Automatic Speech Recognition: Advancements of Networks and Acoustic Modeling in AI

Shlok Kaneria [1] **Ben Mitchell** [1]

Abstract

Speech recognition is invading our lives. It's built into our phones, our game consoles and our smart watches. It's even automating our homes. The primary objective of automatic speech recognition is to build a statistical model to infer the text sequences from a sequence of feature vectors. Recent advancements in network frameworks, such as Connectionist Temporal Classification or Deep Neural Belief Networks have allowed for substantial progress in the scope of Acoustic Modeling. Our focus is mainly on surveying the various networks utilized for speech recognition and comparing them. We also survey real-life applications of many of the networks presented. The paper is meant to be an introduction to acoustic modeling for individuals from various disciplines interested in Automatic Speech Recognition (ASR).

1. Introduction

In recent years, there has been an abundance of progress in the field of automatic speech recognition and acoustic modeling. These progresses have led to the creation of systems that have enabled services such as Google Assistant, Microsoft Cortana, Amazon Alexa, Apple's Siri, and countless others. Many of these achievements are powered by deep learning techniques.

This paper will survey new developments shared by many publications on this topic, as well as discuss core ideas presented in the work. More specifically, in Section 2, we focus our attention to the historical background of the field of speech recognition in general, and how trying to achieve acoustic modeling came about to be over the years. In section 3, we expand on the primary resources that we will be extrapolating information from. Section 4 delves into specific networks presented in depth such as the fundamental Hidden Markov Model, the Gaussian Mixture Model, and deep neural networks that have emerged in more recent studies as viable options for better results. As shown in Figure 1, there are multiple steps required for proper output, and these networks are the backbone for efficient and accurate classification and output.

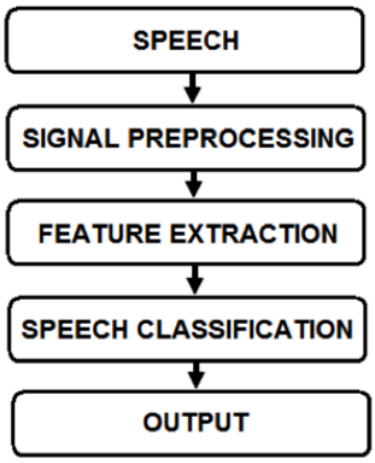

Figure 1. As adapted from (Gevaert et al., 2010), this figure outlines the multiple steps for the purposes of speech recognition, from the input of the acoustic environment, to signal processing, which suppresses irrelevant sources of variation in the audio input, to extracting speech specific features, to finally the classification of these features and relation to the vocabulary set to produce the desired output.

In Section 5, we present the various learning approaches utilized for the most successful deep learning networks, as well as show emerging advances in hybrid frameworks that may not be as well-known or widespread such as connectionist temporal classification (CTC). Due to the recent impact that discoveries of neural networks have caused, Automatic Speech Recognition is now a subject of constantly growing interest.

Section 6 focuses on the various applications that these models and networks are used for currently and could pos-

[1]Department of Computing Sciences, Villanova University, Villanova, PA. Correspondence to: Shlok Kaneria <skaneri1@villanova.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

sibly be used for in the future. Some current technologies that are outlined in this paper includes Bing's Mobile voice search application and advances of Google Assistant and YouTube's speech driven environments.

Finally, We propose core problems to work on and potential future directions to solve them in section 7, where we also summarize the main points that were presented through this survey.

The aim of this work is to generalize the numerous advancements in the scope of speech recognition and provide insight into the most recently discovered frameworks that show promising results to the broader scientific community. This paper will be useful for researchers working in the field of machine learning or individuals looking for an overview and are interested in speech recognition models. This work will provide them with a survey of networks and examples of applications where they have been used. Similarly, scientists who are looking for a consolidated report on the advancements of an extremely dense field can use this as a brief introduction to a niche topic.

This systematic literature review of the progress made in automatic speech recognition will focus on identifying and analyzing the contributions of six key research papers that have been published from 2005 to 2019 in the area of deep neural networks in speech-related applications.

2. Historical Background

Speech is the most efficient way that humans use to communicate with each other. This also means that speech could be a useful interface to interact with machines. For a long time, research on how to improve this type of communication has been done. Some successful examples of using machines to aid with communication includes the invention of the megaphone and telephone. Even centuries ago, people were experimenting on speech synthesis. For example, in the late 18th century, Von Kempelen developed a machine capable of "speaking" words and phrases (Gevaert et al., 2010).

In the early 1970's we begin to see a surge of interest in the field of speech recognition, with sudden advancements such as Dynamic Time Warping (DTW) to handle time variability, a type of distance measure for spectral variability. Then, in the mid to late 1970's, we notice the introduction of the expectation-maximization (EM) algorithm, an iterative method to find maximum likelihood estimates of parameters in statistical models, a key feature for training Hidden Markov Models (HMM). With the EM algorithm, it became possible to develop speech recognition systems for real-world tasks using the richness of Gaussian Mixture Models (GMM) to represent the relationship between HMM states and the acoustic inputs (Hinton et al., 2012).

By the mid 1980's, HMMs had become the dominant technique for all types of automatic speech recognition. In the following years leading into the 90's, many of the current state-of-the-art Large Vocabulary Continuous Speech Recognition Systems (LVCSR), which are hybrids of neural networks and Hidden Markov Models (HMMs), begin to emerge. The majority of the 2000's has consisted of discriminative training to reduce word and "phone" error. In this context, phone shall be defined as the perceptually distinct units of sound in a specified language that distinguish one word from another.

Deep learning has become increasingly popular since the introduction of an effective new way of learning deep neural networks (DNN) in 2006. DNNs have proved very successful for acoustic modeling in speech recognition especially for large-scale tasks, and this success has been based largely on the use of the back-propagation algorithm with rather standard, feed-forward multi-layer neural networks. In addition to improved learning procedures, the main factors that have contributed to the recent successes of deep neural networks have been the availability of more computing power, and the availability of more training data. The initial breakthrough in acoustic modeling was triggered by the use of a generative, layer-by-layer pre-training method for initializing the weights before running the discriminative back-propagation learning procedure, but subsequent research has revealed that generative pre-training is unnecessary when there is a very large amount of labeled training data. Back-propagation can be started from random initial weights given that their scales are carefully determined to prevent the initial error derivatives from being very large or very small (Deng et al., 2013).

More than a year ago, four research groups wrote an overview article called "Deep neural networks for acoustic modeling in speech recognition" in which they presented their shared views on applying DNNs to acoustic modeling in speech recognition. Since then, the four groups and other speech or machine learning groups around the world have done a lot of new work developing new models and learning methods, and performing new evaluation experiments (Deng et al., 2013; Hinton et al., 2012). This overview paper will be extensively referenced in this work as well due in large part to the immense contributions they have outlined for ASR.

3. Related Work

Our work in surveying the field of Automatic Speech Recognition and Acoustic modeling draws information from many fundamental publications dating back to 2005. Riccardi et al. showed promising results to solve the problem of adaptive learning in the context of ASR. The paper proposes an active learning algorithm for this specific pur-

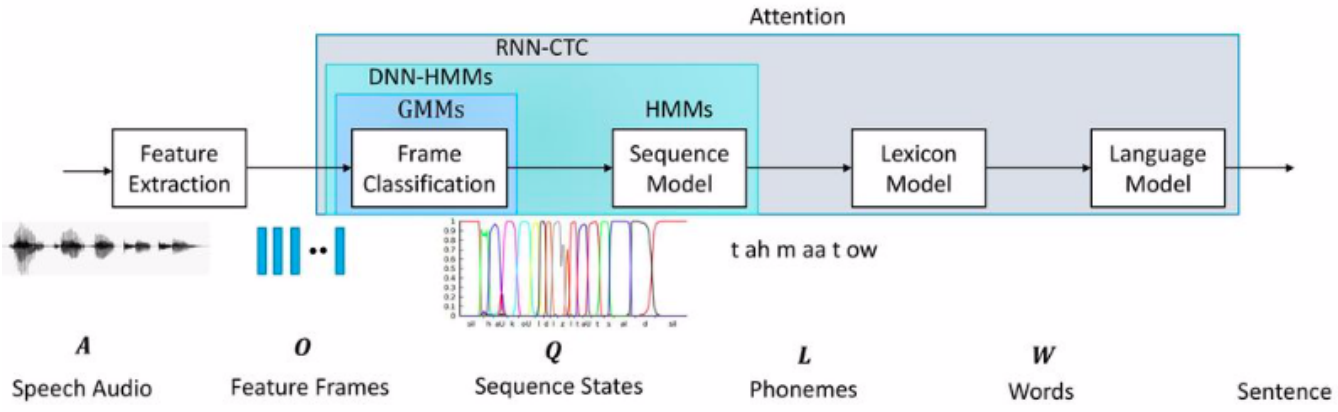

Figure 2. This image accurately sums the overall structure of an ASR network, from the acoustic input, to the application of the hybrid models, which will be described later in this paper, to the end result. Image adapted from (Gevaert et al., 2010; Hinton et al., 2012; Gales & Young, 2007; Gevaert et al., 2010)

pose (Riccardi & Hakkani-Tür, 2005).

Over the years, researchers realized the ease of application for HMMs as a viable solution for the speech recognition problem, and the research for advancements of such a model also became more substantial. In 2010, Gavaert et al. produced a study to investigate speech recognition classification performance. Their main goal of the paper was to analyze how well models such as HMMs and other neural networks fared in terms of handling the biggest issues that many of the previous models have encountered before. Some these issues, as presented in this work, include Speaker Variation, where the same word is pronounced differently by different people due to various attributes such as age, sex, anatomic variations, speed of speech, etc. Background noise and even issues such as the influence of intonation and stress on certain syllables when spoken are omnipresent impediments for many of these models and their structures. Several of these points as addressed in Gavaert then became a foundation for analyzing the proficiency of speech recognition networks (Gevaert et al., 2010).

By 2012 and 2013, there were several new types of networks that showed promising results, resulting in the publication of "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview" by Deng et al. In that paper, Deng provides an brief summary of a multitude of other papers presented at a special session dealing with applications of deep neural networks for ASR in 2013. The technical overview of the papers presented was then organized into five main points on how to improve existing deep learning methods, much of which was based on similar fundamental issues proposed in previous works such as Gavaert et al.:

(1) Better optimization

(2) Better types of neural activation function and better network architectures

(3) Better ways to determine the myriad hyper-parameters of deep neural networks

(4) More appropriate ways to preprocess speech for deep neural networks

(5) Ways of leveraging multiple languages or dialects that are more easily achieved with deep neural networks than with GMMs (Deng et al., 2013).

As previously referenced in the Section 2, perhaps the most influential work in this field comes from a paper written in conjunction with four separate research groups led by Hinton et al., each presenting a singular study regarding deep neural networks for acoustic modeling in speech recognition. This work has been fundamental in reviewing the successes achieved using pretraining, and how this has led to a resurgence of interest in Deep Neural Networks for Acoustic Modeling.

Usually, HMMs and GMMs have been the leading applications for speech recognition systems to show how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. An example structure of a hybrid model network can be seen in Figure 2. This study by Hinton et al. proposes an alternative way to evaluate the same fit, instead by using a feed-forward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output (Hinton et al., 2012).

The papers that followed the survey produced by Hinton

et al. mainly focused on the hybrid models of neural networks and HMMs. This is what is currently used in many of the current state-of-the-art LVCSRs. Most of the systems contain separate components that deal with acoustic modeling, language modeling, and sequence decoding. In their work, "End-to-End Attention Based Large Vocabulary Speech Recognition," Bahdanau et al. investigate a more direct approach in which the HMM is replaced with a Recurrent Neural Network (RNN) that performs sequence prediction directly at the character level (Bahdanau et al., 2016).

In the most recent works that this survey will refer to, the overarching studies revolve around models and frameworks that allow for peak performance such as Connectionist Temporal Classification Frameworks (CTC) and Deep Belief systems. The success of these systems come from the emphasis placed on training of feature vectors and acoustic modeling adaptation. By increasing the number of features and including more layers in the network to increase the number of parameters, we are able to produce Deep Belief Networks. As explained in recent papers, feed forward neural networks offer several potential advantages over state-of-the-art GMMs that have been dominating the ASR field till now:

• Their estimation of the posterior probabilities of HMM states does not require detailed assumptions about the data distribution.

• They allow an easy way of combining diverse features, including both discrete and continuous features.

• They use far more of the data to constrain each parameter because the output on each training case is sensitive to a large fraction of the weights.

As evident by the results produced in papers by Yu et al. and Mohamed et al., it is safe to say that Deep Belief Networks and CTC frameworks are the best structures for ASR as of right now.

4. Types of Networks

The focus of this paper will be mainly assessing the various types of networks that have been used for the purpose of ASR. There have been numerous models that have been proved to be effective over the years, including, but not limited to HMMs, GMMs, RNNs, CNNs, as well as hybrid frameworks and deep belief networks. In the following sections, we will discuss the structure of each of these models and why they have proved to have been effective, as well as compare them to see how certain solutions are touted as better for attention based acoustic modeling.

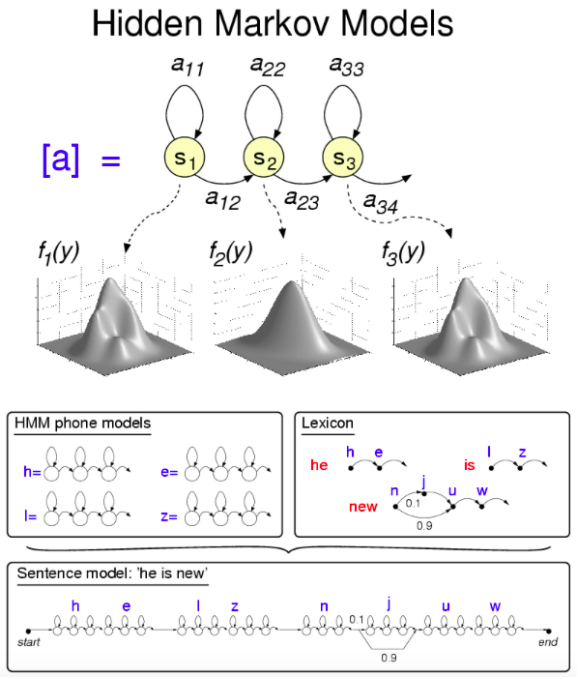

Figure 3. Here we see exactly how the HMM tends to work given an acoustic input. The final sentence model is created by concatenating the correct phone models. Figure adapted from (Gales & Young, 2007).

4.1. Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) provide a simple and effective framework for modeling time-varying spectral vector sequences. As a consequence, almost all present day LVCSR systems are based on HMMs.

Until now, this is the most successful and most used pattern recognition method for speech recognition. It's a mathematical model derived from a Markov Model. For the purposes of ASR, we use a slightly adapted Markov Model. Speech is split into the smallest audible entities. All these entities are represented as states. As a word enters the Hidden Markov Model it is compared to the best suited entity. According to transition probabilities, there exist a transition from one state to another. A state can also have a transistion to it's own state if the sound repeats itself. Markov Models seems to perform quite well in noisy environments because every sound entity is treated separately. If a sound entity is lost in the noise, the model might be able to guess that entity based on the probability of going from one sound entity to another, as seen in Figure 3 (Gevaert et al., 2010).

Most current speech recognition systems use HMMs to deal with the temporal variability of speech, while relying on GMMs to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that

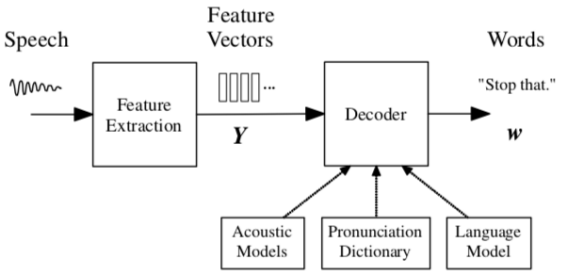

Figure 4. This image shows the architecture of an HMM, from the reception of the auditory input, to analyzing feature vectors, and finally finding the internal states given these observations to produce results.

represents the acoustic input (Hinton et al., 2012).

After the adoption of hybrid models, we slowly begin to see how the DNN/HMM architecture has emerged as the dominant model used in industry. One of the important factors that lead to superior performance in the DNN/HMM hybrid system is its ability to exploit contextual information (Yu & Li, 2018).

HMM ARCHITECTURE

A Markov chain contains all the possible states of a system and the probability of transiting from one state to another. A first-order Markov chain assumes that the next state depends on the current state only. However, in many ASR systems, not all states are observable and we call these states hidden.

The probability of observing an observable given an internal state is called the emission probability. The probability of transiting from one internal state to another is called the transition probability. An HMM is modeled by the transition and the emission probability. For speech recognition, the observable is the content in each audio frame. We can use the Mel Frequency Cepstral Coefficient (MFCC), which is a representation of the short-term power spectrum of a sound, parameters to represent it (Gales & Young, 2007).

Given an HMM model is learned, we can use the forward algorithm to calculate the likelihood of our observations. The main objective is to sum the probabilities of the observations for all possible state sequences, and can be calculated by this formula:

$$p(X) = \sum_{s} p(X, S) = \sum_{s} p(X|S)p(S) \qquad (1)$$

This formula states how $p(X)$, or the probability of observed events, is equal to the sum over all possible time sequences of internal states, which is equal to and can be calculated by finding the summation of the products of the emission probability and the transition probability as described above.

Next, given the HMM model, we must find the internal states given the sequence of observations. This process is called decoding, as depicted in the structure described in Figure 4. If we have an audio clip, the internal states represent the phones. Decoding allows us to find the internal states that maximize the likelihood of observations.

4.2. Gaussian Mixture Model (GMM)

A Gaussian Mixture is a function that is comprised of several Gaussians. The EM algorithm, as explained in Section 2, is used to estimate the parameters of the GMM. The extracted parameters: the means, standard deviations and component weights can be related to each location of several prominent bands of frequency that determine the phonetic quality of a vowel, bandwidths and magnitudes. As the features directly represent the linear spectrum, it is possibly to apply techniques for vocal tract length normalisation and additive noise compensation techniques.

GMM strengths showcase how with enough components, they can model probability distributions to any required level of accuracy, and they are fairly easy to fit to data using the EM algorithm (Hinton et al., 2012).

A huge amount of research has gone into finding ways of constraining GMMs to increase their evaluation speed and to optimize the trade-off between their flexibility and the amount of training data required to avoid serious overfitting.

Despite all their advantages, GMMs have a serious shortcoming, the fact that they are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space.

Since GMMs are non-linear models, it assumes that each data point is generated by a single component of the mixture so it has no efficient way of modelling multiple simultaneous events.

Speech is produced by modulating a relatively small number of parameters of a dynamical system and this implies that its true underlying structure is much lower-dimensional than is immediately apparent in a window that contains hundreds of coefficients. According to Hinton et al., other types of models may work better than GMMs for acoustic modeling if they can more effectively exploit information embedded in a large window of frames (Hinton et al., 2012).

Deep neural networks that have many hidden layers and

are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin (Deng et al., 2013; Hinton et al., 2012; Gevaert et al., 2010).

4.3. Recurrent Neural Network (RNN)

Another key approach that many researchers have used for the purposes of Automatic Speech Recognition includes Recurrent Neural Networks.

The use of recurrent neural networks for acoustic modeling was pioneered by Tony Robinson, but quickly fell out of favor because of the difficulty of training them. Recently, however, RNNs have achieved excellent results at language modeling and the use of multiple hidden layers has allowed them to outperform all other methods on a TIMIT dataset (Deng et al., 2013).

The success of modern RNNs for the purposes of Speech Recognition can probably be explained to a large extent by the elegant way in which they can deal with sequences of variable length.

To obtain a model that uses information from both future frames and past frames, one can pass the input data through two recurrent neural networks that run in opposite directions and concatenate their hidden state vectors. Recurrent neural network of this type are often referred to as bidirectional RNNs.

Finally, it has been shown that better results for speech recognition tasks can be obtained by stacking multiple layers of recurrent neural networks on top of each other (Bahdanau et al., 2016; Deng et al., 2013). This can simply be done by treating the sequence of state vectors as the input sequence for the next RNN in the pile. Two bidirectional RNNs can be stacked on top of each other to construct a deep architecture.

LSTM-RNNs use input, output and forget gates to control information flow. This is so that gradients can be perfected over relatively longer span of time. These networks have been shown to outperform DNNs on a variety of ASR tasks (Yu & Li, 2018).

4.4. Convolutional Neural Network (CNN)

Although convolutional models achieved good classification results, applying them to phone recognition is not straightforward. This is because temporal variations in speech can be partially handled by the HMM component and those aspects of temporal variation that cannot be adequately handled by the HMM can be addressed more explicitly and effectively by hidden trajectory models (Hinton et al., 2012).

For the reason stated above, CNNs, similarly to RNNs, also showed early promise for acoustic modeling but were later abandoned, probably because the convolution was done across time rather than across frequency.

The driving force behind the success of CNNs however, is due to the convolutional layer. The input to the convolution operation is usually a three-dimensional tensor (row, column, channel) for speech recognition.

Because of the translational invariability, CNNs can exploit variable-length contextual information along both frequency and time axes (Bahdanau et al., 2016). If only one convolution layer is used, the translational variability the system can tolerate is limited. To allow for more powerful exploitation of the variable-length contextual information, convolution operations (or layers) can be stacked (Hinton et al., 2012), very similar to the structure of the RNN as described in the section above.

As described, temporal variation is already well-handled by the HMM so convolution across frequency is much more helpful because it provides partial in-variance to changes in the properties of the vocal tract (Bahdanau et al., 2016). Hinton et al. also demonstrated that convolution across frequency was very effective for TIMIT.

Convolutional neural networks are also useful for LVCSR and further demonstrates that multiple convolutional layers provide even more improvement when the convolutional layers use a large number of convolution kernels, such as feature maps.

4.5. Connectionist Temporal Classification Framework (CTC)

Deep CNNs can be used together with RNNs and under frameworks such as a Connectionist Temporal Classification Framework (CTC). CTCs allow for a sequence-to-sequence direct optimization (Yu & Li, 2018), as well as RNNs to predict sequences that are shorter than the input sequence by summing over all possible alignments between the output sequence and the input of the CTC module.

This summation, as mentioned above, is done using dynamic programming in a way that is similar to the forward and backward passes that are used in HMM, and as described in section 4.1. In the CTC approach, output labels are independent, given the alignment and the output sequences. In the context of speech recognition, this means that a CTC network lacks a language model, which greatly boosts the system performance when added to a trained CTC network (Bahdanau et al., 2016).

Speech recognition tasks are a sequence-to-sequence task, which maps the input waveform to a final word sequence or an intermediate phoneme sequence. For the purposes of acoustic modeling, output of word or phoneme sequence is

of utmost importance, instead of the frame-by-frame labeling which the traditional cross entropy training criterion is focused on. Hence, CTCs are introduced to map the speech input frames into an output label sequence. As the number of output labels decreases below the number of input speech frames, CTC path is introduced to force the output to have the same length as the input speech frames by adding blank as an additional label and allowing repetition of labels (Yu & Li, 2018).

The most attractive characteristics of CTC is that it provides a path to end-to-end optimization of acoustic models. The end-to-end speech recognition system is explored to directly predict characters instead of phonemes, hence removing the need of using lexicons and decision trees. This is one step toward removing expert knowledge when building an ASR system. Another advantage of character-based CTC is that it is more robust to the accented speech as the graphoneme (the smallest meaningful contrastive unit in a writing system) sequence of words is less affected by accents than the phoneme pronunciation.

Lastly, the end-to-end optimization strategy is desired, given its simplicity and joint optimization characteristics, if we only need to optimize for the decoding result and have sufficient training data. This has been proven effective with word-based CTCs when trained with hundreds of thousands hours data. However, the one downside to this is that it is not feasible to get that large amount of data for most tasks. (Yu & Li, 2018).

4.6. Deep Belief Networks

Deep Belief Networks (DBN) are capable of making good use of the more detailed information available in this larger input representation. The availability of more layers within a network inherently allows DBNs to perform better among other networks mentioned in this work, as described in the following section of this paper.

The characteristic of a deep belief network is similar to a neural network. The performance of a neural network depends on the structure itself and it is suitable to select the model and size of the network for the data to handle. Deep belief networks have the advantage in speech recognition as it generates the feature learning with a subsequent stage of supervised learning, once the network initialized, the first way using unsupervised and then fine-tuned with the labeled data to train some neuron in initial weight vectors. Deep belief networks have many nonlinear hidden layers to produce posterior of probabilities that take several frames of coefficient as input (Mohamed et al., 2010).

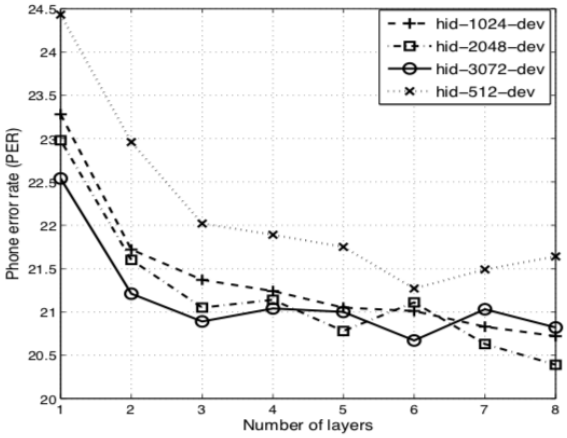

Figure 5. Adapted from (Mohamed et al., 2010), this figure depicts the Phone error rate as a function of the number of layers of a DBN, using 11 input frames.

5. Learning

When GMMs were first used for acoustic modeling, they were trained as generative models using the EM algorithm, and it was some time before researchers showed that significant gains could be achieved by a subsequent stage of discriminative training using an objective function more closely related to the ultimate goal of an ASR system (Hinton et al., 2012).

Currently, the biggest disadvantage of DNNs compared with GMMs is that it is much harder to make good use of large cluster machines to train them on massive data sets (Hinton et al., 2012).

The TIMIT data set provides a simple and convenient way of testing new approaches to speech recognition. The training set is small enough to make it feasible to try many variations of a new method and many existing techniques have already been bench marked on the core test set, so it is easy to see if a new approach is promising by comparing it with existing techniques that have been implemented by their proponents (Gales & Young, 2007).

Mohamed et. al. showed that a DBN-DNN acoustic model outperformed the best published recognition results on TIMIT. Similarly, other publications around that time also achieved a similar improvement on TIMIT by applying state-of-the-art techniques developed for large vocabulary recognition.

Training DBNs of various sizes as mentioned in this paper is quite computationally expensive (Mohamed et al., 2010). All DBNs were pre-trained with a fixed recipe using stochastic gradient decent with a mini-batch size of 128

training cases. Currently DBN-DNN architectures often performed best on the development sets reported. As we can see in Figure 5. the overarching trend of the graph shows how as the number of layers of a given network decreases, the phone detection error rate also tends to decrease. All methods use MFCCs as inputs (Hinton et al., 2012).

When neural nets were first used, they were trained discriminatively. It was only recently that researchers showed that significant gains could be achieved by adding an initial stage of generative pretraining that completely ignores the ultimate goal of the system. The pretraining is much more helpful in deep neural nets than in shallow ones, especially when limited amounts of labeled training data are available. It reduces overfitting, and it also reduces the time required for discriminative fine-tuning with back propagation.

6. Applications

The first successful use of acoustic models based on DBN-DNNs for a large vocabulary task used data collected from the Bing mobile voice search application (BMVS). The task used 24 h of training data with a high degree of acoustic variability caused by noise, music, side-speech, accents, sloppy pronunciation, hesitation, and numerous others as mentioned previously in this paper. The results reported in Hinton et al. demonstrated that the best DNN-HMM acoustic model trained with context-dependent states as targets achieved a sentence accuracy of 69.6 percent on the test set, compared with 63.8 percent for a fairly well trained GMM-HMM baseline (Hinton et al., 2012).

The DNN-HMM training recipe developed for the Bing voice search data was applied unaltered to the Switchboard speech recognition task, a large multi speaker corpus of conversational speech and text, to confirm the suitability of DNN-HMM acoustic models for large vocabulary tasks.

Other notable applications of such networks can be seen through advancements by Google. Google Voice Input, a key feature of many Google products, including Google Assistant, transcribes voice search queries, short messages, e-mails, and user actions from mobile devices. This is a large vocabulary task that uses a language model designed for a mixture of search queries and dictation. Google's full-blown model for this task, which was built from a very large corpus, uses a GMM-HMM model. Similarly, this application goes hand in hand with YouTube's own speech recognition task, which employs the same model and structure (Hinton et al., 2012).

7. Summary

This paper surveyed new developments shared by many publications on the topic of ASR, as well as discuss core ideas presented in these presented works. By delving into specific networks such as the fundamental Hidden Markov Model, the Gaussian Mixture Model, and deep neural networks, we discussed the impact that each of these models have had on the scope of the field. Through our analysis of these papers, we have found that there are promising results presented for CTC frameworks and deep belief networks. These models may be potential avenues that researchers may want to explore further in the future.

This all being said, there is no reason to believe that we are currently using the optimal types of hidden units or the optimal network architectures, and it is highly likely that both the pretraining and fine-tuning algorithms can be modified to reduce the amount of overfitting and the amount of computation. We therefore expect that the performance gap between acoustic models that use DNNs and ones that use GMMs will continue to increase for some time.

References

Bahdanau, Dzmitry, Chorowski, Jan, Serdyuk, Dmitriy, Brakel, Philémon, and Bengio, Yoshua. End-to-end attention-based large vocabulary speech recognition. 2016.

Deng, Li, Hinton, Geoffrey, and Kingsbury, Brian. New types of deep neural network learning for speech recognition and related applications: An overview. 2013.

Gales, Mark and Young, Steve. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2007.

Gevaert, Wouter, Tsenov, Georgi, and Mladenov, Valeri. Neural networks used for speech recognition. *Journal of Automatic Control*, 20:1–7, 2010.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E., rahman Mohamed, Abdel, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N., and Kingsbury, Brian. Deep neural networks for acoustic modeling in speech recognition. pp. 82–97, 2012.

Mohamed, Abdel-Rahman, Dahl, George E., and Hinton, Geoffrey. Acoustic modeling using deep belief networks. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–10, 2010.

Riccardi, Giuseppe and Hakkani-Tür, Dilek. Active learning: Theory and applications to automatic speech recognition. *Transactions on Speech and Audio Processing*, 13, 2005.

Yu, Dong and Li, Jinyu. Recent progresses in deep learning based acoustic models (updated). 2018.