Study of deep learning and CMU sphinx in automatic speech recognition

Conference Paper · September 2017

DOI: 10.1109/ICACCI.2017.8126189

1 author:

Abhishek Dhankar
University of Alberta
1 PUBLICATION **5** CITATIONS

Study of Deep Learning and CMU Sphinx in Automatic Speech Recognition

Abhishek Dhankar
B Tech (CSE) final year, MIT
Manipal University
Manipal, India
abhishekdhankar95@gmail.com

Abstract- Machine learning has proven to be a very effective tool in automatic speech recognition. This paper is an attempt to give a broad overview of the applications of various approaches of machine learning in speech recognition with special reference to deep learning and CMU Sphinx. Deep learning in Speech recognition is a relatively recent development. On the other hand, CMU Sphinx, an open source software has been in use for this purpose for a relatively longer time. CNN, a Deep Learning algorithm learns the invariant features that help it to differentiate between different words and word sequences. CMU Sphinx uses GMM-HMM model to predict the phonemes in the utterance to determine the word or set of continuous words that were spoken.

Keywords- Machine Learning, Deep Learning, CMU Sphinx, Automatic Speech Recognition, CNN, GMM-HMM

I. INTRODUCTION

Automatic Speech Recognition (ASR), which is essentially getting a computer understand spoken words, is a difficult task as pronunciation of each person is different. Also, same person speaks differently in different mental states. Background noise further obscures the speech signals. The variations in the use of devices for sound recording also add to the variability in the speech signals. Despite these obstacles, machine learning has come a long way in solving the above mentioned problems in speech recognition [1].

Machine Learning, in general, has reversed the process by which speech recognition is carried out. Earlier the dominant form of solving recognition problems was to identify the major invariant features of the object to be identified and to code the process of identification of those features so that the recognition is automated. However, with the advent of Machine Learning and specifically Deep Learning, it is no longer absolutely necessary to know the invariant features beforehand. Training data consists of objects to be identified and classified, can be fed to the software, which would then automatically learn the process of recognizing the corresponding object. Language models refer to a system's knowledge of what makes up a possible word, what words are likely to co-occur, and in what sequence[2][3]. Acoustic models include the representation of knowledge about acoustics, phonetics, microphone and environment variability, gender and dialect differences among speakers etc. There are two ways in which CMU Sphinx can be used for speech recognition. The two types of language models are grammar and n-gram model. In both of them, a language model, an acoustic model and a dictionary are required. The difference in the two approaches is due to the different types of language models that can be used. In grammar, the specific syntax of the language to be recognized is specified. The grammar specification approach can be used in command and control systems or in domains where limited vocabulary is required. On the other hand, the n-gram model can be used for a broader range of vocabulary. This model gives the probability of each word occurring in the vocabulary, given that a set of words have already occurred. The dictionary specifies the phonemes in each and every word in order. In this paper attempts have been made to study the existing speech recognition approaches; to identify the most efficient model in terms of accuracy and also to handle the huge volume of data. The deep learning approach has been compared with CMU Sphinx, which is an open source software based on old pattern recognition approach using the concept of Gaussian Mixture Model (GMM).

II. APPROACHES TO SPEECH RECOGNITION

There are three approaches to Automatic Speech Learning:

1. The Acoustic-Phonetic Approach
2. The Pattern Recognition Approach
3. The Deep Learning Approach

1. *The Acoustic-Phonetic Approach*

The following diagram depicts the general process of speech generation and recognition [4].

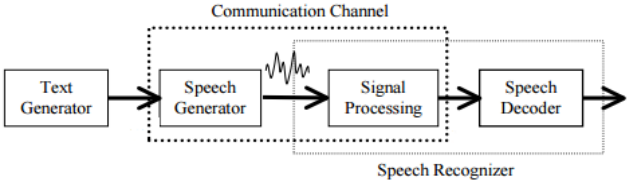

Fig.1. Speech generation and recognition process

The Text generator in the above diagram is our brain and the Speech generator is our vocal tract. The vocal tract is broadly divided into two parts – the larynx and rest of the vocal tract which consists of the oral cavity and the nasal cavity. The larynx consists of two folds on top of the larynx, which vibrate against each other when air passes through the vocal tract. Their vibration determines the pitch of the sound thus produced. This sound produced from the larynx is shaped by the position of articulators into phonemes, which are the basic constituents of speech. The numbers of Phonemes for any particular language are limited. The shape of the mouth is determined by the articulator. Articulator is anything that is involved in directing or obstructing the flow of air to form phonemes. These can be active or passive. An active articulator moves around freely as compared to a passive articulator. A passive articulator remains relatively stationary. For example, the teeth, palate, jaw, etc. Example of an active articulator includes the tongue [5][6][7].

A lot of frequencies are present in the sound produced by the vocal cord. The shape and size of the cavity from the larynx to the lips determines which frequencies will get intensified and which will get dampened. Usually those frequencies near the fundamental or resonant frequency will get intensified while the others will be dampened. The resonant frequencies are also called Formants. These can be observed as peaks in the frequency-magnitude graph. Formants can be reliably used to determine which vowel was spoken. There are a number of formants in each frame of the recording; however, the most important ones are the first three, i.e. F1, F2 and F3. Figure 2 gives an idea of how the formants might determine the vowels heard. The arrangement of the vowels resembles the vowel quadrilateral which is pictorial representation of the relation between the position of the tongue and the vowel spoken [8].

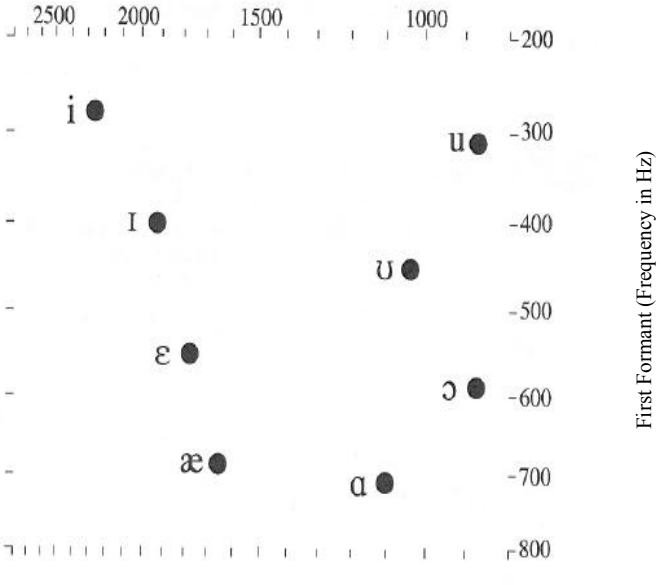

Fig. 2 Formant Chart

Figure 2 shows the frequencies of the first and second formants of eight American English vowels. The scales have been arranged at Bark scale intervals [9].

The inference can be drawn that closer the tongue to the teeth, the higher is the second formant and vice versa. Also, the height of the tongue is inversely proportional to the frequency of the first formant. Such features can help in Acoustic-Phonetic approach.

The consonants are formed by completely and partially obstructing the flow of air. This obstruction is created when the active and passive articulators come close to each other or touch each other [10]. It is by using such spectral measurements that speech recognition using the acoustic phonetic approach is carried out.

Consonants can be of two types; voiced and unvoiced. The difference is based on whether or not the vocal cords were involved in the production of the consonant. This can be helpful in narrowing down the number of possible consonants in the input speech.

Zero Crossing Rate helps in differentiating between voiced and unvoiced consonants. This rate is usually low for voiced consonants and high for unvoiced consonants. To calculate zero crossing rate an open source library in java programming language, TarsosDSP, was used by me for validating the above hypothesis. The program takes inputs from the microphone and thus, the utterances of individual phonemes were taken as input and the rate was calculated. The average zero crossing rate for voiced consonants was found to be lower than unvoiced consonants.

Table 1. Zero crossing rates of voiced and unvoiced consonants

Unvoiced Consonants	Crossing Rate	Voiced Consonants	Crossing Rate
P	0.054668795	b	0.035818007
f	0.041768443	v	0.036322363
Θ	0.028201684	ð	0.03235258
t	0.07752411	d	0.036137223
s	0.0365099	z	0.02969562
ʃ	0.05713588	ʒ	0.035639413
ʧ	0.041762896	ʤ	0.061286777
k	0.06470516	g	0.04236357
Average	0.050284609	Average	0.038701944

The first step in Acoustic Phonetic approach is preprocessing. In this, the features of each and every frame are calculated using Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). These two

measures provide the spectral descriptions of the speech over time.

The next step is feature detection stage, where the spectral measurements are identified for each phonetic unit. There can be features other than spectral measurements, such as, voiced or unvoiced classification, formant location and so on.

The third step is, in which the system finds the stable regions, .i.e. where the spectral or other features remain approximately constant. These stable regions help in segmentation. The segmented regions are then labeled according to the phonetic units whose features most closely match the features of the region. In the final step, the word or word sequence that the identified phonemes match closely are outputted [11].

2. *The Pattern Recognition Approach*

First, features of each and every test case are measured using filter bank analyzer or Discrete Fourier Transform (DFT) analyzer to produce features such as MFCC and LPC coefficients. Thereafter, pattern training is carried out, where multiple test patterns of speech sounds of the corresponding class are used to create a pattern that encapsulates all the invariant features of that class. This created pattern can then be used to identify whether a sample of sound belongs to the class or not.

The next step is pattern classification, in which the unknown test pattern is compared with each class reference, calculated in the previous step, and a measure of similarity between the test pattern and each reference pattern is computed. Mahalanobis distance can be used for the similarity measurements if MFCC and/or LPC coefficients are calculated. However, in that case, it becomes important to compensate for different rates of speaking by methods such as dynamic time warping algorithm.

Lastly, decision logic is used in which the reference pattern similarity scores are used to decide which reference pattern best matches the unknown test parameters.

The second method of pattern recognition is the one that CMU Sphinx applies.

CMU sphinx uses an n-gram language model, an acoustic model and a dictionary, which specifies the phonetic units in the words present in the dictionary. The language model can also be a grammar. Certain versions of CMU sphinx can be given grammar in GRXML format, while others can be given in JSGF format. GRXML is one of the formats for grammar specification prescribed by the World Wide Web Consortium (W3C) standard. The grammar can be used for limited vocabulary, but it increases in complexity as the number of possible sentences and words increase [12].

In the CMU Sphinx approach [13]:

given the acoustic data

$$A = a_1, a_2, a_3 \ldots . a_k \tag{1}$$

given the Word Sequence

$$W = w_1, w_2, w_3 \ldots . w_k \tag{2}$$

The goal is to maximize P(W/A).

According to Bayes' Theorem:

$$P(W \mid A) = \frac{P(A \mid W).P(W)}{P(A)} \tag{3}$$

Where:

P(A |W) = Acoustic model(HMMs)
P(W) = Language model.
P(A) = Constant for a complete sentence.

Conventional speech recognition systems, like sphinx, utilize Gaussian mixture model (GMM) based hidden Markov models (HMMs) to represent the sequential structure of speech signals [14].

3. *The Deep Learning Approach*

This approach has successfully replaced Gaussian mixtures for speech recognition. The Deep Neural Networks or DNNs work well for ASR when compared with GMM-HMM systems and they also outperformed the latter by a large margin in some of the tasks [15].

The major obstacle in using pattern recognition approach is that a detailed analysis of the object or speech has to be carried out to theoretically find the invariant features that would help in an accurate recognition. The problem is that a lot of time and effort need to be invested into mastering these theoretical concepts. Even after that a good recognition rate cannot be guaranteed to any certain degree. Deep Learning solves that problem to a great extent. It employs Convolutional Neural Network (CNN) approach which can automatically learn the invariant features to distinguish and classify the objects. For example, the Figure 3 represents the spectrograms of three different persons saying the words – one, two, three [16].

It can be clearly seen from Figure 3 that these three words can be distinguished. However, by using a CNN approach it is not necessary to theoretically identify the distinguishing features and then code the approach to identify them. CNN can directly learn these features and distinguish between the different words.

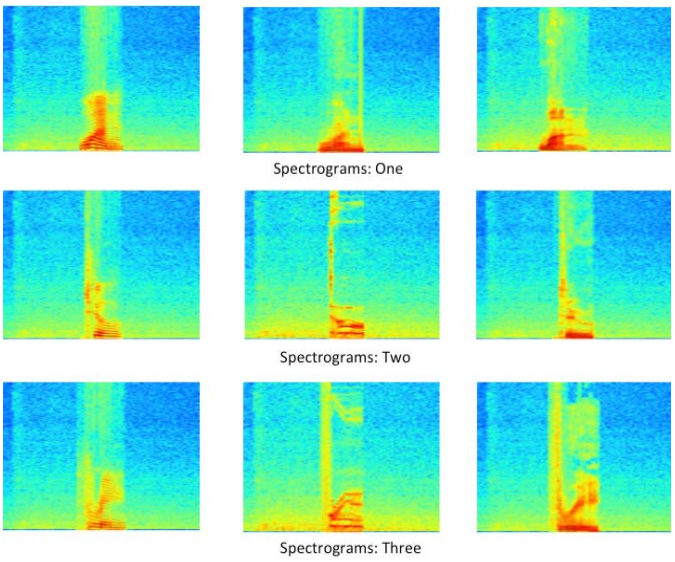

Fig.3. Spectrograms of three persons

The convolutional neural network is given a raw input signal that has been split into a sequence of frames. The CNN then outputs a score for each frame for each class. Such a convolutional layer may be present in a filter stage, followed by a max-pooling layer, the output of which is then fed to the sigmoid function. This filtered signal is then fed to the classification layer, which is usually a multi-layer perceptron [17].

Deep Learning can be implemented using various tools. However TensorFlow seems to be one of the best application methods currently available.

III. DEEP LEARNING VS. CMU SPHINX

The main difference between deep learning and CMU Sphinx implementation is in the amount of data required to train the software. The acoustic model of CMU Sphinx can be adapted to the accent and acoustic conditions of the speaker. This can be accomplished by a one hour recording of the speaker's voice in the acoustic environment that recognition is required. However, in deep learning approach huge amount of data is required. The collection of the data is cumbersome and difficult.

As can be seen from Figure 4, the performance or accuracy of deep learning keeps improving as more and more data is fed. On the other hand, with increase in data the performance of the traditional approaches reaches a plateau and the gap in performance of the two further increases. The performance of deep learning keeps better and better with more data [18].

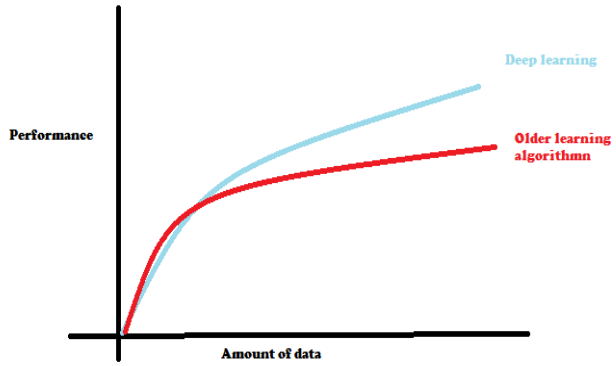

Fig.4. Deep Learning vs. Older learning algorithm [19]

In both deep learning and CMU Sphinx a lot of parameters need to be randomly varied in order to fine tune the code to a particular application. In deep learning, the weights and variances need to be changed to get optimal performance. In CMU sphinx there are parameters like the noise threshold among others that need to be changed according to specific application.

The accuracy of deep learning is directly proportional to the amount of training data available; however same is not necessarily true for CMU Sphinx. In Deep Learning, Convolutional Neural Networks are used. CNN has alternating convolutional and pooling layers. CNN involves convolution, which requires input data to be entered in the form of feature maps. These feature maps can be matrices. The best Deep Neural Network (DNN) model clearly outperforms the best Sphinx model which is GMM based by a large margin in voice recognition [20].

In order to test the two approaches, i.e. CMU Sphinx and Deep Learning, implementations of both were tested against limited number of inputs. These inputs included utterances of calendar dates for example, 5th January, 8th February etc.

IV. EXPERIMENTAL SETUP

CMU Sphinx-3 version was used for testing. The grammar or the possible sequence of words was specified in JSGF format. The grammar represented the language model. The rule <date> was composed of two parts, namely, the day and the month. The first pair of parentheses enclosed the day separated by |, logical or symbol. Similarly, the second pair of parentheses enclosed the months. CMU Sphinx already had a trained acoustic model, for US English which also works well for Indian accent for the limited vocabulary in the grammar model mentioned above. The dictionary provides the breakup of words into phonemes. It was modified to take into consideration different pronunciations of the same word. For example, the word 'first' is sometimes pronounced without the 'r' as 'fst'.

The location of the grammar in the local file system is passed as a parameter to the java code to be used for identifying the date. Thereafter, the program outputs the identified sequence of words i.e. day and the month.

V. RESULTS

CMU Sphinx was tested against limited vocabulary that was described by the following grammar:

public <date> = (First | Second | Third | Fourth | Fifth | Sixth | Seventh | Eighth | Ninth | Tenth | Eleventh | Twelfth | Thirteenth | Fourteenth) (January | February | March | April | May | June | July | August | September | October | November | December);

The results were as follows:

Table 2. Input Vs Output of CMU Sphinx

Input	Correct/Incorrect	Output
5th January	correct	5th January
8th February	correct	8th February
10th March	incorrect	10th August
11th April	correct	11th April
17th June	incorrect	Seventy two
8th July	correct	8th July
4th August	incorrect	12th August
7th September	correct	7th September
2nd October	incorrect	2nd August
3rd November	incorrect	August
6th December	correct	6th December

As per the above results, CMU Sphinx had an accuracy of 54.54%. On the other hand the Deep Learning Model correctly identified all these inputs with 100% accuracy. Hence the accuracy of CMU Sphinx was found to be considerably lower than that of Deep Learning model, which was able to correctly identify all the above utterances. The results showed that deep learning algorithm outperforms the CMU Sphinx by a huge margin in terms of accuracy. The test was conducted for limited vocabulary. For more expansive vocabulary, however, deep learning algorithm requires a large amount of data or a large corpus to show high learning rates in comparison to CMU Sphinx.

VI. CONCLUSION

In this paper different approaches of speech recognition have been described. The accuracy of CMU Sphinx was compared with Deep Leaning model in speech recognition. CMU Sphinx, which is based on Hidden Markov Models broadly uses pattern recognition approach. While both deep learning and pattern recognition approaches are well tested and proven, the former provides an edge over the latter in the form of lesser time and effort required in mastering the theoretical basics and gives a higher recognition rate as well. The Deep Learning based model outperforms CMU Sphinx in terms of lower or nil WER (Word Error Rate) in speech recognition. Deep learning approach is much more accurate and can handle huge data and simultaneously maintain accuracy in speech recognition.

REFERENCES

[1] Preeti Saini, Parneet Kaur, "Automatic speech recognition ; a review", International Journal of Engineering Trends and Technology- Volume4 Issue2- 2013, ISSN: 2231-5381

[2] Robert Mannell, "Phonetics and Phonology" , Department of Linguistics, Macquarie University, Sydney URL http://clas.mq.edu.au/speech/phonetics/phonetics/consonants/place.html

[3] Phonetics Beyond Basics, Last Revised Dec 19, 2002, URL http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

[4] Nitin Indurkhya (Editor), Fred J. Damerau(Editor), Handbook of Natural Language Processing, Second Edition, Chapman & Hall/CRC, Machine Learning and Pattern Recognition Aeries

[5] Indiana University, Bloomington, " How language works" URL http://www.indiana.edu/~hlw/PhonUnits/phonemes.html

[6] University of Birmingham, School of Computer Science, "Natural language processing & applications - Phones and Phonemes" URL https://www.cs.bham.ac.uk/~pxc/nlp/NLPA-Phon1.pdf

[7] Indiana University, Bloomington, Department of Physics "Speech analysis" URL http://www.physics.indiana.edu/~courses/p109/P109fa08/11.pdf

[8] Sami Lemetty, Master's thesis, "Review of Speech Synthesis Technology-Phonetics and theory of speech" Chapter 3, Helsinki University of Technology. URL http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap3.html

[9] University of Arizona, "Introduction to phonetics-formants, spectrograms and vowels" URL http://www.u.arizona.edu/~ohalad/Phonetics/notes/Formants%20Spectrograms%20and%20Vowels.PDF

[10] William F Katz, Phonetics for Dummies, "How consonants are formed : the manner of articulation", URL http://www.dummies.com/education/language-arts/grammar/how-consonants-are-formed-the-manner-of-articulation/

[11] Lawrence Rabiner, Hwang Juang "Fundamentals of Speech Recognition", Prentice Hall, 1993

[12] CMU Sphinx Tutorial for Developers, Open source speech recognition tool kit, URL https://cmusphinx.github.io/wiki/tutorial/

[13] Alex Waibel, Hidden Markov Models, Carnegie Mellon University, URL http://www.cs.cmu.edu/~./15381/slides/381-s17-19.HMM.final.pdf

[14] Stanford University, "Speech Recognition Using Deep Learning Algorithms", URL http://cs229.stanford.edu/proj2013/zhang_Speech%20Recognition%20Using%20Deep%20Learning%20Algorithms.pdf

[15] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition" URL https://www.cs.toronto.edu/~hinton/absps/DNN-2012-proof.pdf

[16] Rohan Raja, "A practical approach to automatic speech recognition using deep learning", July 03, 2016, URL http://algomuse.com/python/a-practical-approach-to-automatic-speech-recognition-using-deep-learning

[17] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional Neural Networks for Speech Recognition, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol.. 22, No.. 10, October 2014

[18] Qi Zhao "AI, Machine Learning and Deep Learning Explained", URL http://blog.operasolutions.com/ai-machine-learning-and-deep-learning-defined

[19] Andrew Ng, "What data scientists should know about deep learning" URL https://image.slidesharecdn.com/andrew-ng-extract-oct2015-nonotes-151124104249-lva1-app6891 /95/ andrew-ng-chief-scientist-at-baidu-30-638.jpg?cb=1448361887

[20]Pavel Kral, Vaclav Matousek, "Text, Speech and Dialogue": 18th International Conference, TSD 2015, Pilsen, Czech Republic, Sept. 14-17, 2015, Proceedings, pp 479-487