A Survey of Machine Learning Approaches for Speech Recognition

Bakhtiar K. Kasi, Riaz Ulamin, Mumraiz Kasi, and Masood Ur Rehman

***Abstract* — Machine learning approaches have been used for a wide range of applications in the recent years. The strength of these approaches mainly lies in their ability to learn from experience. Speech recognition has been area which has gained a lot of popularity in recent years. Siri, Genie, and Cortana are some commonly used examples. The focus of these applications have been to interpret human speech into a set of basic commands for a portable device. In this paper, we present a survey of the some commonly used machine learning approaches for speech recognition. While there are several variations to the speech recognition system, little is known about the challenges associated with each approach. In this paper we present a comparison of the available approaches and highlight the pros and cons of some of the popularly used approaches.**

***Index Terms* — Machine learning, speech recognition, neural networks, hidden markov models**

I. INTRODUCTION

With the recent advancements in technology, Machine Learning techniques have found widespread applications and implementations. The popularity of Machine Learning approaches have extended from Big Data Analysis [1] used for making predictions or calculated suggestions based on large amounts of data to smaller applications and products that we use on daily basis. Some of the most common applications that use Machine Learning techniques includes Netflix, an online movie streaming service that uses machine learning approach for making movie suggestions based on movies already watched in the past [2].

In the recent years, Speech Recognition has been adopted as the *state-of-the-art* technology for on demand user assistance over smart phones and other portable devices. For example, some common applications includes Siri from Apple [3], Kinect [4] and Cortana [5] from Microsoft, and Genie, an android based application. Siri is voice-activated assistant that runs on Apple specific devices and mimics human intelligence. Siri pioneered a higher degree of human conversation ability that helps human interaction with smart phones. Specifically, Siri interprets voice instructions, and, performs the necessary actions. For example, Siri can open apps, search for upcoming movie times and get score updates on ongoing sports events from around the world. Furthermore, Siri can be used to makes calls or send messages to people within the contact list. At it's core Siri involves a number of technologies, including natural language processing, question analysis, data mashups, and machine learning [6].

In this paper, we investigate the specific functionality of Machine Learning in its ability for speech recognition. We compare the different heuristic from Machine Learning that are used for speech recognition and we compare their performance for their applicability in different domains.

Currently much of the research has been focused on approaches that uses acoustic models that uses specific sensors to decode human speech into text. These approaches however assume a simple probabilistic model of speech production whereby the frequency of a specified word is mapped to sequences using the Maximum Posteriori Probability (MAP) function. The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability. We are specifically focused on Machine Learning approaches to speech recognition, which uses some form of recognition technique, that is, either a supervised pattern classification system or an unsupervised pattern classification technique. The supervised machine learning technique is trained with already labeled examples; that is, each input pattern is preassigned with a class label that associates the label with a set of attributes [7]. Unsupervised pattern classifier works in a similar fashion. However, in this case the classifier is not trained with preassigned class examples. The unsupervised approach works by clustering some representation of the input data using implicit groupings within the data, which is commonly referred to as a codebook.

To the best of our knowledge our's is the first study to investigate different Machine Learning approaches to speech recognition. Our specific contributions for this paper includes, i) a survey of the overall machine learning techniques used for speech recognition, ii) comparison of these approaches in terms of the model they employ and their training techniques, and iii) pros and cons of each of these techniques.

The rest of the paper is organized as follows. In Section II, we present our findings on the available *models* for speech recognition within the domain of Machine Learning, we also present the *difference* among these techniques. In Section III, we present the *comparison* of available techniques and their key differences, we also discuss the *pros and cons* of each approach. Finally, we conclude with our findings with a brief outlook into future work in Section IV.

II. PROPOSED METHODOLOGY

The underlying principle of Speech recognition technique is to convert spoken words to a sequence of words. This technique is commonly referred to as Automatic Speech recognition (ASR) or Speech to Text (STT). In practice, there are several applications to this approach as speech

Bakhtiar Kasi and Masood Ur Rehman are with the Department of Computer Engineering, Riaz Ulamin and Mumraiz Khan Kasi are with the Department of Computer Sciences, Balochistan University of information Technology, Engineering & Management Sciences (BUITEMS), Quetta. Email: bakhtiark asi@gmail.com,riazulamin@gmail.com,mumraizkk@gmail.com, masood.ba zai@gmail.com, Manuscript received on Feb 11, 2017 revised on April 28, 2017 and accepted on May 14, 2017

recognition applications normally ranges from cell phones to telephony and extends into home appliances as well. Furthermore, these approaches are widely used for data entry purpose and have some advanced applications as well for example, within aircrafts, and for air traffic controllers, and medical transcriptions as well [8].

A. *Training Techniques*

Several techniques have been proposed to train a speech recognition system built upon Machine Learning approach. Broadly speaking the techniques are either based on the use of human speech for training purposes or leveraging the meta model information about language models within a specific language. Furthermore, the level of complexity involved in the training step differs with the nature of training as well.

One such technique is dependent on a *trainer*, whose goal is to speaks words from within the vocabulary in order to train the system. This form of training is in line with the principles of *supervised machine learning*, where a human who acts as the speaker/trainer reads sections of text into the speech recognition system (normally the vocabulary is limited to less than 1,000 words). The **pros** of such system are that the human speech frequency of spoken words is directly mapped into the training system and therefore, one can expect a high degree of accuracy. However, the **cons** of such system are that it takes a lot of time for the trainer to train the system and if the speech is not clear or the accent is not clear enough, then certain words could easily be confused with others. Another disadvantage is the limitation on the vocabulary, that the total words allowed per training session, as a vocabulary rich language could easily mislead the training process.

A speaker independent system on the other hand has a large vocabulary of words and is trained by using sub word models. One such system built on the principle of *trainer independent* training is SPHINX [9]. SPHINX introduced a *linear predictive coding* technique which provides speaker independence. The training methodology of SPHINX was based on multiple codebooks of fixed-width parameters, and an enhanced recognizer with carefully designed models and word-duration modeling [9]. They also introduced two sub-word speech units in order to deal with co-articulation in continuous speech, an issues which was not previously addressed. The overall evaluation of SPHINX showed that it attained word *accuracies* of 71%, 94%, and 96%, for the traditional approach and the two new sub-word speech units, on a 997-word task.

B. *Models*

In this Section, we discuss the specific models that can be applied from within Machine Learning to transform text into speech. But, first we must understand how the whole process works in principle.

The overall components of a speech recognition system are presented in Figure 1. Typically a speech recognition system comprises of three main components, which performs functions such as *Speech Digitization*, *Acoustic Analysis*, and *Linguistic Interpretation*. The sequence of actions are as follows:

Digitization: The digitization process involves the conversion of analogue signal into a digital representation. This is a popular technique involving Signal Processing [10], in which a series of numbers are generated for an analogue signal such as speech. The resultant numbers show discrete set of points or samples that corresponds to analogue signals.

Acoustic Analysis: After the digitization of speech has been performed, the next phase involves the analysis of digitized sound in order to map sound frequencies to the corresponding words. The model that corresponds to this mapping is created by applying statistical models onto the digital representation of sound and also mapping the textual representation of the speech with the digital representation. Therefore, as a result of the acoustic analysis, commonly confused words such as *the* and *there* may be distinguished with a higher accuracy.

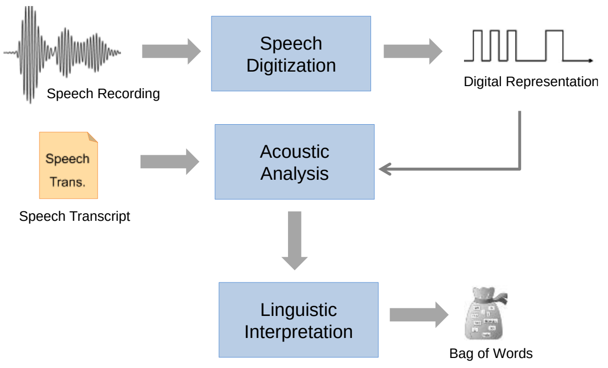

Fig. 1. Components of a Speech Recognition System.

Linguistic Interpretation: As with grammar of any language, the semantics of a language denote the meanings associated with the syntax of language itself. In speech recognition systems, the semantics comes into affect within the linguistic interpretation phase, whereby the statistical mapping of acoustic analysis is used to associate meanings with the spoken words.

Both acoustic modeling and language modeling are important parts of modern statistically based speech recognition algorithms. As explained in Section II-A, normally the process is accompanied by a trainer or the training is based on sub-word analysis. The statistical approach that employ machine learning techniques uses transcribed speech recordings for training and apply statistical processes to search through the space of all possible solutions, and pick the statistically most likely one [11].

Next we will discuss the commonly used (popular), machine learning models for speech recognition.

1) *Hidden Markov Models*

Hidden Markov Models (HMM) were first used by Baker for automatic speech recognition [12]. Markov Models can be described at any time, as being in one of a set of N distinct states $\{S_1, S_2, \ldots, S_N\}$. A process within the model can undergo a change of state (possibly back to the same state) according to the probabilities associated with the state in the model. Figure 2 shows a sample representation of the states in an HMM. At each state the decision to whether stay in the same state or moves to another state is determined by the probability associated with each transition. Let $\{p_{ij}\}$ denote the probability p of moving from state i at time t to the state j at time $t+1$. Hence for the same reason the sum of probabilities is composed of the probability that it stays at the same state or moves to some other state in sequence, which is equal to 1.0.

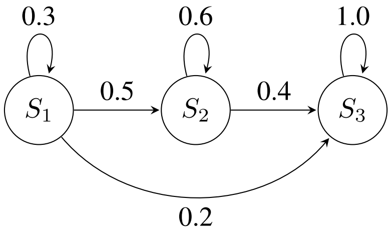

Fig. 2 Sample States of a Hidden Markov Model.

In speech recognition, a Hidden Markov Model is trained for each word within the vocabulary. However, for large models HMM is used to represent a *phoneme* instead. For example words such as *cat, bat, pat, sat* differs only by their initial phonemes: $/c/,/b/,/p/,/s/$. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution. Thus, accounting for the differences within the pronunciation of certain words just by changing the phonemes. A Hidden Markov model for a sequence of words/phonemes is therefore made by concatenating the individual trained Hidden Markov Models generated separately for the words/phonemes [9].

For example, the three self loops in Figure 2, model three parts of a phoneme, and the lower transitions explicitly model duration of one, two, or three frames. Instead of assigning a unique output probability density function to each transition, each phoneme is assigned three distributions representing the beginning, middle, and end of the phoneme. Each of these three distributions are shared by several transitions [13].

2) *Neural Networks*

Waibel et al [14] used Time-Delay Neural Networks (TDNN) for acoustic modeling with great success. Since then, neural networks have been used in many aspects of speech recognition such as phoneme classification, isolated word recognition, and speaker adaptation.

As compared to HMM, neural networks makes no assumptions about feature statistical properties. With multiple layers and the interconnections between units in each layer makes the network a good candidate algorithm for learning complex patterns including speech recognition. When used to estimate the probabilities of a speech feature segment, neural networks allow discriminating training in a natural and efficient manner. The approach performs better than HMM. In a laboratory study a performance evaluation of TDNN used a vocabulary of 5,240 common Japanese words. The words were spoken in isolation by three Japanese speakers and the study compared performance of HMM with TDNN. The results showed that the recognition rate of TDNN was 98 percent correct and HMM was 94 percent correct.

III. RESULTS

We present our findings about the strengths and weaknesses of the models in the domain of machine learning for speech recognition. The DARPA [15] project was the first of it's kind to evaluate the effectiveness of trainer dependent and trainer independent approaches for speech recognitions. The DARPA project was based on naval resource management and contained a database that is intended for use in designing and evaluating algorithms for speaker-independent, and speaker-adaptive speech recognition system. The data contains read sentences of about 1000-word task vocabulary, which represents 21,000 recorded utterances from 160 talkers with a variety of dialects [15].

Some of the challenges discussed in DARPA includes the, i) Training time: a resource consuming activity that when done in a trainer dependent fashion requires a lot of human hours, and must be done in a controlled fashion; where background noise is minimized to the extent possible. Additional, special care must be take to account for different variations of dialect and as well as the pronunciation of words, ii) Vocabulary limitations: one of key limitations of such systems is the extent to which the system can be trained, keeping in view the rich vocabulary of a particular language, normally such systems has a limited set of vocabulary and in particular, contains only words that are frequently used within that language, iii) Language models: For a trainer independent system, the rules of the language governs rules of grammar specifically about the semantics of language. Hence in such systems, it is important to incorporate the model governing the overall language. These models are not readily available and needs input from domain experts, usually linguists.

Another training dependent system that has been popularly used is known as SWITCHBOARD [16]. The SWITCHBOARD is an automated telephone-based speech recognition system, that was created by training data from about 2,500 conversations by 500 speakers from around the U.S.A. It has over an hour of speech from each of the speakers. Fifty "target" speakers participated at least 25 times, which adds up to more than an hour of speech gathered over a period of several weeks. The projects mainly used for automated routing calls to appropriate sections/departments. Voice recognition used in SWITCHBOARD helps in directing callers to target departments.

Finally, we also surveyed a system named SPHINX [17], which started off as a research project at the Carnegie Mellon University and was aimed at automated speech recognitions. SPHINX has now offerings of full fledged speech recognition applications, as well as support of light-weight applications for portable devices such as an android based application. A basic assumption of SPHINX is that the statistical models which describe a particular

TABLE I. CHARACTERISTICS OF SOME POPULAR SPEECH RECOGNITION SYSTEMS.

System	Model Used	Trainer Dependent	Trainer Independent
DARPA	HMM	21,000 recorded utterances from 160 talkers	9,120 recorded sentences
SWITCHBOARD	HMM	2,500 conversations by 500 speakers from around the U.S. with speech transcriptions	-
SPHINX Pocketsphinx (Android Based) Sphinx4 (Desktop App.)	HMM/N-Gram	-	Model's must be available for different languages
Waibel et al [14]	TDNN	-	5,240 words from Japanese vocabulary.

language should be available. However, SPHINX also maintains a set of freely available models trained for various acoustic conditions and various performance requirements.

IV. DISCUSSION AND CONCLUSION

An overview of some of the most popular systems in speech recognition is provided in Table I. As discussed earlier the systems have either used a trainer dependent mechanism for training purposes or have relied on statical models of language, or some form of acoustic analysis in a trainer independent system. The predominant model has been HMM for most of the systems and especially the trainer dependent systems. However, N-Grams has been a popularly used technique for a trainer independent system. The training techniques of a trainer dependent system has mainly relied on the voice recordings of trainers over a period of time where the focus has been to incorporate as many dialects and accents as possible, in order to achieve a high accuracy. However, the performance largely has been affected with inflicted noise at the time of recordings. The trainer independent system on the other hand has used N-Grams in some cases and other statistical models which augments with existing models of a language that are normally prepared by domain experts. The only exceptions to the use of HMM, was the use of Time Delay Neural-Network (TLDN) as used by Waibel et al [14]. However, the difference in performance of HMM and TLDN was found to be comparable, as the overall accuracy of TLDNN was 98 percent whereas the rate obtained for HMM was 94 percent correct.

While there are various approaches to speech recognition, the approach used by each of these approaches have differed mainly by the training techniques. The Hidden Markov Models (HMM) has been the predominantly used model in most of the systems. The training mechanism in each case have several pros and cons including: training time, the availability of language model, ability to cope with dialects, and the limitation on the total number of words in language vocabulary. These factors play together the overall applicability of a specific model in a specific domain. For example, the android based lightweight model of Sphinx (*Pockectsphinx*) uses only a fraction of the language model that is being used in full blown version of Sphinx4.

In future work, we would like to extend our study by incorporating evaluation of the popularly available techniques using a unified experimental approach. Most of the existing approaches have been evaluated with different datasets and have not been tested on a common dataset/trainer set/vocabulary. For example, Waibel et al [14] evaluated their approach using Japanese vocabulary, DARPA used very specific naval resource management datasets, and SWITCHBOARD used voice recording from within U.S.A only. Therefore, it is difficult to compare their performance in regard to each other. It would be interesting to see how, all these and other approaches perform on a single benchmark under similar conditions.

REFERENCES

[1] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," *SIGKDD Explor. Newsl.*, vol. 14, no. 2, pp. 1–5, Apr. 2013. [Online]. Available: http://doi.acm.org/10.1145/2481244.2481246

[2] M. Szomszor, C. Cattuto, H. Alani, K. O?Hara, A. Baldassarri, V. Loreto, and V. D. Servedio, "Folksonomies, the semantic web, and movie recommendation," 2007, event Dates: 3-7th, June 2007. [Online]. Available: https://eprints.soton.ac.uk/264007/

[3] "iOS - Siri." [Online]. Available: http://www.apple.com/ios/siri/

[4] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[5] "Cortana - Meet your personal assistant - Microsoft - Global." [Online]. Available: https://www.microsoft.com/en/mobile/experiences/cortana/

[6] J. R. Bellegarda, "Spoken language understanding for natural interaction: The siri experience," in *Natural Interaction with Robots, Knowbots and Smartphones* Springer, 2014, pp. 3–14.

[7] T. M. Mitchell, "Machine learning," *Burr Ridge, IL: McGraw Hill*, vol. 45, 1997.

[8] L. R. Rabiner, "Applications of speech recognition in the area of telecommunications," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, Dec 1997, pp. 501–510.

[9] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, Jan 1990.

[10] S. Smith, Digital signal processing: a practical guide for engineers and scientists. Newnes, 2013.

[11] B. H. Juang and L. R. Rabiner, "Automatic speech recognition - A brief history of the technology development," *Elsevier Encyclopedia of Language and Linguistics*, 2015.

[12] J. Baker, "The dragon system–an overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 24–29, 1975.

[13] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[14] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, Mar 1989.

[15] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The darpa 1000-word resource management database for continuous speech recognition," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, Apr 1988, pp. 651–654 vol.1.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Mar 1992, pp. 517–520 vol.1.

[17] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Mountain View, CA, USA, Tech. Rep., 2004.