# Big Data and Speech Recognition

Yuchen Liu
I523-HID:213
1750 N Range Rd, Apt D302
Bloomington, IN 47408
liu477@iu.edu

## ABSTRACT

Nowadays, Speech Recognition is becoming more and more important. Many technology companies are trying to use Big Data to develop more efficient and accurate algorithm for Speech Recognition. Nowadays, Deep learning can be described as the foundation of Speech Recognition. Deep learning algorithms such as RNN and CNN often need to supported by large amount of data – Big data. Before Big Data and deep learning, the word error rate was 24 percent. Recently, IBM published a paper where the word error rate was below 5.5 percent. In August, Microsoft speech recognition system has reached a 5.1 percent error rate.

## KEYWORDS

i523,HID213,Big Data, Speech Recognition

## 1 INTRODUCTION

Speech recognition, also known as automatic speech recognition, is a sub-field of computational linguistics. It can help computer to translate speech that we are spoken to text. [13]

Speech recognition is a branch of pattern recognition, and it belongs to the field of signal processing science. It also has a very close relationship with phonetics, linguistics, mathematical statistics and neuro-biology. The purpose of speech recognition is to let the machine "understand" the language of human use. The word "understand" here including two meanings: one is verbatim to understand non-translated text into written language; the other is if the speech contained the request or ask, the computer will make the right response.

## 2 PROBLEMS OF SPEECH RECOGNITION

The definition of Speech Recognition has been raised for many years. Before 1980s, the most serious problem of Speech Recognition is the limited choice of algorithms. In 1952, the United States ATT Bell Labs developed the first electronic computer-based voice recognition system Audrey[12], which can identify 10 English figures, the accuracy rate of 98 percent .In the 1960s, the two major areas of speech recognition is linear Predictive coding and dynamic time specification techniques.

In the late 1960s, the Hidden Markov Model was proposed by Leonard E. Baum. HMM is a major breakthrough in the history of speech recognition. The error rate of speech recognition is greatly reduced. [11] HMM can be used in many areas. For example, in our daily life, we always want to predict the weather according to the current weather situation. One way is to assume that each state of the model depends only on the previous state. This assumption is called the Markov hypothesis, and this assumption can greatly simplify the problem. Obviously, this assumption is also a very bad assumption, which resulting in a lot of important information are lost. When it comes to the weather, the Markov hypothesis is described as assuming that if we know the weather information for some days before, then we can predict the weather today. Of course, this example is somewhat unrealistic. However, such a simplified system can be beneficial to our analysis, so we usually accept such assumptions, because we know that such a system allows us to get some useful information, although not very accurate.

In 1980s, artificial neural networks have been introduced into speech recognition.[8] Neural Network have many advantages. First, the neural network is non-linear. The neural networks that are interconnected by non-linear neurons are non-linear. The neural networks that are interconnected by non-linear neurons are non-linear in nature. Second, neural network is contextualinformational. The specific structure of the neural network and the state of excitation represent knowledge. Each neuron in the network is potentially affected by the global activity of all other neurons in the network. Therefore, the neural network will naturally be able to handle contextual information. However, because of the lack of high quality speech data and the lack of computational power, Neural Network still have a high error rate on Speech Recognition.

## 3 WHY BIG DATA IS IMPORTANT

Andrew Ng has predicted that as speech recognition goes from 95 percent accurate to 99 percent accurate, it will become a primary way that we interact with computers.[4]In recent years, the idea of Big Data has been brought out. As the amount of data and computational power both increase, neural network is widely used in speech recognition tasks. Big Data becomes the answer of the problem of Speech Recognition.

There are more than 7000 different languages in this world and different people who speak the same language have different accent. Therefore, a large amount of data is required in order to make the Speech Recognition result accurate. A recent Google research paper shows that "Large language models have been proven quite beneficial for a variety of automatic speech recognition tasks".[3] In the paper, the researchers found that data sets and larger language models will bring fewer errors predicting the next word based on the words that precede it. they also found that increasing the model size by two orders of magnitude will reduces the world error rate by 10 percent relative." The world error is 24 percent for Speech Recognition.

In March 2017, IBM announced that they are reaching a new record of Speech Recognition of 5.5 percent error rate. In order to get the goal, IBM combined LSTM (Long Short Term Memory) and WaveNet language models with three strong acoustic models.[9] The first acoustic models is a six-layer bi-direction LSTM model with multiple feature inputs. The second acoustic models is trained

with speaker-adversarial multi-task learning. For the third model, it not only learns from positive examples but also learn more the negative examples. So it gets smarter and smarter. It also performs better when similar speech patterns are appeared.

In August 2017, Microsoft then announced that they improve the Speech Recognition accuracy to 5.1 percent error rate. Basically, they improved the recognizer's language model by using the entire history of a dialog session to predict what is likely to come next, effectively allowing the model to adapt to the topic and local context of a conversation.[10] The data amount that they are using is huge and the improvement of the amount of data that they use result in a better result.

## 4 CURRENT ALGORITHM AND BIG DATA

Xuedong Huang, who leads Microsoft's Speech and Language Group, said that "People are speak with oxygen. Big data is just like the oxygen to speech recognition, there must be large data in order to make the algorithm accurate." He also said that for Speech Recognition, there are two things that are most important. One is data and the other is algorithm.[6]

A variety of neural network learning methods are in fact similar, basically through the gradient descent method to find the best parameters of the model. Then find the optimal model by deep learning. Nowadays, there are two different kind of neural network that is widely used in Speech Recognition field. One is RNN(Recurrent Nerural Network) and the other is CNN(Convolutional Neural Network).

The most important idea of RNNs is to make use of sequential information. In a traditional neural network we always assume that all inputs and outputs are independent of each other. But for many situations it is not true. For example, If there is a sentence said "Tom broke the glass, Mr.Peter critized ( )." In this situation, we should know that we need to put the name "Tom" in the blank. If you want to predict the next word in a sentence you'd better know which words came before it. RNNs are called recurrent because we can bring the information form the previous sentence to the next sentence, which the output being depended on the previous computations. In order to train a muti-layer RNN model, a large amount of data is required. Only if we let the model seen different combinations of text and different ways to talk, the accuracy of the Speech Recognition can be acceptable. Both Microsoft and IBM use a specific RNN architecture in their research, The model is called LSTM (Long short-term memory). The concept LSTM was invented by Hochreiter & Schmidhuber in 1997. It was invented to solve the Long-Term Dependencies of RNN.[2]

CNN (Convolutional Neural Network) is very similar to DNN. They both build up by neurons and have a cost function. However, the input that CNN take is different from the Ordinary Neural Network. CNN architectures make an assumption that the inputs are images, which allows us to encode certain properties into the architecture. When a computer sees an image, it will see an array of pixel values. Depending on the resolution and size of the image, it will see a 32 x 32 x 3 array of numbers.For Speech Recognition, we can put a plot of the audio data as the input of the CNN model. The experimental results from Microsoft show that CNNs reduce the error rate by 6 percent-10 percent compared with DNNs on the TIMIT phone recognition and the voice search large vocabulary speech recognition tasks.[1]

## 5 CURRENT PRODUCT AND BIG DATA

Speech recognition is becoming more and more important in our daily life. It is almost build into all our electronic devices. For example: phones, smart watches, computers and game consoles. It is even automating our home. People are becoming more and more familiar with talking to electronic devices. Many IT Giant is working on different algorithms to make the recognition result better by using Big Data. They also producing many interesting product: Apple Siri, Amazon Echo, Microsoft Cotana etc.

### 5.1 Apple and Siri

Although Apple is the most profitable technology companies, in the field of Big Data, it still catching up. However, they have increased the pace of entering Big Data. Recently, Big Data expert Bernard Marr analyzed how Apple used Big Data.[7]

In the mobile market, Apple is a powerful presence. They have been actively encouraging developers to develop applications based on user data monitoring and sharing. An obvious example is that they recently announced a partnership with IBM to promote healthy mobile application development. They also developed a number of air travel, banking and insurance applications with IBM. Apple Watch's launch accelerated the process. Many commentators believe that smart watches may be the ultimate device for wearing devices. With more sensors, it can collect more data for more extensively analysis.

The most typical Big Data Application for Apple is Siri. In terms of speech recognition, the most common personal voice assistant in the United States has a really good performance. At 95 percent accuracy, Siri surpassed all of its Silicon Valley Silicon Valley giants. The most impressive part about siri is that it can perform really good in so many different languages. Even on some dialects. That is because of the use of Big Data. Only if the amount of data is large enough for the training process, apple can get this great result.

### 5.2 Google and Google Now

Google is the founder of the Big Data age. Its Big Data technology architecture has always been the study and research by other Internet companies. Google provides Big Data analysis intelligent applications such as customer emotional analysis, fraud analysis, product recommendation and Speech Recognition. Those Big Data applications has brought Google 23 million in revenue each day.

According to Mary Meeker's annual Internet Trends Report,[5] Google's speech recognition model Google Now has achieved a 95 percent word accuracy rate on May 2017 for English, which is really similar of what Apple have. This current rate is also known as the threshold for human accuracy. This accomplishment is based on the huge data that Google could get every data. As the world mostly used search engine, Google gains more than 20PB of data everyday. This huge amount of data helps the model performance better.

## 6 CONCLUSIONS

In conclusion, Big Data is an indispensable part of Speech Recognition. Without Big Data, the existing powerful algorithms such as CNN and RNN will not work. These algorithms has gained huge successes in a broad area of applications now. Such as speech recognition, face identification, and computer vision. In today's world, the size of data that we can use is huge. Big Data also means big opportunities. Also, the huge amount of data also bring huge challenges to harnessing data and information. As the volume of data keeps getting bigger, using the proper algorithm plays a key role to increase the accuracy of Speech Recognition and other data analytic solutions.

## REFERENCES


[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (October 2014), 1533–1545. https://www.microsoft.com/en-us/research/publication/convolutional-neural-networks-for-speech-recognition-2/

[2] Rowel Atienza. 2017. LSTM by Example using Tensorflow. (18 March 2017). Retrieved October 2, 2017 from https://medium.com/towards-data-science/lstm-by-example-using-tensorflow-feb0c1968537

[3] Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. 2012. *Large Scale Language Modeling in Automatic Speech Recognition.* Technical Report. Google.

[4] Adam Geitgey. 2016. Machine Learning is Fun Part 6: How to do Speech Recognition with Deep Learning. (24 December 2016). Retrieved October 2, 2017 from https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a

[5] APRIL GLASER. 2017. Googlefis ability to understand language is nearly equivalent to humans. (31 May 2017). Retrieved October 2, 2017 from https://www.recode.net/2017/5/31/15720118/google-understand-language-speech-equivalent-humans-code-conference-mary-meeker

[6] Microsoft Asia Research Institute. 2017. Huang Xuandong: How does Microsoft use artificial intelligence to do voice recognition? (24 April 2017). Retrieved October 2, 2017 from http://www.msra.cn/zh-cn/news/features/xuedong-huang-talk-20170424

[7] Bernard Marr. 2015. How Apple Uses Big Data To Drive Success. (22 May 2015). Retrieved October 2, 2017 from http://www.datasciencecentral.com/profiles/blogs/how-apple-uses-big-data-to-drive-success

[8] Margi Murphy. 2015. Everything you need to know about deep learning and neural networks. (19 August 2015). Retrieved October 2, 2017 from https://www.techworld.com/data/why-does-google-need-deep-neural-network-deep-learning-3623340/

[9] George Saon. 2017. Reaching new records in speech recognition. (07 March 2017). Retrieved October 2, 2017 from https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/

[10] Catherine Shu. 2017. Microsoftfis speech recognition system hits a new accuracy milestone. (20 August 2017). Retrieved October 2, 2017 from https://techcrunch.com/2017/08/20/microsofts-speech-recognition-system-hits-a-new-accuracy-milestone/

[11] Asta Speaks. 2011. History and Theoretical Basics of Hidden Markov Models. (19 Apirl 2011). Retrieved October 2, 2017 from https://www.intechopen.com/books/authors/hidden-markov-models-theory-and-applications/history-and-theoretical-basics-of-hidden-markov-models

[12] Asta Speaks. 2014. Audrey: The First Speech Recognition System. (14 October 2014). Retrieved October 2, 2017 from https://astaspeaks.wordpress.com/2014/10/13/audrey-the-first-speech-recognition-system/

[13] Wikipedia. 2004. Speech Recognition. (March 2004). Retrieved October 2, 2017 from https://en.wikipedia.org/wiki/Speech_recognition


## 7 BIBTEX ISSUES

## 8 ISSUES

### 8.1 Writing Errors

Spelling errors

### 8.2 Structural Issues

Acknowledgement section missing

### 8.3 Details about the Figures and Tables

Capitalization errors in referring to captions, e.g. Figure 1, Table 2