Speech recognition techniques, past to future

A literature study

Authors
Niklas Qvarforth, nahlen@kth.se
Erik R Svensson, ersvenss@kth.se

Abstract

In this report we have investigated progress and techniques from the 1980's to present time in the field of automatic speech recognition (ASR). This includes Hidden Markov Models (HMM), Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) and hybrids of these. In the early stages of speech recognition HMM was the most effective model in machine learning for speech recognition. It was later combined with the ANN into the ANN-HMM hybrid model that performed better than all previous models. As processing power allowed for more layers in ANN, the Deep Neural Networks improved further upon this hybrid model. The last few years, the CNN has become the leading model surpassing ANN models in terms of performance and the CNN-HMM model is one of the most efficient models in speech recognition.

1 Introduction

Speech is the most efficient way of communication we know so far. Humans and also animals use sound to communicate and to transfer knowledge between each other. Humans have the skill to interpret sound from birth and we keep up relying on it throughout our lifetime [17]. This way of communication and possibility of easy communication is also the way we would want to communicate with machines. The problem is that machines are not under conscious control and to interpret sounds is a complex task. Speech is also very different in form of accents, emotions, gender and so on [17]. Automatic Speech Recognition (ASR) is very important in the field of computer interaction, it uses the process and relevant technology to convert speech signals into sequence of words by using computers and implemented algorithms [17]. In present time the newest technology can understand and interpret thousands of words under functional environment [17]. Speech signals give informations about both the identity of a speaker and the content of the speech [17]. To be able to handle speech in machine learning different methods have evolved since the introduction of speech recognition in the 1950s. This report first present the general concepts used in speech recognition over this time period HMM, ANN, RNN and CNN. Then the time-line are examined, starting from were Hidden Markov Models were introduced in the early 1980s up until present day were Neural Networks are used in combinations with other techniques to receive the best results.

2 Background

2.1 HMM, Hidden Markov Models

A Markov Model is a stochastic model of various states. It consists of a number of states that it can be in, a starting probability of in which state the model will start in, and a transition probability for each state that tells the probability of going from that state to another in each time-step.

A Hidden Markov Model adds another lay atop the regular MM. Instead of being able to see which state the model is currently in with absolute certainty there is an observation probability that says that in each state there is a probability of making a certain observation. The number of possible observations also need not be the same as the number of states.

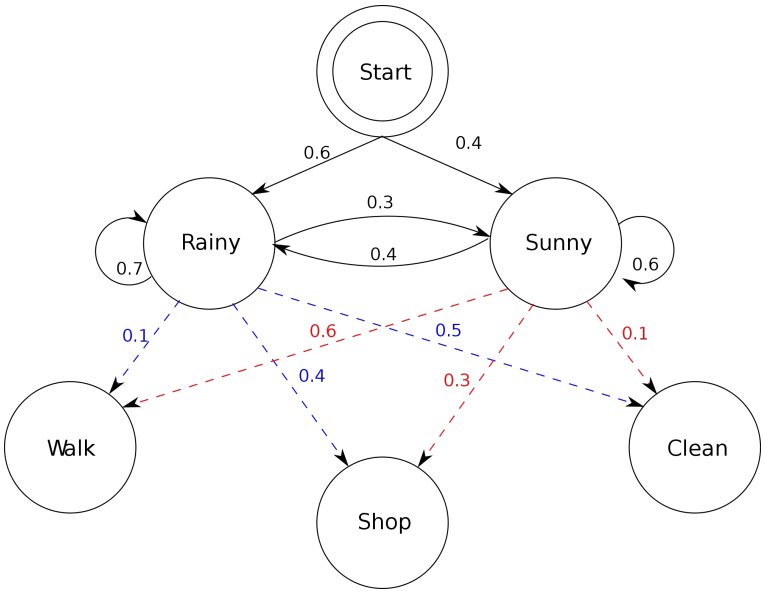

Figure 1: Visualization HMM

The way a HMM learns is that given one or several sequences of observations it tries to fit the transition and observation likelihoods so that it maximizes the probability of the given sequence of observation. This will however only find a local maximum and is thus sensitive to initialization.

Then given a properly trained HMM one can use algorithms such as forward-backward to find the maximum probability of sequence of states given a sequence of observations. This however is a local maximum, and in most problems of interest the optimization surface will be very complex and have many local maxima. [22]

2.2 ANN, Artificial Neural Networks

The human brain consists of billions of nerve cells or *neurons*, these communicate with electrical signals that are short lived impulses in the voltage of the cell wall membrane [16]. The inter-neuron connections are mediated by electrochemical junctions called synapses, these are located on branches of the cell called dendrites. Each neuron receives thousands of connections from other neurons and are constantly receiving a multitude of incoming signals. These signals does eventually reach the cell body, in the cell body the signals are integrated or summed in some way, and if the resulting signal exceeds some threshold the neuron will generate an impulse in response to other neurons [16]. This is transmitted via a fiber called

an axon. When determining weather an impulse should be produced or not, some incoming signals produce an inhibitory effect and tend to prevent generation of the impulse, while others are excitatory and promote impulse generation[16]. The distinctive processing ability of each neuron is then supposed to reside in the type "excitatory or inhibitory" and strength of its synaptic connections with other neurons [16].

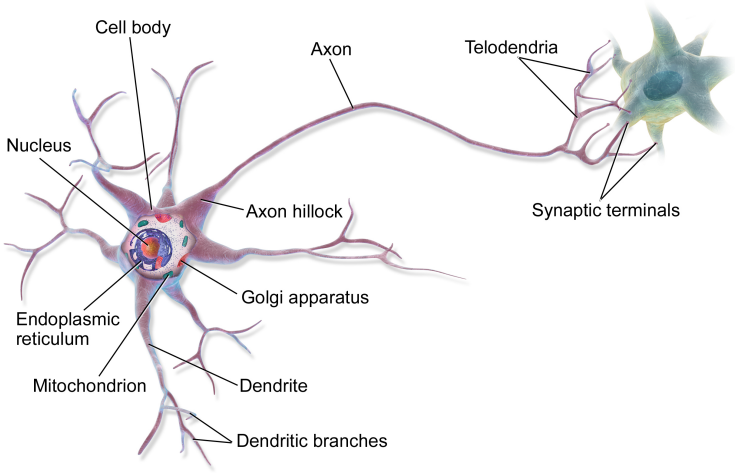

Figure 2: Visualization of biological network

In the same way the architecture and style of processing are used and incorporated when designing an artificial neural network [16]. When creating a artificial neural network the neurons become nodes in our system. The synapses are modeled by a single number or weight so that every input is multiplied by this weight before forwarded to to the "cell body". In the "body" the signals are summed up together by simple arithmetics, like addition to supply node activation. Output is produced depending if it exceeds a threshold or not (output could be 0 or 1) [16]. The term network is used for any type of system that use artificial neurons, from a single node to a large collection of nodes. In real neurons the strength of the synapse signal may be modified so that the behavior of each neuron can change or adapt to its input stimulus [16]. In an artificial environment this is equivalent to the altering of the weights values [16]. In this way the neural network can (using learning techniques) adapt to certain patterns, and given lots of training data the process can be repeated, the network is then trained until it is converged [16]. The result should be ready to take on unknown data [16].

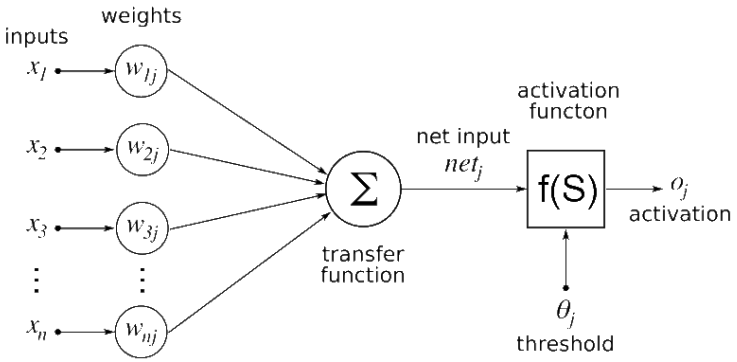

Figure 3: Artificial neural network

2.3 RNN, Recurrent Neural Networks

A Recurrent neural Network is specialized on handling sequences. Mostly sequences in time. It does this by taking and outputting a form of hidden state for each step in evaluation of the sequence. This hidden state functions as the the memory of the network wherein it remembers the previous inputs into the network. As this hidden state is dependent on previous iterations of the network, to calculate it's gradient it become necessary to back-propagate through these previous steps. [14]

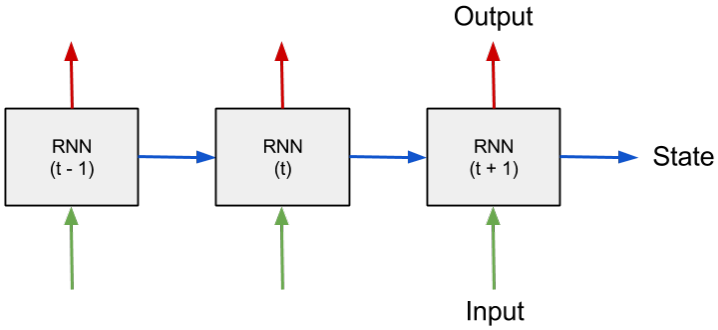

Figure 4: Artificial neural network

2.4 CNN, Convolutional Neural Networks

Convolutional Neural Networks "CNN", are a specialized kind of neural network (ANN) for processing data that has a known grid-like topology [6]. Examples include time-series data, which can be sought of as a 1D grid taking samples at regular time intervals, and sound data which can be thought of as a 2D grid with features [6]. Convolutional networks has played an important role in history of deep learning and are as mentioned above about ANN also a successful application of insights obtained by studying the brain and applied to machine learning [6]. CNN were one of the first deep models to perform well, and were some of the first neural networks to solve commercial applications. The interest in CNN's and deep learning began when [3], won the ImageNet object recognition challenge, but convolutional neural networks had been used to win other machine learning and computer vision contests for years earlier, giving less impact [6]. Convolutional networks were one of the first working deep networks trained with back-propagation [6]. General back propagation were considered to have failed and why CNN's were a success is unclear. Ideas as to why CNN's were working better were that they had more computational efficiency than fully connected networks[6], or that recent failures could be dependent on psychological factors that made practitioners not to fully try experiments (they simply did not believe they should work). CNN's are most efficient on a two dimensional topology, to process one-dimensional data, recurrent neural networks (RNN) perform better [6]. A convolutional network are simply neural networks that instead of a matrix multiplication uses a convolution instead of matrix multiplication in at least one of the layers [6]. The convolution between two functions $f(x * w)$ can be described of the integral [6]:

$$s(t) = f(x * w)(t) = \int_0^t x(a)w(t-a)da$$

This function is used to give a smoothed estimate of the measurement we want to obtain. In convolutional neural networks the $\mathbf{x}$ usually refers to the *input* and the $\mathbf{w}$ to the *kernel* the output is sometimes referred to as the *feature map*. when working with data on a computer the time is discretized and the discrete convolution function can be written as [6]:

$$s(t) = f(x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a).$$

In machine learning applications, the input is usually a multidimensional array of data, and the kernel is usually a multidimensional array of parameters that are adapted by the learning algorithm [6]. Because each element of the input and the kernel needs to be stored separately, we assume that all functions are zero in all values except the finite set of points for which we store values. Thus one can implement an infinite summation, as a summation over a finite number of array elements [6]. Convolutions are often used over more than on axis (two-dimensional I and two-dimensional W) at once and has a commutative property [6].

$$s(i,j) = f(I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n).$$

$$s(i,j) = f(I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n).$$

The commutative property of convolution arises when flipping the kernel relative to the input, in the sense that as m increases, the index to the input increases, but the input to the kernel decreases. This commutative property is usually good for writing proofs but not so important for implementation. Many neural networks implements instead a function called **cross-correlation [6]:**

$$s(i,j) = f(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n).$$

This is the same as convolution but without flipping the kernel. It is rare in machine learning implementations that convolutions is the alone function, instead it is used combined with other functions [6]. Discrete convolution can be viewed as multiplication by a matrix, but the matrix has several entries constrained to be equal to other entries. Any neural network algorithm that works with matrix multiplication should work with convolution [6].

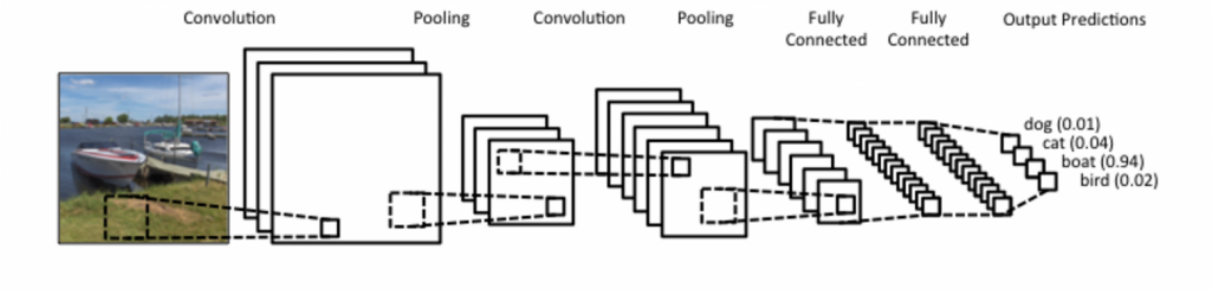

Figure 5: Convolutional neural network

Convolution presents three important ideas which is presumed to leverage improvement to machine learning, sparse interactions, parameter sharing and equivalent representations. Convolutions also allows for inputs of variable size [6]. The main advantage of CNN's is their accuracy in recognition problems, the drawbacks are mainly, slow training, computational expensive and the need of a lot of training data (parameters). Also the CNN's needs supervised training, at least on a high level.

3 Method

We used a combination of Google Scholar to find and KTH library to access scientific articles on the subject of ANN, CNN and speech recognition. By going through reports regarding speech recognition techniques from different time periods in the research development, we aimed to see if there has been any progress and what techniques has been involved. We also wanted to conclude what is the best technique today and what is supposed to be the future of speech recognition. While there are many flaws and different abbreviations to these methods described we go through the general use of successful concepts, giving an possible way to overlook the progress made in the area of Speech Recognition.

4 Results

Reading the reports in a chronological order one can see how various forms of HMM has been the dominant model in speech recognition. Starting around the 1980s the models of this time went from being more intuitive template-like approach to become a more rigorous statistical framework [2]. The approach of the HMM was understood in a few laboratories but the methodology was not complete until the mid 1980s. After the publication of [12, 15] the HMM became the preferred method for speech recognition and became the the standard for the upcoming two decades, along with a steady stream of refinements and improvements to the technology [2]. The stochastic process of the HMM was able to model the intrinsic variability of the speech and also the structure of the spoken language. As speech could vary a lot from instance to instance because of differences in noise, pronunciation and similar. This gave the effect that even the same words could have a very different input to the model. The HMM takes this into account, as it creates the probabilistic model based upon the variabilities of the acoustic realizations in the utterance [2]. With this model an algorithm called *Baum-Welsh* is used. This algorithm returns an indication from the model of the likelihood that this model represents the word spoken, based upon the input from the utterance.

Another technique that emerged in the late 1980's was the ANN (Artificial Neural Network). Neural networks was really first introduced in the 1950's but but failed to produce any results [2]. In the late 1980't it was again introduced, this time together with a function called error back-propagation, later in development this technique will have great impact in speech recognition. The main reason for the attention in the 80's was the capability to approximate any function to arbitrarily precision. Attempts with this method was successful with phonemes of a few words, but problems of speech recognition was that of handling the variations of a speakers utterances, and up to this time NN had not showed to be able to handle this kind of problems [2].

Researchers started to try to integrate the HMM model within the NN. In the 1990's several innovations took place also in the field of pattern recognition. This was mainly because a change of paradigm which was the result of changing the problem set. From like before following Baye's framework with estimations of the parameters and distributions of data, to become an optimization problem of parameters [2]. This optimization problem involved minimization of the recognition error. The objective of an recognizer was now to be able find the least recognition error instead of fitting the distribution function. This change spawned a lot of new techniques for example the Support Vector Machines (SVM) which became very popular to study among scientists [2]. In the mid 1990's a tool called the FSM (finite state machine) library, was created. This was a finite state network approach in a unified transducer framework. This library has been a major component of almost all modern speech recognition and understanding systems [2].

When statistical methods received success in the field of ASR, it receiver interest from DARPA (Defense advanced research agency). This resulted in that new systems spawned like SPHINX CMU [18], BABYLOS (BBN) [11] and DECIPHER (SRI) [20]. The SPHINX system was especially successful and was able to successfully integrate the statistical model of the HMM method [2]. With the support of DARPA a lot of different fields and tasks were investigated and evaluated int the 90's and into the 20'th century. One

of the most challenging tasks proposed by DARPA was the Switchboard, were the speaker were supposed to be conversational and spontaneous [2]. The result was the conclusion that this problem was far more complicated to solve than the regular task driven speech approach. A result which also emerged from this proposal was that the error rate decreased when bigger training sets were used, this is something also later seen in present neural networks [2]. I the beginning of the 20'th century the Neural networks model began to get more impact in speech recognition techniques. The emerge of deep learning has become the most significant advances in the field of speech recognition [23] and all models tend to become different abbreviations of neural networks.

This led to the ANN-HMM hybrid model becoming the most efficient model in speech recognition for its time. Surpassing its peers with a reduced error rate of 30% [6, 5, 7]. These ANN-HMM hybrid models were later improved upon into DNN-HMM models as processing power allowed for larger number of layers in the ANN model. These models surpassed other HMM such as the GMM-HMM [4, 3]. Experiments with DNN-HMM hybrids on the TIMIT database has yielded an almost 30% error rate [4, 3].

Later on as different versions of RNN was being experimented with it managed to become the leading model within speech recognition, outperforming other state of the art models at the time [13].

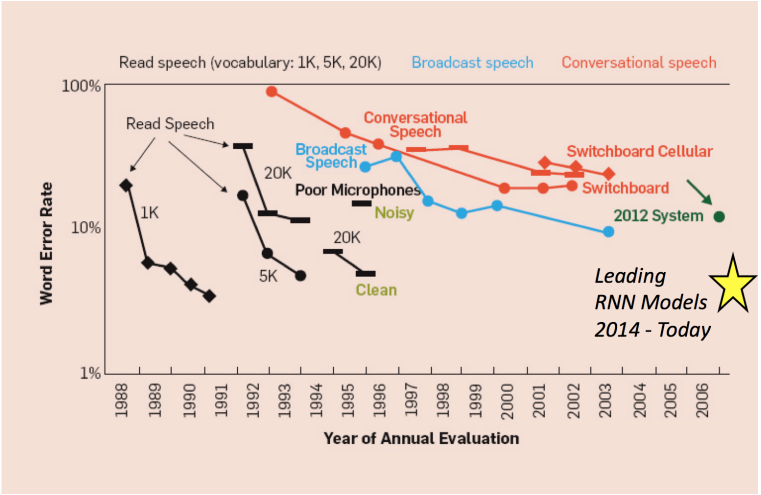

Figure 6: Techniques timeline [19, 1]

CNN began within the field of image recognition and computer vision where it had proved itself [8]. Experiments were made to test if this could be applied to the spectral feature image of speech. CNN-HMM models proved able to achieve a better result than DNN-HMM with the same number of NN layers. And has also given an overall lower error rate of 6-10% compared to previous models. Resulting in a 20% error rate on the TIMIT database [9, 10, 8].

Experiments with CNN on raw speech data has yielded some promising results and may indicate that CNN can come to replace much of the signal processing currently done to speech data [21]. Which might also suggest that a version of NN may in the future be the dominant model in speech recognition.

5 Discussion and Conclusions

HMM has remained the dominant model used for speech recognition for decades. With the rise of ANN it was combined into various NN-HMM hybrid models that performed better than previous models. The NN part of these hybrid models have developed from the early ANN, to DNN to currently CNN, each step improving the results of the previous.

As HMM is stochastic models it will always have a degree of uncertainty that will probably be hard to overcome. NN doesn't seem to have that problem as much as with more and more neurons and proper training it looks to be moving towards a more deterministic way of recognizing patterns. With the major obstacle of simply adding more and more neurons to NNs is the available processing power, which is increasing exponentially, it stands to reason that NN may very well make most other models of machine learning, especially stochastic ones, obsolete.

It seems as ANN, especially CNN, is becoming the most effective method of speech recognition. While HMM is still used to great effect in these new methods, as CNN was shown to have the potential to work on its own, it might very well find itself outdated.
Recently, (ASR) has achieved major breakthroughs and greatly improved performance. Applications, including smartphone assistants like Siri, Cortana, Google Now and products such as Amazon Echo and Kinect Xbox One, have started to become part of our daily life. All these with high standard and performance using new techniques of ASR [23].

References

[1] Haamiz Ahmed. *Microsoft Beats Everyone, Nearly Perfects Speech Recognition With New Tech.* URL: https://propakistani.pk/2016/09/15/microsoft-beats-everyone-nearly-perfects-speech-recognition-new-tech/.

[2] Biing-Hwang Juang et al. "Automatic speech recognition–a brief history of the technology development". In: *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1 (2005), p. 67.

[3] Geoffrey Hinton et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[4] George E Dahl et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition". In: *IEEE Transactions on audio, speech, and language processing* 20.1 (2012), pp. 30–42.

[5] Giulia Bernardis et al. "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems". In: (1998).

[6] Ian Goodfellow et al. *Deep Learning.* http://www.deeplearningbook.org. MIT Press, 2016.

[7] Joao Neto et al. "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system". In: *Fourth European Conference on Speech Communication and Technology.* 1995.

[8] Ossama Abdel-Hamid et al. "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition". In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE. 2012, pp. 4277–4280.

[9] Ossama Abdel-Hamid et al. "Convolutional Neural Networks for Speech Recognition". eng. In: *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* 22.10 (Oct. 2014), pp. 1533–1545. ISSN: 2329-9290.

[10] Ossama Abdel-Hamid et al. "Exploring convolutional neural network structures and optimization techniques for speech recognition." In: *Interspeech.* Vol. 2013. 2013, pp. 1173–5.

[11] Richard Schwartz et al. "The BBN BYBLOS continuous speech recognition system". In: *Proceedings of the workshop on Speech and Natural Language.* Association for Computational Linguistics. 1989, pp. 94–99.

[12] Stephen E Levinson et al. "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition". In: *The Bell System Technical Journal* 62.4 (1983), pp. 1035–1074.

[13] Tomáš Mikolov et al. "Recurrent neural network based language model". In: *Eleventh Annual Conference of the International Speech Communication Association.* 2010.

[14] Denny Britz. *Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs.* URL: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/.

[15] J. D. Ferguson. "Hidden Markov analysis: an introduction". In: *Proc. of the Symposium on the Applications of Hidden Markov Models to Text and Speech* (), pp. 8–15.

[16] Kevin Gurney. *An Introduction to Neural Networks.* Bristol, PA, USA: Taylor & Francis, Inc., 1997. ISBN: 1857286731.

[17] Bhushan C Kamble. "Speech Recognition Using Artificial Neural Network–A Review". In: *IEEE trans* (2016).

[18] Kai-Fu Lee. "On large-vocabulary speaker-independent continuous speech recognition". In: *Speech communication* 7.4 (1988), pp. 375–379.

[19] *Microsoft's Voice Recognition Technology Almost as Accurate as Humans.* URL: https://news.developer.nvidia.com/microsofts-voice-recognition-technology-almost-as-accurate-as-humans/.

[20] Hy Murveit. "SRI's DECIPHER System". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '89. Philadelphia, Pennsylvania: Association for Computational Linguistics, 1989, pp. 238–242. DOI: 10.3115/100964.100990. URL: https://doi.org/10.3115/100964.100990.

[21] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert. *Analysis of cnn-based speech recognition system using raw speech as input*. Tech. rep. Idiap, 2015.

[22] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[23] Zixing Zhang. "Deep learning for environmentally robust speech recognition: An overview of recent developments". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 9.5 (2018), p. 49.

6 Appendix

Comment 1
The models are described good in overall. But it might be better to give an explanation of RNNs in the background section (RNN are only mentioned in the figure 5 and one sentence in section 2.3).

Adjustments:
We have added a section about RNN.

Comment 2
I did not find any clear mistakes. It is covered so much about the neurons, instead it might be better to mention about deep neural networks, back propagation and recurrent neural networks.

Adjustments:
We have tried to adjust this. Still we think that the basis of the neural network is an important path and all other techniques builds upon this so its motivated to have it there. Also it does not take away anything important from other parts except it gives the text more depth.

Comment 3
Extended parts of the report, including definitive claims, came with just one reference. This reference often was to other overall field assessment work, and not to first order research papers.

Adjustments:
To find information about the history of ASR is hard to do if not allowed to look at other overall field assessment reports and would have been way out of bounds of this project to look for first order sources for all time periods. Instead we have tried to do something more excessive as to look for a period of time that are longer 1980 to present. We feel that it is motivated to look in historical reviews to find information about history, not have to find actual first order source of every time period.

Comment 4
"Especially the final claim that hybrid HMM methods may be left behind is not supported at all".
Adjustment:
Changed the wording in that it is a likely possibility given that one test with CNN showed that NN might work in speech recognition on its own. It's also part of the discussion as the conclusion and discussion are in the same section. As it is part of a discussion it does not need to be supported by references other than the logical assumptions presented in the argument based on the background knowledge of both HMM and NN.

Comment 5
"When presenting the HMM, ANN and CNN the explanations vary wildly in depth, with the CNN being very in depth. A similar level of explanation should be provided throughout. None of the presented background methods is explained how is used in the context of speech recognition specifically. The metaphor of the human brain to explain a neural network is trite in this point in time, and does more harm than good since the differences between a brain and a NN are much more than the superficial similarities. The historical overview of the methods used lacks a clear narrative. There is a vague progression between different architectures and methodologies used, but it could be highlighted even more, and a timeline following the accompanying graph could be put forward. Lines 143 to 146 describing the initial failure of ANNs and the introduction of backpropagation should be rewritten making clear what the problem with ANNs was and how backprop solved it".

Adjustments:
We have tried to adjust the narrative path, thus we don't agree with the reviewer about the comparison

because we think that it gives more depth, also we don't know what was intended with comment about back propagation.