LARGE RECEPTIVE FIELD CONVOLUTIONAL NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

by

Salar Jafarlou

APPROVED BY SUPERVISORY COMMITTEE:

John H.L. Hansen, Chair

Carlos Busso

P Rajasekaran

To my father,

the most true man I know

LARGE RECEPTIVE FIELD CONVOLUTIONAL NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

by

SALAR JAFARLOU, BS

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN

ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

December 2020

ACKNOWLEDGMENTS

My sincere thanks to Dr. John Hansen for his motivation, support and guidance in my research. His support has been an invaluable asset to my education and professional development at UT Dallas. I also thank him on behalf of many of his international students, for being a huge support thorough the challenging academic path.

I wish to thank the other committee members - Professors Hansen, Busso and Rajasekaran for their valuable time and expert opinions on my thesis. Special thanks to my colleague Soheil Khoram for the suggestions and help throughout the project. And many thanks to my friends Vinay and Chunlei who helped develop the reverberated data and helped me get comfortable on the Speech toolkit.

Many thanks to all my colleagues at CRSS- Shahram, Midia, Aditya, Fahimeh and Wei for a great productive work environment, professional support and sharing knowledge in the lab.

I gratefully acknowledge The University of Texas at Dallas for their financial support of this research project.

Finally, I thank my parents and my brother, all I have in life. I could not have successfully completed my master's degree amidst my hectic work life and personal life without the encouragement of my brother and to him and my parents I dedicate this work.

August 2020

LARGE RECEPTIVE FIELD

CONVOLUTIONAL NEURAL NETWORKS

FOR ROBUST SPEECH RECOGNITION

Salar Jafarlou, MS
The University of Texas at Dallas, 2020

Supervising Professor: John H.L. Hansen, Chair

Despite significant efforts over the last few years to build a robust automatic speech recognition (ASR) systems for different acoustic settings, the performance of the current state-of-the-art technologies significantly degrades in noisy reverberant environments.

Convolutional Neural Networks (CNNs) have been successfully used to achieve substantial improvements in many speech processing applications including distant speech recognition (DSR). However, standard CNN architectures were not efficient in capturing long-term speech dynamics, which are essential in the design of a robust DSR system. In this thesis, we address this issue by investigating variants of large receptive field CNNs (LRF-CNNs) which include *deeply recursive networks*, *dilated convolutional neural networks*, and *stacked hourglass networks*. To compare the efficacy of the aforementioned architectures with the standard CNN for Wall Street Journal (WSJ) corpus, we use a hybrid DNN-HMM based speech recognition system. Then in order to evaluate the system performances in reverberated environments (the case for distant speech recognition) we evaluated the system in both simulated and realistic reverberated environments. For the former, we used realistic room impulse responses (RIRs) to simulate the reverberated versions from a clean channel. Finally, for realistic reverberation settings, we used UTD-Distance corpus to evaluate our system. Our experiments show that

with fixed number of parameters across all architectures, the large receptive field networks show consistent improvements over the standard CNNs for both clean and distant speech. Amongst the explored LRF-CNNs, stacked hourglass network has shown improvements with a **8.9**% relative reduction in word error rate (WER) and **10.7** % relative improvement in frame accuracy compared to the standard CNNs for distant simulation setups. Stack of hourglass also gave a **13.68** % and **12.90** % relative reduction for 1 m and 3 m distanced microphones respectively. For 6 m far microphones recursive networks were the one with the most WER gain of **7.46** %. This thesis is a study on a set of unsupervised techniques achieved by modifications on acoustic modeling component of the HMM-based ASR engine for robustness in reverberate environments. These techniques showed a consistent improvements in both simulated and realistic settings and demonstrates a track of research in the field of alternative acoustic modeling structures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Distant Speech Recognition (DSR) is a technology that uses distant microphone(s) to accomplish natural human-machine interfaces. Recent years have seen the application of DSR in consumer devices, such as Amazon Echo, Google Home, smart TVs, etc. Due to the existence of background noise, multiple overlapping speakers and reverberation, building a robust DSR system has become a challenging task for present speech systems. Broadly speaking, a DSR system can be split into two sub-tasks: (i) a front-end speech enhancement system, and (ii) a back-end automatic speech recognition (ASR) system modification which can be designed to operate on speech recordings from either a single distant microphone or multiple distant microphones. A DSR system, engineered using multiple distant microphones, use advanced front-end microphone array processing techniques that yield in a substantially reduced word error rate (WER) compared to systems engineered using a single distant microphone. In a considerable portion of real life applications a fixed length array of microphones is not practical, especially considering the fact that quality of array-based preprocessings are dependent on the distance of microphones from each other. These kind of limitations in front-end enhancements makes back-end systems a better choice for some applications. Most back-end state-of-the-art ASR systems used in a DSR system typically divide the recognition task into three sub-tasks: (i) feature extraction, (ii) acoustic modeling, and (iii) language modeling, which are optimized independently to achieve the best performance.

Over the years, steady attempts by speech community researchers have helped in optimizing the aforementioned building blocks of the ASR system. Feature extraction, a process of extracting discriminative characteristics from speech signals to accurately classify linguistic content has been extensively studied, leading in features such as Mel-filterbank cepstral coefficients (MFCCs) and perceptual linear prediction coefficients (PLPs) providing optimum

efficiency for many speech-related systems. Similarly, extensive studies in natural language processing (NLP) have shown that recurrent neural network-based language models (RNN-LMs) generate accurate probability distributions over word sequences, helping an ASR system to decrease prediction errors. For acoustic modeling, researchers have used Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) for more than a decade. Later, studies in this area have shown that acoustic models based on fully connected deep neural network (FC-DNNs) outperformed the conventional GMM-HMM systems. In addition, significant improvements were also made by replacing fully connected DNNs with convolutional neural networks (CNNs) because of their effectiveness in capturing local (short-term) dependencies of speech signals. This leads to significant improvements in WER for speech recordings from a close-talk microphone. Consequently, CNNs do not efficiently capture global (long-term) dependencies which make them less effective in designing a DSR system.

CNN is a multi-layer stacked neural network which includes convolutional layers, non-linearities, and pooling layers(in some frameworks) Krizhevsky, Sutskever, and G. E. Hinton 2012. Convolutions in different layers of the standard CNN consider current and few neighboring inputs from a previous layer to produce a single output. As the number of layers in this network increases the region of the input space (the first layer of the network) that affects a neuron in a particular layer of the neural network also increases. This region is well recognized in CNN architecture as the receptive field. In general, any neuron of any layer can be investigated for its receptive field. Nonetheless, this term is commonly used to describe the region of an input that impacts a specific network output. Therefore, we can say that the receptive field of a CNN is a measure of its temporary learning capacity that increases linearly with the number of layers and the size of a convolution kernel used in a CNN. In CNNs, it is evident that the receptive field size can be increased in the following ways: (i) stacking more layers (increasing the depth of the network), (ii) sub-sampling (introducing pooling after convolutions, having a lower stride), and (iii) increasing kernel size (dilating the

convolutional kernel). Although the expansion of the receptive field significantly increases the number of parameters, it is beneficial in capturing global and local dependencies which are crucial for building a DSR system.

The goal of this paper is to explore the efficiency of DSR systems built using hybrid DNN-HMM and large receptive field networks for acoustic models. We perform a thorough analysis on the design of these networks and on the relationship between receptive field size and the number of parameters of the networks.

CHAPTER 2

BACKGROUND

2.1 Approaches in Capturing Long-term Dynamics

In this section, we discuss the past and present research work relevant to capturing long-term dependencies in speech. There are two distinct approaches to address long-term dependencies in a speech signal: (i) in feature extraction component by modifying this component in a way that the output features are dependent on a large portion of input signal, or (ii) using acoustic models that can learn the long-term dependencies given a short-term speech features Peddinti, Povey, and Khudanpur 2015a.

2.2 Feature Extraction Approach

Initial efforts from researchers in speech and audio processing were exclusive to explore feature extraction strategies to address this issue. For instance, (i) TRAPs, a feature extraction technique which replaced standard spectral patterns with long-term temporal patterns of spectral energies Hermansky and Sharma 1999, (ii) A wavelet-based multi-scale spectro-temporal feature extraction technique which consider multiple time and spectral resolutions tuned to capture fast and slow changes in modulation patterns Mesgarani, Shamma, and Slaney 2004, and more recently (iii) Features from deep scattering spectrum, which extend standard MFCCs by calculating multiple-orders of modulation spectrum coefficients with the use of wavelet cascades Andén and Mallat 2014; Yousefi, Khorram, and Hansen 2019. These long-term speech dynamics capturing features showed reasonable performance improvements when tailored to a specific task (or) speech from a particular acoustic environment. These feature extraction techniques can not be generalized for all acoustic conditions because it needs the expertise to tune parameters in the extraction process to compensate for the distortions induced by an acoustic condition on speech which are inconsistent and change

swiftly. It was therefore found that the best approach to address long-term speech dynamics capturing problem may be to seek for alternative strategies for acoustic modeling rather than the feature extraction. Later, acoustic modeling strategies were researched in great detail to deal with this problem.

2.3 Acoustic Model Approach

With advances in machine learning, FC-DNNs learning strategies were adapted to build robust state-of-the-art acoustic models that can statistically map an acoustic sound precisely to its corresponding transcript. Although FC-DNNs have shown significant improvements over GMM-HMM-based acoustic modeling, their temporal modeling capabilities were limited as they operate on the information from a fixed-size sliding window of acoustic frames. This made them unsuitable for handling long-term dependencies. Subsequently, recurrent neural networks (RNNs), a progression to FC-DNNs with cyclic connections over time, were able to collect and store information for an arbitrary number of neighboring acoustic frames, showing their capacity to capture long-term dependencies Greff et al. 2017. Several RNN architectures have since been explored for acoustic modeling (e.g., GRUs Wu and King 2016, LSTMs Graves, Mohamed, and G. Hinton 2013; Ghorbani, Bulut, and Hansen 2018, BLSTMs Graves, Jaitly, and Mohamed 2013, RNMs Baskar et al. 2017). Training RNN is usually performed through a time-expansion operation where the input at time '$t + 1$' relies on the output at time 't'. Due to this time-expansion operation, parallelization of training routines for these networks becomes quite challenging even with techniques such as sequence batching and distributed optimization. In addition to the challenges in training phase, these architectures generally are difficult to run in inference mode too because of the state (memory) they need to build from the beginning of the input; yielding to a larger memory requirements on device.

2.3.1 Long-term Dynamics in End2End

End to End systems have recently gained a lot of attention especially in the academic community. These systems are essentially built of a consecutive deep layers which are designed to transfer the acoustic features directly to words/characters sequences. There are a huge variety of different architectures like RNN-based Encoder-Decoder Chorowski et al. 2014 , to Transformers Dong, S. Xu, and B. Xu 2018 and also different training criteria like CTC S. Kim, Hori, and Watanabe 2017 and many other examples. Figure 2.1 demonstrates structure

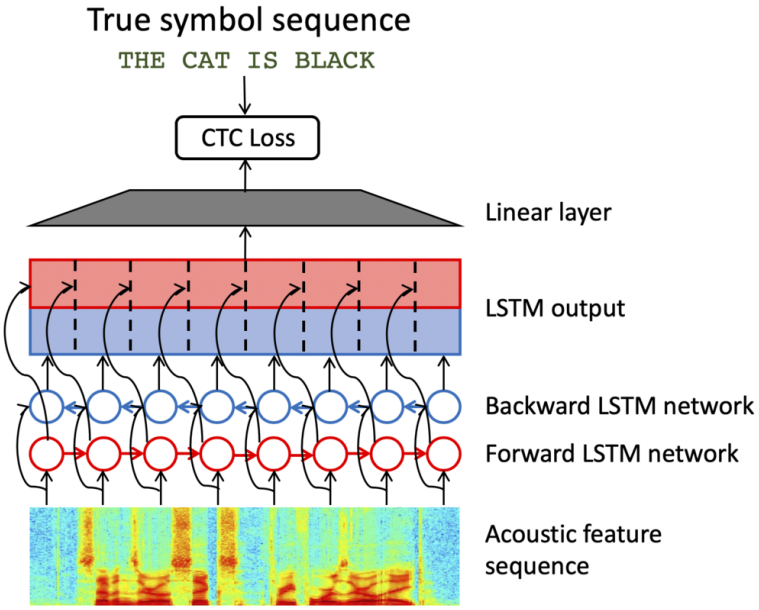

Figure 2.1. End to End system architecture. This is a simplified version of a RNN-based e2e speech recognition system where the features are directly mapped to the desired text through a couple of consecutive recurrent neural networks. CTC is a loss function that yields to a comparingly accurate transcription with the cost of high computation complexity.

of a simple RNN-based ASR model trianed with CTC loss. This system leverages a couple of forward and backward RNNs (LSTM in particular) which take the acoustic features and output the final words/characters. In these architectures (and also Encoder-Decoders) the temporal dependencies are being captured and modeled in the recurrent units. The recurrent layers take acoustic inputs and through a set of weightings, update their hidden states and go

on to the next time step. Theoretically these models are capable of capturing any length of temporal dynamics in input signal. This modeling capacity makes these models perform even better sometimes compared to their older alternatives. At the same time more robustness toward different errors especially in reverberation. One of the key aspects of RNN-based systems yielding this performance is the maximum possible receptive field they have, which in fact is the total length of utterance. It means that these models are consuming the hidden state to output every prediction label. Despite of their performance, the drawbacks that these models are suffering from, are the lack of scalability, their huge dependency on the GPU-based calculations and difficulty of online decoding. These factors are resistances in leveraging them in industry as dominant as their alternative HMM-based models.

2.3.2 Long-term Dynamics in HMM-based

End to end systems have a well-established and more industrially dominant alternatives: HMM-based systems. These systems approach to the automatic speech recognition with a less data driven way, meaning different components are trained separately to do specific task, then combine them to make a final decoding graph. In the heart of their decode, there is an Hidden Markov Model (HMM) assumed and trained which are in charge of modeling acoustic behaviour of each uttered phoneme.

While decoding the acoustic model will be trained in a way to predict the states of this HMM given the input features. Early versions of these systems were GMM-HMM where a Guasian Mixture Model (GMM) (a generative model) was trained with a separate GMM for each phoneme in the vocabulary. As introduction of deep learning the researcher started to change the decoding problem to a discriminate version in order to use the modeling capabilities of Deep Neural Networks, resulting in what is now known as DNN-HMM systems 2.2. Historically, these systems proposed by a Convolutional Neural Networks (CNN) as the DNN component due to their compatibility to the problem. But the mentioned modeling

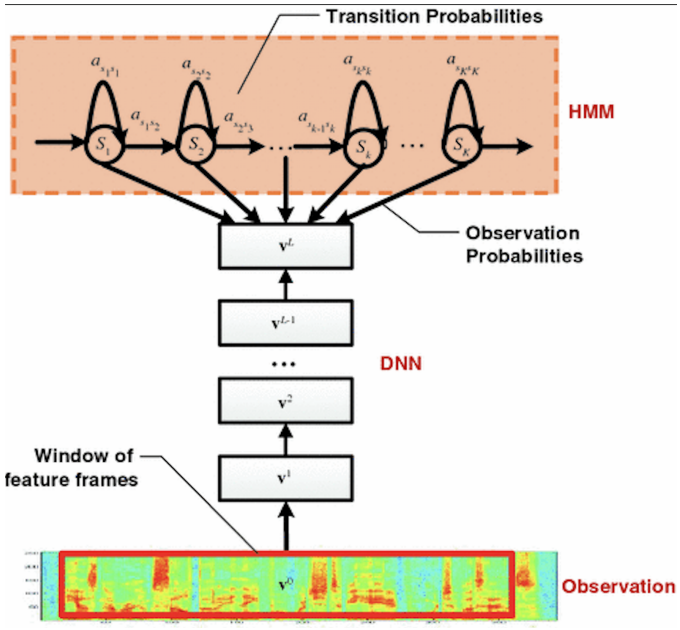

Figure 2.2. HMM-based automatic speech recognition

capabilities enhancement came with a cost; these models by structure were limited to a certain observation length of the input to generate the every output instance. This length are named receptive field and is going to be introduced in more details in the methods but it must be mentioned the larger this receptive field gets, the more context our classifier has. Gradually scientists realized that one of the key components in the performance and robustness of these systems is the receptive field and tried to enhance it in different ways.

2.4 Attempts to Capture Long-term Dynamics in CNN-based Acoustic Models

Convolutional neural networks are one of many other machine learning strategies adapted for acoustic modeling to handle the long-term dependencies in ASR. Similar to RNNs, CNNs have also shown significant improvements in ASR performance over FC-DNNs Hau and K. Chen 2011; Abdel-Hamid, L. Deng, and D. Yu 2013; Ghorbani, Khorram, and Hansen 2019; Ghorbani and Hansen 2018. Recent research also shows that the use of residual connections can train deeper CNN architectures in a more efficient way compared to RNNs He et al. 2016.

Thus, deep CNNs with restricted local connectivity and weight sharing were successfully used in document recognition LeCun et al. 1998. Researchers have studied various variants of CNNs that use the concept of large receptive field to build robust systems in the areas of human pose recognition Newell, K. Yang, and J. Deng 2016, face expression recognition J. Yang, Liu, and K. Zhang 2017, human speech emotion recognition Khorram, Aldeneh, et al. 2017; Khorram, McInnis, and Provost 2019, signature verification Al-Jarrah and Arafat 2014 and also in many machine learning applications associated with super-resolution image processing. Tiled-CNNs that learns rotational and scale-invariant features over time has proven to perform better than traditional CNN for small time-series data Ngiam et al. 2010. CNNs have also been used for speech dereverberation applications in multiple configurations and have successfully demonstrated their ability to learn the long-term effects of reverberation on speech Ernst et al. 2018; Yousefi, Shokouhi, and Hansen 2018. In addition, Dilated CNNs have also proven their abilities to learn relevant information from a bigger context F. Yu and Koltun 2015a. Therefore, we focus on studying the large-receptive field networks for acoustic modeling, especially for distant speech recognition.

CHAPTER 3

METHODS

In this section, we discuss the working principles of standard CNN, dilated CNNs, and stacked hourglass network. We compute and compare the receptive field size of the mentioned networks to better understand the increase/decrease in the performance of each network.

3.1 Standard CNN

As mentioned in the previous sections, CNNs can be considered as a variation of regular feed-forward networks. In CNNs, *weight sharing* is normally achieved by sliding a linear filter throughout the output of the previous layer. In these architectures, every layer's output is generated by sliding a convolving window over the output of previous layer. Normally in many signal processing tasks, between every two convolution layer we put a pooling layer as well, which basically reduces the resolution of the output of layers before delivering to the next layer. Although in some cases, including acoustic modeling of HMM-based ASR, we need the output resolution to be equal to the input resolution.
Assuming each convolutional layer uses a linear filter of kernel width 'W', we can compute the receptive field of a CNN network with 'L' layers as follows:

$$RF_{standard} = L(W - 1) + 1 \tag{3.1}$$

where RF_{CNN} is the RF size of the standard CNN. This equation is showing how much of the input layer is involved in generating one prediction of the final output. It is evident from this equation that the RF size increases linearly with respect to both W and L. As a quick forward to Table 4.2 in standard 10 layers of CNN with each window length of 8 yields a receptive field of 71 on the output. Although, as we are going to see, in some architectures it is not this straightforward to calculate the receptive fields and we can only track the growth of it with respect to different parameters of architecture.

As the RF size increases, the number of learning parameters also increases linearly, making the network not effective for tasks where large receptive fields are required. This is while we always prefer smaller networks over larger ones for different reasons including generalization, training/inference speed and efficiency. The linear growth of output receptive field to the number of parameters (W and L) makes standard CNNs inefficient for our purpose.

Also, due to the linear relationship between RF size and network complexity, it is difficult to find an optimal point in these architectures. This is our main motivation to investigate further architectures to find a better trade-off of RF and network complexity.

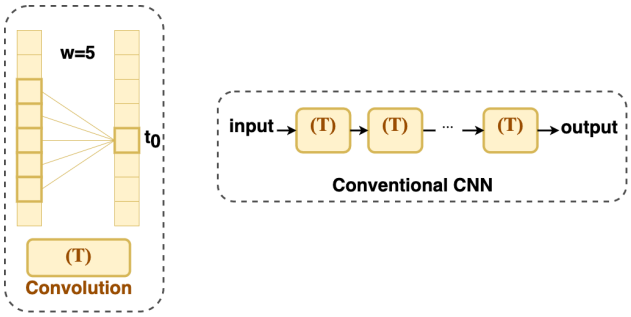

Figure 3.1. Convolution architecture. The figure layer illustrates one layer of standard convolution kernel, while the right one regards to the whole network made of these convolution layers.

3.2 Dilated Networks (DIL-Net)

Networks that use dilated convolutions have shown to be effective in many tasks, including image segmentation F. Yu and Koltun 2015b, speech synthesis Oord et al. 2016 and ASR Sercu and V. Goel 2016. Dilated networks, provide an effective technique for increasing the RF size without causing a significant rise in the number of learning parameters. In a dilated network, the convolutional filter (kernel) is obtained by inserting (fix) zeros between the regular filter samples. This method expands the filter in time at the expense of lower resolution; making the filter sparse when compared to a standard CNN convolutional filter. Inserting zeros

and sparse filters all basically mean skipping some values of the input in each layer. This technique somehow resembles the pooling technique with the difference that in this version the input resolution conserves until the final output.

Figure 3.2 (top) shows an example of the dilated convolution filter with the dilation factor of $d = 3$. A dilated convolutional filter is simply obtained by inserting $(d - 1)$ zeros symmetrically between successive filter coefficients. With this definition, a dilated network is generally constructed by stacking N dilated convolutional layers with a 2^n (for n'th layer) dilation factor for each layer. Considering the pooling metaphor for dilation, it is like increasing the pooling window and stride both exponentially. We usually put a couple of standard convolution layers as *preprocessing subnet* at the beginning. This *preprocessing subnet* is a feature processing block with the highest resolution as shown in Figure 3.2(middle and bottom). The receptive field of this dilated network can be computed as follows:

$$RF_{dilated} = (L + (2^{L-1} - 1))(W - 1) + 1 \tag{3.2}$$

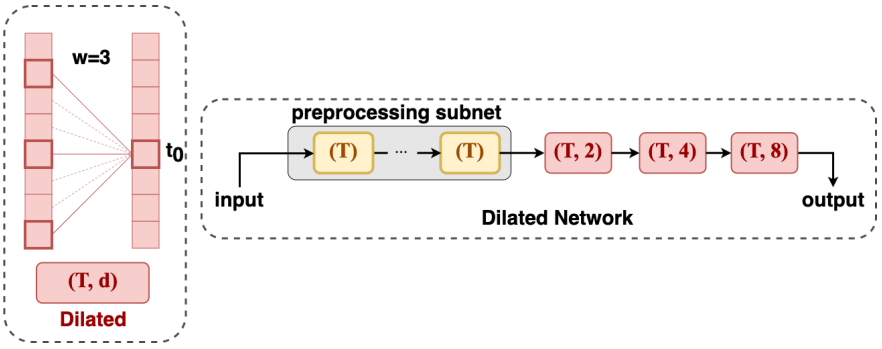

Figure 3.2. Dilated architecture. The left figure illustrates one layer of dilated convolution kernel. In this special form of convolution some values are being skipped and number of these skipped values are determined by the dilation value. The right figure regards to the whole network made of these dilated convolution layers.

L is the number of layers and W is the width of the convolutional layers. The RF size grows exponentially, while the number of parameters grows linear to the number of layers.

For instance, the dilated network shown in 4.2 with 7 layers of dilated layers and window width of 5 and dilation of 4 yields to a receptive field of 421 which is considerably more than even deeper standard CNN network (71). This is a big advantage in terms of receptive field to model complexity. This means by adding one layer we achieve a much bigger receptive field growth in comparison of linearity of normal convolution. Of course this is done with the cost of losing high-resolution information especially as we move to the deeper layers, but as we will see in the results, this is a better trade off for acoustic modeling task.

3.2.1 Time Delayed Neural Network (TDNN)

A variant of dilated networks is achieved by inserting zeros asymmetrically between successive filter coefficients Peddinti, Povey, and Khudanpur 2015b. This network is commonly known as time-delay neural network (TDNN). Figure 3.3(c) shows a single layer of TDNN with asymmetric dilations. Each layer in a TDNN can have different dilation values $d_{l,1}$ and $d_{l,2}$. The asymmetric dilation characteristic of TDNN makes it more flexible and gives the network a better learning capacity compared to dilated networks. On the contrary, $(d_{l,1}, d_{l,2})$ hyper-parameters are extremely data-dependent and can only be tuned by empirical studies to optimize the efficiency of the networks. The RF size of a TDNN can be computed as follows:

$$RF_{tdnn} = 1 + \sum_{l=1}^{L} (d_{l,1} + d_{l,2}) \tag{3.3}$$

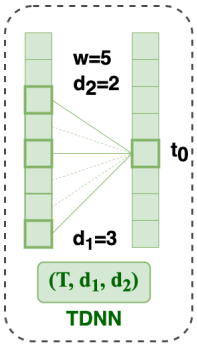

Figure 3.3. Time-delayed architecture. The figure layer illustrates one layer of time-delayed convolution kernel. In this type of convolution, we skip some values in the input layer and only pass certain values. In this type of convolution the number of skipping values could be asymmetric between passing ones.

3.3 Stacked Hourglass Network (HG-Net)

Stacked hourglass structure (HG-Net) was initially designed to solve facial landmark localization J. Yang, Liu, and K. Zhang 2017 and human pose estimation Newell, K. Yang, and J. Deng 2016 which required parallel process of both high-resolution (local view) and low-resolution (global view) versions of an image in Oliva and Torralba 2006. This property is equivalent to processing short-term and long-term temporal dynamics of the speech signal.

HG-Net is built using a stack of hourglass networks to processes both short-term and long-term temporal dependencies in parallel, see Figure 3.4(bottom). As shown in Figure 3.4(center), each hourglass unit in an HG-net contains 'L'-layers with two sub-networks: (1) a down-sampling network; (2) an up-sampling network in each layer. Figure 3.4(top) shows the convolutions involved in the down/up sampling networks. The down-sampling network generates low-resolution representations of the input, and the up-sampling network converts the representations learned from low-resolution to high-resolution signals.

The down-sampling network consists of a series of convolutions and max-pooling layers. The max-pooling layer reduces the resolution of the signal and increases the RF of the network. Various pooling operations can be used instead of max-pooling. The up-sampling network consists of a series of up-pooling and convolutional layers. This network combines

all the representations learned from different resolutions of input. In addition, the hourglass network exploits a specific skip connection mechanism that connects representations that can allow us to leverage many layers for down-sampling and up-sampling networks without having the vanishing gradient problem. Therefore, we can down-sample the input signal to a low resolution to achieve a large RF.

Assuming W_d, P_d, L_d to be the filter size of convolutions, pooling and number of layers in a down-sampling network[1], the RF size of a down-sampling network can be computed as:

$$RF_{down} = L_d(W_d + P_d - 1) - 1 \tag{3.4}$$

RF size of the stacked hourglass network (HG-Net), $RF_{stacked-hg}$, can be approximately calculated as:

$$RF_{stacked-hg} \approx S \times (RF_{down} * 2^L) \tag{3.5}$$

where S, L denotes the number of hourglass units in an HG-Net and number of layers in each hourglass unit. This shows that $RF_{stacked-hg}$ exponentially increases with L. RF size can be efficiently increased by using more layers in the down-sampling and up-sampling networks as well.

[1]The number of layers in the down-sampling and up-sampling networks must be equal in the hourglass network

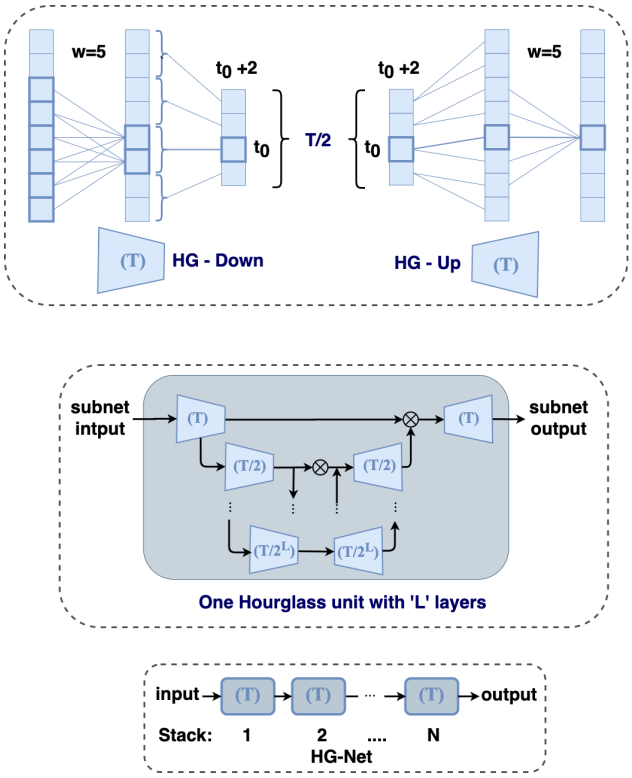

Figure 3.4. Hourglass architecture. The top figure illustrates the Down and Up components that are made with convolution layer and down-sampling or up-sampling layers. The figure in the middle shows an hourglass structure made of combining these up and down components in parallel. the lower figure is the final stack of hourglass network architecture which is constructed by a stack of hourglass components.

3.4 Deeply Recursive Network (REC-Net)

Deeply recursive neural network (REC-Net) is first proposed by Kim et al. as an image super-resolution method J. Kim, Kwon Lee, and Mu Lee 2016. The idea is to use a big network with a large number of layers and allow different layers to share their learnable parameters. REC-Net is a stack of recursive subnetworks, as it is shown in 3.5 (left) . Each recursive network contains a series of convolutional layers that all of them share the same weights. In the recursive subnetwork, increasing the number of layers will increase the RF size without increasing the number of parameters. REC-Net can provide a large RF with a small number of parameters.

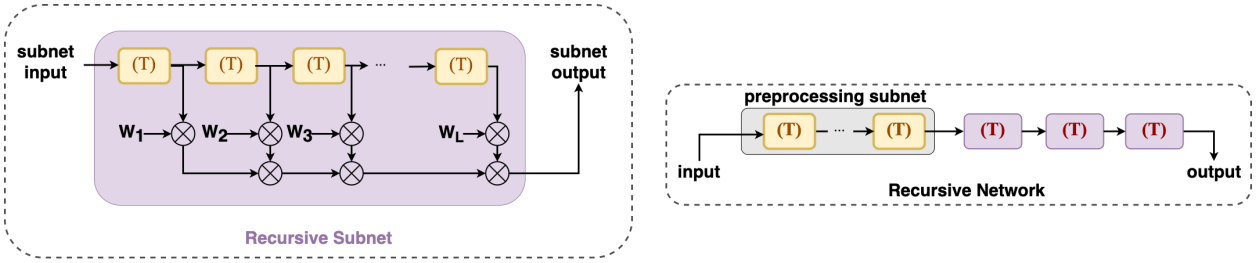

Figure 3.5. Hourglass architecture. The top figure illustrates the Down and Up components that are made with convolution layer and down-sampling or up-sampling layers. The figure in the middle shows an hourglass structure made of combining these up and down components in parallel. the lower figure is the final stack of hourglass network architecture which is constructed by a stack of hourglass components.

REC-Net has a number of problems: (1) to capture a large RF, we must use a large stack of identical layers in the recursive subnetwork. Training this structure is difficult and may lead to a vanishing/exploding gradient problem. To solve this problem, authors in J. Kim, Kwon Lee, and Mu Lee 2016 proposed a skip-connection strategy shown in Figure 3.5 where the output of the recursive subnetwork is obtained through a weighted average of the output of all layers in the recursive subnetwork; (2) training REC-Net is computationally expensive (in both time and memory requirements) since this network requires a large number of identical layers to capture long-term dependencies.

Unlike the conventional network, all these big receptive field networks provide an efficient way to increase the size of the receptive field without causing a significant rise in the number of learning parameters. Thus, we compare the efficiency of these networks with the conventional network by setting the number of learning parameters to be the same across all networks.

CHAPTER 4

EXPERIMENTS

4.1 Simulated Reverberation

Reverberation in distant speech recordings can be simulated by convolution of the audio signals with a room impulse response from a point source to a receiver location in a room. These RIRs are in charge of modeling acoustic behaviour of the room. The environment's shape, the microphone's position and the reflection ratio of different sound frequencies, defined by the textures and so on are all modelled in the final RIR. One common method of synthesising natural audio is by applying the RIR over the source clean sound. RIRs are highly sensitive to changes in receiver position, speaker position or positions of different obstacles in the room Kuttruff 2016.

Assuming the RIRs do not change over a small instances of time corresponding to a particular source and receiver positions, We use a set of 325 real RIRs composed of three databases: the RWCP sound scene database Nakamura et al. 2000, the REVERB challenge database Kinoshita et al. 2013 and the Aachen impulse response database Jeub, Schafer, and Vary 2009 and clean speech signals from WSJ corpus to simulate the distant speech recordings. All these RIR datasets are collected from real environments.

We have reverberated the clean data in order to measure our proposed methods' robustness against this particular types of error. While standard CNNs trained to have a limited receptive field in comparison to large receptive field networks we expect modified architectures to be more robust. A model with a larger receptive field is trained observing a larger portion of input so it could be more robust toward the repetitions of the same signal.

4.1.1 Analysis

In this section, we study perceptual and objective speech quality measures such as signal-to-noise (SNR), perceptual evaluation speech quality (PESQ), Itakura-Saito (IS) and cepstral

distance (CD) that can quantify the degradation in the speech caused due to the reverberation.

Cepstral Distance (CD)

The Cepstral Distance is a measure of the log-spectral based distance between a clean speech sample and a degraded/test speech sample. The Cepstrum is a time-domain version of the log-spectrum of a speech sample. The process of inverting the logarithmic fourier transform of the speech signal segregates the speech excitation source and the spectral transformation experienced during speech production from the vocal tract and lip radiation effects. The Cepstrum is computed for the clean and degraded speech signals using the Levison-Durbin recursion and the cepstral distance between the two signals is calculated as follows and normalized to limit the output range:

Itakura-Saito (IS)

As noted earlier, LPC is a well known all-pole filter based speech model and can be used to model a given speech signal as shown in Eq-4.1, where p is the order of the all-pole filter, $a_x(k)$ are the LPC coefficients and G_x is a filter gain used for the excitation signal. The Itakura Saito measure is a speech metric which is similar to LLR but also includes the gain differences (G_c & G_d) between all-pole models of the clean and degraded speech signals. Itakura-Saito is defined by Eq-4.2,

$$x(n) = \sum_{k=1}^{p} a_x(k)x(n-k) + G_x e(n) \tag{4.1}$$

$$d_{IS}(a_c, a_d) = \frac{G_c}{G_d}\left(\frac{a_d^T R_a a_d}{a_c^T R_a a_c}\right) + \log\frac{G_d}{G_c} - 1. \tag{4.2}$$

Here $e(n)$ is the modeling error residual used as the excitation signal, which is usually considered to be modeled as a zero-mean and unit variance white noise.

Perceptual Evaluation of Speech Quality (PESQ)

While there are many objective speech quality measures, perceptual evaluation of speech quality (PESQ) has most recently been used extensively by researchers in speech processing to validate speech enhancement algorithms versus other LPC-based measures Barnwell III 1979. PESQ compares a clean speech signal with either a degraded or an enhanced version of an input degraded signal processed by speech enhancement and attempts to predict the perceived quality that would be assigned by humans in a subjective listening testRix et al. 2001. PESQ creates numerous delayed versions of the clean speech and processes the transformation of the clean signals to test signals which is analogous to the psychophysical visualization of speech signals by humans. PESQ takes into account time-alignment, level alignment, loudness scaling, and other factors to suppress minor time, amplitude or frequency changes between the clean and degraded/enhanced signals. However, drastic or abrupt changes in degraded/enhanced signals with respect to time-delayed versions of the clean signal are shown to reduce the output PESQ quality score. Thus, a higher PESQ score represents greater speech quality, with a range in values from -0.5 to 4.5.

Although SNR measure is an objective measure mostly designed to measure level of additive noise, we can still leverage it in our case. Cepstrum-based comparisons are equivalent to comparisons of the smoothed log spectra of the signals. Table 4.1 shows the simulated distant speech signals generated using real recordings of RIRs in various acoustic environments are heavily distorted w.r.t clean speech signals from WSJ corpus. This is not the case for the IS and PESQ since they are designed as a more subjective metrics. We calculated these measures to have an intuition toward how distorted makes the signal each environment (room). We also observe the SNR value does not efficiently represent the distortion added by reverberation. Among these measures CD is more efficiently modeling the reverberation. It also fits to our subjective understanding of the environments, for instance we expect the

Table 4.1. Objective Quality Measures for simulated distant speech signals w.r.t. clean speech signals from WSJ corpus.

Data	**SNR(dB)**	**PESQ**	**IS**	**CD**
Lecture Hall	-3.18	1.52	8.11	6.22
Office Room	-3.09	1.62	3.5	5.3
Meeting Room	-2.58	2.33	7.98	4.84
Stairway	-2.69	1.87	12.06	5.59
Average	**-2.88**	**1.83**	**7.912**	**5.48**

recording would be highly distorted in the stairway comparing to meeting room. While CD discriminates these two environments, their SNR values are way close to eachother.

4.2 Realistic Reverberation

The UTD-Distant Reverb corpus consists of two environments; (i) a highly reverberant space (Racquetball court), and (ii) classroom. For this study, we consider only environment-#2 (Classroom). We collected this portion of the corpus in a mildly reverberant space with an average reverberation time (RT60) of approximately 400 ms. The Figure 4.1 demonstrates recording setup for Racquetball environment which is identical to classroom setup. The corpus includes a total of 6 hours of recordings from three different native-English speakers, two male, and one female. In this work we only used one male and one female speakers. All speakers had no prior history of speech or hearing limitations. The recordings were made using a combination of (i) close-talk microphone (CTM) - the subject wears a headset, (ii) single distant microphones on a stand placed at 1m, 3m, and 6m away in tandem from the subject, and (iii) a four-microphone linear array placed at 6m away from the subject 4.2. Each subject/speaker also wore a portable naturalistic audio capturing device (LENA) which is capable of recording and storing up to 16 hours of recordings.

These LENA based wearable audio acquisition devices help us capture and analyze naturalistic variabilities in the audio stream from a speech/signal processing standpoint. This corpus also incorporates recordings from four volunteers wearing LENA devices standing stationary at fixed locations from the subject. These volunteers at designated times move collectively either clockwise/counter-clockwise around the subject allowing the LENA devices to capture distant-speech based on their location. UBI-sense, placed at four corners of the reverberant space tracks the volunteers, equipped with RFID tags, in the reverberant space. The speech in all microphones was captured at a 44.1kHz sampling rate using an 8-channel multi-channel TASCAM system, to ensure all the audio is synchronized. It is noted that the recordings from wearable devices are time-aligned afterward, to all other stationary microphones.

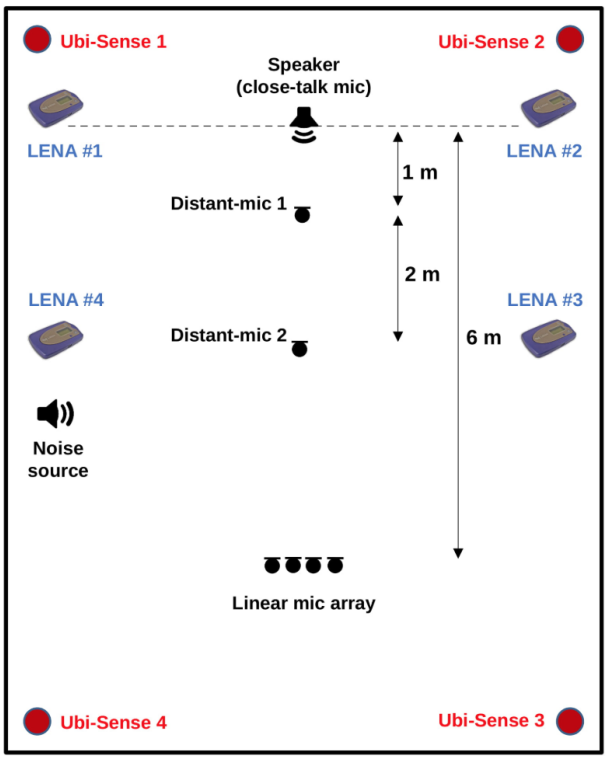

Figure 4.1. UTD-Distant Reverb Data Collection - Room Setup Diagram (Classroom)

The UTD-Distant Reverb corpus can be broadly classified into four major categories: (i) clean prompted (IEEE sentences) & spontaneous speech with all volunteers stationary at their locations, (ii) clean prompted & spontaneous speech with volunteers moving around the target/speaker in the reverberant space, (iii) noisy prompted & spontaneous speech with all volunteers stationary at their locations, and (iv) noisy prompted & spontaneous speech with volunteers moving around the speaker. In this work we are focused on the first category which is the clean prompted utterance of IEEE sentences. All speech samples used in our experiment are naturalistic recordings with reverb/noise with no simulated signals. Therefore, in our experiments, since the close-talk microphone contains limited reverberation effects, an additional speech dereverberation algorithm **dereverb** is used to suppress any remaining early reflections. This enhanced version of the close-talk recordings is used as the clean reference speech to validate/test against degraded speech signals. A robust speech activity detector, combo-SAD **ComboSAD** is applied to the close-talk microphone signal in order to

Figure 4.2. UTD-Distant Reverb Data Collection - Microphone and Speaker arrangement (Classroom)

distinguish speech frames from silent/reverberant frames. Each of two speakers speak 712 utterances consisting 1 hour of speech data.

Figures 4.3 and 4.4 demonstrate spectrogram of the same utterance over different microphones. We can see how speech frequencies are preserved even in distanced microphones of the classroom setup while the tend to get much fuzzier in the Racquetball court environment. As the distance of the microphone from the speaker increases, we see more degrade in the discriminative frequencies of the audio. The spectrogram of 6m microphone shows a better frequency lines because of the higher power microphone it is (the array mic) but because of it's closer to noise source as we will see we have lowest WER performance over it. We can

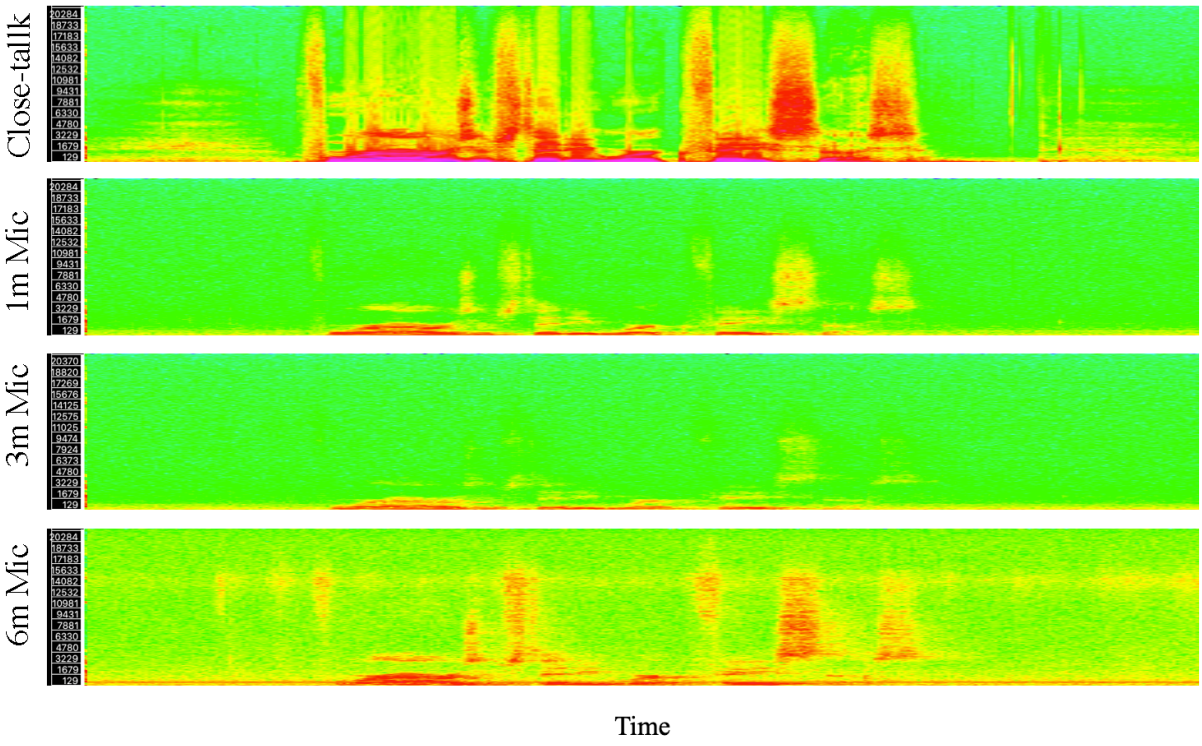

Figure 4.3. Spectrogram of close-talk and distanced microphones recordings of an utterance in Classroom

also clearly see the reverberation effect in Racquetball as horizontal stretch of the frequencies over the spectrogram.

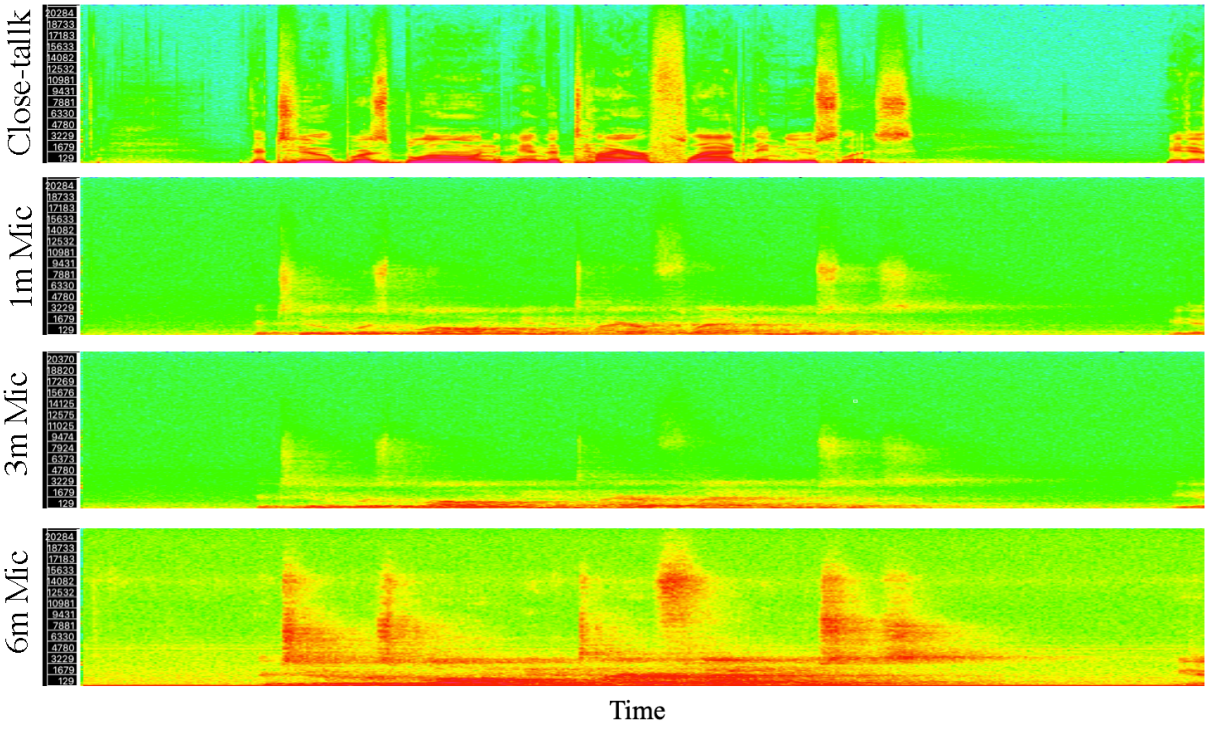

Figure 4.4. Spectrogram of close-talk and distanced microphones recordings of an utterance in Racquetball Court

4.3 ASR Engine

We used Wall Street Journal (WSJ) dataset to evaluate the performance of the large RF convolutional networks explained in the previous section. The training data consists of 80 hours of speech both telephone and microphone speech, the bulk of which is in English. All wideband audio is downsampled to 8kHz. The evaluation is performed on the Eval93 subset of the WSJ. The Dev93 subset of the WSJ is used to tune the parameters across all networks. We used 40-dimensional Mel-filterbank (MFB) features normalized with Cepstral Mean and Variance Normalization (CMVN) as the input features of the networks Khorram, Jaiswal, et al. 2018; B. Zhang, Khorram, and Provost 2019. We also implemented Feature space Maximum Likelihood Linear Regression (FMLLR) transformation in our initial experiments, but it did not yield performance improvements. Since the main focus of this paper is on the effect of large RF covering, we did not explore the effect of speaker normalization methods

(e.g., i-vectors) in our experiments. We trained our models up to 20 epochs using the Adam optimizer ($\alpha = 0.001$). Our initial experiments showed that ReLu activation function outperforms other activations and therefore we applied ReLu in the intermediate layers of the networks. We employed softmax for the output layer. We also implemented a discriminative softmax (AMSoftmax) Wang et al. 2018 that did not improve the results.

We trained a triphone model with 3392 states in four iterations and used it as the HMM component of the DNN-HMM pipeline ASR. No language model refinement was applied in the decoding phase. We used Kaldi Povey et al. 2011 implementation of HMM and we implemented all the networks using the TensorFlow Abadi et al. 2016 open-source library. We performed hyper-parameter tuning by leveraging two well-known measures: frame accuracy (Acc) and cross-entropy (CE). In addition to these measures, we also report word error rate (WER) of all networks.

For the standard CNN, we evaluated all the networks with the kernel size of $w = 3$ and 5, and the number of layers ranging from $L = 3$ to 10. $W = 5$ and $L = 10$ performed the best in both validation accuracy and WER. As we used raw MFB features, we considered a stack of standard convolutional layers (with 3 layers) as the *preprocessing sub-network* in DIL-Net and REC-Net (Figure 1(g), (h)). We implemented DIL-Net as shown in Figure 1(e). Our DIL-Net contained 3 and 4 dilation layers, with the dilation factor ranging exponentially from 2 to 8 (i.e., $d = (2, 4, 8)$). Skip connections were applied to this structure, but they did not lead to better performance. For REC-Net, we used 5 layers of inner convolutions and 5 recursive sub-networks. For HG-Net, we validated for the number of stacks $S = 1$ to 5, convolutional kernel size $W = 3$ and 5 and number of layers $L = 3$ and 5. Parameters of $S = 5, W = 5$, and $L = 3$ achieved the best performance in terms of WER. For consistency of comparisons, we used the same number of kernels (512 kernels) for all the convolutional layers.

4.4 Results

After understanding of each proposed architecture and ASR experiment setup, in this section we are going to present the results and discuss what attributes were effecting.

4.4.1 Empirical Experiment

Next, we run an elementary empirical experiment using a standard CNN with one layer of convolution to comprehend "how long?" is actually long enough to capture the long-term dynamics in distant simulated speech signals, We train this single layer standard CNN for various receptive field sizes[1] using the simulated speech signals, see Figure 4.5. It is evident from this experiment that the accuracy increases with an increase in RF size. However, having a greater RF size than required neither hurts nor improves the system's performance. Thus, for our experiments, we fix the number of parameters across all the networks based on the optimal RF size determined from this experiment.

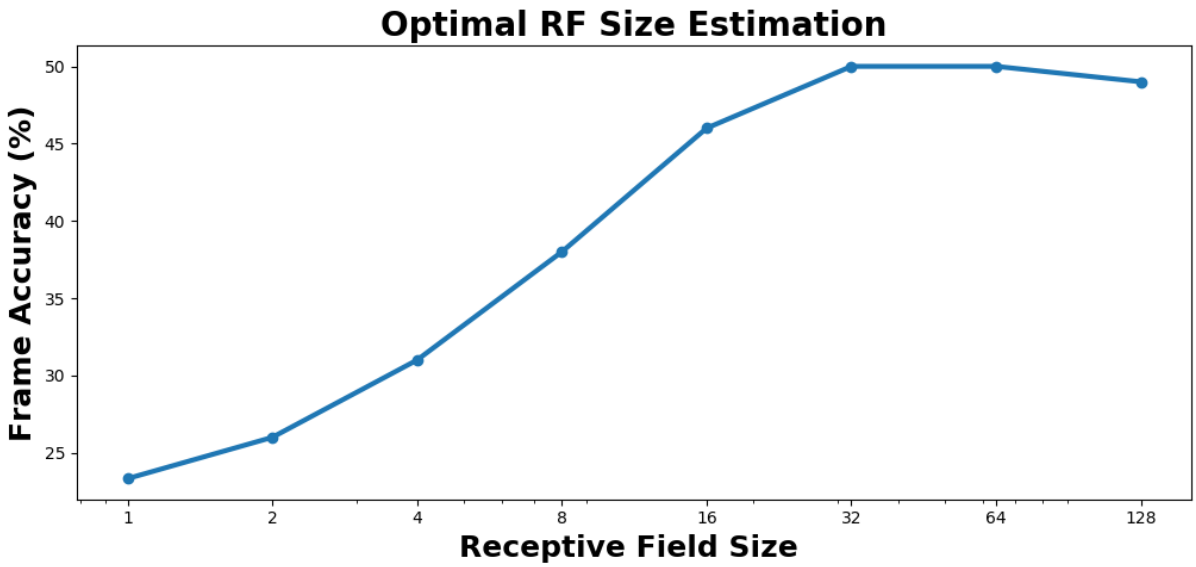

Figure 4.5. Optimal Kernel Size for capturing long-term dynamics in simulated distant speech

[1]For a single layer standard CNN, kernel size will be the same as the receptive field size

4.4.2 Frame Accuracy Performance

Furthermore, for better understanding of LRF networks, we test the frame accuracies obtained by all the networks on Dev93 for various architectures, see Table 4.2. It shows the performance of all LRF networks. We observe a linearly growth trend in standard CNN's efficiency (in terms of validation frame accuracy) with increased kernel size and number of layers, in other words, RF size. There is also one important observation that in standard CNN the accuracy of prediction is still increasing by training a bigger network i.e. the highest accuracy of network is achieved by the largest one. This indicates that to achieve the better accuracy we still need to increase our model parameters.

Table 4.2. Performance of Standard CNN and large receptive field networks for different configurations.

Network	**Architecture**	**Acc(%)**
Standard CNN	W:5 L:6	55.09
	W:5 L:8	57.92
	W:5 L:10	60.52
	W:6 L:10	60.47
	W:7 L:10	**61.80**
Dilated Net	W:5 d:2 L:5	56.59
	W:5 d:2 L:7	58.41
	W:5 D:4 L:7	**64.17**
	W:5 D:4 L:9	63.65
	W:5 D:8 L:7	61.42
Recursive Net (RIN/ROUT:3)	EMBD Layer:1	60.23
	EMBD Layer:2	**61.45**
	EMBD layer:3	60.55
Hourglass CNN	HG:1 W:5 L:3	63.21
	HG:3 W:3 L:3	65.98
	HG:3 W:3 L:5	67.25
	HG:3 W:5 L:5	**67.55**
	HG:5 W:3 L:5	67.48
	HG:5 W:5 L:5	67.01

Unlike the standard CNNs, the LRF networks showed optimal performance over all the variations tested in their architectures for a specific RF size. It can, therefore, be expressed that having a large receptive field customized to distortion levels in speech can enhance the efficiency of a system; LRF networks can achieve this at a reduced computational expense than standard CNNs. As we pointed out earlier, the later three architectures all achieve their best performances in the middle sized versions indicating that these models achieve receptive field saturation sooner than the other architectures.

4.4.3 WER Performance

We observe that all LRF networks have minor relative improvements in performance compared to the standard CNNs for clean speech signals. However, for distant speech signals, where the reverberation introduces smearing effects in both time and frequency, we see higher relative improvements using the LRF networks compared to a standard CNN for a fixed number of parameters in order to reduce the architectural complexity, see Table 4.3. The reason this architectures are performing better is all related their coverage of the input. We know that reverberation affect is a result of multiple repetition of the signal through out time. These repetitions are due to the reflections of the environment. In fact architectures with larger receptive field brings robustness by covering all of these variants of the actual signal into account by covering a larger portion. In these architectures, network has access to a wide added versions of the signal while predicting the output label. Even in the clean test data, large receptive fields naturally bring a more context in classifying the senone. This is the same idea of tying monophones constructing triphones in order to bring context to phone classification. Now with these modified networks this context is brought directly from the input.

This indicates the importance of capturing the long-term dynamics for distant speech recognition. Although the dilated networks have the best WER for clean speech, it can

Table 4.3. WER and frame accuracies of LRF networks for clean and simulated distant speech versions of Eval93 (with fixed number of parameters ≈ 25600).

Network	**Frame Acc. (%)**		**WER(%)**	
	Clean	**Reverb**	**Clean**	**Reverb**
Standard CNN	71.39	60.48	8.13	18.31
Dilated Net	74.16	64.17	**7.25**	17.52
Recursive Net	73.61	65.17	7.54	17.13
Hourglass Net	**75.43**	**67.00**	7.98	**16.68**

be argued that the architectures chosen in this comparison study were forced to have the same number of learning parameters instead of being the best in their respective category. Nonetheless, the best WER performance for simulated the distant speech was achieved by Hourglass network.

Next, we test our system in real recorded data of UTD Distance corpus. We trained our model with the WSJ train set without any simulated reverberation included so the method is totally unsupervised. Table 4.5 and 4.4 shows the mean of Word Error Rates (WER) across two speakers for racquetball court and classroom respectively. As we can see in Table 4.5 the reverberation level for the distance microphones in the court is so high (T60 ≈ 9000) that the ASR engine is not capable of transcribing anything. For the close-talk microphone though, we can see generally better performance on the LRF networks comapring to standard CNNs.

Table 4.4. WER of LRF networks for UTD-Distance - Classroom (with fixed number of parameters ≈ 25600).

Mic. Distance	**Standard CNN**	**Dilated**	**Recursive**	**Hourglass**
Close-talk	12.39	**11.92**	12.52	12.01
Mic 1m	20.75	18.63	18.12	**17.91**
Mic 3m	21.80	19.89	19.74	**19.08**
Mic 6m	40.98	38.31	**37.92**	39.59

In contrast to racquetball court, the reverberation level in the classroom is more reasonable (T60 ≈ 425) that makes the trade-off in alternative ASR network solutions possible. For close

Table 4.5. WER of LRF networks for UTD-Distance - Racquetball (with fixed number of parameters $\approx$ 25600).

Mic. Distance	**Standard CNN**	**Dilated**	**Recursive**	**Hourglass**
Close-talk	13.21	**12.99**	13.05	13.51
Mic 1m	110.04	99.63	120.02	115.63
Mic 3m	99.82	121.09	99.74	118.78
Mic 6m	99.82	121.09	99.74	118.78

talk microphones in both datasets, dilated network is performing the best. Although as we move away from the source of speech the Hourglass and Recursive networks are performing better as for microphone 1m and 3m, Hourglass network has 13.68 % and 12.90 % relative reduction in WER. While for the further 6 m microphone the best performance is achieved by Recursive network with 7.46 % relative reduction. we can also see that in all of these DNN-based ASR engines the WER generally is growing non-linear to the distance from the speaker. This trend suggests reprocessing dereverberation and feature engineering modules for applications with a distance more than a certain value more.

As we mentioned earlier we avoided most of the prepossessing and language modeling expansion in this section, in order to solely analyse the role that suggested LRF networks in reverberation environments.

Another observation from the results is the overall growth of WER for these architectures and in general for HMM-based engines 5. The growth rate to the WER to the distance from the speaker is not a linear relationship. This means for first few meters have the largest effect in the performance.

Table 4.6 shows the top 10 words with highest number of substitution errors. We selected the 1m microphone and hourglass network as distance microphone in this case. As we can see almost all of these words convey no important information and the main reason of their highest frequency in errors is their exponentially high number of appearances in the ground truth data. Although these words are easier to distinguish by the lexicon and language model

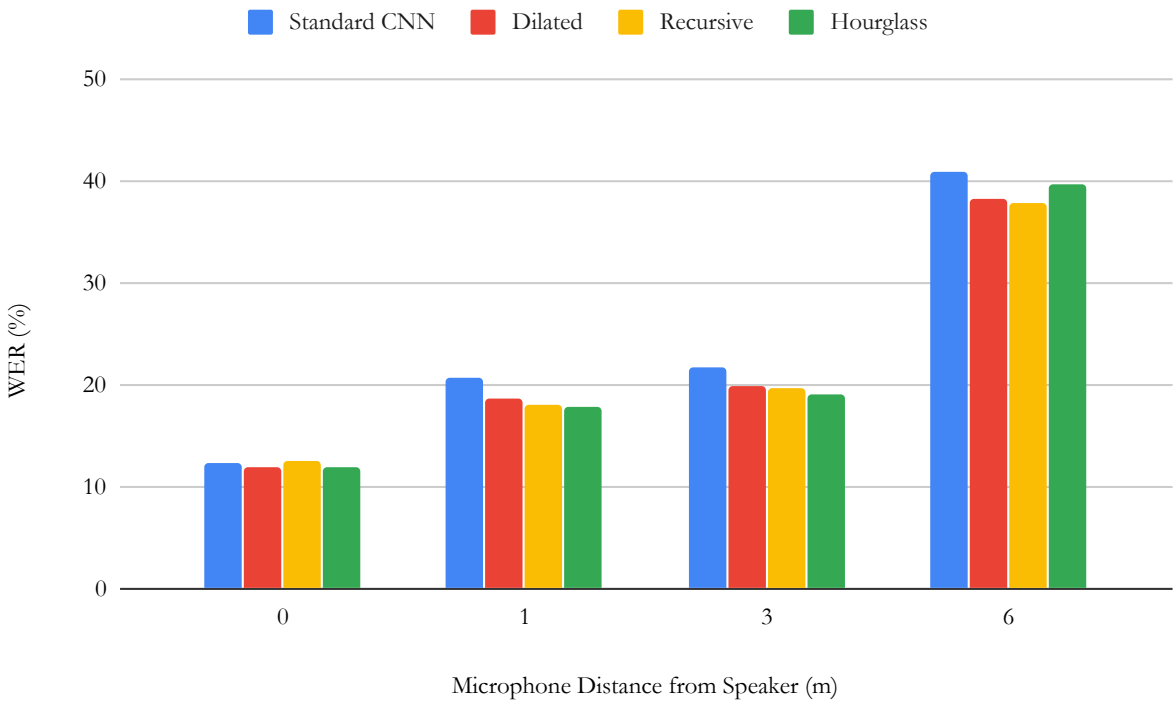

Figure 4.6. WER of Different Acoustic Models to the Distance from Speaker

they tend to mix up with the next words and in our evaluation metric it is considered as a substitution error.

Table 4.6. Words with the highest substitution errors. First and Second columns are distribution over close talk microphone and distance microphone and the last column is the top high-frequency words of Wikipedia

Close Talk Mic		**Distance Microphone**		**Wikipedia**
Word	**Errors #**	**Word**	**Errors #**	**Word**
was	10	and	9	the
and	10	is	8	be
with	9	inc	8	to
on	9	was	6	of
is	8	on	6	and
in	8	they	4	a
by	5	when	4	in
when	5	with	4	that
its	5	are	3	have
inc	5	s	3	I

CHAPTER 5

CONCLUSION

This work highlights the importance of capturing long-term temporal dependencies of the speech signal in distant speech recognition systems. We begin by understanding the importance of the receptive field and its role in convolutional neural networks. Although previously there were architectures proposed to capture these long-term dependencies there were many other variations that had not been investigated. In this work we proposed different solutions addressing capturing long-term dependency problem and used them as acoustic modeling component of HMM-based ASR systems.

Avoiding recurrent or attention-based architectures, DNN-HMM systems are highly effective in terms of scaling and decoding speed. It makes these architectures highly dominant over their recent alternatives. Although they are very prone to be over trained on a specific domain and perform poor in others. One of these domain mismatch is the reverberation. A solid way to robust these architectures against reverberation is large receptive field networks as acoustic models. By having a LRF network in acoustic model, we are basically providing large context for generating the output labels. Having larger context have proven to be effective in many different aspects among which is triphone modeling which is the larger context in generating the final labels. In order to study the effect of these LRF networks, we compared performance of a conventional CNN with dilated and variants of large receptive field networks.

We used clean speech signals from WSJ corpus to simulate distant speech signals with real recordings of RIRs. We observed that these architectures are modeling acoustic behaviour of the clean signals even better than standard CNNs. We also did a empirical study on different configurations of each architecture and their performance in predicting desired labels. It was shown how standard CNNs require larger number of parameters to meet the performance of their alternative architectures.

Later, we did analyze the impacts of reverberation on speech using quality measures such as SNR, PESQ, Itakura-Saito and cepstral distance. We also studied convolutional CNNs with various receptive field size to better understand its impact on distant speech. This was another contribution of this thesis is to investigate the robustness effect of these architectures against reverberation.

Using the optimal RF size, we then compared the LRF networks constraining the parameters to find that hourglass network performs ≈ 2 % and ≈ 9 % relatively better compared to standard CNNs for clean and simulated distant speech signals. We also observed constant improvement in the performance of all distanced microphones in the realistic UTD-Distance corpus environment. It worth mentioning all of the best performances of clean tests are achievable by standard CNNs but with more number of parameters and considering the drawbacks of training a large network, the LRF networks are preferable. On the other hand the robustness these LRFs are providing make them completely better alternatives for standard CNNs. We conclude from this study that we can improve our ASR model and robust it against reverberation by having a better acoustic modeling component of HMM-based architectures. The next step in this study track would investigate the performance of each architecture in different level of reverberation. Then we could look for dynamically selecting better model for decode an audio given the type/level of reverberation it has.

BIBLIOGRAPHY

Abadi, Martın et al. (2016). "Tensorflow: A system for large-scale machine learning". In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283.

Abdel-Hamid, Ossama, Li Deng, and Dong Yu (2013). "Exploring convolutional neural network structures and optimization techniques for speech recognition." In: *Interspeech*. Vol. 2013, pp. 1173–5.

Andén, Joakim and Stéphane Mallat (2014). "Deep scattering spectrum". In: *IEEE Transactions on Signal Processing* 62.16, pp. 4114–4128.

Barnwell III, Thomas P (1979). "Objective measures for speech quality testing". In: *The Journal of the Acoustical Society of America* 66.6, pp. 1658–1663.

Baskar, Murali Karthick et al. (2017). "Residual memory networks: Feed-forward approach to learn long-term temporal dependencies". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4810–4814.

Chorowski, Jan et al. (2014). "End-to-end continuous speech recognition using attention-based recurrent NN: First results". In: *arXiv preprint arXiv:1412.1602*.

Dong, Linhao, Shuang Xu, and Bo Xu (2018). "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5884–5888.

Ernst, Ori et al. (2018). "Speech Dereverberation Using Fully Convolutional Networks". In: *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 390–394.

Ghorbani, Shahram, Ahmet E Bulut, and John HL Hansen (2018). "Advancing Multi-Accented Lstm-CTC Speech Recognition Using a Domain Specific Student-Teacher Learning Paradigm". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 29–35.

Ghorbani, Shahram and John HL Hansen (2018). "Leveraging Native Language Information for Improved Accented Speech Recognition". In: *Proc. Interspeech 2018*, pp. 2449–2453.

Ghorbani, Shahram, Soheil Khorram, and John HL Hansen (2019). "Domain Expansion in DNN-based Acoustic Models for Robust Speech Recognition". In: *workshop on automatic speech recognition and understanding*.

Graves, Alex, Navdeep Jaitly, and Abdel-rahman Mohamed (2013). "Hybrid speech recognition with deep bidirectional LSTM". In: *workshop on automatic speech recognition and understanding*. IEEE, pp. 273–278.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *ICASSP*, pp. 6645–6649.

Greff, Klaus et al. (2017). "LSTM: A search space odyssey". In: *IEEE trans. on neural networks and learning systems* 28.10, pp. 2222–2232.

Hau, Darren and Ke Chen (2011). "Exploring hierarchical speech representations with a deep convolutional neural network". In: *UKCI 2011 Accepted Papers*, p. 37.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hermansky, Hynek and Sangita Sharma (1999). "Temporal patterns (TRAPS) in ASR of noisy speech". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. Vol. 1. IEEE, pp. 289–292.

Al-Jarrah, Omar and Ahmad Arafat (2014). "Network Intrusion Detection System using attack behavior classification". In: *2014 5th International Conference on Information and Communication Systems (ICICS)*. IEEE, pp. 1–6.

Jeub, Marco, Magnus Schafer, and Peter Vary (2009). "A binaural room impulse response database for the evaluation of dereverberation algorithms". In: *16th International Conference on Digital Signal Processing*, pp. 1–5.

Khorram, Soheil, Zakaria Aldeneh, et al. (2017). "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition". In: *arXiv preprint arXiv:1708.07050*.

Khorram, Soheil, Mimansa Jaiswal, et al. (2018). "The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild". In: *arXiv preprint arXiv:1806.10658*.

Khorram, Soheil, Melvin McInnis, and Emily Mower Provost (2019). "Jointly Aligning and Predicting Continuous Emotion Annotations". In: *IEEE Transactions on Affective Computing*.

Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee (2016). "Deeply-recursive convolutional network for image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645.

Kim, Suyoun, Takaaki Hori, and Shinji Watanabe (2017). "Joint CTC-attention based end-to-end speech recognition using multi-task learning". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4835–4839.

Kinoshita, Keisuke et al. (2013). "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech". In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 1–4.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kuttruff, Heinrich (2016). *Room acoustics*. Crc Press.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Mesgarani, Nima, Shihab Shamma, and Malcolm Slaney (2004). "Speech discrimination based on multiscale spectro-temporal modulations". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, pp. I–601.

Nakamura, Satoshi et al. (2000). "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition". In: *Proceedings of 2nd ICLRE*. Citeseer.

Newell, Alejandro, Kaiyu Yang, and Jia Deng (2016). "Stacked hourglass networks for human pose estimation". In: *European Conference on Computer Vision*. Springer, pp. 483–499.

Ngiam, Jiquan et al. (2010). "Tiled convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1279–1287.

Oliva, Aude and Antonio Torralba (2006). "Building the gist of a scene: The role of global image features in recognition". In: *Progress in brain research*, pp. 23–36.

Oord, Aaron van den et al. (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*.

Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur (2015a). "A time delay neural network architecture for efficient modeling of long temporal contexts". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

— (2015b). "A time delay neural network architecture for efficient modeling of long temporal contexts". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Povey, Daniel et al. (2011). *The Kaldi speech recognition toolkit*. Tech. rep. IEEE Signal Processing Society.

Rix, Antony W et al. (2001). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE, pp. 749–752.

Sercu, Tom and Vaibhava Goel (2016). "Dense prediction on sequences with time-dilated convolutions for speech recognition". In: *arXiv preprint arXiv:1611.09288*.

Wang, Feng et al. (2018). "Additive Margin Softmax for Face Verification". In: *arXiv preprint arXiv:1801.05599*.

Wu, Zhizheng and Simon King (2016). "Investigating gated recurrent networks for speech synthesis". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5140–5144.

Yang, Jing, Qingshan Liu, and Kaihua Zhang (2017). "Stacked hourglass network for robust facial landmark localisation". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 79–87.

Yousefi, Midia, Soheil Khorram, and John HL Hansen (2019). "Probabilistic permutation invariant training for speech separation". In: *arXiv preprint arXiv:1908.01768*.

Yousefi, Midia, Navid Shokouhi, and John HL Hansen (2018). "Assessing Speaker Engagement in 2-Person Debates: Overlap Detection in United States Presidential Debates." In: *Interspeech*, pp. 2117–2121.

Yu, Fisher and Vladlen Koltun (2015a). "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122*.

— (2015b). "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122*.

Zhang, Biqiao, Soheil Khorram, and Emily Mower Provost (2019). "Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5871–5875.

BIOGRAPHICAL SKETCH

A remarkable academic performance at highschool in mathematics and science led in his admission into the ECE Department of the University of Tehran (UT), the earliest, largest and highly prestigious university in Iran. His great interest in computers coupled with his math skills motivated him to choose Computer Science and Engineering as his undergraduate major. In the first two years, he took different introductory courses to build a strong base of programming and algorithm design. At the University of Tehran, he had the chance to take excellent courses such as Data Structures that sparked his interest in data extraction and interpretation. Thus, he joined NLP Laboratory in the University of Tehran to work on a machine translation project. His involvement in that project was mainly focused on developing a program to automatically crawl audio and textual documents from Google Translate. Their results are available at www.faraazin.ir. Moreover, his Algorithms Design and Analysis course in second year significantly broadened his perspective on data science. He showcased his talent and interest in this topic by achieving the top grade in the class, and the instructor advised him to collaborate in some ongoing data management projects in his Social Network group.

Within the last year of his stay in Iran, he founded a start-up called *Zistiha* along with another friend. The technical side of the start up was done solely by him. The project was to create a website to synchronously hold online exams for high school students. This was his first industrial level project on web applications and he designed and implemented in RIA structure containing separated front-end and back-end application components from interaction with students to payment modules and so on (implemented under Angular and Django frameworks). He was in charge of maintaining and upgrading the website for almost three months and did the documentations for the next group while leaving Iran. Now the website is still up and running under zistiha.ir domain.

He started his master's studies in the CRSS (Center of Robust Speech Systems) lab at UT Dallas under supervision of Professor John Hansen. He chose this destination to get a deeper insight toward signal processing and machine learning techniques, specifically deep learning. His master's research project was essentially on robusting automatic speech recognition (ASR) systems toward reverberations and noise exposed by separating the microphone from the speaker (scenarios like Google home, Amazon Alexa). This robustness can be applied in different components of ASR system like acoustic modeling (phoneme classification given features) or front-end (raw signal to features mapping). In one of the projects, he compared large receptive field convolutional neural network (CNN) architectures as acoustic modeling component. Two of the architectures in this project were leveraged for the first time in ASR systems and shown to be effective in robusting the system toward reverberation. The results of this project has been published in ASRU workshop .

CURRICULUM VITAE

Salar Jafarlou

Deep Learning Scientist | ML Engineer | Software Engineer

RESEARCH PROJECTS

Active Learning for Speech Recognition

- Bayesian risk extraction from DNN-HMM decoder lattice
- Extraction of posteriograms from the TDNN acoustic model
- Propose novel full-CNN and RNN-CNN models to predict the Bayesian risks directly from posteriograms skipping lattice generation (Pytorch)
- Speeding up Bayesian risk computation by about 4 times with skipping the lattice generation and confusion matrix extraction

Image Processing Networks as Acoustic Model

- Investigation of large receptive field networks in image processing and feasibility of leveraging them as acoustic modeling
- Implementation of Stack of Hourglass, Time Delayed Recursive Neural Network structures
- Comprehensive study on phoneme recognition performance and nal word error rate of decoding
- Write up of the paper submitted to ASRU-2019 Workshop

Traumatic Brain Injury Detection in Mice Using EEG Signal

- Preprocess and noise elimination on EEG signals gathered from mice
- ResNet-like deep architectures implementation specified for detecting TBI in mice from raw signal and frequency band features
- Write up of the report paper for MDPI Journal

MultiView Deep CCA for Automatic Speech Recognition

- Investigation of Canonical Correlation Analysis (CCA) structures and design of deep neural network CCA structure
- Beamforming and cross-correlation-based synchronization
- Implementation and training of deep-CCA network (Tensroflow)

Sentiment Analysis on a Microblogs Documents

- Data Augmentation and train Word2Vec feature generator
- Design, implement and train a Convolutional Neural Network for polarity detection

Expert System for Detecting ADHD in Children

- Review ADHD diagnosis process, domain experts corporation to design survey
- Feature selection & algorithm testing, online reinforcement learning implementation of Random Forests algorithm
- Implementation and deployment of portal for survey & online ADHD diagnosis expert system

INTERNSHIP

ASAPP Inc. Jan 2020 - April 2020

Speech and NLP company in silicon valley with cutting the edge technologies offering wide range of services from voice transcription to text sentiment analysis. As a researcher I was in charge of a project on applying deep active learning methods across different speech systems to conduct a semi supervised learning experiment. The project was established and the codes and outputs were delivered in less than 3 months.

EDUCATION

2017 – 2020 **Master of Science**
ELECTRICAL ENGINEERING - SIGNALS
University of Texas at Dallas
GPA: 3.6

2012 – 2017 **Bachelor of Engineering**
ECE - COMPUTER ENGINEERING
University of Tehran, Iran
GPA: 3.5

SELECTED COURSES

GRADUATE	Machine Learning - Big Data - Multi-Modal Signal Analysis - Statistical Methods in Data Science
UNDERGRADUATE	Multi-Media - Data Structures - Algorithms Design and Analysis
ONLINE	Deep Learning (1-3) Micro Economics - Capital Markets I

CODING SKILLS

CODING	Python, Java, C++, C
FULL STACK	Django, ReactJS, NodeJS, Angular
DEEP LEARNING	Pytorch, Keras, TensorFlow, Kaldi
MACHINE LEARNING	Scikit Learn,SciPy,Weka
BIG DATA	GCP, AWS, Apache Spark, Apache

SOFTWARE ENGINEER

Nar-Co Iran - 2016

- Implementation a module in C++ to communicate with a cheap through serial port
- Lead software engineer of a group in the software product of the company for final end user in C#

Zistiha.ir

Calgary - 2019

- Full-stack development of a website for holding online, simultaneous exams and correction and rankings
- Different technologies were leveraged: Django, Angular, Bootstrap, MySQL

PUBLICATIONS

Salar Jafarlou,S. Khorram, J. Hansen. "Large Receptive Field Convolutional Networks for Continuous Speech Recognition" *ASRU-2019*

Vinay Kothapally, Xia Wei, **Salar Jafarlou**, J. Hansen "Large Receptive Field Convolutional Networks for Continuous Speech Recognition" *Interspeech-2020* (in progress)

Manoj Vishwanath,**Salar Jafarlou**, Ikhwan Shin, Miranda M. Lim, Nikil Dutt, Amir M. Rahmani and Hung Cao "Investigation of Machine Learning Approaches for Traumatic Brain Injury Classification via EEG Assessment in Mice" *MDPI*