P.p1 to compare these methods and show

Design, Architecture

p. p1 {margin: 0.

0px 0. 0px 0. 0px 0. 0px; font: 10. 0px Helvetica}p.

p2 {margin: 0. 0px 0. 0px 0. 0px 0. 0px; font: 14. 5px Helvetica}p.

p3 {margin: 0. 0px 0. 0px 0.

0px 0. 0px; font: 12. 0px Helvetica}p. p4 {margin: 0. 0px 0. 0px 0. 0px 0. 0px; font: 9.

0px Helvetica}span. s1 {font: 24. 0px Helvetica}span.

s2 {color: #002486}Overview of Three Dierent Structures of Artificial Neural Networks forSpeech Recognitions1736880Abstract Automatic speech recognition (ASR) is the translation, through some methodologies, of humanspeech into text by machines and plays an importantrole nowadays. In this research reviewwe examine three di erent artificial neural networkarchitectures that are used in speech recognitionfield and we investigate their performancein di erent cases. We analyze the state-of-artdeep neural networks (DNNs), that have evolvedinto complex structures and they achieve significantresults in a variety of speech benchmarks.

Afterward, we explain convolutional neural networks(CNNs) and we explore their potential inthis field. Finally, we present the recent researchin highway deep neural networks (HDNNs) thatseem to be more flexible for resource constrainedplatforms. Overall, we critically try to comparethese methods and show their strengths and limitations. Each method has its benefits and applicationsand from them we try to draw someconclusions and give some potential future directions.

I. IntroductionMachine Learning (ML) is a field of computer sciencethat gives the computers the ability to learn throughdi erent algorithms and techniques without being programmed. ASR is closely related with ML because it usesmethodologies and procedures of ML 1 , 2 , 3 .

ASR hasbeen around for decades but it was not until recently thatthere was a tremendous development because of the advancesin both machine learning methods and computerhardware. New ML techniques made speech recognitionaccurate enough to be useful outside of carefully controlledenvironments, so it could easily be deployed in many electronicdevices nowadays (i. e. computers, smart-phones)and be used in many applications such as identifying andauthenticating a user via of his/ her voice. Speech is the most important mode of communicationbetween human beings and that is why from the early partof the previous century, e orts have been made in orderto make machines do what only humans could perceive.

Research has been conducted through the past five decadesand the main reason was the desire of making tasks automatedusing machines 2 . Many motivations using di erenttheories such as probabilistic modeling and reasoning, pattern recognition and artificial neural networks a ectedthe researchers and helped to advance ASR. The first single advance in the history of ASR occurredin the middle of 70’s with the introduction of theexpectation-maximization (EM) 4 algorithm for traininghidden Markov models (HMMs). The EM technique gavethe possibility to develop the first speech recognition systemsusing Gaussian mixture models (GMMs). Despite allthe advantages of the GMMs, they are not able to modele ciently data that lie on or near a nonlinear surface in thedata space (i. e. sphere).

This problem could be solved byartificial neural networks because they can capture thesenon-linearities in the data but the computer hardware ofthat era did not allow us to build complex neural networks. As a result, in the beginning most speech recognition systemswere based on HMMs. Later the neural network andhidden Markov model (NN/ HMM) hybrid architecture 5 was used for ASR systems. After 2000s and over the lastyears the improvement of computer hardware and the inventionof new machine learning algorithms made possible thetraining for DNNs. DNNs with many hidden layers havebeen shown to achieve comparable and sometimes muchbetter performance than GMMs in many di erent databases(with speech data) and in a range of applications 6 .

Afterthe huge success of DNNs, researchers try other artificialneural architectures such as recurrent neural networks withlong short-term memory units (LSTM-RNNs) 7 , deepbelief networks and CNNs, and it seems that each one ofthem has its benefits and weaknesses. In this literature review we present three types of artificialneural networks (DNNs, CNNs, and HDNNs). Weanalyze each method, we explain how they are used fortraining and what are their advantages and disadvantages. Finally we compare these methods in the context of ASR, identifying where each one of them is more suitable andwhat are their limitations. Finally, we draw some conclusionsfrom these comparisons and we carefully suggestsome probable future directions.

II. MethodsA. Deep Neural NetworksD NNs are feed-forward artificial neural networks withmore than one layer of hidden units. Each hiddenlayer has a number of units (or neurons) each of whichInformatics Research Review (s1736880) takes all outputs of the lower layer as input and passes themthrough a linearity.

After that we apply a non linear activationfunction (i. e. sigmoid function, hyperbolic tangentfunction, some kind of rectified linear unit function (ReLU8 , 9), or exponential linear unit function (ELU 10)) forthe final transformation of our initial inputs. Sometimes, fora multi-class classification problem, the posterior probabilityof each class can be estimated using an output softmaxlayer. For the training process of DNNs we usually use theback propagation technique 11 . For large training sets, itis typically more convenient to compute derivatives on amini-batch of the training set rather than the whole trainingset (this is called stochastic gradient descent). As cost functionwe often use the cross-entropy (CE) in order to have acomparison meter between the output of the network andthe actual output but the choice of the cost function actuallydepends on the case.

The di culty to optimize DNNs with many hiddenlayers along with overfitting problem force us to use pretrainingmethods. One such a popular method is to usethe restricted Boltzmann machines (RBMs) 12 . If weuse a stack of RBMs then we can construct a deep beliefnetwork (DBN) (you should not be confused with dynamicBayesian network). The purpose of this is to add an initialstage of generative pretraining. The pretraining is veryimportant for DNNs because it reduces overfitting and italso reduces the time required for discriminative fine-tuningwith propagation. DNNs in the context of ASR play a major role. Manyarchitectures have been used by di erent research groups inorder to gain better and better accuracy in acoustic models.

You can see some methodologies in the article 6 that itpresents some significant results and shows that DNNs ingeneral achieve higher speech recognition accuracy thanGMMs on a variety of speech recognition benchmarks suchas TIMIT and some other large vocabulary environments. The main reason is that they take advantage from the factthat they can handle the non-linearities in the data and sothey can learn much better models comparing to GMMs. However, we have to mention that they use many model parametersin order to achieve a good enough speech accuracyand this is sometimes a drawback.

Furthermore, they arecomplex enough and need many computational resources. Finally, they have been criticized because they do not preservesome specific structure (we can use di erent structuresuntil we achieve a significant speech accuracy), theyare di cult to be interpreted (because they have not somespecific structure) and they possess limited adaptability (weuse di erent approaches for di erent cases). Besides allof these disadvantages they remain the state-of-the-art forspeech recognition the last few years and they have givenus the most reliable and consistent results overall. B.

Convolutional Neural NetworksConvolutional neural networks (CNNs) can be regardedeas DNNs with the main di erence that instead of usingfully connected hidden layers (as it happens in DNNs; fullconnection with all the possible combinations among thehidden layers) they use a special network structure, whichconsists of convolution and pooling layers 13 , 14 , 15 . Basicrule is that the data have to be organized as a numberof feature maps in order to be passed in each convolutionallayer. One significant problem we have when we want totransform our speech data in feature maps concerns frequencybecause we are not able to use the conventionalmel-frequency cepstral coe cient (MFCC) technique 16 .

The reason is that this technique does not preserve the localityof our data (in the case of CNNs), although we wantto preserve locality in both frequency and time. Hence, asolution is the use of mel-frequency spectral coe cients(MFSC features) 15 . Our purpose with MFSC technique is to form the inputfeature maps without loosing the property of locality in ourdata. Then we can apply the convolution and pooling layerswith their respective operations to generate the activationsof the units in those layers. We should mention that eachinput feature map is connected to many feature maps andthe feature maps share the weights. Thus, firstly, we usethe convolution operation to construct our convolutionallayers and afterwards, we apply the pooling layer in orderto reduce the resolution of the feature maps.

This processcontinues depending on how deep we want to be our network(maybe we could achieve higher speech accuracy withmore layers on this structure or maybe not). You can seethe whole process and the usage of convolution and poolinglayers in the paper 15 . Moreover, as it happens for DNNswith RBMs, there is a respective procedure CRBM 17 forCNNs that allow us pretraining our data in order to gainin speech accuracy and reduce the overfitting e ect.

In thepaper 15 , the authors also examine the case of a CNNwith limited weight sharing for ASR (LWS model) and theypropose to pretrain it modifying the CRBM model. CNNs have three major properties: locality, weightsharing, and pooling. Each one of them has the potentialto improve speech recognition performance. These propertiescan reduce the overfitting problem and they can addrobustness against non-white noise. In addition, they canreduce the number of network weights to be learned.

Bothlocality and weight sharing are significant factors for theproperty of pooling which is very helpful in handling smallfrequency shifts that are common in speech signals. Theseshifts may occur from di erences in vocal tract lengthsamong di erent speakers 15 . In general, CNNs seem tohave a relative better performance in ASR taking advantagefrom their special network structure. C.

Highway Deep Neural NetworksH DNNs are depth-gated feed-forward neural networks18 . They are distinguished from the conventionalDNNs for two main reasons. Firstly they use much lessmodel parameters and secondly they use two types of gatefunctions to facilitate the information flow through the hiddenlayers. Informatics Research Review (s1736880) HDNNs are multi-layer networks with many hiddenlayers.

In each layer we have the transformation of theinitial input or of the previous hidden layer with the correspondingparameter of the current layer (they are combinedin a linear way) followed by a non-linear activation function(i. e. sigmoid function). The output layer is parameterizedwith the parameter and we usually use the softmax functionas the output function in order to obtain the posterior probabilityof each class given our initial inputs. Afterwards, given the target outputs, the network is usually trained bygradient descent to minimize a loss function such as crossentropy(CE function). So, we can see that the architectureand the process are the same as in DNNs that we describedin subsection of DNNs . The di erence from the standard DNNs is that highwaydeep neural networks (HDNNs) were proposed to enablevery deep networks to be trained by augmenting the hiddenlayers with gate functions 19 .

This augmentation happensthrough the transform and carry gate functions. The firstscales the original hidden activations and the latter scalesthe input before passing it directly to the next hidden layer18 . Three main methods are presented for training, thesequence training, the adaptation technique and the teacherstudenttraining in the papers 18 , 20 , 21 . Combining thesemethodologies with the two gates it is demonstrated howimportant role the carry and the transform gate play in thetraining. The main reason is that the gates are responsibleto control the flow of the information among the hiddenlayers.

They allow us to achieve comparable speech recognitionaccuracy to the classic DNNs but with much lessmodel parameters because we have the ability to handle thewhole network through the parameters of the gate functions(which are much less comparing to the parameters of thewhole network). This outcome is crucial for platforms suchas mobile devices (i. e. voice recognition on mobiles) due tothe fact that we have not many disposal resources in thesedevices. D. Comparison of the MethodsThese methods, that we presented, have their benefits andlimitations. In general, DNNs behave very well and inmany cases they have enough better performance comparedto GMMs on a range of applications. The main reason isthat they take advantage from the fact that they can handlemuch better the non linearities in the data space.

On theother hand, their biggest drawback compared with GMMsis that it is much harder to make good use of large clustermachines to train them on massive data 6 . As far as the CNNs are concerned, they can handlefrequency shifts which are di cult to be handled withinother models such as GMMs and DNNs. Furthermore, it isalso di cult to learn such an operation as max-pooling instandard artificial neural networks. Moreover, CNNs canhandle the temporal variability in the speech features aswell 15 . On the other hand, the fine-tuning of the poolingsize (carefully selection of pooling size) is very importantbecause otherwise we may cause phonetic confusion, especiallyat segment boundaries.

Despite the fact that CNNsseem to have better accuracy than DNNs with less parameters, computationally are more expensive because of thecomplexity of the convolution operation. HDNNs are considered to be more compact than regularDNNs due to the fact that they can achieve similarrecognition accuracy with many fewer model parameters. Furthermore, they are more controllable than DNNs andthis is because through the gate functions we can control thebehavior of the whole network using a very small numberof model parameters (the parameters of the gates). Moreover, HDNNs are more controllable because the authorsin paper 18 show that simply updating the gate functionsusing adaptation data they can gain considerably in speechrecognition accuracy. We cannot conclude much for theirgeneral performance because they are a recent proposaland it is needed more research to see their overall benefitsand limitations.

However, the main idea is to use them inorder to have comparable ASR accuracy with DNNs andsimultaneously to reduce the model parameters. III. ConclusionsOverall , we can say that DNNs are the state-of-thearttoday because they behave very well on a rangeof speech recognition benchmarks. However, other architecturesof artificial neural networks such as CNNs haveachieved comparable performance in the context of ASR. Besides that, research continues to be conducted in thisfield in order to find new methods, learning techniquesand architectures that will allow us to train our data setsmore e ciently. This means less parameters, less computationalpower, less complex models, more structured models. Ideally we would like to have a whole general model thatcovers a lot of cases and not many di erent models thatapplied in di erent circumstances.

On the other hand this isprobable di cult, so just distinct methodologies and techniquesfor di erent cases may be our temporary or uniquesolution. In this direction, HDNNs or other methods maybe used to deal with specific cases.

Many future directions have been suggested the lastfew years for research in order to advance ASR. Someprobable suggestions are to use unsupervised learning orreinforcement learning for acoustic models. Another potentialdirection is to search for new architectures or specialstructures in artificial neural networks or inventing newlearning techniques and at the same time improving ourcurrent algorithms.