ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 5(J), pp. 27195-27198, May, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Review Article

ASR- A REVIEW

Ameya Parkar and Pooja Kamble

Department of MCA, VESIT

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0905.2208

ARTICLE INFO

Article History:

Received 12th February, 2018
Received in revised form 9th March, 2018
Accepted 26th April, 2018
Published online 28th May, 2018

Key Words:

Deep learning, HMM, LSTM, ASR, acoustic modelling.

ABSTRACT

Deep learning is a branch of machine learning, which have made a big impact in different fields like Automatic Speech Recognition, Image processing, Bioinformatics, Face Detection, etc. In this paper, we will study how deep learning is important and how it is implemented to develop speech recognition applications. The paper also illustrated models that are used traditionally to recognize the human voice. We will also have a look at the efficiency of these recognition applications.

INTRODUCTION

Artificial neural networks are developed to build the recognition, decision-making ability in a machine. A human has the ability to learn because of a large number of neural networks in the brain. The artificial neural network contains the input layer, hidden layer, and output layer. A hidden layer is where all the processing is performed and a final output is delivered.

The neural networks require some learning algorithms. The older learning algorithms failed to achieve better performance. Deep learning is a field where efficient algorithms are built with artificial neural networks to train them with more data. Deep learning provides a huge amount of data, which helps neural networks to learn more from different learning algorithms. The invention of deep learning increased the performance of the neural networks and the growth of artificial intelligence area.

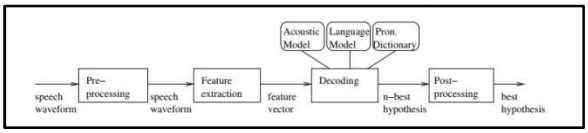

Figure 1 Structure of SRS

Automatic speech recognition (ASR) develops functionality which takes human voice as input and converts it into text format. This output in text format is used for the further task completion by the computer/machine. The processing in between contains different modeling approaches to deliver the accurate result. The speech recognition is widely used in different fields to automate or to ease the human work.

Speech Recognition

Speech recognition is the field that develops methodologies that enable the recognition and converting to text format by computers [1].

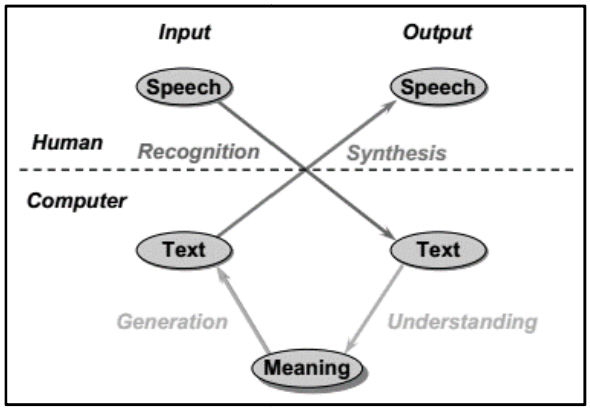

Figure 2 Communication and Spoken Language

Speech recognition applications allow people to give the command to search, make an appointment, play a song, make a

*Corresponding author: **Ameya Parkar**
Department of MCA, VESIT

phone call etc tasks by only using the human voice. It will process the voice and perform prediction analysis using deep learning algorithms.

Earlier speech recognition system used Hidden Markov Models (HMM) with feed forward artificial neural networks. Today, the speech recognition system widely use LSTM (Long-term-short-memory) which is a deep learning method.

Speech recognition is based on two important models- Acoustic and Language model. The Hidden Markov Models were used commonly in different systems. The Figure(1) below shows how the input voice is processed and how models are used to generate more accurate translation or recognized words from the computer.

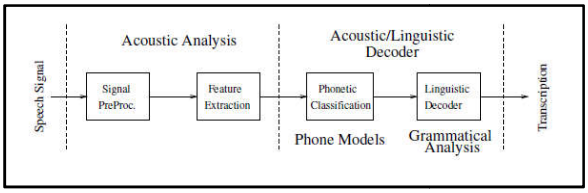

Figure 3 Structure of ASR

Hidden Markov Model theory is based on doubly embedded stochastic processes.

Each model can visit different states x and accordingly generating an output o in subsequenttime steps.

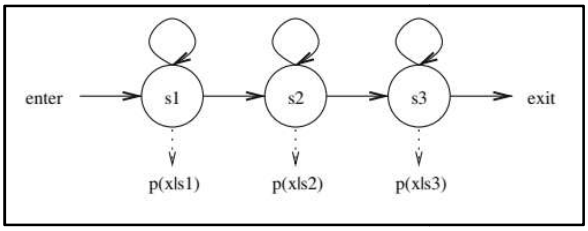

Figure 4 A three-state hidden Markov model

When implementing HMM, we face three problems - The training problem, single word recognition problem and continuous speech recognition problem.

ASR[2] is the first stage in an overall human/computer interaction pipeline that also includes Voicebox's related Natural Language Understanding (NLU) and Text-to-Speech (TTS)technologies. Voice box's advanced ASR module is a multi-stage pipeline that uses techniques from machine learning, graph theory, traditional grammar development, and statistical analysis of large sets to form high-confidence transcriptions of audio input.

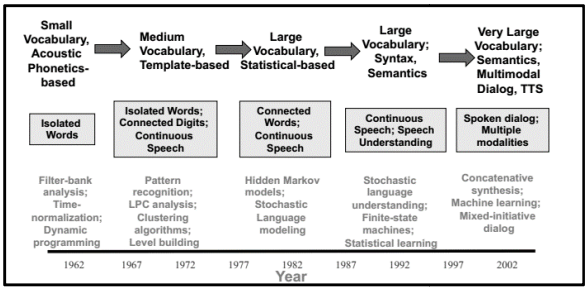

Figure 5 ASR Milestone

Acoustic Modeling

Acoustic modeling is used to distinguish the audio signal and the phonemes, one of the unit of sound. It uses audio recordings and the same in written format to generate the statistical representation of sound for each word. This model learns the given set of recordings and accordingly matches when input voice is given to it.

The model represents the sound signal characteristics. It is responsible for extracting information from the speech signal for further processing.

The process is divided in four blocks: Signal preprocessing, feature extraction, phonetic classification and linguistic decoder. The first two blocks in acoustic analysis phase are responsible for producing acoustic observations. These observations are vector coefficients which are used to generate speech signal characteristics.

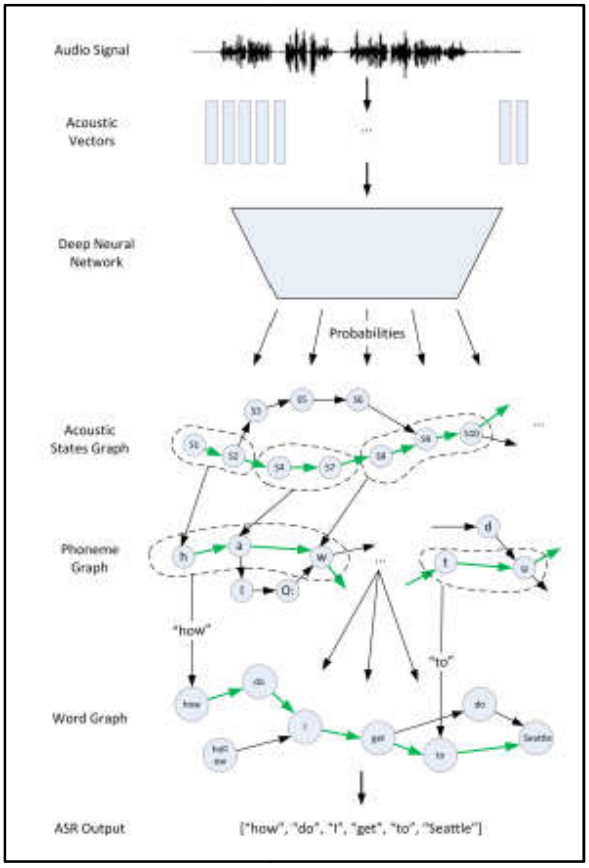

Figure 6 ASR Processing Pipeline [2]

The Acoustic/Linguistic decoder is considered as the core part of ASR. It is responsible for classifying strings of observation units frames into classes of language units. There are many algorithms to perform such task, but there are two classes of methods which are basically used in all speech recognition system. The two basic approaches used are deterministic approaches also called dynamic time warping algorithms. The second class used is stochastic approaches based on statistical methods. They are essentially Hidden Markov models (HMMs), which are employed in the system, and artificial neural networks.

Language Modeling

Language modeling uses different applications like speech recognition, machine translation, part-of-speech tagging, parsing, handwriting recognition, information retrieval and other applications. A language model can provide probability of various words and phrases to articulate and generate accurate prediction necessary about voice. The variation in pronunciation and voice of words and phrases make different words spell different sound similar but have completely different meaning or even know any specific meaning. There the prediction analysis may fails to output the correct word spoken.

In American English, some words may sound differently. This may produce wrong results and to remove this ambiguity, language model is used with pronunciation model and acoustic model. The role of language model is to check most likelihood phrase or word for the given voice input signal.

Performance

The performance of these algorithms depend upon different factors such as amount of learning data available, speed, accuracy, read and spontaneous speech , level of confusion in pronunciation of words etc. The accuracy depends upon the error rate in algorithm. Also the error rate may increase when there are more vocabulary size and when there is more confusion about the pronunciation of word/phrase.

The efficiency of the algorithm depend upon the amount of sample data made available for learning different signals and creating its statistical data using models.

Far-Field Techniques

The term "far-field" refers to situations in which the ASR's microphone is relatively far from the speaker. Far-field situations are common in home appliance scenarios where the appliance is situated somewhere in the same room as the user, but not necessarily close to them. The main challenges in far-field ASR are reverberation and echoes. Reverberation is when the microphone records slightly time-delayed copies of the user's speech signal, as that signal bounces off of walls and other surfaces in the surrounding environment. Echos are similar, but refer to removing a known audio signal—for example, whatever song the system is currently playing—from the microphone input so as not to contaminate the user's speech signal.

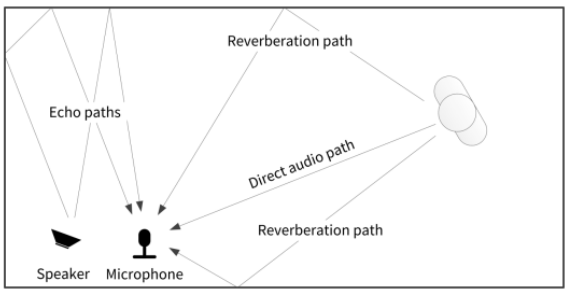

Figure 7 Path of Far Field

Problems in ASR

1. Robustness – graceful degradation, not catastrophic failure.
2. Portability – independence of computing platform.
3. Adaptability – to changing conditions such as different mic, background noise, new speaker, new task domain, new language.
4. Language Modelling – is there a role for linguistics in improving the language models?
5. Confidence Measures – better methods to evaluate the absolute correctness of hypotheses.
6. Out-of-Vocabulary (OOV) Words – Systems must have some method of detecting OOV words, and dealing with them in a sensible way.
7. Spontaneous Speech – disfluencies such as filled pauses, false starts, hesitations, ungrammatical constructions remain a problem.
8. Prosody –Stress, intonation, and rhythm convey important information for word recognition and the user's intentions (e.g., sarcasm, anger).
9. Accent, dialect and mixed language – non-native speech is a huge problem, especially where code-switching is commonplace.

Applications of Speech Learning

Automated Identification

It is used to authenticate a person by checking his/her voice as credentials. It is helpful to maintain confidentiality. As it is difficult to create same voice signals same as someone, which hence improves the security issues in the business world.

Buying Products and Services with the Sound of Your Voice

Since it is all made easy for people, they can give command to device and it will perform that specific task on behalf of him/her. It is automatic and human don't need to even use UI of his phone but only need to give commands and wait for the results.

Healthcare

It automates the process of admitting the patient and saving his/her data quickly in no time.

Cars and Other Vehicles

Speech devices take driver's voice as input and display routes, maps, etc.

Help for disabled

These applications will be most useful for physically disabled people. They can use their voice to get the work done.

CONCLUSION

In this paper author studied about speech recognition and how the models play their role to get the accurate output in less time. It also states that the performance is totally depend upon the number of input signals given for learning. Deep learning algorithms is important as it allows more data to process with great response time and performance as compared to traditional algorithms. The goal of ASR should be to accurately and efficiently convert a speech signal into a text message independent of the device, speaker or the environment. The applications area of focus is automation of complex operator based tasks, such as customer care, dictation, form filling applications, provisioning of new services, customer help lines, e-commerce.

References

1. R.E. Gruhn *et al*., Statistical Pronunciation Modeling for Non-Native Speech Processing, Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_2, Springer-Verlag Berlin Heidelberg.
2. https://www.voicebox.com/wp-content/uploads/2017/05/Automatic-Speech-Recognition-Overview-and-Core-Technology.pdf
3. https://ieeexplore.ieee.org/document/4767902/
4. https://www.inf.ed.ac.uk/teaching/courses/asr/
5. G&Y: MJF Gales and SJ Young (2007). The Application of Hidden Markov Models in Speech Recognition, Foundations and Trends in Signal Processing, 1 (3), 195-304.
6. S Young (1996). A review of large-vocabulary continuous-speech recognition, IEEE Signal Processing Magazine 13 (5), 45-57.
7. R&H:SRenals and T Hain (2010). Speech Recognition, in Computational Linguistics and Natural Language Processing Handbook, A Clark, C Fox and S Lappin (eds.), Blackwells, chapter 12, 299-332.
8. G Hinton *et al* (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine, 29(6):82-97.
9. S Young (2008). HMMs and Related Speech Recognition Technologies, in Springer Handbook of Speech Processing, J Benesty, MM Sondhi and Y Huang (eds), chapter 27, 539-557.
10. https://www.ncbi.nlm.nih.gov/pubmed/19173117
11. https://www.sciencedirect.com/science/article/pii/S0167639307000404
12. http://www.esna.com/services/education/RH5_OL_Documents/Feature_Description_Guide/RH5_OL_FG/OL_ASR/Automatic_Speech_Recognition_ASR.htm
13. https://arxiv.org/abs/1803.02551
14. https://www.ibm.com/blogs/watson/2017/06/automatic-speech-recognition-are-all-tests-comparable/
15. https://www.microsoft.com/en-us/research/publication/lstm-time-and-frequency-recurrence-for-automatic-speech-recognition/

How to cite this article:

Ameya Parkar and Pooja Kamble.2018, ASR- A Review. *Int J Recent Sci Res.* 9(5), pp. 27195-27198.
DOI: http://dx.doi.org/10.24327/ijrsr.2018.0905.2208

****