On top of this beam search is used to get the output probabilities of the top best prediction of words and from that, the one with the highest probabilities is selected.

M. Radfar et.al [2] focuses on understanding the semantic information from audio signals for converting speech-to-text. The architecture consists of a self-attention mechanism by which the semantic context is extracted from the audio signal. This end-to-end transformer computes the correlation between all input vector pairs and thus the model knows where to attend to infer semantic information embedded in the audio signal. The model achieves higher accuracy and reduction in size compared to two other competitive models i.e

RNN and RNN-Sincnet. This is a hierarchical model which can be used for variable length domains. It is also highly parallelizable which makes it a good candidate for on-device speech language understanding.

George Saon et.al[3] have presented a set of improvements in language modelling, LSTM, and acoustic to English SwitchBoard system resulting in better word error rate. When recurrent and convolutional neural networks are combined the resulting accuracy improves on a variety of test sets. In this paper, they have experimented with ResNet and LSTM with several layers. For language models, various types of models are used such as word-lstm, char lstm, etc. The best one achieves a word error rate of 5.5%.

Ladislav Mo˘sne et.al[4] noticed that for real-world speech recognition applications there can be noise in the data and to make the system robust to noise is a challenge. Although large vocabulary speech recognition is of high accuracy by applying neural networks, it requires thousands of hours of time which is time consuming and expensive to collect, and its performance under a noise environment may still suffer. The clean data is augmented with noise and then both the data i.e clean and clean+noisy is passed to the two identical architectures each of which contains LSTM layers and softmax and then the error is calculated and backpropagated. It gives the word error rate of 10.1% and 28.7% on clean and noisy test sets. This is achieved by keeping the softmax temperature to 2.

W. Xiong et. al[5] proposed a speech recognition system that uses neural network-based acoustic and language modeling to improve the word error rate. The combination of ResNet architecture and RNNLM, achieves a word error rate of 6.2%. Since this model is costly to compute, that's why they have parallelized training. The CNN model uses window sizes of the input audio whereas LSTM processes one frame at a time. This also includes noise, laughter, and silence. The N-best hypotheses are then rescored using a combination of the large N-gram LM and several RNNLMs.We found the best results with an RNNLM configuration that had a second, non-recurrent hidden layer. This produced lower perplexity and word error than the standard, single-hidden-layer RNNLM architecture. The number of out-of-set words is recorded for each hypothesis and a penalty for them is estimated for them when optimizing the relative weights for all model scores.

Amin Fazel et.al[6] noticed that End-to-End Automatic Speech Recognition models have shown superior performance over the traditional hybrid ASR models. But training an end-to-end ASR requires a lot of data which is not only expensive but may also raise dependency on production data. It consists of a multi-context text-to-speech engine to generate synthetic speech, and an RNN model for speech recognition. The model for the text-to-speech engine is trained and evaluated independently from the speech recognition model. The training data are sampled from a combination of real speech recordings and TTS based synthetic speech audio. The ratio between real recordings and synthetic audio seen during RNN training is optimized with sampling weights. This method well mixes the real and synthetic data in each batch so that the ASR model sees both data. In addition, data augmentation is applied at both audio level and feature level for RNN training.

Zhen Huang et. al[7] proposed that very deep CNN's can achieve state-of-the-art results in speech recognition, but are difficult to train. The most popular way to train very deep CNN's is to skip connection with batch normalization. This paper introduces a new activation function called SELU(scaled exponential linear unit) which can reduce training time without affecting the accuracy of the system. By using this activation function, there is no need for both skip connection and batch normalization. Using SELU in ResNet can achieve the same or lower word error rate and it also reduces latency.

Andrew Y. Ng et. al[8] proposed an architecture which is significantly simpler than traditional speech systems, which rely on laboriously engineered processing pipelines does not need hand-designed components to model background noise, reverberation, or speaker variation, but instead directly learns a function that is robust to such effects. The model consists of simple RNN layers and language models for predicting the text. They have also applied regularization, model parallelism, and striding to make the system more robust.

Wenyong Huang et.al[9] proposed an End-to-End Speech Recognition which achieves competitive performance on libre speech dataset with 3.6% WER on the test set. The architecture described in this paper consists of three blocks. Each block is composed of three convolution layers followed by a unidirectional transformer. Self-attention is used to focus attention on input while producing output. During decoding, the self-attention of the unidirectional transformer attends to all history content, which increases the computational cost. Due to this, they have limited the history content of self-attention with a fixed sized window which in turn makes the computational cost of each step constant.

Anmol Gulati et.al[10] achieved the best of the worlds by studying how to combine convolutional neural networks and transformers. A conformer block is composed of four modules stacked together, i.e, a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module in the end. The proposed Conformer block contains two Feed Forward modules sandwiching the Multi-Headed Self-Attention module and the Convolution module. They found that the convolution module stacked after the self-attention module works best for speech recognition. The model exhibits better accuracy with fewer parameters than previous work on the LibriSpeech dataset, and achieves a new state-of-the-art performance at 1.9%/3.9% for the test set.

Table 1: Summary of Literature Survey

Sr.No	Authors	Approach	Advantages	Disadvantages
1	William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals	Using Bidirectional LSTM and attention mechanism for predicting the character sequence, one character at a time.	Out of vocabulary words are automatically handled as the output is a character and not a word. On top of this, beam search is used for predicting the correct sequence of words.	The model achieves 14.1% WER on the google voice search task. The error can be reduced by using convolutional neural networks.
2	Awni Hannun, Andrew Y. Ng, Carl Case et.al	The model is fairly simple to understand and implement. The model uses RNN and n-gram language model for getting accurate predictions.	Additional noise and strides are added to the input, to get the model to generalize better.	A bidirectional model and CNN can be added to improve the performance of the system.
3	Martin Radfar, Athanasios Mouchtaris, Siegfried Kunzmann	This model consists of embedding layer, encoder, decoder and attention. The SLU gives domain, intent and slots given and input.	The accuracy of this model is greater than CNN-RNN architectures.	Since, the word embeddings are learned with the model, there is a chance that the model may not output the spoken word, if the word has not appeared in the training data.
4	George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas et.al	IBM' ASR consists of Resnet Bi-directional LSTM and language model for speech recognition. It uses n-gram language model for getting accurate prediction	Improved our bidirectional LSTMs through speaker adversarial training and we replaced VGG nets with ResNets which resulted in low WER	Since the language model is n-gram, if new word appears and n-gram model doesn't recognize than it can output some other word.

5	Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote et.al	When we don't have enough data to train a ASR end-to-end, we can use synthetic data. We can create synthetic data by converting text to speech and then we can use ASR for speech to text	Synthetic data can be created when there is less training data. It can also reduce cost of collecting data and processing.	As we are using text to generate speech, we cannot find the context in the text.Homographs are spelled the same way but they differ in meaning and usually in pronunciation
6	W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer et.al	The ASR is made using CNN, Bidirectional LSTM and language model. RNNLM is used which reduces the WER	The combination of convolutional, language model and lstm gave a performance boost.	The use of language model has some limitations as well. Out of vocabulary word can be missed out.
7	Ladislav Mo˘sner, Minhua Wu, Anirudh Raju et.al	Introducing noise in the dataset and comparing it with the original value to see how robust the ASR is if it is subjected to noise.	The speech recognition system improved by this method which can be used for real world applications	There is the parameter called temperature which needs to be changed based on the use case of the speech recognition
8	Zhen Huang, Tim Ng, Leo Liu, Henry Mason, Daben Liu et.al	Scaled Exponential linear unit(SELU) activation function is introduced to induce self-normalization.	It can converge very deep networks without skip connection and batch normalization. It can reduce latency of the network significantly.	It is a method to reduce latency in the speech recognition so there are no drawbacks.
9	Wenyong Huang, Wenchao Hu, Yu Ting Yeung, Xiao Chen et.al	This is an end-to-end speech recognition system which is streamable which have low latency. The ASR consists of convolutional layers followed by a unidirectional transformer. It also has a better word error rate of 3.6% on test set.	This architecture achieves better WER on test set without using language model. Some of the speech recognition system works offline but this can work while streaming with low latency.	The dataset used to build this system is libre speech, but in input speech there can also be words that are not present in the training dataset. Because of this new words cannot be detected unless it contains in the training set.
10	Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar et.al	The speech recognition described in this paper consists of convolutional and transformer blocks. It introduces a new block called as conformer which helps in extracting information from audio.	The combination of convolutional and transformer proved to be best for speech recognition.	It mainly focuses on how convolutional layers can improve the system. The out of vocabulary words might not get converted to text.

III. Conclusion

The Listen Attend and Spell architecture identifies out of vocabulary words by generating characters instead of a whole word. It is an attention-based neural network can directly transcribe acoustic signals to characters. When recurrent and convolutional neural networks are combined the resulting accuracy improves on a variety of test sets. For language models, various types of models are used such as word-lstm or char lstm which achieves better word error rate. For real world speech recognition applications there can be noise in the data, to tackle this clean voice data is augmented with noise and then the models are trained so that it robust to noise as well. Training and end-to-end ASR requires a lot of data so a synthetic voice dataset can be generated and the model architecture can be trained on synthetic data as well as clean data to get better output. The architecture of the speech to text system containing deep CNN layers can be difficult to train, so a new activation function was introduced named SELU which reduces the training time without affecting accuracy.

IV. Acknowledgement

We are sincerely grateful for having Dr. Suvarna Pansambal as our guide and Head of Computer Engineering Department, for giving us the opportunities and time to conduct and present research on the topic. Research would have seemed difficult without their motivation, constant support and valuable suggestions.

References

[1] William Chan et.al Navdeep Jaitly, Quoc V. Le, Oriol Vinyals "Listen Attend and spell", IEEE ICASSP, 20 Aug 2015, Shanghai, China.

[2] Martin Radfar, Athanasios Mouchtaris et.al "End-to-End Neural Transformer Based Spoken Language Understanding", Interspeech, October 25-29 2020, Shanghai, China.

[3] George Saon, Hong-Kwang J. Kuo et.al "English conversational speech recognition", Interspeech, 2015, New York.

[4] Ladislav Mo˘sne, Minhua Wu, et.al "Improving Noise Robustness Of ASR", ICASSP, 2019, USA.

[5] W. Xiong, J. Droppo et.al, "THE Microsoft 2016 Conversational Speech Recognition System", IEEE ICASSP, Volume V2, Jan 2017, USA.

[6] Amin Fazel, Wei Yang et.al, "SynthASR: Unlocking Synthetic Data for Speech Recognition", Interspeech, August 30 - September 3, 2021, Brno, Czechia

[7] Zhen Huang, Tim Ng et.al "SNDCNN: Self-Normalizing Deep CNNs With Scaled Exponential Linear Units For Speech Recognition", ICASSP, May 4-8 2020, Spain

[8] Andrew Y. Ng et.al,"Deep Speech: Scaling up end-to-end speech recognition",CoRR, December 19,2014, 1195 Bordeaux Avenue, Sunnyvale CA 94086 USA.

[9] Wenyong Huang et.al, "Conv-Transformer Transducer: Low Latency, Low Frame Rate, Streamable End-to-End Speech Recognition", Interspeech 2020, October 25-29, China.

[10] Anmol Gulati et.al, "Conformer: Convolution-augmented Transformer for Speech Recognition", Interspeech 2020, October 25-29, China.