Vehicle Navigation Using Natural Spoken Language

Neethu V M[1], Shemitha P A[2],
[1]*M.tech Student, Department of Computer Science and Engineering, IES College of Engineering, Chittilappilly, Thrissur, Kerala*
[2]*Assistant Professor, Department of Computer Science and Engineering, IES College of Engineering, Chittilappilly, Thrissur, Kerala*

***Abstract*- Vehicle navigation using natural spoken language is a navigation dialogue system we can place this system in our vehicle and we can use it while we are driving. We can communicate it with our spoken language. This system is mainly based on two parts, first is ASR (Automatic Speech recognition) and NLP (Natural Language Processing). ASR is the process of converting speech to text. For this we have to consider a decoder to convert speech to text by using lexicon dictionary, language model and acoustic files. Next part is NLP, it takes the output text from ASR and analyzes this text to check whether the text is based on navigation oriented or not. And after all this process using RNN/LSTM (Recurrent Neural Network/ Long Short Term Memory) find the semantic representation of text. Then process the data for direction with map and speech.**

***Index Terms*- Automatic Speech Recognition, Natural Language Processing, Deep Neural Network, Recurrent Neural Network.**

I. INTRODUCTION

Driving is a difficult task because of the need to make correct decisions rapidly. Each decision a driver makes is important since it directly impacts traffic safety. Maneuvers involving changes in speed and steering wheel angle are the most influential factors concerning safety. Any change in speed or steering wheel can make the driving situation unsafe. Here integrate artificial intelligence with conversational environment and applying this expertise into a vehicle environment. So the auto makers and suppliers can create a highly customizes solutions that will offer personalized experience for both drivers and passengers while keeping hands on the wheel and eyes on the road. Driving is a complex decision-making task. Drivers must understand the relations that exist between their vehicle and the environment. Based on this understanding, drivers make appropriate decisions and perform reliable changes in vehicle movement, such as stopping, turning right, changing lanes, etc.

The artificially intelligent automotive assistant does more than just listen. Navigation dialogue system platform uses AI and natural language understanding to go way beyond interpreting simple commands. It understands human speech and meaning to analyze what's being said and deliver a response. Its entire stack is powered by innovative machine learning and contextual reasoning to create an AI platform optimized for the connected car. Voice is the most efficient, natural and safest way to connect the driver and passengers to the intelligent car. AI elevates the car above mere transportation and transforms it into an intelligent automotive assistant giving the driver instant access to information, services and content.

We know that, speech contains lots of information's that can be extracted with different levels of difficulties. The most basic from is the audio signal exiting a human's mouth. It is nothing else than vibrations of air pushed out of the lungs and formed by the vocal cord. All speakers have a specific frequency of vocal cord which the listener defects as tone and pitch of the voice. This is how we distinguish information can be extracted to be rewritten as a sequence of phonemes, which are then joined into words and sentences. Finally we can apply NLP on these transcripts to extract the meaning of what was said. Speech is analog, our computers are digital, sampled and quantized, and so they are discrete in time and value. Both sampling and quantization introduce errors in the signal

In the case of NLP (Natural Language Processing) stage, frist step is POS tagging, there is open source NLP libraries to done POS tagging called Stanford CoreNLP. NLP is used to analyze text, allowing machines to understand how human speak. This

human-computer interaction enables real word applications like automatic text summarization, sentiment analysis etc. After POS tagging next process is context extraction, i.e. Identify POI and literals. Literals mean antonyms and synonyms to check actual meaning other than POI. RNN/LSTM algorithms are neural network algorithm, which is used to check the words are exactly same and it check with previous mostly used words and correct word matching.

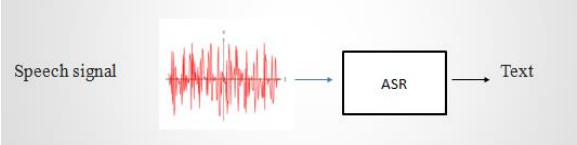

Figure 1. Automatic Speech Reognition

II. LITERATURE SURVEY

In 2013, Bando et al. (2013) proposed a new framework that can automatically translate driving data into sequences of "drive topics" in natural language. In this framework, brake pressure, opening rate, steering wheel angle, and velocity are considered as physical features. Each of these had different frequencies for each word. They used Latent Dirichlet Allocation (LDA). This is used to cluster extracted driving situation as data based on the existing frequency of physical behavioral features that are observed in each driving sequence. The distribution of the physical behavioral features included in each drive topic was used for automatic driving word labeling. The result of this step is a small number of drive topics, such as "accelerating" and "high speed", assigned to each driving words. Although DAA and LDA are completely unsupervised methods, this framework creates human-understandable tags. Being independent of any human-created tags is one of the greatest benefits of this method.

The term "double articulation structure" was first presented in order to analyze a speech stream. A speech stream data possesses a dual layer of information that can be decomposed into several meaningful linguistic units, and each unit can be divided into meaningless elements. Meaningless elements called phonemes are at the lowest level of speech organization. Morphology, syntax, and semantics give the meaning to phonemes and they are the higher levels of speech organization. In order to understand long-term human action, it has to be decomposed into short-term chunks. To extract long-term human action chunks, Taniguchi, T., and Nagasaka, S. (2011) presented a DDA framework. Which include a language model called Nested Pitman- Yor (NPYLM), and a stochastic model called sticky Hierarchical Dirichlet Process Hidden Markov Model (sHDP-HMM). Sticky HDP-HMM is an augmented version of HDPHMM (Fox, E. B. et al. 2007 in which the number of states is not predefined. Figure 2 shows the graphical representation of sticky HDP-HMM, which is an improvement over normal HDP-HMM.Parameter k in Figure [2] is the transition weight which is responsible for controlling rapid switching. If we set k = 0, sticky HDPHMM algorithm behaves same as normal HDP-HMM. DAA assumes that human action is a continuous series of smallest "meaningless" data time-series, and the smallest "meaningful" units as sequences of the meaningless elements. This structure can find and connect several short-term segments of human action. If we look at our spoken language, it also has a double articulation structure. For example, a sentence can be decomposed into single letters. Then single letters can be chunked into words. Letters individually do not have any meaning, however words do. Human action time-series have the same pattern. It is obvious that at each time, point data do not have any meaning individually, but when some of them come together, they form a meaningful segment. By expanding this idea, several successive segments of time series data form a meaningful sequence of a specific human action. A sequence of a specific meaningful human action is assumed to be a word.

III. SYSTEM ARCHITECTURE

Fig 2 is the system architecture. It mainly consists of two parts, ASR and NLP. In ASR, speech to text process is done by using feature extraction, speech recognition decoder. In NLP part, the ASR output text is first send to POS tagging and then do context extraction. After this processing RNN/LSTM neural network algorithms are doing to get the actual semantic representation of the input text. There is another two parts other than ASR and NLP, i.e. Accident distress call and Real time traffic prediction using Twitter. Accident distress call is an application

on the dialogue system, which works when we are in an accident we can press the emergency button on the system then an automatic call and SMS will send to the predefined numbers that we have stored early.

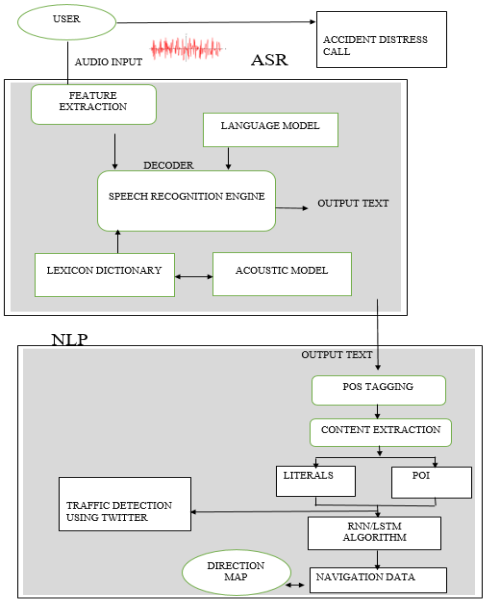

Figure 2 . System Architecutre

Real time traffic prediction using twitter is another application that we can predict early whether there is traffic whiles us on the go. It is done by analyzing the tweets from twitter. And the location of the vehicle is located by using GPS. Vehicle Navigation System using Natural Spoken Language is a voice based human interface to understand driver's spoken language in a natural way. Navigation dialogue system is placed in the vehicle so the navigation dialogues are typically employed while a user may be driving, on-the-go, or in other environments. It also has the capability to search for simple locations on Map, acting as an assistant, talk with the driver and totally act as a guide for driving. To find a destination, driver may either speak out a point-of-interest (POI), business name, specify the exact address, and number of a street. Then the system will respond back with the actual result using direction map. Navigation dialogue system is a voice based GPS methodology.

Voice is accessed as digital waves and ASR converts these waves into text. Actual page numbers and other running heads will be modified when the publications are assembled. The task of an Automatic speech recognition (ASR) system is to map spoken utterances (a speech signal) into sequences of words. For this purpose the incoming speech signal is first segmented into regions containing speech and silence. Development of speech recognition system comprises of two phases: training and recognition. Simply speaking, in the training phase, one or more acoustic patterns (or templates) of linguistics units such as the words (or phonemes) in the recognition vocabulary are derived. In the recognition process, the task is to match an incoming speech with the stored acoustic patterns.

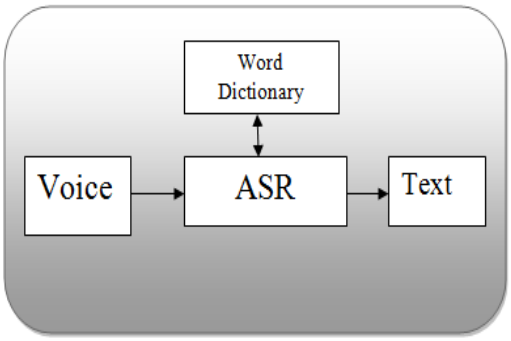

Figure 3. Speech to Text Conversion

An acoustic model is created by taking a large database of speech corpus and using special trained algorithm called SVM. SVM is a most powerful method for pattern classification. SVM maps the inputs into a high dimensional space and then distinguishes the classes with a hyper plane. The application of SVM can be speaker and language recognition. It is a two class classifier or also called binary classifier.

NLP is the second part after the completion of ASR part. In this we have to retrieve the meaningful data from the input text. It checks whether the extracted data is navigation oriented or not. For this it uses deep neural network algorithm for the semantic representation of whole sentence.

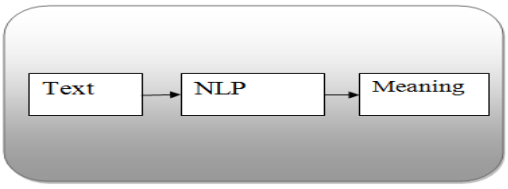

Figure 4. Natural Language Processing

The dialogue system also contains user profile and emergency numbers. GPS is used to locate the vehicle. We can set custom messages to send SMS and call. When we are in accident we can press the emergency button on this application, then calls and

SMS will automatically transfer to predefined numbers. Precise location of accident is send to all predefined contacts. Traffic is a major issue in many cities. Social media is an active site which has many followers, using the traffic related tweets tries to control traffic. So here is an android application to show and suggest graphical route format of traffic area using text mining and NLP. Classify traffic related tweets coming from which area in latitude and longitude format and find the route from GUI map and select route.

IV. RNN/LSTM ALGORITHM

Recurrent Neural Network/Long Short Term Memory (RNN/LSTM) is a Deep Neural Network (DNN) method. We will use RNN/LSTM; it has a memory that influences future predictions. RNN/LSTM model sequentially takes each word in a sentence, extract its information and embeds it into a semantic vector. It has the ability to capture long term memory. RNN/LSTM accumulates increasingly richer information as it goes through the sentence, when it reaches the last word; the hidden layer of the network provides a semantic representation of the whole sentence. The main idea of using RNN/LSTM algorithm is for sentence embedding and is to find low dimensional semantic representation by sequentially and recurrently processing each word in a sentence and mapping it into low dimensional vector. LSTM-RNN model sequentially take each word in a sentence, and extract its information and embeds into a semantic vector.

VI. EXPERIMENT RESULTS

Speech recognition

Vehicle speech corpus is used for both speech recognition and natural language processing. Word Error Rate (WER) is measured to calculate the speech recognition result.

$$WER = (S+D+I) / N$$

Where N is total number of words used and S, D, I are substitutions, deletions, and insertions. If the SNR of audio is 5~60 db then 25% of WER results will occurs. Noise is the main reason for increasing WER. It may be due to the car window was open or any other reasons. Most of the errors are caused by deletions and substitutions. The reason for substitution error is noise. Major cause of deletion error is some drivers may have a higher speaking rate, thus increasing the difficulty for ASR system. Another reason for noise error is when the drivers are under stress; their speech may be less fluent and difficult to understand. So their speech may contain lots of additional words or some words may be missing. To improve the ASR performance, reduce the noise impact maximum.

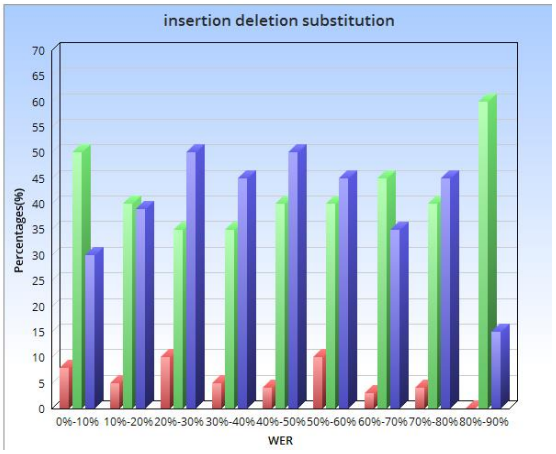

Figure 5. Percentages of insertions, deletions, and substitutions in errors

Language Understanding

Language understanding part is based on the NLP stage. Natural language processing is used to measure whether the data from ASR is related to navigation or not. For this we use the WER results. For this we use 80%~90% of samples are used for training and remaining 20%~10% are used for testing. This will ensure that NLP does not contain ASR errors. Sentiment analysis is done by using this result for this we use an algorithm called Deep Neural Network (DNN). Almost all errors like noise, distant speech problems are solved by using this algorithm.

VII. CONCLUSION

Vehicle navigation system using natural spoken language is a voice based human interface. Here mainly focused on natural language processing with RNN/LSTM architecture. Automatic Speech recognition, sentiment analysis and context level extractions are the main process in this system. For the high performance we need to reduce the WER and use good performance NLP. To ensure the safe driving it is necessary to provide efficient and user

friendly interaction between driver and dialogue system. WER is used to check performance of speech recognition. RNN/LSTM architecture is a new method to increase the performance of NLP. It will overcome most of the distant speech problems and noise.

REFERENCES

[1] Yang Zheng, Yongkang Liu, and John H.L, Hansen, Fllow,"Navigation Orientated Natural Spoken Language Understanding for Intelligent Vehicle Dialogue", IEEE Intelligent Vehicles Symposium (IV) June 11-14, 2017, Redondo Beach, CA, USA

[2] Bando, T., Takenaka, K., Nagasaka, S., & TaniguchiT."Unsupervised drive topic finding from driving behavioral data". In Intelligent Vehicles Symposium (IV), 2013 IEEE (pp. 177-182). IEEE june 2013.

[3] Tadahiro Taniguchi, Shogo Nagasaka, "Double Articulation Analyzer for Unsegmented Human Motion using Pitman-Yor Language Model and Infinite Hidden Markov Model" , SI International IEEE 2011.

[4] adahiro Taniguchi, Shogo Nagasaka, Kentarou Hitomi, Naiwala P. Chandrasiri, "Sequence Prediction Of Driving Behavior Using Double Articulation Analyzer" IEEE Transactions On Systems, Man, And Cybernetics: Systems, VOL. 46, No. 9, September 2016

[5] Bando, T., Takenaka, K., Nagasaka, S., & Taniguchi, T. "Automatic drive annotation via multimodal latent topic model". In Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference (pp. 2744-2749). IEEE 2013.

[6] Beauchemin, S. S., Bauer, M., Kowsari, T., & Cho, J. "Portable and scalable vision-based vehicular instrumentation for the analysis of driver intentionality". Instrumentation and Measurement, IEEE Transactions on, 61(2), 391-401, 2012.

[7] www.nuance.com/mobile/automotive/dragon-drive.html, "Bringing intelligence to the driver, to the road and the beyond" 2016.

[8] Alexandre Alahi, Kratarth Goel , Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, Silvio Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces", IEEE Conference on Computer Vision and Pattern Recognition 2016

[9] Oluwatobi Olabiyi, Eric Martinson, Vijay Chintalapudi, Rui Guo "Driver Action Prediction Using Deep (Bidirectional) Recurrent Neural Network", Intelligent Computing Division Toyota InfoTechnology Center USA, jan 2017

[10] Ryunosuke Hamada, Takatomi Kubo, Kazushi Ikeda, Zujie Zhang, Tomohiro Shibata, Takashi Bando ,Masumi Egawa, "Towards Prediction Of Driving Behavior Via Basic Pattern Discovery With Bp-Ar-Hmm" Nara Institute of Science and Technology, Japan IEEE 2013

[11] Mark Gales,.M., & Young, S, The Application of Hidden Markov Models in Speech Recognition, Ebook 2008.

[12] Yuxi Li, Deep Reinforcement Learning: An Overview, EBook 2017

[13] Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S." A sticky HDP-HMM with application to speaker diarization". The Annals of Applied Statistics, 1020-1056 (2011).