Automated Urdu Speech Recognition System using Deep Learning

Author

WAQAS RASHED

Regn Number

Fall 2015-MS (CSE) 00000118445

Supervisor

Dr. Arslan Shaukat

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JULY 2019

Automated Urdu Speech Recognition System using Deep Learning

Author

WAQAS RASHED

Regn Number

Fall 2015-MS (CSE) 00000118445

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Software Engineering

Thesis Supervisor:

Dr. Arslan Shaukat

Thesis Supervisor's Signature: _____

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JULY 2019

Declaration

I certify that this research work titled "*Automated Urdu Speech Recognition System using Deep Learning"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

WAQAS RASHED

Fall 2015-MS (CSE) 00000118445

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student
WAQAS RASHED
Fall 2015-MS (CSE) 00000118445

Signature of Supervisor
Dr. Arslan Shaukat

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

I would also like to express special thanks to my supervisor Dr. Arslan Shaukat for his help throughout my thesis.

I would also like to pay special thanks to Dr.Usman Akram for his tremendous support and cooperation. Each time I got stuck in something, he came up with the solution. Without his help I wouldn't have been able to complete my thesis. I appreciate his patience and guidance throughout the whole thesis.

I would also like to thank Dr. Sajid Gul Khawaja and Dr. Wasi Haider Butt for being on my thesis guidance and evaluation committee. I am also thankful to my friend Zeeshan Asif for his support and cooperation.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

A computer needs to be able to understand what you said, before it can even understand what you mean and it encompasses Natural Language Processing (NLP). There are limited open source speech recognition systems are available with close to human level performance and it is particularly true for the Urdu language. In this thesis we worked on the development of an Urdu language Automatic Speech Recognition (ASR) system based on Deep Learning. Mozilla's DeepSpeech which is an open source implementation of TensorFlow, Optimized RNN, CTC and Bi-directional LSTM used to train and build acoustic model for Urdu speech recognition. A language model is constructed and trained to represent Urdu alphabets and common vocabulary and its binary file is created with help of KenLM tools to feed it in the system for estimation and decoding purposes. This method makes training a speech recognition system a lot simpler and it does not require many complex neural network layers or knowledge about a language to train. The system is trained using virtual machine on Google's cloud platform with GPU support. Based on WER (Word Error Rate) the number of nodes in the core layers of neural network is optimized acquired on data set, having concern to GPU memory limits. Our study lead to an Urdu language acoustic model which has been trained based on data set we have collected. We collected almost 390K speech instances of total 1008 sentences from 400 male and female students. Desktop and mobile applications are developed to automatically record and collect on the cloud the spoken audio files along with their transcripts. Language model is constructed with these commonly used Urdu sentences. Data set is separated into three categories including training, validation and testing data with a split of 70%, 20% and 10% respectively—the best results in form of WER we get from the system is less than 10%.

Key Words: *Deep Learning, Urdu ASR, DeepSpeech, LSTM, RNN, CTC*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

The thesis investigates using deep learning based speech recognition system for Urdu language. The model is trained with DeepSpeech which is deep learning based speech recognition framework using the data acquired mostly from voices of male and female university students. It mainly consists of a deep neural network and a hybrid CTC/attention based encoder-decoder architecture.

This chapter describes the problem and methodology for solving it. It outlines the basic differences between using the traditional and deep learning based approach for speech recognition. We define the problem, set a hypothesis for solving it and set target goals for validating the results.

1.1 Problem

The thesis investigates using deep learning based speech recognition for Urdu language. Deep learning based method has not been previously researched on Urdu language.

The current research on speech recognition for Urdu language is done using the traditional approach. Traditional speech recognition systems are currently based on very complex components. These components are acoustic models based on hidden Markov models, Gaussian mixture models, deep neural networks, n-gram and neural network based language models, complicated training and decoding algorithms. Deep learning based speech recognition replaces all these different components in the traditional pipeline with a single deep learning based deep recurrent neural network (RNN).

1.2 Objectives

The objective for this thesis is to test a deep recurrent network model with hybrid CTC/attention based encoder-decoder architecture for Urdu speech recognition. The model is trained using about 75 hours of Urdu speech recordings from approximately 400 unique speakers.

The results are evaluated using word error rate and character error rate on n-best lists. Word error rate measures the accuracy by comparing the exact match of words. Character error rate compares each character individually, which usually gives a correctness score better than with word error rate. These scores are calculated for deep learning based speech recognition system and then combined with the traditional approach.

The goal is to achieve a word error rate below 10% and a character error rate below 2% for only using deep learning based speech recognition with better results than traditional approach.

1.3 Methodology

The model is trained using different deep learning based speech recognition systems. These systems all include a deep neural network. The difference is the method used for scoring the output of a neural network.

Around 390K speech utterance of total 1008 sentences from 400 male and female students are collected. Different applications are developed for automatically record and collect the speech files along with their transcripts on the cloud. A language model is created with commonly used Urdu sentences. Data set is separated into categories including 70% of training, 20% of validation and 10% of testing data. Model id trained using DeepSpeech and results are extracted in term of WER.

1.4 Thesis Outline

The first chapter describes the problem and methodology for solving Urdu language deep learning based speech recognition. It outlines the basic differences between using the traditional and deep learning based approach for speech recognition. We define the problem, set a hypothesis for solving it and set target goals for validating the results.

The second chapter gives a detailed overview of the main component used for using an deep learning based speech recognition system, e.g. feature extraction, RNN, LSTM, CTC, attention based encoder-decoder and evaluation metrics.

The third chapter introduces previous related works that are related to this thesis' problem. These alternative works have used deep learning based speech recognition for other languages. The only Urdu language related work concentrates on the traditional method for speech recognition.

The fourth chapter specifies the solution part of the thesis. It gives a detailed overview of the methods used with training the deep learning based speech recognition system. The deep learning based methods and DeepSpeech architecture is explained.

The fifth chapter is about the experiments. It describes the data used in the experiments, why the DeepSpeech framework is chosen, the used model for training, the results of the experiments and the analysis of the results.

The sixth chapter draws conclusion on how the deep learning based speech recognition system performs on Urdu language. The objectives and hypothesis are analyzed whether they were achieved as initially planned.

CHAPTER 2: RELATED BACKGROUND

This chapter gives a detailed overview of the main component used for deep learning based speech recognition system, e.g. feature extraction, RNN, LSTM, CTC, attention based encoder-decoder and evaluation metrics.

2.1 Deep Learning Based Speech Recognition

Speech recognition is the process of automatically extracting the word conveyed by speech wave. Traditional speech recognition systems are currently based on very complex components. These systems use components such as hidden Markov models, Gaussian mixture models, deep neural networks, n-gram and neural network based language models. Deep learning based speech recognition replaces all these different components in the traditional pipeline with a single deep learning based deep recurrent neural network.

Deep learning based speech recognition does not know anything about words or how they are used. It tries to guess each letter on a given audio and combine those to form words. This is different from the traditional approach where vocabulary and n-grams are used. The traditional approach then tries to calculate probabilities of what a person might have said compared to a given vocabulary and n-grams.

The main benefit of using deep learning based speech recognition is that is simplifies the process of training and deployment. Because of the fact that deep learning based does not need a vocabulary or n-gram, it can be used with different languages more easily, when only training data is available. It will also simplify deployment on mobile devices, because it does not use a typical n-gram language model, which takes a lot of disk space.

Deep learning based speech recognition system has one important downside. Even though training the system is a lot simpler than in traditional approach, it needs a lot of training data. The system will have to learn different characteristics of a language by itself and that is why it needs to have a sufficient amount of data. Usually, the system needs to have thousands of hours of data to train on. It is possible to train with less data, but the results will not be as good as the system is capable of achieving.

Deep learning based speech recognition is mainly based on deep recurrent neural network. A deep recurrent neural network alone is not enough for a speech recognition system. The first

attempts used CTC, but it is incapable of learning the language and needs language model to clean up common mistakes. Instead of CTC, attention-based models were tested and they proved to be outperforming previous models, due to the ability of learning all components of a speech recognizer. Both methods still have some benefits over one another and recent works have started using a hybrid CTC/attention based architecture, which is also used in this thesis.

2.2 Feature Extraction

The first step in recognizing speech from audio is to extract features. Feature extraction is for removing background noise, emotion and all other useless information to get only the components that can be used for identifying linguistic content. This pre-processing is needed for making neural network's processing easier to recognize text from audio [1].

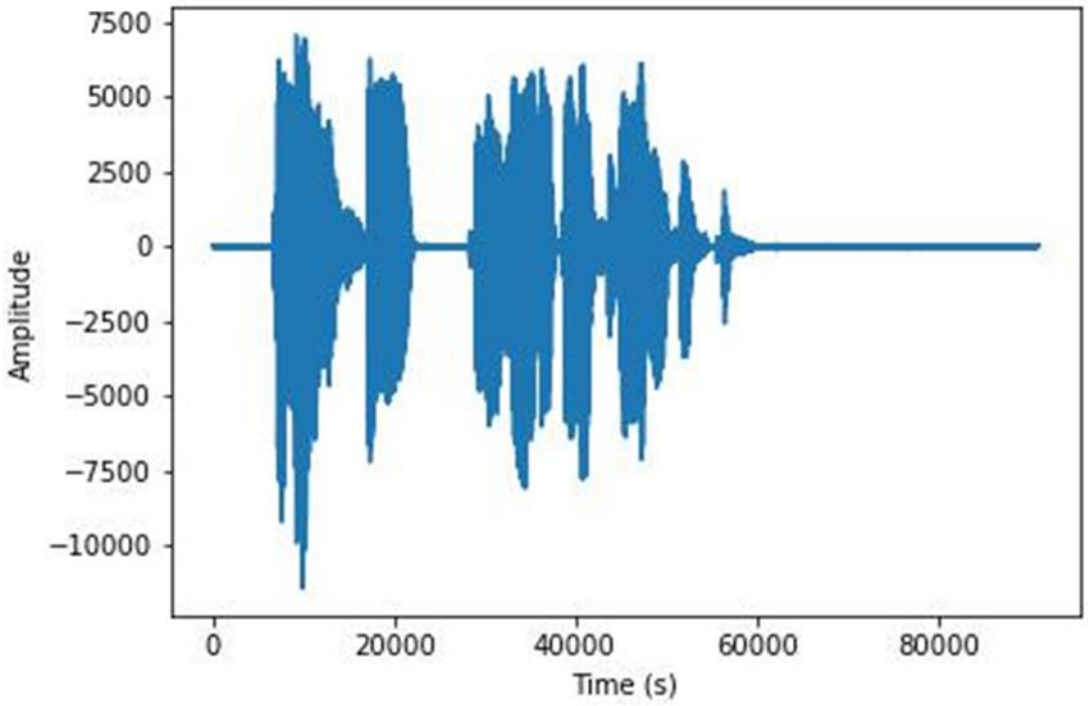

Figure 2.1: An audio signal wave [1]

For extracting features from audio, the signal is usually framed into 20-40ms frames. Each frame is selected after 10ms, which means that the next frame will have some of its contents from the previous frame. All of these frames still contain redundant information that require filtering. To remove all sudden endings of an audio signal in each frame, a window function is used, such

as the Hamming window. The Hamming window function also allows to counteract the assumption made by the discrete Fourier transform that the signal is infinite and to reduce spectral leakage.

By using discrete Fourier transform, each complex sound wave is broken apart into simple sound waves. It takes a windowed signal as input and outputs a complex number for each frequency band. Each of those sound waves' contained energy is added up to get a score of how important each frequency band is. To better visualize the output of this process, a spectrogram is created using each frame's contained energy scores.

After using the discrete Fourier transform, the spectrogram still has too much information. Triangular filters are applied on a Mel-scale to the power spectrum for extracting frequency bands. The Mel-scale's purpose is to mimic the non-linear human ear perception of sound. Lower frequencies are filtered with narrower and higher frequencies with wider bands. Each of those filters collect energy from a number of frequency bands in discrete Fourier transform.

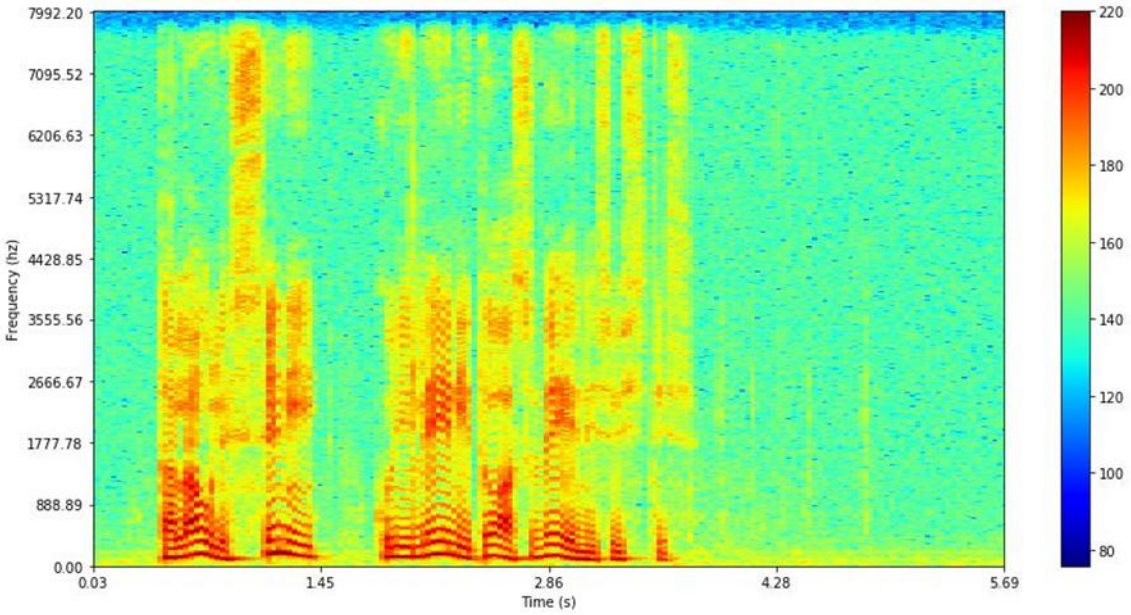

Figure 2.2: Spectrogram of an audio clip [1]

The spectrogram in Figure 2 visualizes the patterns of low and high pitch frequency ranges. This data is better for a neural network to process, because it can find patterns more easily. That is why this is the actual representation of audio data that gets fed into the neural network. The neural network will then try to figure out the best possible letter for each of these 20ms frames.

2.3 Recurrent Neural Network (RNN)

RNN is a type of deep learning model that works best for handling sequential information. RNN assumes, that all inputs and outputs are dependent on each other, unlike the traditional neural network. It keeps a memory of previous outputs and passes those as inputs from one step of the network to the next (Figure 3). This way the network can have a deeper understanding of the statement [2].

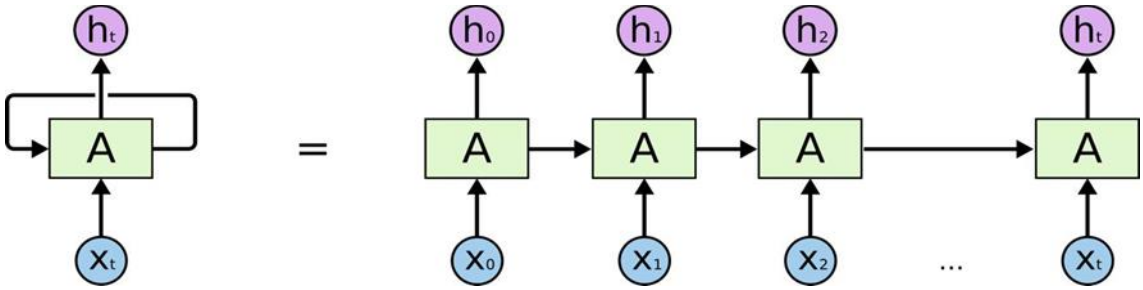

Figure 2.3: An unrolled RNN [3]

The above figure shows a chunk of neural network (A), that takes (Xt) and previous output as inputs and outputs a value (ht). The recurrence allows the network to pass information from one step to the next [3]. This is the basic workflow of a RNN, but it is often used with bidirectional to get more accurate results.

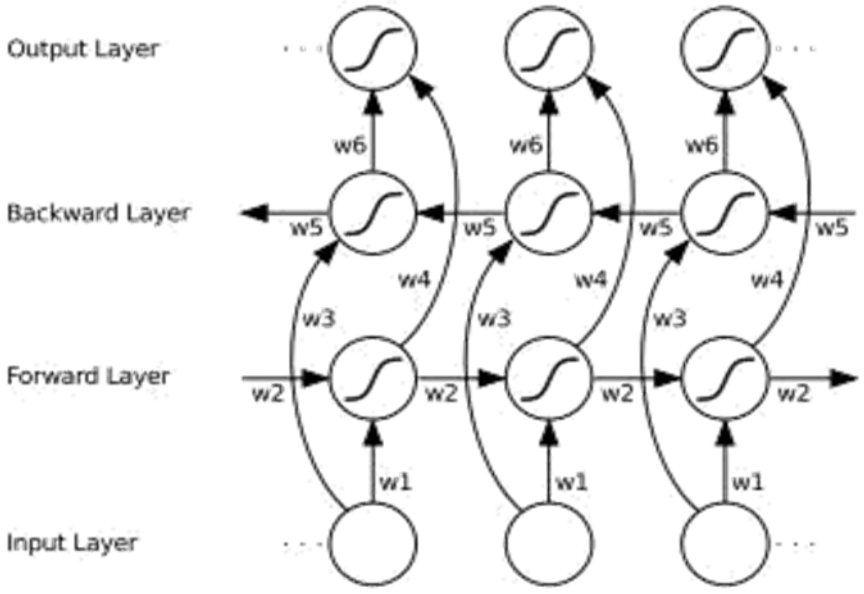

Figure 2.4: Bidirectional RNN (BRNN) [4]

It is often beneficial to know some information about the future as well as past context. Usually, when classifying a letter in a word, it is beneficial to know about the previous and next letters. This is something that bidirectional RNNs will help to solve. The idea behind BRNN is to have two hidden layers (Figure 4), one for forward and one for backward layer. This way the output layer will have both the past and future context for every point in the input sequence [4].

Standard RNN does not always perform very well. The problem is that RNNs cannot preserve memory from far away in the sequence. RNN makes predictions based on the most recent sequences. This means that the context about the start of a sentence might be lost while predicting the end of the sentence. To solve this problem, RNN is often used with long short-term memory (LSTM) architecture to have the context of a whole sentence always available in memory [5].

2.4 Long Short-Term Memory (LSTM)

Standard RNN architectures have a problem with multiple hidden layers. When passing information from one hidden layer to another, the information might get lost, if there are many layers. LSTM handles this kind of situation and enables RNN to preserve memory throughout the whole learning process.

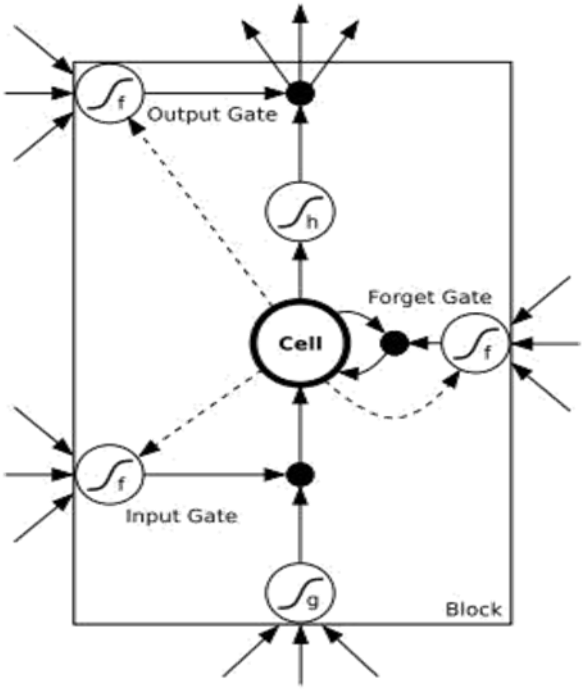

Figure 2.5: LSTM memory block with one cell [4]

LSTM architecture consists of memory blocks, which are all recurrently connected to each other. Each memory block contains at least one self-connected memory cell [6]. The memory cell allows information to be stored in, written to or read from. It also decides, which information to store and when to allow reading, writing and erasing. This is done using input, output and forget gates that open and close as shown in Figure 5.

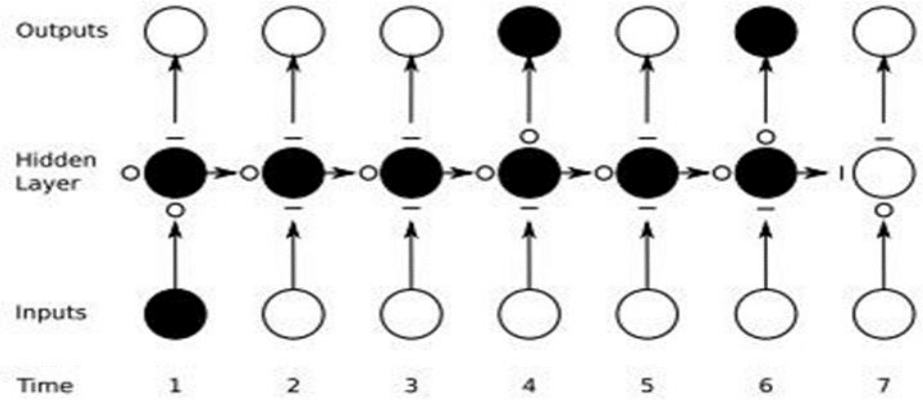

Figure 2.6: Preservation of gradient information by LSTM [4]

Figure 6 shows the preservation of gradient information by LSTM. Input state, forget and output gates are shown below, to the left and above the hidden layer respectively. (O) represents an entirely open gate and (-) is for a closed gate. Looking at the diagram, LSTM memory cell can preserve information as long as the input gate is closed and forget gate is open. Getting information from output gate does not have any affect memory cell's data.

When using LSTM architecture with bidirectional RNN gives bidirectional LSTM (BLSTM) [4]. Using BLSTM allows to preserve information from the past as well as from the future. This is important, when the understanding of past and future context is needed to find the correct next word in any time.

For better understanding of BLSTM, it can be explained with a simple speech recognition example. Let us say, we need to detect the next word for a sentence starting with "I will go to ". Currently the only available information about the sentence's context is in the past. Finding the correct next word can be difficult, when there are almost limitless possibilities. Now, the BLSTM allows to get context from the future as well. When the sentence continues with "and learn machine learning", the detection for the missing word becomes simpler, because of the extra context about the whole sentence [7].

2.5 Connectionist Temporal Classification (CTC)

People talk with very different rates of speed which makes training an ASR system a lot more difficult. That is why the alignment between characters in the transcript and audio is always unknown. One way of solving this problem is to manually align all characters to their location in the audio. The major downside is that it's very time consuming when dealing with large datasets. Another option is to use connectionist temporal classification (CTC) which has become a very popular among RNNs [8].

CTC is a type of neural network output and associated scoring function. It is used with RNNs to handle sequential problems. CTC sums over the probability of all possible alignments between the input and the output [4]. Assuming that an input has a length greater than the actual word's length, one option for solving the problem is to collapse all repeating characters.

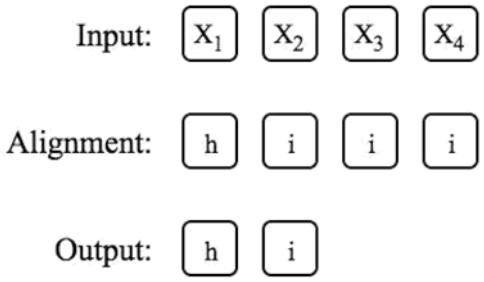

Figure 2.7: Merging repeated characters

Collapsing all repeating words is not the best way of tackling this problem, because it will remove all repeating characters even when there should be a repetition. That is why CTC uses a token called blank (here referred as _). This token is always removed from the output, but used in alignment process.

Figure 2.8: Merging repeated characters with blank token

Firstly, CTC merges all repeating characters and secondly, removes all blank tokens. The remaining output will be 'hello' not 'helo', which would be the output without the token.

There are a lot of possible ways of character alignment for every input. CTC loss functions combines all alignments where the output is the same. It then calculates the score for each of these combinations and sums over all scores. While decoding, a character with the highest score for each time step is picked. After that, the duplicate characters are merged and blank tokens removed to get the final output [9].

2.6 Attention Based Encoder-Decoder Architecture

Encoder-decoder architectures are mostly used to deal with sequences where the input and output length size in unknown. Both encoder and decoder are RNNs. The encoder transforms the input to a higher level representation where the length size is fixed. The decoder then uses this representation and generates output sequences.

When dealing with a simple encoder-decoder architecture, the decoder generates a transcription based on the last hidden state from the encoder. This is not a reasonable approach, because when dealing with sentences containing many words, the encoder will have to encode every information into a single vector. The decoder must then produce a valid output only based on this single vector. While decoding, the decoder has to consider information from the beginning of the sentence and when dealing with RNNs, it is known that long-range information might get lost.

Attention based encoder-decoder architecture solves the problem of encoding everything into one single vector. The attention mechanism allows decoder to get information from all parts of the source sentence at every step of output generation. The model will learn by itself what information is important and should be considered. Each decoder output does not depend on the last vector anymore, but instead on a weighted combination of all the input states [10].

2.7 Evaluation Metrics for Speech Recognition

Encoder-decoder architectures are mostly used to deal with sequences where the input and output length

2.7.1 Word Error Rate (WER)

WER is a method for calculating the performance of a speech recognition software. This is not always easy to measure, because the correct input length can be different from the detected value length.

$$WER = \frac{S + D + I}{N} \quad \textbf{(2.1)}$$

Equation 1 shows the equation for calculating WER. S shows the number of substitutions, D is for deletions, I is for insertions and N is for words in the reference [11].

There are three possibilities for an automated speech recognition (ASR) software to make mistakes that WER will calculate:

1) **Deletion** – ASR system deletes a word
   Correct input: Machines can think
   ASR result: Machines think
2) **Insertion** – ASR system inserts an unneeded word
   Correct input: Machines can think
   ASR result: Machines can not think
3) **Substitution** – ASR system substitutes a correct word with an incorrect one Correct input: Machines can think
   ASR result: Machines can learn

WER can sometimes give very unreasonable results when dealing with compound words. Sentences like “water melon tastes good” and “watermelon tastes good” are both very well understandable. But the calculated WER would be 50%, which is unfair considering that the actual mistake is only adding an unnecessary white space. This is where character error rate will give more adequate results.

2.7.2 Character Error Rate (CER)

Another method for calculating the performance of ASR is character error rate. CER is calculated with the minimum number of operations necessary to transform the original text into ASR output. The smaller the number, the more accurate both texts are.

The equation for calculating CER is the same for WER as shown in Equation 1. But for CER, N is for the total number of characters and the minimal number of character substitutions as S, deletions as D and insertions as I, required for transforming original text into automatic transcription [12].

White space and case are also important for CER. While contiguous white spaces are usually considered as one, a word pair “auto mobile” with more than one space between them still gives an accuracy of 10%. When comparing words with different case like “Hello World” and “hello world”, CER sees them as substitutions and calculates an accuracy of 18%.

CER is most commonly used when dealing with languages that have difficult declensions. When the original reference in Urdu is “koerast” and ASR recognizes it as “koeras”, the WER would be 100%, but CER is only 14%. For these kind of languages, where a word has many different cases, WER might show a bit unfair results compared to CER. Although the result of WER is high, the word is still readable and in the meaning of the sentence would still be understandable.

CHAPTER 3: LITERATURE REVIEW AND RELATED WORK

The problem of using deep learning based speech recognition for Urdu language has not been investigated in any earlier research.

This chapter introduces previous related works that are related to this thesis' problem. These alternative works have used deep learning based speech recognition for other languages. The only Urdu language related work concentrates on the traditional method for speech recognition.

Asadullah et al. [13] presented a method for development of Automatic Speech Recognition for Urdu isolated words. Research is built on Urdu language comprising 250 words of medium vocabulary speech corpus using the open source toolkit Sphinx. Using this approach Mel Frequency Cepstral Coefficients (MFCC) features are fetched and created a Hidden Markov Model (HMM) to accomplish speech recognition. Reported accuracy is 78.2% for both experimentations of 100 and 250 words separately. Results recommend that improved speech recognition accurateness has been accomplished with this method as the preceding results stated 70.69% on same corpus.

Syed Abbas Ali et al. [14] proposed Urdu to English language speech interpreter consuming Deep Neural Network. ASR section in suggested pipeline is using deep neural network which is modest as matched to old-fashioned ASR which needs difficult engineering like feature mining and huge resources such as phoneme lexicon. Proposed technique displays extraordinary accurateness when the input to system is recorded voice and deprived performance with the real time voice input. On HTTP request per input text shaped English transformation for Text to Text conversion using TextBlob toolkit of Python. The Output attained with an interval of 30 seconds. Reported error rate of proposed system varies from 10% to 21%.

Awni Hannun et al. [15] presented a speech recognition method develop with end-to-end deep learning. Design is meaningfully naiver than old-fashioned speech methods which depend on painfully engineered handling pipelines which be likely to achieve poorly while in loud surroundings. This approach is a well improved RNN training method which uses several GPUs and an innovative data fusion methods that permit resourcefully acquire huge amount of diverse data for the training task. System is named DeepSpeech which overtakes formerly available results on commonly studied Switchboard Hub5'00 while reaching 16.0% error on complete test dataset.

DeepSpeech likewise handles very noisy surroundings enhanced than state-of-the-art extensively used commercial speech recognition systems.

Hazrat Ali et al. [16] presented the construction of Urdu language Medium Vocabulary Speech Corpus. These Corpus includes 250 isolated words with digits and most commonly vocalized words of Urdu language. These words are carried from 5000 commonly words among 19.3 M words of the Urdu language. Particular words have been spoken by 50 utterers in a noise free workplace. Speakers includes male and female, teenagers and old peoples.

Dario Amodei et al. [17] showed that end-to-end deep learning method could be implemented to recognize both English and Mandarin Chinese speech which are two very diverse languages. Because it substitutes intact pipelines of hand engineered mechanisms with neural networks end-to-end learning permits to hold a varied selection of speech data comprising loud environments pronunciations and altered languages. Use of HPC methods allowing investigates formerly take many weeks and now in days. This permits to reiterate more rapidly to classify bigger architectures and procedures. Resulting numerous circumstances the system is modest with transcription of language when benchmarked on ordinary datasets.

Eric Battenberg et al. [18] performed an experiential evaluation between RNN Transducer, CTC and responsiveness Seq2Seq representations of end-to-end speech recognition. It is stated that deprived of any language model, RNN Transducer and Seq2Seq models both overtake greatest stated CTC models with a language model on the prevalent Hub5'00 standard. For their internal assorted dataset these developments last. RNN Transducer representations scored with language model afterward beam search beats our finest CTC models. Outcomes streamline speech recognition pipeline so that decoding can further articulated only as neural network actions.

S Shaleva et al. [19] examined the concrete issues in emerging a speech to text system by deep neural networks. Development of Russian language speech recognition constructed on DeepSpeech is described. The use of language models with numerous concentrated sizes of word arrangements and designated the model that presented the greatest WER. Result is an acoustic Russian language model which is trained constructed on data set including audio and captions from video clips of YouTube. Language model was constructed built on the texts of captions and openly accessible corpus of Russian language from articles of Wikipedia's. Resulting system verified on dataset containing audio recordings of Russian literature accessible on voxforge.com and the best WER 18% verified by the system.

Ossama Abdel-Hamid et al. [20] proposed to put CNN in speech recognition inside the context of NNHMM a hybrid model which would be use native sifting in frequency domain and max pooling to standardize speaker adjustment to accomplish advanced multi speaker speech recognition enactment. A pair of native sifting layer and max pooling layer added at lowermost end in neural network to regularize phantom differences in speech signals. Experiments showed the suggested CNN structural design is assessed in a speaker autonomous speech recognition consuming ordinary TIMIT datasets. Tentative results demonstrated that suggested CNN technique can accomplish above 10% comparative error decline in central TIMIT test datasets when associating with consistent NN using identical hidden layers. Results showed the finest result of suggested CNN model enhanced than formerly available results on same TIMIT datasets that uses pre trained deep NN models.

Tara N. Sainath et al. [21] took benefit of complementarity LSTMs, DNNs and CNNs by merging them into single combined design. Suggested architecture which called CLDNN on a variability of huge vocabulary tasks that fluctuating from 200 to 2,000 hours of data. It is found that CLDNN delivers 4-6% comparative enhancement in WER above than LSTM which is the robust of the all three distinct models.

Huda Sarfraz et al. [22] presented the construction of language and acoustic models for Urdu speech recognition consuming CMU Sphinx which is an open source Toolkit. Three prototypes have been developed one by one with the accumulation of Urdu voice data of two utterers per pass for which one model consuming voice data one from 41 male utterers solitary, from 40 female utterers solitary and one with both male and female utterers with a total of 81 utterers. Best results presented in term of WER is 29.1%.

3.1 Hybrid CTC/attention-based deep learning based architecture

In machine translation, where word order for input and output can be different, the attention-based encoder-decoder works fairly well. It allows nonsequential alignments between each element of the output sequence and acoustic encoder network generated hidden states for each frame of acoustic input.

But for speech recognition, word order is the same for input and output except some small within-word deviations that may happen.

Another problem is the different lengths of input and output sequences. The difference in length comes from each speaker's speaking rate and writing system. That makes it difficult for the ASR to track the alignment between input and output. The attention mechanism could solve all these problems, but for better results, a CTC-based alignment will be used for training the model [41].

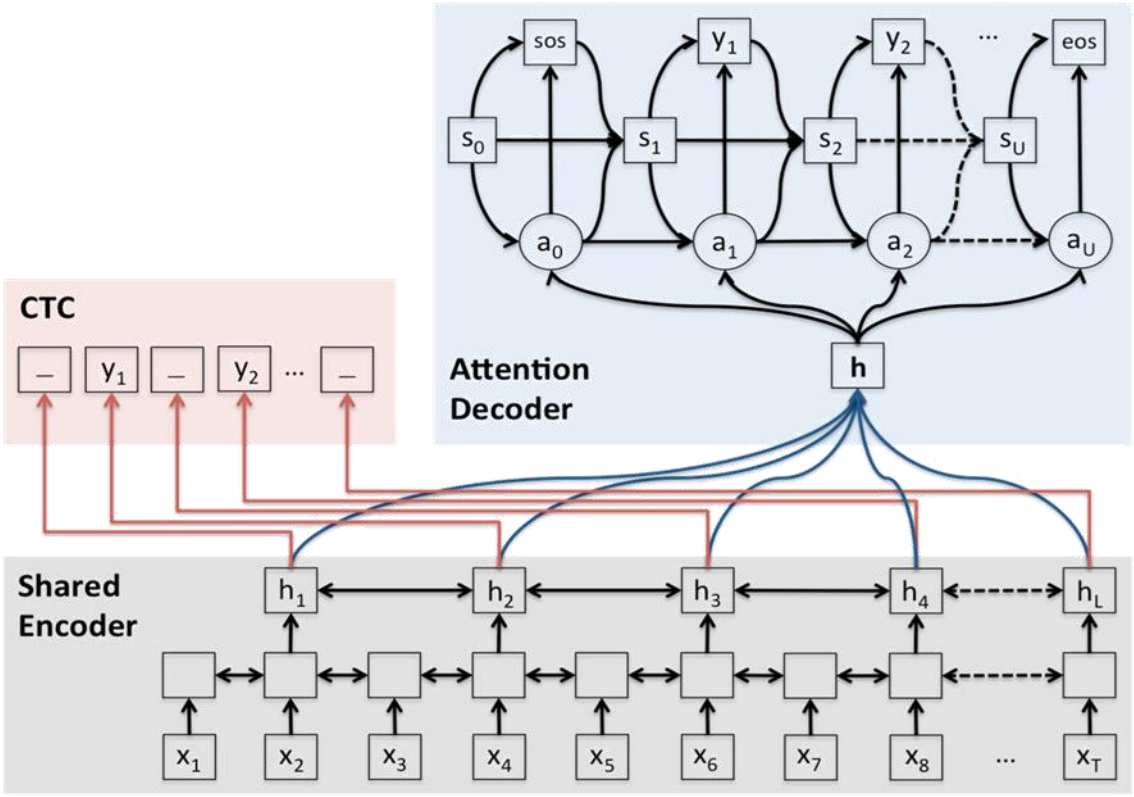

Figure 3.1: Hybrid CTC/attention-based deep learning based architecture [33]

Hybrid CTC/attention-based deep learning based architecture, as shown in Figure 9, solves both the word order and alignment problem between input and output. This architecture uses a CTC objective function as an auxiliary task for training the attention model encoder. The BLSTM encoder network is shared between CTC and attention model.

The decoding process uses both attention-based and CTC scores. Because of the fact that CTC and attention-based decoder computes scores differently, combining them is nontrivial. A rescoring/one-pass beam search algorithm is used to combine those scores. The outcome of this would eliminate all irregular alignments.

Using this joint architecture, the learning process of the network is quicker and it works better in noisy conditions or with long sentences. The forward-backward algorithm of CTC

enforces monotonic alignment between speech and label sequences. This helps to acquire more accurate alignments in noisy conditions. Using CTC as an auxiliary task also improves the speed in estimating alignments without the aid of rough estimates. That way the estimations for alignments in long sequences are not solely dependent on data-driven attention methods [33].

3.2 Combining n-best lists with ROVER method

N-best list is generated by the ASR system and it contains a list of likely possibilities for input sentence which is sorted by the best score. Each possibility is different and has a score of how sure the system is in its correctness. N-best list allows to combine multiple different ASR systems to achieve better results.

For combining multiple n-best lists, a ROVER system is used. The first step for this system is to align all hypothesis transcripts from ASR systems to get one word transition network (WTN). It firstly creates WTNs for all ASR system outputs to be able to combine them.

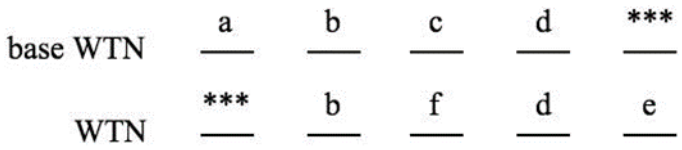

Figure 3.2: WTNs alignment

A base WTN is chosen from which the composite WTN is developed. All other WTNs are aligned according to the base WTN as shown in Figure 10. For example, if there are 3 different systems, a base WTN is chosen and then one of the remaining WTN is aligned with the base WTN to form a new base WTN. The process is repeated with all other remaining WTNs to eventually get one final composite WTN.

When the final composite WTN is found, a voting module is used to find the best scoring word sequence. The voting module finds the occurrences of each word and accumulates them.

The Equation show how the voting is performed. The number of occurrences of word type *w* is accumulated in correspondence set *i* in the array *N(w,i)*. To scale the frequency of occurrence to unity, the array is then divided by the number of combined systems *Ns*.

The measured confidence scores for word w create an array C(w,i). The parameter α is trained to be the trade-off between using word frequency and confidence scores.

The voting can be done in three different ways. When setting the α parameter to 1, the information about confidence scores become irrelevant. This way the voting is made by frequency of occurrence. When training the parameter α a priori on the training data, the voting will use confidence scores to find either average or maximum confidence scores. The parameter α can be trained by quantizing the parameter space into a grid of possible WER values and then exhaustively searching for the lowest WER [42].

3.3 Deep Learning Based Speech Recognition

In machine translation, where word order for input and output can be different, the attention-based encoder-decoder works fairly well. It allows nonsequential alignments between each element of the output sequence and acoustic encoder network generated hidden states for each frame of acoustic input.

Deep Speech 2: Deep learning based Speech Recognition in English and Mandarin was created in 2015 to show the possibilities of implementing deep learning based speech recognition on very different languages [26]. The system consists of three main components:

- RNN with one or more convolutional input layers
- Multiple recurrent layers and one fully connected layer
- CTC

For training the models, this research uses 11940 hours of labeled speech, which contains 8 million utterances, for English model and for Mandarin, there are 9400 hours of labelled speech, which contains 11 million utterances.

The trained model's WER for English language is comparable with human WER, when the audio is clearly understandable. In these cases, the WER differs between 3-13% using different datasets. When testing with accented or noisy audio, the WER becomes understandably bigger. The difference between human level and the trained model becomes clearer when dealing with accented or noisy audio.

The results for Mandarin language show that deep learning based speech recognition can give better results than an average human speaker. When transcribing short voice-query like utterances, the trained system for Mandarin language works better than human level performance. The system achieves a WER of 3.7% for 100 random utterances labelled by a committee of 5 and

5.7% for 250 utterances labelled by a single person. A typical Mandarin Chinese speaker achieves approximately 4% for committee labelled utterances and 9.7% for utterances labelled by an individual.

Listen, attend and spell (LAS) is research done in 2015 and has a key improvement over previous deep learning based CTC models. LAS uses a neural network, that transcribes speech utterances to characters. The system has two components: a listener and a speller, which are both jointly learned. The listener is a pyramidal recurrent network encoder that uses filterbank spectra for inputs. The speller is an attention-based recurrent network decoder that sends out characters as outputs [40].

Without using a language model or a dictionary, LAS achieves a WER of 14.1% on a subset of the Google voice search task. The result is not as good as the traditional DMM-HMM models, but still quite good for a system that has not been fully researched and developed.

There have also been many other recent researches about deep learning based speech recognition using LAS, such as [28] [29] [30] [31] [32].

Joint CTC/attention decoding for deep learning based speech recognition is another research for deep learning based speech recognition created in 2017 [33]. Previous works on deep learning based ASR systems have used either CTC or attention architecture. This research has created a deep learning based speech processing framework called DeepSpeech which proposes a hybrid CTC/attention architecture to utilize both advantages in decoding [34].

The testing is done on spontaneous Japanese and Mandarin Chinese datasets. For getting better results, the train set is expanded by linearly scaling the audio lengths by factors of 0.9 and 1.1. It eventually achieved a WER of 29.9% which is better than systems using only CTC.

Using CTC in deep learning based speech recognition is also researched by many others, such as [35] [36] [37] [38] [39].

3.4 Hybrid CTC/attention-based deep learning based architecture

Recent improvements in Urdu LVCSR is a paper from 2010 by Agha Ali Raza which uses the traditional method for solving Urdu language speech recognition [23] [24].

This paper describes a speech-to-text transcription system for semi-spontaneous speech. The system is based on the Kaldi framework and uses deep neural network based hidden Markov models (DNN-HMM) as main acoustic models. For restoring the final lattices, the system uses

neural network based phone duration models, which gives significant improvements over the basic DNN-HMM architecture.

For training the model, over 100 hours of speech was transcribed and used. The audio contains various speakers and no special processing has been made with it. This system achieves WER of 17.9% on broadcast conversations and 26.3% on conference speeches.

CHAPTER 4: MATERIAL AND METHODS

This chapter specifies the solution part of the thesis. It gives a detailed overview of the methods used with training the deep learning based speech recognition system. The deep learning based architecture DeepSpeech is explained.

4.1 Flow Diagram

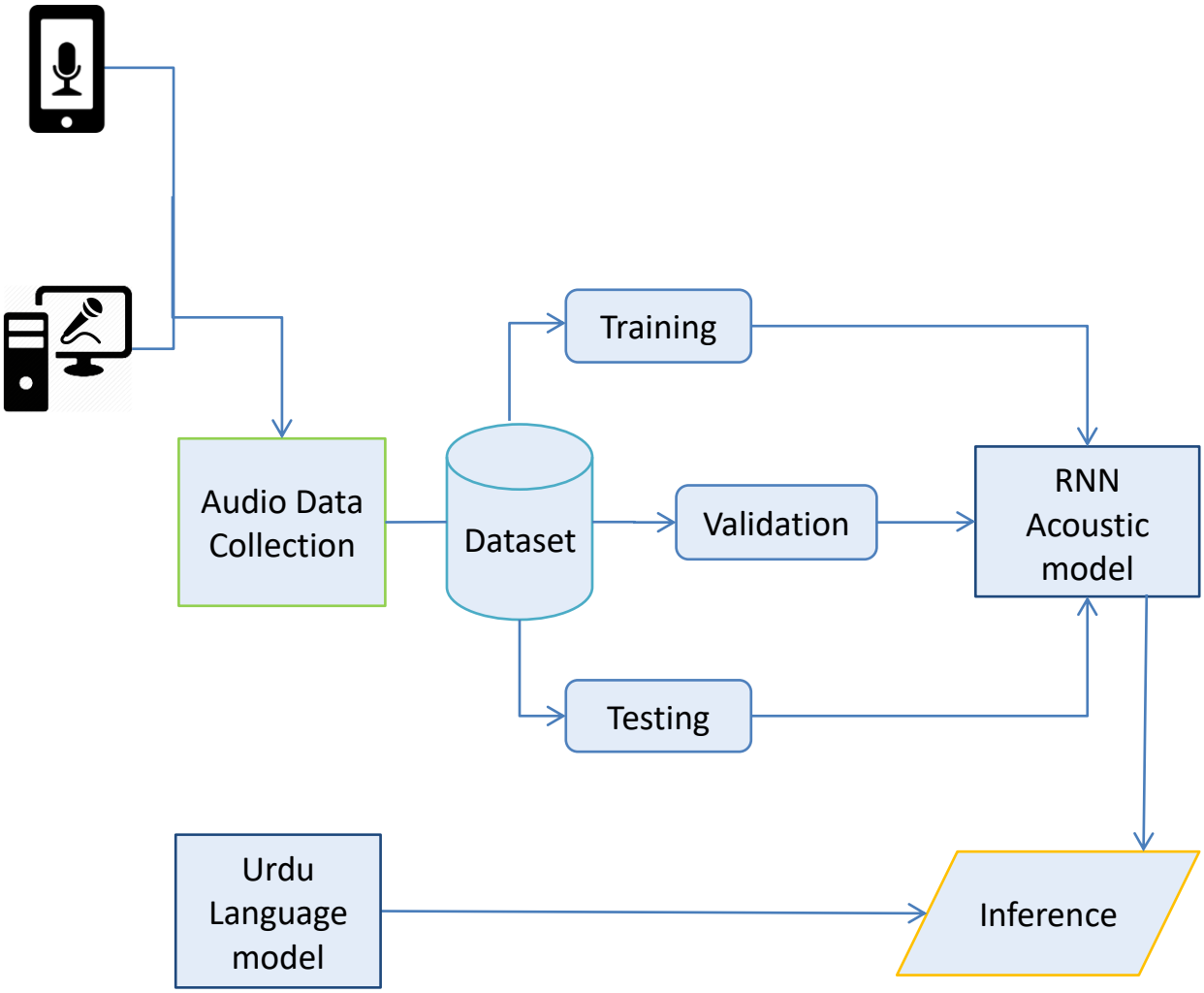

Figure 4.1: Flow Diagram

4.2 Data Collection

Initially we've used the dataset presented by Center for Speech and Language Technologies which consists of 70 minutes of speech data involving 708 Urdu sentences. Problem with this data was it's a single speaker speech data and we need to collect multiple instances from multiple speakers both male and female to train it with our Deep Neural Network for better accuracy.

We developed a desktop application and feed it with that 708 sentences to record the voices of multiple speakers and distributed it to more than 400 different university students both male and female. Data collection back from students was a difficult task there for we gave an option in application to upload recorded voice data to cloud so, we could collect all data from one place.

Almost 300K instances of speech data was collected but there is a big issue with it, because the sentences were very difficult to read and speak, most of the data was incorrect and didn't match with their transcriptions. Also it was not easy for students to stick with their desktops/laptops all the time to record their sentences and voice quality was also not very good.

Table 4.1: Sample transcriptions (CSaLT)

Urdu Transcription	No.
نیلم نے سالگرہ پر ہیڈ سیسموگراف اسود قریشی کے ماتھے پر اینٹھن اور غم کی آتشیں رو محسوس کی	1
حاجی مجاہد بلگرامی مخزن اور غزوہ کے ایک ارب قارئین میں انتہائی صادق اور جینوئن قاری تھے	2
سامعین انفارمیشن کی گھن گرج سنیں تو ویزے کی رپورٹ میں پوشیدہ ایک محدود ایل وی ڈومیسٹک پیکج ہے	3
کمیونسٹ لوگوں نے تنگ ہونے کے باوجود کئی شعبوں میں تندہی سے اپنے کیریئر کو مزین کر لیا	4
ٹرانسفارمر پر مڈنائٹ میں شاہین گدھ اور عقاب سمیت چیسٹ کے بل سرعام سینکڑوں برڈ بیٹھتے ہیں	5
کڑوے قہوے کا شیدائی اصغر کاشمیری باغبانی سیکھنے والا پانچواں منیجر ہے	6
پٹھوں اور سرکا حادثاتی درد بام سے فوراً سلجھائیں پھر میٹھے اور کولیسٹرول سے بچیں	7
کیولری ڈیم کی موزونیت سمجھانے اور ادارے کے ارکان کی تربیتی خدمات کے لیے حبیب نے تحریک چلائی	8
اقتصادی معاملات کی تفہیم اور فرماں بردار نوجوانوں کی بریفنگ گفتگو میں سرفہرست لکھیں	9
سجاد ریسکیو معلومات کھنگالنے کے بعد جانے کیوں بلاسوچے اچانک سوئمنگ پر راغب ہے	10
حامد سب ڈیلروں کے سامنے ان ویڈیوز کے بدلے میں اپنے اعزہ کو ایک ماچس بھی نہیں بخشے گا	11
تم ڈائیلوشن کے عمل سے بچنے کے لیے نتھنوں سے گاؤ کیونکہ جبڑا سوجنے سے تلخی اور سوجن ہو گی	12
شجر کی بیرونی شاخ پر گیارہ ڈوڈیاں توازن کی درست مثال ہیں جو سنگھاسن بیضے کے مترادف سمجھی جاتی ہیں	13
ڈیڈ پچز پر کیچ میچ کا رخ پھیریں تو نتیجے میں بیشتر لوگ خوشی سے رو کر بھی قیامت ڈھا دیتے ہیں	14
مملکتیں بجھی اور بے ڈھب موجوں یا صیہونی ڈاج سے رعشہ زدہ نیشن کو ٹرانسپیرنسی نہیں دے سکتی	15
ریاست کے چوالیس تعلقوں میں صیغہ بسائیں تو روہی اور متھرانی زمین مزارعوں کو پسند ہے	16
سرویئر نے انستھیزیا اور مالیخولیا میں پائیریسی سے بچاؤ کے تھیم میں خود کو منوا لیا ہے	17
زیست کے چنگل کے اس فیز میں جیون ایک افلاس میں پھنسی پرنم آنکھ کی طرح ہوتا ہے	18

19	برسراقتدار اہل دانش سے قسم یا حلف کی گٹھڑی اٹھوا لو تو فرض اولی چُھپ جاتا ہے
20	روضہ کی دیسی ساختہ فیبرکس کا ذریعہ انٹرفیس میں درج ہے
21	مونا اس سفر میں الجھ کر سنبھلی اور ٹھہرنے میں کامیاب رہی
22	اخوندزادہ نے ساتویں سنچری نئے طرزعمل اور استقلال سے بنائی اور اؤٹ ہو گیا
23	ایکسیئن نے اسی موشن میں ایمبیسی کی تحویل میں گھسنے کی کوشش کی
24	ہوشربا کڑوی تبدیلیاں عیاں اور سازباز کی جڑیں ملغوبہ انداز کی ہیں
25	ننانوے ملین متاثرین سے پوچھے بغیر آپ کا تعاون معاف نہیں ہو سکتا
26	وفاق میں منظم آتشزدگی پر چھیانوے بااصول ملازمین کا استفسار بالکل چاہیے
27	حملوں سے مجید صاحب کا بیس بیڈروم کا گھر اتھل پتھل ہو گیا ہے
28	کانگرس اسی کوڈ میں انسان کی موزوں قدروقیمت ظاہر کرتی ہے
29	عابد اسے جھڑک کر بھیجے اور ریٹائرمنٹ کے بعد اس کی کفالت سے مستغنی ہو جائے
30	بپھرے ہوے لوگوں کو چھوئیں اور ننگے الفاظ سے سنگباری نہ کریں
31	ہمارا پیغام ہے نماز پڑھو کاجو کھاؤ اور اپنی عمریں علم کے لیے صرف کرو
32	خانقاہوں میں زیوں حال نشئی غلیل چوری کرنے کے لیے چھڑیاں اٹھائے کھڑے ہیں
33	ویسٹ فن لینڈ میں باؤلر جھٹکے سے زیب کو کچو کے لگاتا رہا اور رنز بٹورتا رہا
34	کانگریس حکومت یہ پورا قومیائے تو خدوخال ابھریں اور پَلوں میں دھبے مٹیں
35	چوزوں کا جھلس کر بیضوی لاشیں بن جانا صلاحیت یا رولز کی کوتاہی ہے
36	نیم بحرانی مورخین کی تجرد میں قید ایک خواستگار کی توپن ہے
37	ایڈونچر میں گیَر لگانا گھوڑے دوڑانے سے آسان اور سیف سمجھو
38	زیڈ گھمن نے دو مرلہ پر سنگھاڑے اگائے تو بچہ کھلکھلا کر پھدکتا ہوا وہاں آ گیا
39	شہباز کو ہیجڑوں کے شوکیسوں کی پہچان ہے
40	یورپی نوبل شیئرنگ ایمبولینس کے نشر کے لیے یہ پیچیدہ ساعت ہے
41	بوڑھوں کے لیے نل کی دیکھ بھال ایک تذلیل ہے اور ناقابل افورڈ ہے
42	رینگتے ہوے اس مہم میں سوشلزم کے لیے بغاوت کا علم بلند کیا
43	مواضعات میں بنگالی مندری کا تھیوا اور سوئچ بیچنے جایا کرتے ہیں
44	پھونک سے بھینس کے تھنوں کی سرخی قدرے رخصت ہوئی
45	صحبتوں میں گالف اور آنکھ مچولی جوق در جوق کھیلتے ہیں
46	وعدے وعید کے بعد شوکت کی نمک خواری رائگاں جانا پرلطف ہے
47	میت کو روڈ سے قبرستان بھجوائیں اور درود پڑھنے کی کوشش کریں
48	دوغلا شخص جونیجو ہے تو سزا کے لیے پھانسی نامزد کرو
49	گردش کے اوپر گائنی وارڈ میں ریسورس ٹریبیونل ہے
50	تبسم جمخانہ میں کیموتھراپی کی توضیحات پر بحثتا ہے

We decided to create simple readable and speak able Urdu transcriptions and for speaker convenience we also developed a mobile application to record and collect speech data. We created a list of 100 random sentences and after that 200 commonly used sentences in daily life. Almost 90K instances of speech data was collected of these two lists. We verified the data and discarded the voice data with bad quality and wrongly spoken voices.

Table 4.2: Sample transcriptions (Collected)

Urdu Transcription	No.
اللہ سب سے بڑا ہے	1
محمد صلی اللہ علیہ و آلہ وسلم اللہ کے رسول ہیں	2
اسلام علیکم صبح بخیر	3
وعلیکم السلام	4
خوش آمدید کیا حال ہے	5
میں ٹھیک ہوں شکریہ اور آپ کیسے ھیں	6
اچھا سب کچھ ٹھیک ہے	7
بہت بہت شکریہ	8
کوئی نئی خبر نہیں	9
دوست سنو کوی بات نہیں	10
مجھے آپ کی بہت کمی محسوس ہوی	11
شب بخیر پھر ملیں گے خدا حافظ	12
لگتا ہے میں کھو گیا ہوں	13
کیا میں آپ کی مدد کر سکتا ہوں	14
کیا آپ میری مدد کر سکتے ہیں	15
وہ میری مدد کو نہ آسکے	16
باتھ روم واش روم کہاں ہے	17
میڈیکل سٹور کہاں ہے	18
سیدھے جا کر دائیں بائیں مڑجائیں	19
میں کسی کو ڈھونڈ رہا ہوں	20
یہ لڑکی کس کو ڈھونڈ رہی ہے	21
یہ لڑکا بہت نیک اور شریف ہے	22
برائے مہربانی کچھ دیر انتظار کیجئے	23
برائے مہربانی ہولڈ کیجیے	24
یہ والا کتنے کا ہے	25
معذرت چاہتا ہوں ذرا سنیے	26
معاف کرنا دوست	27
میرے ساتھ چلیے	28
کیا آپ انگریزی یا اردو بول سکتے ہیں	29
جی بالکل صرف تھوڑی سی	30
آپ کا نام کیا ہے	31
میرا نام وقاص ہے	32
محترم محترمہ مس مسٹر سب موجود ہیں	33
آپ سے مل کر خوشی ہوئی	34
آپ بہت مہربان ہیں	35
آپ کا تعلق کہاں سے ہے	36
میرا تعلق پاکستان اور امریکہ سے ہے	37
میں پاکستانی اور امریکی ہوں	38
آپ کہاں رہتے ہو	39

40	میں اپنے گھر رہتا ہوں رہتی ہوں
41	کیا آپ کو یہاں آکر اچھا لگا
42	یہ ایک اچھا ملک ہے
43	دنیا بہت خوبصورت ہے
44	آپ کیا کام کرتے ہیں
45	میں جاب اور بزنس کرتا ہوں
46	میں مینیجر اور تاجر ہوں
47	مجھے اردو اور انگریزی اچھی لگتی ہے
48	میں کافی عرصے سے یہ زبان سیکھ رہا ہوں
49	دن مہینے سال لگتے ہیں
50	واہ بہت خوب بہت اچھے
51	آپ کی عمر کتنی ہے
52	میری عمر تیس سال ہے
53	مجھے اب جانا ہے اجازت چاہتا ہوں
54	جلدی واپس آؤں گا میرا انتظار مت کرنا
55	یہ گھڑی آپ کو اچھی لگے گی
56	مجھے گنتی سناؤ
57	ایک دو تین چار پانچ چھے سات آٹھ نو دس گیارہ بارہ
58	بیس تیس چالیس پچاس ساٹھ ستر اسی نوے سو
59	ہزار لاکھ کروڑ ارب کھرب
60	اللہ کا فضل اور شکر ہے
61	سالگرہ مبارک ہو مزے کرو
62	نیا سال مبارک ہو
63	میں ایک دن یہاں جانا چاہتا ہوں چاہتی ہوں
64	سب کو میرا سلام کہنا
65	اللہ رحم کرے ہم سب پر
66	شب بخیر اور سہانے خواب
67	معذرت چاہتا ہوں یا چاہتی ہوں
68	معاف کرنا مجھے
69	کوئی بات نہیں معافی کی ضرورت نہیں
70	کیا آپ اسے دوبارہ اور آہستہ کہہ سکتے ہیں
71	برائے مہربانی اپنا کام لکھیے
72	میں سمجھا نہیں
73	مجھے نہیں معلوم
74	مجھے اندازہ نہیں تھا
75	اسے انگلش میں کیا بولتے ہیں
76	یہ کیا ہے سب کچھ خراب ہے
77	پریشان نہیں ہوں سب ٹھیک ہو جائے گا
78	مجھے اس کی مشق کرنی چاہیے
79	اچھا ہو یا برا ہے وہ معمولی سا
80	بڑا چھوٹا سب برابر ہے

81	مجھے گاڑی چاہیے آج اور ابھی
82	آنے والا کل گزرے ہوئے کل سے بہتر ہے
83	ہاں یا نہ میں جواب دو
84	یہ لیجیے ہوگیا سارا کام
85	کیا یہ آپ کو پسند آیا
86	مجھے تو بہت پسند آیا
87	مجھے بھوک اور پیاس لگ رہی ہے
88	میں کام کرتا ہوں صبح میں شام میں رات میں
89	یہاں بھی بارش ہو رہی ہے اور وہاں بھی
90	جلدی کرو یار
91	سچ میں تم پاگل ہو
92	یہ ایک لڑکے اور لڑکی کی کہانی ہے
93	میں اور آپ اسے مل کر سنتے ہیں
94	ابھی کیا وقت ہوا ہے
95	تمہارے منہ پر بارہ کیوں بجے ہیں
96	مجھے یہ دیجیے آدھے جملے ہو گے
97	میں اپنے امی ابو سے پیار اور محبت کرتا ہوں
98	یہ میرے دادا دادی نانا نانی ہیں
99	تم سب میرے بہن بھائی ہو
100	یہ میرے چچا چاچی تایا تائی ہیں
101	یہ میرے خالہ خالو ماموں ممانی ہیں
102	یہ بیوی اور شوہر کے آپس کا معاملہ ہے
103	میرے ساس اور سسر بہت اچھے ہیں
104	تمہارا سسرال تو قریب ہی ہے
105	بیٹا ہو یا بیٹی اولاد نیک ہونی چاہیے
106	والدین کی عزت اور خدمت کرو
107	میری طبیعت خراب ہے
108	میرے لیے دعا کیجئے
109	وہ بلڈ پریشر اور شوگر کا مریض ہے
110	مجھے ڈاکٹر کی ضرورت ہے
111	مجھے انصاف چاہیے
112	سیاستدانوں نے ہمیں ٹرک کی بتی کے پیچھے لگایا ہے
113	لائٹ چلی گئی
114	لائٹ آگئی
115	مجھے مختلف پھلوں کے نام بتاؤ
116	سیب انگور کیلا کنو آم انار چکوترا کھجور
117	مجھے کھانے کھانا اور پانی پینا ہے
118	میرے پاس سب سواریاں ہیں
119	سائکل موٹر سائیکل کار بس ٹرک
120	ہوائی جہاز بحری جہاز بیرون ملک سفر کے لئے استعمال ہوتے ہیں
121	مجھے سبزیوں کے نام بھی یاد ہیں

122	آلو گوبھی مٹر گاجر شلجم مولی کھیرا کدو ٹینڈے بھنڈی توری پالک
123	جنگل میں جانور اور چرند پرند ہیں
124	بلی کتا شیر گدھا الو بندر لومڑی چڑیا طوطا مور
125	باہر شدید گرمی ہے
126	آج تو کافی ٹھنڈ ہے
127	پاکستان میں سب ھی موسم پائے جاتے ہیں
128	موسم گرما سرما بہار خزاں
129	طوفانی بارش برس رہی ہے
130	برف باری بھی پڑ رہی ہے
131	سٹوڈنٹ موبائل پر فیس بک اور واٹس ایپ یوز کر رہے تھے
132	میں فارغ وقت میں لیپ ٹاپ یوز کرتا ہوں وہ ٹی وی دیکھتا ہوں
133	چائے پینے کا موڈ ہو رہا ہے
134	ادھر کا کھانا بہت مزے کا ہے
135	مسلمان مصیبت میں گھبرایا نہیں کرتے
136	گریپ فروٹ جوڑوں کے درد میں مفید ہے
137	جسم سے فاسد مواد کا اخراج کرتا ہے
138	وزن کم کرنے میں مدد کرتا ہے
139	بے خوابی میں مبتلا لوگ اس کا استعمال ضرور کریں
140	کھانسی اور گلے کی خراش میں کھانا فائدہ مند ہے
141	معدے کی تیزابیت دور کرتا ہے
142	موٹاپے کو کم کرنے میں مدد گار ہے
143	چھاتی کے کینسر میں فائدہ مند ہے
144	لبلبے کے سرطان میں مفید ہے
145	شریانوں کی رکاوٹ کو دور کرتا ہے
146	اسٹرابری دانتوں اور ہڈیوں کی کمزوری کو دور کرتی ہے
147	سوجن میں کمی لاتی ہے
148	نزلہ زکام کا خاتمہ کرتی ہے
149	کینسر کے خلاف قوت مدافعت کو بڑھاتی ہے
150	امراض چشم میں فائدہ مند ہے
151	اس کے کھانے سے بلڈ پریشر نارمل رہتا ہے
152	کولیسٹرول کی سطح کو نارمل رکھتی ہے
153	جھریاں پڑنے سے روکتی ہے
154	جوڑوں کے درد اور گنٹھیا میں مفید ہے
155	پیاس کو کنٹرول کرتی ہے
156	گاؤں اور شہر ایک ہو رہے ہیں
157	سکول کالج یونیورسٹی دفتر بینک سب کھل گئے ہیں
158	لوگ مہنگائی سے تنگ آکر خودکشی کرنے پر مجبور ہیں
159	ڈاکٹر اور مریض میں تلخ کلامی ہوئی
160	مولوی مفتی قاری مسجد میں بیٹھے ہیں
161	اپنے اساتذہ کا احترام کریں
162	پولیس غنڈہ گردی پر اتر آئی

163	عدالت میں وکیل اور جج بیٹھے ہیں
164	اللہ بہتر جانتا ہے
165	میرے سپروائزر بہت محنتی ہیں
166	لڑکوں کے نام عبداللہ علی عمر ابوبکر عثمان ہیں
167	لڑکیوں کے نام عائشہ صبا مریم زینب ہیں
168	انشاءاللہ کامیابی تمہارے قدم چومے گی
169	نماز ہم پر فرض ہے
170	فجر ظہر عصر مغرب عشاء
171	نزدیک کوئی قریبی ہسپتال ہے
172	اکرم صاحب کا بیٹا انجینئر اور بیٹی ڈاکٹر ہے
173	جو ہوتا ہے بہتر ہوتا ہے
174	ہینڈ بیگ اور سوٹ کیس یاد سے اٹھا لینا
175	ریموٹ نہیں مل رہا
176	میری جرابیں نہیں مل رہی
177	کمپیوٹر ہارڈ وئیر اور سافٹ وئیر سے چلتا ہے
178	درخت اور پھول فضا کو صاف اور شفاف رکھتے ہیں
179	شہری فیکٹری کے دھویں اور گردوغبار سے تنگ ہیں
180	کچھ عقل کرو
181	مجھے مرغی کا گوشت اور انڈے پسند ہیں
182	عید کا چاند نظر آگیا ہے
183	ہم اس بار بکرے اونٹ گائے کی قربانی کریں گے
184	بچے عیدی مانگتے ہیں
185	عید مبارک خیر مبارک
186	سڑکیں پل اور انڈر پاس بن رہے ہیں
187	کراچی لاہور فیصل آباد اور ملتان بڑے شہر ہیں
188	پنجاب سندھ خیبر پختونخوا گلگت بلتستان پاکستان کے صوبے ہیں
189	اللہ ہمارے ملک کو قائم و دائم رکھے
190	پاکستان زندہ باد
191	پاکستان پائندہ باد
192	اسلام آباد خوبصورت شہر ہے
193	اور پاکستان کا دارالحکومت ہے
194	آنکھ کان ناک بال چہرے کا حصہ ہیں
195	ٹانگیں بازو سر دھڑجسم کا حصہ ہیں
196	سر درد کر رہا ہے
197	تعاون کے لئے آپ کا شکریہ
198	اللہ حافظ خدا حافظ اللہ نگہبان
199	یہ آخری جملہ ہے آپ کا شکریہ
200	اللہ ایک ہے اور اس کا کوئی شریک نہیں

4.3 Data Preparation

Data preparation is required to bring collected data in the style allowed by DeepSpeech. Two commonly used formats are LibriSpeech and CommonVoice. We created utility application to bring our collected data to that formats which includes conversion of files, moving files, renaming files and some validation tasks such as audio files are not corrupted or empty, names of audio files and transcription files are matching etc.

4.3.1 LibriSpeech:

For the voice data we collected in first phase we used LibriSpeech format. It requires separate sets for training, validation and testing. Each audio file in these three sets must have its own transcription file. Audio file must be in WAV format, single channel (mono), sample rate must be 16000 Hz with depth of 16 bit and transcription must be in TXT file format. So, if we have 300K audio files of speech data we must have their respective transcription in same quantity.

4.3.2 CommonVoice:

For the voice data collected in second phase we used CommonVoice format. It is the most convenient dataset format for DeepSpeech. Same like LibriSpeech it requires separate sets for training, validation and testing. Against each set a CSV file is created in which information regarding audio file physical path, audio file size and related Urdu transcription is written. For CommonVoice audio files must be in WAV file format, single channel (mono) and 16000 Hz sample rate with depth of 16 bit. CSV's are built with the help of our utility application.

Table 4.3: Datasets LibriSpeech vs CommonVoice

	LibriSpeech	**CommonVoice**
Audio files	300 K	90 K
Length	460 hrs.	75 hrs.
Sample rate	16000 Hz	16000 Hz
Channel	Mono	Mono
Depth	16 bit	16 bit

Verified	No	Yes
Collection Method	Desktop Application	Mobile Application
Online Storage	Dropbox	Google Cloud

Table 4.4: Datasets Distribution

Dataset	Training	Validation	Testing
Split	70%	20%	10%

Table 4.5: Speakers Distribution

	Male	**Female**	**Total**
Speakers	165	235	400
	41.25%	58.75%	

4.4 Language Model

DeepSpeech implements a probabilistic language model to increase the precision of speech recognition. This Urdu language model is basically a dataset that comprises approximation probabilities of word sequences in Urdu language. Every sequence sorts in length from 1 to N and assigned to it a number which relates to the probability of that sequence. Word arrangements that comprise of N words are termed n-grams.

The extreme number N that relates to the size of a word arrangement describes the measurement of that model, if the size of an arrangement surpasses model's measurement or arrangement in a dataset not originated, probability of that arrangement is predictable by reference to probability of smaller arrangements it comprises of. Making queries to language model DeepSpeech uses KenLM toolkit.

4.4.1 KenLM

To make language model and trained it with KenLM [25] we created an Urdu alphabet file which contain one Urdu character per line and an empty space to represent spaces. We also created

an Urdu vocabulary file in while we put whole Urdu transcriptions, one line per transcription. Both files are in TXT format. Kenlm employs smoothing technique to adjust n-grams to make better estimation of most probable Urdu sentences.

In execution KenLM attested to be more time and memory proficient. The trigram language model for Urdu is constructed and compiled it into binary file format for quick loading. Assessment was done by multiple Urdu sentences with constructed language model and confident scores was achieved.

4.5 DeepSpeech

DeepSpeech is an open source speech to text engine based on Baidu's DeepSpeech research paper using model trained by machine learning practices. DeepSpeech uses Google's TensorFlow to mark the operation easier. Pre built binaries for execution implication with a qualified model can be installed with pip command. Proper arrangement using a virtual setting is suggested.

A pre qualified English model is accessible for consumption and can be downloaded. Presently only 16 kHz, 16-bit and mono channel WAVE audio files are maintained in the Python. When everything installed you can use deepspeech binary to prepare speech to text on short roughly 5 second lengthy audio records:

```
pip install deepspeech
deepspeech --model models_test/output_graph_test.pbmm --alphabet
models_test/alphabet_test.txt --lm models_test/lm_test.binary --trie
models_test/trie_test --audio my_audio_file_test.wav
```

Instead faster inference could be done using a supported NVIDIA GPU on Linux. To run deepspeech on a GPU connect the GPU precise package:

```
pip install deepspeech-gpu
deepspeech --model models/output_graph_test.pbmm --alphabet models/alphabet_test.txt
--lm models/lm_test.binary --trie models/trie_test --audio my_audio_file_test.wav
```

4.5.1 Architecture

DeepSpeech network consists of five layers in which the input is fed into first three fully connected layers which trailed by a bidirectional RNN layer and last network layer is a fully connected layer for output. It requires audio dataset along with their transcripts as input. The output of DeepSpeech network is matrix of character’s likelihoods over time means network for each time outputs single likelihood for each character in the alphabet which denotes the probability of

character consistent to what's in the audio being said at that time step. CTC loss function reflects all arrangements at similar time of audio to transcription permitting to exploit likelihood of accurate transcription projected deprived of disturbing arrangement. Lastly training is done using Adam optimizer. Architecture is presented with the help of diagram in Figure 11 and 12.

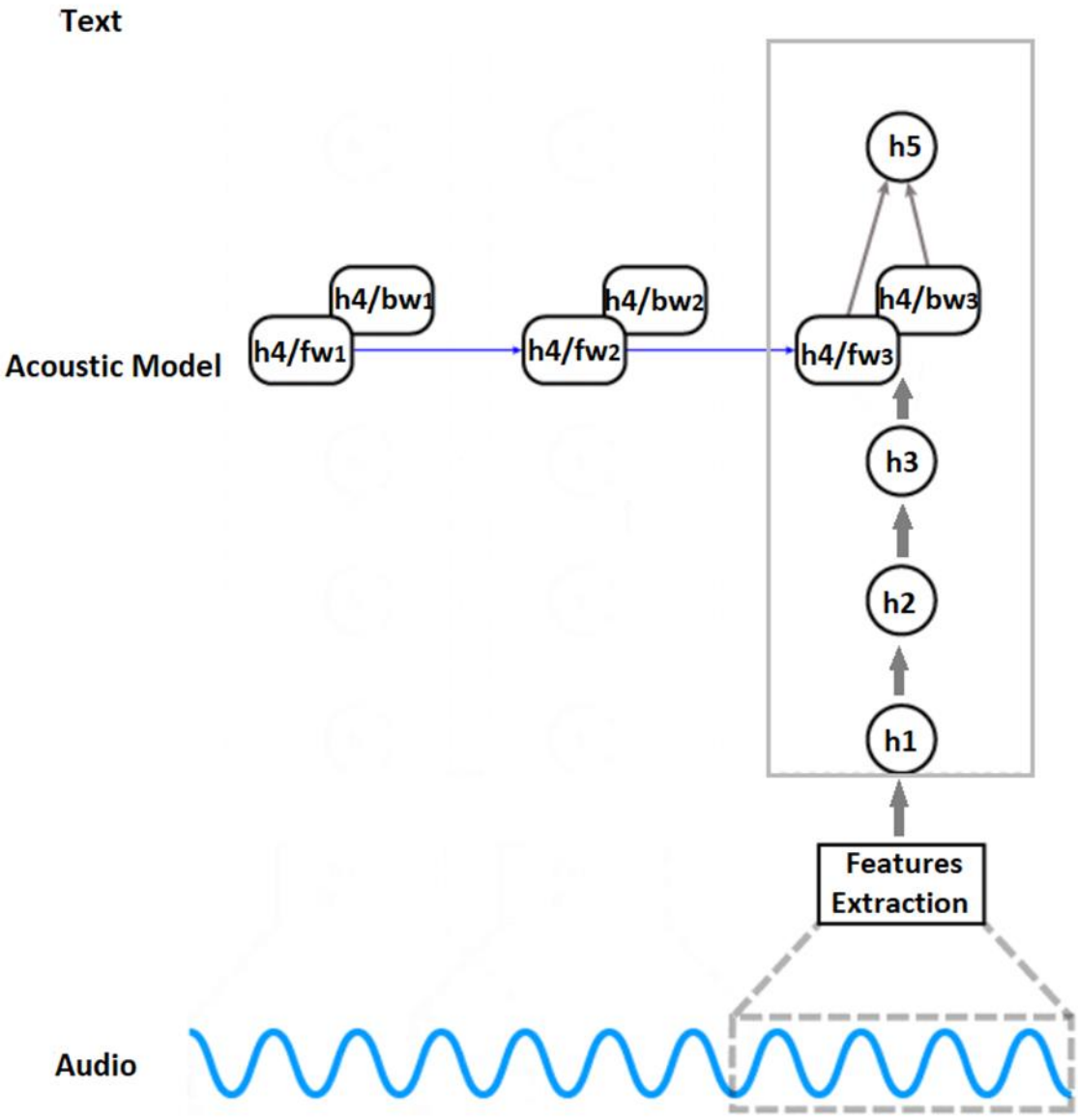

Figure 4.2: Architecture – Training network

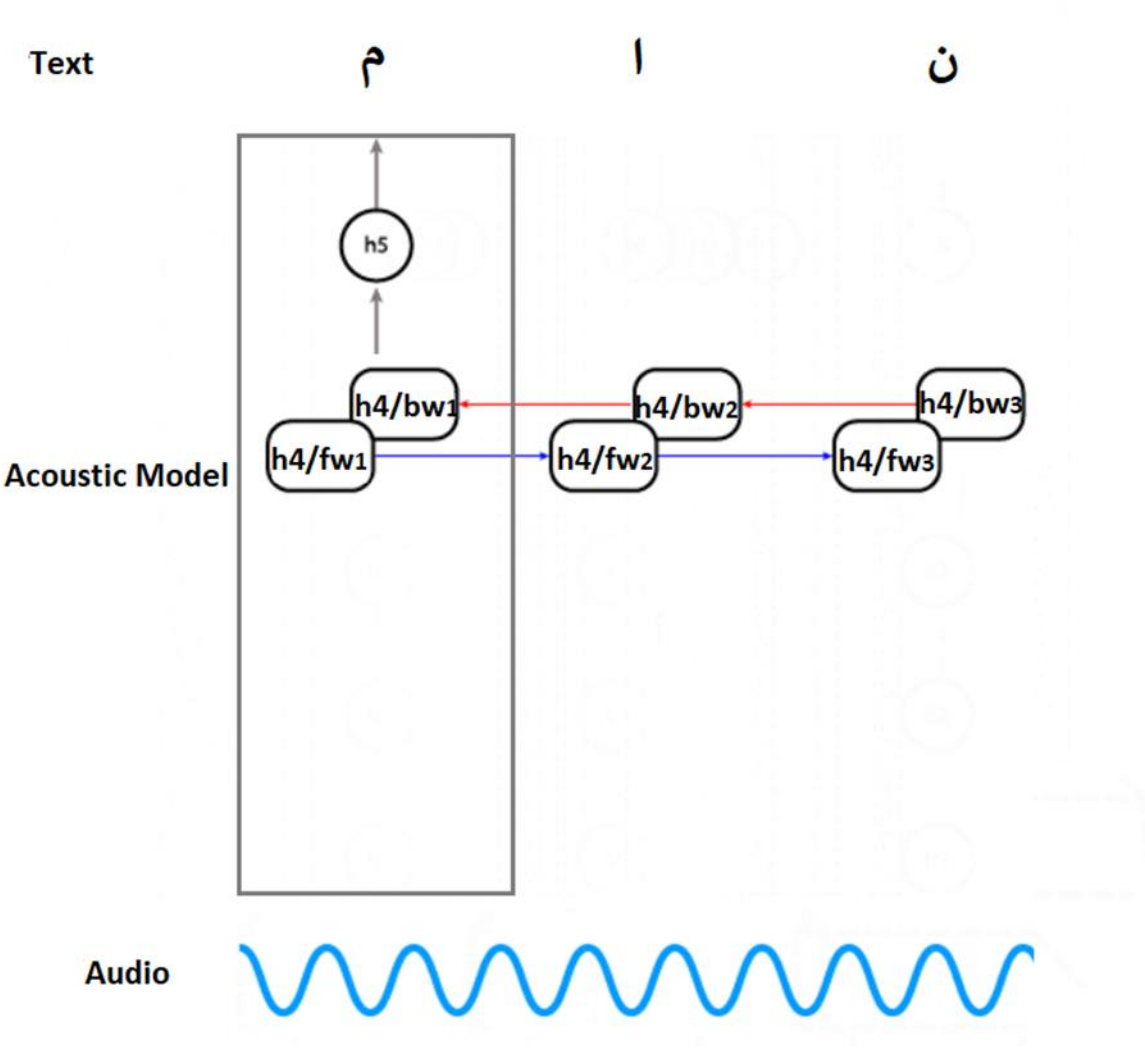

Figure 4.3: Architecture – Decoding step

4.5.2 Features Extraction

Features extraction assumes a key job in ASR frameworks. Features are special identifiers by which a discourse can be perceived. Complete discourse can't be handled, in light of the fact that it requires computational costly preparing which could bring about failure of ASR framework. Features are a diminished trademark portrayal of a discourse articulation. There are some notable features in writing for discourse acknowledgment, for example, Linear Predictive Coding (LPC), Perceptual Linear Predictive Coding (PLP) and Mel Frequency Cepstral Coefficients (MFCC) which is utilized by DeepSpeech.

4.5.3 Requisites

- Python 3.6
- Git Storage for large files
- Linux or Mac environment

4.5.4 Receiving code

Install Git Large File Storage either through a package manager or manually if available on your system. Then duplicate the DeepSpeech source normally:

```
git clone https://github.com/mozilla/DeepSpeech
```

4.5.5 Installing Training Prerequisites

Install the prerequisite dependences using `pip`:

```
cd DeepSpeech
pip install -r requirements.txt
```

You'll likewise need to introduce the ds_ctcdecoder Python bundle. ds_ctcdecoder is required for deciphering the yields of the deepspeech acoustic model into content. You can utilize util/taskcluster.py with the - decoder banner to get a URL to a twofold of the decoder bundle fitting for your stage and Python rendition:

```
pip install $python util/taskcluster.py --decoder
```

This order will download and introduce the ds_ctcdecoder bundle. In the event that you lean toward structure the pairs from source. You can abrogate the stage with - curve in the event that you need the bundle for ARM7 --arch arm or ARM64 --arch arm64.

4.5.6 Recommendations

On the off chance that you have a skilled (NVIDIA, at any rate 8GB of VRAM) GPU, it is profoundly prescribed to introduce TensorFlow with GPU support. Preparing will be essentially quicker than utilizing the CPU. To empower GPU support, you can do: Please guarantee you have the required CUDA reliance.

```
pip uninstall tensorflow
pip install 'tensorflow-gpu==1.14.0'
```

4.5.7 Training a model

The focal (Python) content is DeepSpeech.py in the undertaking's root catalog. For its rundown of order line alternatives, you can call:

```
./DeepSpeech.py --helpfull
```

To get the yield of this in a somewhat better-arranged way, you can likewise look into the choice definitions top DeepSpeech.py.

For executing pre-arranged preparing situations, there is a gathering of accommodation contents in the canister envelope. The vast majority of them are named after the corpora they are designed for. Remember that the other discourse corpora are huge, on the request for several gigabytes, and some aren't free. Downloading and preprocessing them can take quite a while, and preparing on them without a quick GPU (GTX 10 arrangement prescribed) takes much more.

In the event that you experience GPU OOM blunders while preparing, take a stab at decreasing the clump size with the - train_batch_size, - dev_batch_size and - test_batch_size parameters. As a basic first model you can open a terminal, change to the registry of the DeepSpeech checkout and run:

```
./bin/run-ldc93s1.sh
```

This content will prepare on a little example dataset called LDC93S1, which can be overfitted on a GPU in no time flat for exhibit purposes. From here, you can adjust any factors concerning what dataset is utilized, what number of preparing cycles are run and the default estimations of the system parameters. Feel likewise allowed to pass extra (or superseding) DeepSpeech.py parameters to these contents. At that point, simply run the content to prepare the adjusted system.

Each dataset has a relating merchant content in container/that can be utilized to download (if it's uninhibitedly accessible) and preprocess the dataset. See canister/import_librivox.py for a case of how to import and preprocess an enormous dataset for preparing with DeepSpeech.

On the off chance that you've run the old merchants (in util/shippers/), they could have evacuated source documents that are required for the new merchants to run. All things considered, essentially evacuate the removed organizers and let the shipper concentrate and procedure the dataset sans preparation, and things should work.

4.5.8 Checkpointing

During preparing of a model alleged checkpoints will get put away on plate. This happens at a configurable time interim. The reason for checkpoints is to permit interference (additionally on account of some startling disappointment) and later continuation of preparing without losing long stretches of preparing time. Continuing from checkpoints happens consequently by simply (re)starting preparing with the equivalent - checkpoint_dir of the previous run.

Know anyway that checkpoints are legitimate for a similar model geometry they had been produced from. As such: If there are blunder messages of specific Tensors having contrary measurements, this is in all probability because of a contradictory model change. One normal way out is wipe all checkpoint records in the checkpoint catalog or transforming it before beginning the preparation.

4.5.9 Exporting model for inference

On the off chance that the - export_dir parameter is given, a model will have been traded to this catalog during preparing. In the event that you need to explore different avenues regarding the TF Lite motor, you have to send out a model that is perfect with it, at that point utilize the - export_tflite banners. In the event that you as of now have a prepared model, you can re-send out it for TFLite by running DeepSpeech.py again and determining the equivalent checkpoint_dir that you utilized for preparing, just as passing - export_tflite - export_dir/model/trade/goal.

4.5.10 Making mmap-able model for inference

The output_graph.pb model record produced in the above advance will be stacked in memory to be managed when running induction. This will bring about additional stacking time and memory utilization. One approach to maintain a strategic distance from this is to legitimately peruse information from the circle.

TensorFlow has tooling to accomplish this: it requires building the objective/tensorflow/contrib/util:convert_graphdef_memmapped_format (doubles are delivered by our TaskCluster for certain frameworks including Linux/amd64 and macOS/amd64), use util/taskcluster.py apparatus to download, indicating tensorflow as a source and convert_graphdef_memmapped_format as ancient rarity. Delivering a mmap-capable model is as straightforward as:

```
$ convert_graphdef_memmapped_format --in_graph=output_graph.pb --
out_graph=output_graph.pbmm
```

Upon sucessfull run, it should report about transformation of a non-zero number of hubs. In the event that it reports changing over 0 hubs, something isn't right: ensure your model is a solidified one, and that you have not connected any contrary changes (this incorporates quantize_weights).

4.5.11 Continuing training from a release model

On the off chance that you'd like to utilize one of the pre-prepared models discharged by Mozilla to bootstrap your preparation procedure (move adapting, calibrating), you can do as such by utilizing the - checkpoint_dir banner in DeepSpeech.py. Determine the way where you downloaded the checkpoint from the discharge, and preparing will continue from the pre-prepared model.

For instance, in the event that you need to tweak the whole diagram utilizing your very own information in my-train.csv, my-dev.csv and my-test.csv, for three ages, you would something be able to like the accompanying, tuning the hyperparameters as required:

```
mkdir fine_tuning_checkpoints
python DeepSpeech.py --n_hidden 2048 --checkpoint_dir path/to/checkpoint/folder --
epochs 3 --train_files my-train.csv --dev_files my-dev.csv --test_files my_dev.csv --
learning_rate 0.0001
```

4.6 Acoustic Model

After creating Urdu language model we trained our data with DeepSpeech architecture to create acoustic model. Resulting acoustic model for Urdu speech could be used for retraining, inferences and could be converted into light version to be used for mobile applications.

4.6.1 Training Parameters

For training multiple parameters need to be provided including training, validation and testing dataset paths, language model path including Urdu alphabet file and binary language model file, CSV's path containing transcriptions, checkpoint directory path where checkpoints are saved after some epochs so training process could be resume in case of any interruption, summary directory path where training summary is saved, export directory path where acoustic model is exported after training completion, batch size of training, validation and testing dataset which can

be set according to GPU's available memory which in our case was 50 for training and validation and 30 for testing.

4.6.2 Training Command

Python DeepSpeech.py --train_files urdu_voice_train.csv --dev_files urdu_voice_dev.csv --test_files urdu_voice_test.csv --checkpoint_dir checkpoints_urdu_v2 --train_batch_size 50 --dev_batch_size 50 --test_batch_size 30 --export_dir exportmodel_urdu_v2 --validation_step 1 --early_stop True --summary_dir summary_urdu_v2 --summary_secs 10 --n_hidden 512 --epoch 201 --earlystop_nsteps 201 --estop_mean_thresh 0.1 --estop_std_thresh 0.1 --dropout_rate 0.30 --default_stddev 0.046875 --learning_rate 0.0001 --report_count 100 --use_seq_length False --alphabet_config_path urdu_lm/alphabet.txt --lm_binary_path urdu_lm/lm.binary --lm_trie_path urdu_lm/trie

4.6.3 Hyper Parameters

Some hyper parameters needed to define for the fine tuning of training process for our Urdu acoustic model. These parameters including validation step which was 1 in our case mean for each training step validation step is executed to adjust network weights.

Other parameters including number of epochs which was 500, learning rate 0.0001, dropout rate 0.30, standard deviation 0.046875 and early stop is set to true which helped to escape training process for over fitting, early stop mean threshold 0.1 and early stop standard deviation 0.1. These parameters helped us to achieve WER less than 10%.

4.7 Training Platform

We used Google’s cloud platform for training because training related tasks needs huge computing power and sometime needs multiple GPU’s with extensive memory. We created a Linux virtual machine with following technical parameters:

- CPU: 8 vCPUs
- RAM: 32 GB
- GPU: 1 x NVIDIA Tesla P100 with 16 GB of memory
- Physical Memory: 160 GB
- OS: Ubuntu

CHAPTER 5: EXPERIMENTAL RESULTS

We've trained our Urdu acoustic model and found different results for that training process. We run total 500 epochs and calculate WER, CER and step loss. WER is considered as accuracy rate in speech recognition. Minimum WER means more accuracy and vice versa. While training till 350 epochs WER decreases but after that WER increases gradually, which means our model is fully trained on 350 epochs and after that model is moving to the state called overfitting.

5.1 Test Results

The best result for testing dataset we achieved is 9.75% WER after training 350 epochs. Results are presented in Table 5.1.

Table 5.1: Avg. Test WER, CER and Step Loss

Epochs	Testing (Avg.)		
	WER	CER	Loss
100	13.88%	2.16%	17.35%
200	12.34%	1.87%	16.65%
350	9.75%	1.38%	13.23%
500	10.58%	1.55%	14.43%

5.1.1 Speaker Dependent Test Results

These are speaker dependent test results which we have got from the model while training ends. As we have separated test data from the dataset, this data does not exist in training and validation datasets but similar speaker have voices in all datasets. Therefore we consider it speaker dependent test results. Results are presented in Table 5.2.

Table 5.2: Speaker Dependent Test Results

WER	Urdu Transcription		No.
	اللہ سب سے بڑا ہے	Original	1
11.76%	اللہ سب سے بڑا	Inference	
	محمد صلی اللہ علیہ و آلہ وسلم اللہ کے رسول ہیں	Original	2
6.52%	محمد صلی اللہ علیہ و آلہ وسلم اللہ کے رسول	Inference	

3	Original	اسلام علیکم صبح بخیر	
	Inference	اسلام علیکم صبح بخیر	0.00%
4	Original	وعلیکم السلام	
	Inference	وعلیکم السلام	0.00%
5	Original	خوش آمدید کیا حال ہے	
	Inference	خوش آمدید حال ہے	20.00%
6	Original	میں ٹھیک ہوں شکر یہ اور آپ کیسے ھیں	
	Inference	میں ٹھیک ہوں شکر یہ	45.71%
7	Original	اچھا سب کچھ ٹھیک ہے	
	Inference	اچھا سب کچھ ٹھیک ہے	0.00%
8	Original	بہت بہت شکریہ	
	Inference	بہت بہت شکریہ	0.00%
9	Original	کوئی نئی خبر نہیں	
	Inference	کوئی نئی خبر نہیں	0.00%
10	Original	دوست سنو کوی بات نہیں	
	Inference	دوست کوی بات نہیں	19.05%
11	Original	مجھے آپ کی بہت کمی محسوس ہوی	
	Inference	مجھے آپ کی کمی ہوی	35.71%
12	Original	شب بخیر پھر ملیں گے خدا حافظ	
	Inference	پھر ملیں گے خدا حافظ	25.00%
13	Original	لگتا ہے میں کھو گیا ہوں	
	Inference	لگتا ہے میں گیا ہوں	17.39%
14	Original	کیا میں آپ کی مدد کر سکتا ہوں	
	Inference	کیا میں آپ کی مدد کر سکتا	10.34%
15	Original	کیا آپ میری مدد کر سکتے ہیں	
	Inference	کیا آپ میری مدد کر سکتے ہیں	0.00%
16	Original	وہ میری مدد کو نہ آسکے	
	Inference	وہ میری مدد کو نہ آسکے	0.00%
17	Original	باتھ روم واش روم کہاں ہے	
	Inference	باتھ روم کہاں ہے	33.33%
18	Original	میڈیکل سٹور کہاں ہے	
	Inference	میڈیکل سٹور کہاں ہے	0.00%
19	Original	سیدھے جا کر دائیں بائیں مڑجائیں	
	Inference	سیدھے جا کر دائیں مڑجائیں	16.13%
20	Original	میں کسی کو ڈھونڈ رہا ہوں	
	Inference	میں کسی کو رہا ہوں	25.00%
21	Original	یہ لڑکی کس کو ڈھونڈ رہی ہے	
	Inference	یہ لڑکی کس کو ڈھونڈ رہی	7.69%
22	Original	یہ لڑکا بہت نیک اور شریف ہے	
	Inference	یہ لڑکا بہت نیک اور شریف ہے	0.00%
23	Original	برائے مہربانی کچھ دیر انتظار کیجئے	

	Inference	برائے مہربانی کچھ انتظار کیجئے	11.76%
24	Original	برائے مہربانی ہولڈ کیجیے	
	Inference	برائے مہربانی ہولڈ کیجیے	0.00%
25	Original	یہ والا کتنے کا ہے	
	Inference	یہ والا کتنے کا ہے	0.00%
26	Original	معذرت چاہتا ہوں ذرا سنیے	
	Inference	معذرت چاہتا ہوں سنیے	16.67%
27	Original	معاف کرنا دوست	
	Inference	معاف کرنا دوست	0.00%
28	Original	میرے ساتھ چلیے	
	Inference	میرے ساتھ چلیے	0.00%
29	Original	کیا آپ انگریزی یا اردو بول سکتے ہیں	
	Inference	کیا آپ انگریزی بول سکتے ہیں	22.86%
30	Original	جی بالکل صرف تھوڑی سی	
	Inference	جی بالکل صرف تھوڑی سی	0.00%
31	Original	آپ کا نام کیا ہے	
	Inference	آپ کا نام کیا ہے	0.00%
32	Original	میرا نام وقاص ہے	
	Inference	میرا نام وقاص ہے	0.00%
33	Original	محترم محترمہ مس مسٹر سب موجود ہیں	
	Inference	محترم محترمہ مس سب موجود	24.24%
34	Original	آپ سے مل کر خوشی ہوئی	
	Inference	آپ سے مل کر خوشی ہوئی	0.00%
35	Original	آپ بہت مہربان ہیں	
	Inference	آپ بہت مہربان ہیں	0.00%
36	Original	آپ کا تعلق کہاں سے ہے	
	Inference	آپ کا تعلق کہاں سے ہے	0.00%
37	Original	میرا تعلق پاکستان اور امریکہ سے ہے	
	Inference	میرا تعلق پاکستان سے ہے	32.35%
38	Original	میں پاکستانی اور امریکی ہوں	
	Inference	میں پاکستانی امریکی ہوں	14.81%
39	Original	آپ کہاں رہتے ہو	
	Inference	آپ کہاں رہتے ہو	0.00%
40	Original	میں اپنے گھر رہتا ہوں رہتی ہوں	
	Inference	اپنے گھر رہتا ہوں	40.00%
41	Original	کیا آپ کو یہاں آکر اچھا لگا	
	Inference	کیا آپ کو یہاں آکر اچھا لگا	0.00%
42	Original	یہ ایک اچھا ملک ہے	
	Inference	یہ ایک اچھا ملک ہے	0.00%
43	Original	دنیا بہت خوبصورت ہے	
	Inference	دنیا بہت خوبصورت ہے	0.00%

44	Original	آپ کیا کام کرتے ہیں	
	Inference	آپ کیا کام کرتے ہیں	0.00%
45	Original	میں جاب اور بزنس کرتا ہوں	
	Inference	میں جاب اور بزنس کرتا ہوں	0.00%
46	Original	میں مینیجر اور تاجر ہوں	
	Inference	میں مینیجر اور تاجر ہوں	0.00%
47	Original	مجھے اردو اور انگریزی اچھی لگتی ہے	
	Inference	مجھے اردو اور اچھی لگتی	32.35%
48	Original	میں کافی عرصے سے یہ زبان سیکھ رہا ہوں	
	Inference	میں کافی یہ زبان سیکھ رہا	29.73%
49	Original	دن مہینے سال لگتے ہیں	
	Inference	دن مہینے سال لگتے ہیں	0.00%
50	Original	واہ بہت خوب بہت اچھے	
	Inference	واہ بہت خوب بہت اچھے	0.00%
51	Original	آپ کی عمر کتنی ہے	
	Inference	آپ کی عمر کتنی ہے	0.00%
52	Original	میری عمر تیس سال ہے	
	Inference	میری عمر تیس سال ہے	0.00%
53	Original	مجھے اب جانا ہے اجازت چاہتا ہوں	
	Inference	مجھے اب جانا ہے اجازت	32.26%
54	Original	جلدی واپس آؤں گا میرا انتظار مت کرنا	
	Inference	جلدی واپس آؤں گا میرا انتظار مت کرنا	0.00%
55	Original	یہ گھڑی آپ کو اچھی لگے گی	
	Inference	یہ گھڑی آپ کو اچھی لگے گی	0.00%
56	Original	مجھے گنتی سناؤ	
	Inference	مجھے گنتی سناؤ	0.00%
57	Original	ایک دو تین چار پانچ چھے سات آٹھ نو دس گیارہ بارہ	
	Inference	ایک دو تین چار چھے سات آٹھ دس گیارہ	26.53%
58	Original	بیس تیس چالیس پچاس ساٹھ ستر اسی نوے سو	
	Inference	بیس چالیس پچاس ساٹھ ستر سو	31.58%
59	Original	ہزار لاکھ کروڑ ارب کھرب	
	Inference	ہزار لاکھ کروڑ کھرب	17.39%
60	Original	اللہ کا فضل اور شکر ہے	
	Inference	اللہ کا فضل شکر ہے	18.18%
61	Original	سالگرہ مبارک ہو مزے کرو	
	Inference	سالگرہ مبارک ہو مزے کرو	0.00%
62	Original	نیا سال مبارک ہو	
	Inference	نیا سال مبارک ہو	0.00%
63	Original	میں ایک دن یہاں جانا چاہتا ہوں چاہتی ہوں	
	Inference	میں ایک دن یہاں جانا چاہتا ہوں	25.00%
64	Original	سب کو میرا سلام کہنا	

	Inference	سب کو میرا سلام کہنا	0.00%
65	Original	اللہ رحم کرے ہم سب پر	
	Inference	اللہ رحم کرے ہم سب پر	0.00%
66	Original	شب بخیر اور سہانے خواب	
	Inference	شب بخیر اور خواب	27.27%
67	Original	معذرت چاہتا ہوں یا چاہتی ہوں	
	Inference	معذرت چاہتا ہوں	46.43%
68	Original	معاف کرنا مجھے	
	Inference	معاف کرنا مجھے	0.00%
69	Original	کوئی بات نہیں معافی کی ضرورت نہیں	
	Inference	کوئی بات نہیں معافی کی ضرورت نہیں	0.00%
70	Original	کیا آپ اسے دوبارہ اور آہستہ کہہ سکتے ہیں	
	Inference	کیا آپ اسے دوبارہ اور کہہ سکتے ہیں	14.63%
71	Original	برائے مہربانی اپنا کام لکھیے	
	Inference	برائے مہربانی اپنا کام لکھیے	0.00%
72	Original	میں سمجھا نہیں	
	Inference	میں سمجھا نہیں	0.00%
73	Original	مجھے نہیں معلوم	
	Inference	مجھے نہیں معلوم	0.00%
74	Original	مجھے اندازہ نہیں تھا	
	Inference	مجھے اندازہ نہیں تھا	0.00%
75	Original	اسے انگلش میں کیا بولتے ہیں	
	Inference	اسے انگلش میں کیا بولتے ہیں	0.00%
76	Original	یہ کیا ہے سب کچھ خراب ہے	
	Inference	یہ کیا ہے سب کچھ خراب ہے	0.00%
77	Original	پریشان نہیں ہوں سب ٹھیک ہو جائے گا	
	Inference	پریشان نہیں ہوں سب ہو جائے گا	14.71%
78	Original	مجھے اس کی مشق کرنی چاہیے	
	Inference	مجھے اس کی مشق کرنی چاہیے	0.00%
79	Original	اچھا ہو یا برا ہے وہ معمولی سا	
	Inference	اچھا ہو یا برا ہے وہ معمولی	10.00%
80	Original	بڑا چھوٹا سب برابر ہے	
	Inference	بڑا چھوٹا سب برابر ہے	0.00%
81	Original	مجھے گاڑی چاہیے آج اور ابھی	
	Inference	مجھے گاڑی چاہیے ابھی	25.93%
82	Original	آنے والا کل گزرے ہوئے کل سے بہتر ہے	
	Inference	آنے والا کل گزرے ہوئے سے بہتر	17.14%
83	Original	ہاں یا نہ میں جواب دو	
	Inference	ہاں یا نہ میں جواب دو	0.00%
84	Original	یہ لیجیے ہوگیا سارا کام	
	Inference	یہ لیجیے ہوگیا سارا کام	0.00%

85	Original	کیا یہ آپ کو پسند آیا	
	Inference	کیا یہ آپ کو پسند آیا	0.00%
86	Original	مجھے تو بہت پسند آیا	
	Inference	مجھے تو بہت پسند آیا	0.00%
87	Original	مجھے بھوک اور پیاس لگ رہی ہے	
	Inference	مجھے بھوک اور پیاس لگ رہی ہے	0.00%
88	Original	میں کام کرتا ہوں صبح میں شام میں رات میں	
	Inference	میں کام کرتا ہوں صبح شام	40.00%
89	Original	یہاں بھی بارش ہو رہی ہے اور وہاں بھی	
	Inference	یہاں بھی بارش ہو رہی ہے وہاں بھی	11.11%
90	Original	جلدی کرو یار	
	Inference	جلدی کرو یار	0.00%
91	Original	سچ میں تم پاگل ہو	
	Inference	سچ میں تم پاگل ہو	0.00%
92	Original	یہ ایک لڑکے اور لڑکی کی کہانی ہے	
	Inference	یہ ایک لڑکے اور لڑکی کی کہانی ہے	0.00%
93	Original	میں اور آپ اسے مل کر سنتے ہیں	
	Inference	میں اور آپ اسے مل کر سنتے ہیں	0.00%
94	Original	ابھی کیا وقت ہوا ہے	
	Inference	ابھی کیا وقت ہوا ہے	0.00%
95	Original	تمہارے منہ پر بارہ کیوں بجے ہیں	
	Inference	تمہارے منہ پر بارہ کیوں بجے ہیں	0.00%
96	Original	مجھے یہ دیجیے آدھے جملے ہو گے	
	Inference	مجھے یہ دیجیے آدھے جملے ہو گے	0.00%
97	Original	میں اپنے امی ابو سے پیار اور محبت کرتا ہوں	
	Inference	میں اپنے امی ابو سے پیار کرتا ہوں	21.43%
98	Original	یہ میرے دادا دادی نانا نانی ہیں	
	Inference	یہ میرے دادا دادی نانا نانی ہیں	0.00%
99	Original	تم سب میرے بہن بھائی ہو	
	Inference	تم سب میرے بہن بھائی ہو	0.00%
100	Original	یہ میرے چچا چاچی تایا تائی ہیں	
	Inference	یہ میرے چچا چاچی تایا	26.67%
101	Original	یہ میرے خالہ خالو ماموں ممانی ہیں	
	Inference	یہ میرے خالہ خالو ماموں ہیں	18.18%
102	Original	یہ بیوی اور شوہر کے آپس کا معاملہ ہے	
	Inference	یہ بیوی اور شوہر کے آپس کا معاملہ ہے	0.00%
103	Original	میرے ساس اور سسر بہت اچھے ہیں	
	Inference	میرے ساس اور سسر بہت اچھے ہیں	0.00%
104	Original	تمہارا سسرال تو قریب ہی ہے	
	Inference	تمہارا سسرال تو قریب ہی ہے	0.00%
105	Original	بیٹا ہو یا بیٹی اولاد نیک ہونی چاہیے	

	Inference	بیٹا ہو یا بیٹی اولاد نیک ہونی	16.67%
106	Original	والدین کی عزت اور خدمت کرو	
	Inference	والدین کی عزت اور خدمت کرو	0.00%
107	Original	میری طبیعت خراب ہے	
	Inference	میری طبیعت خراب ہے	0.00%
108	Original	میرے لیے دعا کیجئے	
	Inference	میرے لیے دعا کیجئے	0.00%
109	Original	وہ بلڈ پریشر اور شوگر کا مریض ہے	
	Inference	وہ پریشر اور شوگر کا مریض ہے	12.50%
110	Original	مجھے ڈاکٹر کی ضرورت ہے	
	Inference	مجھے ڈاکٹر کی ضرورت ہے	0.00%
111	Original	مجھے انصاف چاہیے	
	Inference	مجھے انصاف چاہیے	0.00%
112	Original	سیاستدانوں نے ہمیں ٹرک کی بتی کے پیچھے لگایا ہے	
	Inference	ہمیں ٹرک کی کے پیچھے لگایا ہے	36.17%
113	Original	لائٹ چلی گئی	
	Inference	لائٹ چلی گئی	0.00%
114	Original	لائٹ آگئی	
	Inference	لائٹ آگئی	0.00%
115	Original	مجھے مختلف پھلوں کے نام بتاؤ	
	Inference	مجھے مختلف پھلوں کے نام بتاؤ	0.00%
116	Original	سیب انگور کیلا کنو آم انار چکوترا کھجور	
	Inference	سیب انگور کیلا کنو آم انار کھجور	17.07%
117	Original	مجھے کھانے کھانا اور پانی پینا ہے	
	Inference	مجھے کھانے کھانا اور پانی پینا ہے	0.00%
118	Original	میرے پاس سب سواریاں ہیں	
	Inference	میرے پاس سب سواریاں ہیں	0.00%
119	Original	سائکل موٹر سائیکل کار بس ٹرک	
	Inference	سائکل موٹر سائیکل کار بس ٹرک	0.00%
120	Original	ہوائی جہاز بحری جہاز بیرون ملک سفر کے لئے استعمال ہوتے ہیں	
	Inference	ہوائی جہاز جہاز بیرون ملک سفر کے استعمال ہوتے ہیں	15.52%
121	Original	مجھے سبزیوں کے نام بھی یاد ہیں	
	Inference	مجھے سبزیوں کے نام بھی یاد ہیں	0.00%
122	Original	آلو گوبھی مٹر گاجر شلجم مولی کھیرا کدو ٹینڈے بھنڈی توری پالک	
	Inference	آلو گوبھی مٹر گاجر مولی کھیرا کدو ٹینڈے توری	26.67%
123	Original	جنگل میں جانور اور چرند پرند ہیں	
	Inference	جنگل میں جانور اور چرند پرند ہیں	0.00%
124	Original	بلی کتا شیر گدھا الو بندر لومڑی چڑیا طوطا مور	
	Inference	بلی کتا شیر بندر چڑیا طوطا	42.22%

125	Original	باہر شدید گرمی ہے	
	Inference	باہر شدید گرمی ہے	0.00%
126	Original	آج تو کافی ٹھنڈ ہے	
	Inference	آج تو کافی ٹھنڈ ہے	0.00%
127	Original	پاکستان میں سب ھی موسم پائے جاتے ہیں	
	Inference	پاکستان میں سب ھی موسم ہیں	25.71%
128	Original	موسم گرما سرما بہار خزاں	
	Inference	موسم گرما سرما بہار خزاں	0.00%
129	Original	طوفانی بارش برس رہی ہے	
	Inference	طوفانی بارش برس رہی ہے	0.00%
130	Original	برف باری بھی پڑ رہی ہے	
	Inference	برف باری بھی پڑ رہی ہے	0.00%
131	Original	سٹوڈنٹ موبائل پر فیس بک اور واٹس ایپ یوز کر رہے تھے	
	Inference	سٹوڈنٹ موبائل پر فیس بک اور کر رہے	33.33%
132	Original	میں فارغ وقت میں لیپ ٹاپ یوز کرتا ہوں وہ ٹی وی دیکھتا ہوں	
	Inference	میں فارغ وقت میں کرتا ہوں وہ ٹی وی دیکھتا ہوں	19.30%
133	Original	چائے پینے کا موڈ ہو رہا ہے	
	Inference	چائے پینے کا موڈ ہو رہا ہے	0.00%
134	Original	ادھر کا کھانا بہت مزے کا ہے	
	Inference	ادھر کا کھانا بہت مزے کا ہے	0.00%
135	Original	مسلمان مصیبت میں گھبرایا نہیں کرتے	
	Inference	مسلمان مصیبت میں نہیں کرتے	23.53%
136	Original	گریپ فروٹ جوڑوں کے درد میں مفید ہے	
	Inference	فروٹ جوڑوں کے درد میں ہے	26.47%
137	Original	جسم سے فاسد مواد کا اخراج کرتا ہے	
	Inference	جسم سے فاسد مواد کا کرتا ہے	18.18%
138	Original	وزن کم کرنے میں مدد کرتا ہے	
	Inference	وزن کم کرنے میں مدد کرتا ہے	0.00%
139	Original	بے خوابی میں مبتلا لوگ اس کا استعمال ضرور کریں	
	Inference	میں لوگ اس کا استعمال ضرور کریں	30.43%
140	Original	کھانسی اور گلے کی خراش میں کھانا فائدہ مند ہے	
	Inference	کھانسی اور گلے کی میں کھانا فائدہ مند ہے	11.11%
141	Original	معدے کی تیزابیت دور کرتا ہے	
	Inference	معدے کی دور کرتا ہے	29.63%
142	Original	موٹاپے کو کم کرنے میں مدد گار ہے	
	Inference	موٹاپے کو کم کرنے میں ہے	25.00%
143	Original	چھاتی کے کینسر میں فائدہ مند ہے	
	Inference	چھاتی کے کینسر میں فائدہ مند ہے	0.00%
144	Original	لبلبے کے سرطان میں مفید ہے	
	Inference	کے میں مفید ہے	42.31%
145	Original	شریانوں کی رکاوٹ کو دور کرتا ہے	

	Inference	کی رکاوٹ کو دور کرتا ہے	22.58%
146	Original	اسٹرابری دانتوں اور ہڈیوں کی کمزوری کو دور کرتی ہے	
	Inference	دانتوں اور ہڈیوں کی کمزوری کو دور کرتی ہے	16.00%
147	Original	سوجن میں کمی لاتی ہے	
	Inference	سوجن میں کمی لاتی ہے	0.00%
148	Original	نزلہ زکام کا خاتمہ کرتی ہے	
	Inference	نزلہ زکام کا خاتمہ کرتی ہے	0.00%
149	Original	کینسر کے خلاف قوت مدافعت کو بڑھاتی ہے	
	Inference	کینسر کے خلاف قوت کو بڑھاتی ہے	18.92%
150	Original	امراض چشم میں فائدہ مند ہے	
	Inference	امراض چشم میں فائدہ مند ہے	0.00%
151	Original	اس کے کھانے سے بلڈ پریشر نارمل رہتا ہے	
	Inference	اس کے کھانے سے بلڈ پریشر نارمل رہتا ہے	0.00%
152	Original	کولیسٹرول کی سطح کو نارمل رکھتی ہے	
	Inference	کی سطح کو نارمل رکھتی ہے	26.47%
153	Original	جھریاں پڑنے سے روکتی ہے	
	Inference	جھریاں پڑنے سے روکتی ہے	0.00%
154	Original	جوڑوں کے درد اور گنٹھیا میں مفید ہے	
	Inference	جوڑوں کے درد اور میں مفید ہے	20.00%
155	Original	پیاس کو کنڑول کرتی ہے	
	Inference	پیاس کو کنڑول کرتی ہے	0.00%
156	Original	گاؤں اور شہر ایک ہو رہے ہیں	
	Inference	گاؤں اور شہر ایک ہو رہے ہیں	0.00%
157	Original	سکول کالج یونیورسٹی دفتر بینک سب کھل گئے ہیں	
	Inference	سکول کالج دفتر بینک کھل گئے ہیں	29.55%
158	Original	لوگ مہنگائی سے تنگ آکر خودکشی کرنے پر مجبور ہیں	
	Inference	لوگ مہنگائی سے تنگ آکر کرنے پر مجبور ہیں	14.89%
159	Original	ڈاکٹر اور مریض میں تلخ کلامی ہوئی	
	Inference	ڈاکٹر اور مریض میں تلخ کلامی ہوئی	0.00%
160	Original	مولوی مفتی قاری مسجد میں بیٹھے ہیں	
	Inference	مولوی مفتی قاری مسجد میں ہیں	17.65%
161	Original	اپنے اساتذہ کا احترام کریں	
	Inference	اپنے اساتذہ کا احترام کریں	0.00%
162	Original	پولیس غنڈہ گردی پر اتر آئی	
	Inference	پولیس پر اتر آئی	38.46%
163	Original	عدالت میں وکیل اور جج بیٹھے ہیں	
	Inference	عدالت میں وکیل اور جج بیٹھے ہیں	0.00%
164	Original	اللہ بہتر جانتا ہے	
	Inference	اللہ بہتر جانتا ہے	0.00%
165	Original	میرے سپروائزر بہت محنتی ہیں	
	Inference	میرے سپروائزر بہت محنتی ہیں	0.00%

166	Original	لڑکوں کے نام عبداللہ علی عمر ابوبکر عثمان ہیں	
	Inference	لڑکوں کے نام علی عمر ابوبکر عثمان	26.67%
167	Original	لڑکیوں کے نام عائشہ صبا مریم زینب ہیں	
	Inference	لڑکیوں کے نام عائشہ صبا مریم ہیں	13.51%
168	Original	انشاءاللہ کامیابی تمہارے قدم چومے گی	
	Inference	کامیابی تمہارے قدم چومے گی	25.00%
169	Original	نماز ہم پر فرض ہے	
	Inference	نماز ہم پر فرض ہے	0.00%
170	Original	فجر ظہر عصر مغرب عشاء	
	Inference	فجر ظہر عصر مغرب	23.81%
171	Original	نزدیک کوئی قریبی ہسپتال ہے	
	Inference	نزدیک کوئی قریبی ہسپتال ہے	0.00%
172	Original	اکرم صاحب کا بیٹا انجینئر اور بیٹی ڈاکٹر ہے	
	Inference	اکرم صاحب کا بیٹا اور بیٹی ڈاکٹر ہے	18.60%
173	Original	جو ہوتا ہے بہتر ہوتا ہے	
	Inference	جو ہوتا ہے بہتر ہوتا ہے	0.00%
174	Original	ہینڈ بیگ اور سوٹ کیس یاد سے اٹھا لینا	
	Inference	ہینڈ بیگ اور سوٹ کیس یاد سے اٹھا لینا	0.00%
175	Original	ریموٹ نہیں مل رہا	
	Inference	ریموٹ نہیں مل رہا	0.00%
176	Original	میری جرابیں نہیں مل رہی	
	Inference	میری جرابیں نہیں مل رہی	0.00%
177	Original	کمپیوٹر ہارڈ وئیر اور سافٹ وئیر سے چلتا ہے	
	Inference	ہارڈ اور سافٹ وئیر سے چلتا ہے	28.57%
178	Original	درخت اور پھول فضا کو صاف اور شفاف رکھتے ہیں	
	Inference	درخت اور پھول فضا کو صاف رکھتے ہیں	20.93%
179	Original	شہری فیکٹری کے دھویں اور گردوغبار سے تنگ ہیں	
	Inference	شہری فیکٹری کے دھویں اور گردوغبار سے تنگ ہیں	0.00%
180	Original	کچھ عقل کرو	
	Inference	کچھ عقل کرو	0.00%
181	Original	مجھے مرغی کا گوشت اور انڈے پسند ہیں	
	Inference	مجھے مرغی کا گوشت اور انڈے پسند ہیں	0.00%
182	Original	عید کا چاند نظر آگیا ہے	
	Inference	عید کا چاند نظر آگیا ہے	0.00%
183	Original	ہم اس بار بکرے اونٹ گائے کی قربانی کریں گے	
	Inference	ہم اس بار بکرے گائے کی قربانی کریں گے	11.90%
184	Original	بچے عیدی مانگتے ہیں	
	Inference	بچے عیدی مانگتے ہیں	0.00%
185	Original	عید مبارک خیر مبارک	
	Inference	عید مبارک خیر مبارک	0.00%
186	Original	سڑکیں پل اور انڈر پاس بن رہے ہیں	

	Inference	سڑکیں پل اور پاس بن رہے ہیں	15.63%
187	Original	کراچی لاہور فیصل آباد اور ملتان بڑے شہر ہیں	
	Inference	کراچی لاہور فیصل آباد اور ملتان بڑے شہر ہیں	0.00%
188	Original	پنجاب سندھ خیبر پختونخوا گلگت بلتستان پاکستان کے صوبے ہیں	
	Inference	پنجاب سندھ پاکستان کے صوبے ہیں	47.37%
189	Original	اللہ ہمارے ملک کو قائم و دائم رکھے	
	Inference	اللہ ہمارے ملک کو قائم رکھے	20.59%
190	Original	پاکستان زندہ باد	
	Inference	پاکستان زندہ باد	0.00%
191	Original	پاکستان پائندہ باد	
	Inference	پاکستان پائندہ باد	0.00%
192	Original	اسلام آباد خوبصورت شہر ہے	
	Inference	اسلام آباد خوبصورت ہے	16.00%
193	Original	اور پاکستان کا دارالحکومت ہے	
	Inference	اور پاکستان کا دارالحکومت ہے	0.00%
194	Original	آنکھ کان ناک بال چہرے کا حصہ ہیں	
	Inference	آنکھ کان ناک کا حصہ ہیں	28.13%
195	Original	ٹانگیں بازو سر دھڑجسم کا حصہ ہیں	
	Inference	بازو سر کا حصہ ہیں	40.63%
196	Original	سر درد کر رہا ہے	
	Inference	سر درد کر رہا ہے	0.00%
197	Original	تعاون کے لئے آپ کا شکریہ	
	Inference	تعاون کے لئے آپ کا شکریہ	0.00%
198	Original	اللہ حافظ خدا حافظ اللہ نگہبان	
	Inference	اللہ حافظ خدا حافظ اللہ	23.33%
199	Original	یہ آخری جملہ ہے آپ کا شکریہ	
	Inference	یہ آخری جملہ ہے آپ کا شکریہ	0.00%
200	Original	اللہ ایک ہے اور اس کا کوئی شریک نہیں	
	Inference	اللہ ایک ہے اس کا کوئی شریک نہیں	11.11%
Testing (Avg.)			**9.75%**

5.1.2 Speaker Independent Test Results

These are also speaker independent test results which we have managed to collect by recordings of specified sentences which do not exist in training, validation and test datasets. We have done inference of all these voice data with already prepared model and got 14.21% Avg. WER. As these voice data do not exist in all data sets we consider it speaker independent test results. Results are presented in Table 5.3.

Table 5.3: Speaker Independent Test Results

WER	Urdu Transcription		No.
	اللہ سب سے بڑا ہے	Original	1
29.41%	اللہ سب سے	Inference	
	محمد صلی اللہ علیہ و آلہ وسلم اللہ کے رسول ہیں	Original	2
36.96%	محمد صلی اللہ علیہ و آلہ وسلم	Inference	
	اسلام علیکم صبح بخیر	Original	3
20.00%	اسلام علیکم بخیر	Inference	
	وعلیکم السلام	Original	4
0.00%	وعلیکم السلام	Inference	
	خوش آمدید کیا حال ہے	Original	5
50.00%	خوش آمدید	Inference	
	میں ٹھیک ہوں شکر یہ اور آپ کیسے ھیں	Original	6
45.71%	میں ٹھیک ہوں شکر یہ	Inference	
	اچھا سب کچھ ٹھیک ہے	Original	7
26.32%	اچھا سب کچھ ہے	Inference	
	بہت بہت شکریہ	Original	8
0.00%	بہت بہت شکریہ	Inference	
	کوئی نئی خبر نہیں	Original	9
0.00%	کوئی نئی خبر نہیں	Inference	
	دوست سنو کوی بات نہیں	Original	10
23.81%	دوست سنو کوی بات	Inference	
	مجھے آپ کی بہت کمی محسوس ہوی	Original	11
14.29%	مجھے آپ کی کمی محسوس ہوی	Inference	
	شب بخیر پھر ملیں گے خدا حافظ	Original	12
25.00%	پھر ملیں گے خدا حافظ	Inference	
	لگتا ہے میں کھو گیا ہوں	Original	13
17.39%	لگتا ہے میں گیا ہوں	Inference	
	کیا میں آپ کی مدد کر سکتا ہوں	Original	14
24.14%	کیا میں آپ کی کر سکتا	Inference	
	کیا آپ میری مدد کر سکتے ہیں	Original	15
0.00%	کیا آپ میری مدد کر سکتے ہیں	Inference	
	وہ میری مدد کو نہ آسکے	Original	16
0.00%	وہ میری مدد کو نہ آسکے	Inference	
	باتھ روم واش روم کہاں ہے	Original	17
33.33%	واش روم کہاں ہے	Inference	
	میڈیکل سٹور کہاں ہے	Original	18
0.00%	میڈیکل سٹور کہاں ہے	Inference	
	سیدھے جا کر دائیں بائیں مڑجائیں	Original	19
19.35%	سیدھے جا کر دائیں مڑجائیں	Inference	

20	Original	میں کسی کو ڈھونڈ رہا ہوں	
	Inference	میں کسی کو رہا ہوں	25.00%
21	Original	یہ لڑکی کس کو ڈھونڈ رہی ہے	
	Inference	یہ لڑکی کو ڈھونڈ رہی	23.08%
22	Original	یہ لڑکا بہت نیک اور شریف ہے	
	Inference	یہ لڑکا بہت نیک اور شریف ہے	0.00%
23	Original	برائے مہربانی کچھ دیر انتظار کیجئے	
	Inference	برائے مہربانی کچھ دیر انتظار کیجئے	0.00%
24	Original	برائے مہربانی ہولڈ کیجیے	
	Inference	برائے مہربانی ہولڈ کیجیے	0.00%
25	Original	یہ والا کتنے کا ہے	
	Inference	یہ والا کتنے کا ہے	0.00%
26	Original	معذرت چاہتا ہوں ذرا سنیے	
	Inference	معذرت چاہتا ہوں ذرا سنیے	0.00%
27	Original	معاف کرنا دوست	
	Inference	معاف کرنا دوست	0.00%
28	Original	میرے ساتھ چلیے	
	Inference	میرے ساتھ چلیے	0.00%
29	Original	کیا آپ انگریزی یا اردو بول سکتے ہیں	
	Inference	کیا آپ بول سکتے ہیں	45.71%
30	Original	جی بالکل صرف تھوڑی سی	
	Inference	جی بالکل صرف تھوڑی سی	0.00%
31	Original	آپ کا نام کیا ہے	
	Inference	آپ کا نام کیا ہے	0.00%
32	Original	میرا نام وقاص ہے	
	Inference	میرا نام ہے	31.25%
33	Original	محترم محترمہ مس مسٹر سب موجود ہیں	
	Inference	محترم محترمہ مس سب موجود	24.24%
34	Original	آپ سے مل کر خوشی ہوئی	
	Inference	آپ سے مل کر ہوئی	23.81%
35	Original	آپ بہت مہربان ہیں	
	Inference	آپ بہت مہربان ہیں	0.00%
36	Original	آپ کا تعلق کہاں سے ہے	
	Inference	آپ کا تعلق کہاں سے ہے	0.00%
37	Original	میرا تعلق پاکستان اور امریکہ سے ہے	
	Inference	میرا تعلق اور امریکہ سے ہے	23.53%
38	Original	میں پاکستانی اور امریکی ہوں	
	Inference	میں پاکستانی امریکی ہوں	14.81%
39	Original	آپ کہاں رہتے ہو	
	Inference	آپ کہاں رہتے ہو	0.00%
40	Original	میں اپنے گھر رہتا ہوں رہتی ہوں	

	Inference	میں اپنے گھر رہتا ہوں	30.00%
41	Original	کیا آپ کو یہاں آکر اچھا لگا	
	Inference	کیا آپ کو آکر اچھا لگا	17.86%
42	Original	یہ ایک اچھا ملک ہے	
	Inference	یہ ایک اچھا ملک ہے	0.00%
43	Original	دنیا بہت خوبصورت ہے	
	Inference	دنیا بہت ہے	42.11%
44	Original	آپ کیا کام کرتے ہیں	
	Inference	آپ کیا کام کرتے ہیں	0.00%
45	Original	میں جاب اور بزنس کرتا ہوں	
	Inference	میں جاب اور بزنس کرتا ہوں	0.00%
46	Original	میں مینیجر اور تاجر ہوں	
	Inference	میں مینیجر اور تاجر ہوں	0.00%
47	Original	مجھے اردو اور انگریزی اچھی لگتی ہے	
	Inference	مجھے اردو اور انگریزی اچھی لگتی ہے	0.00%
48	Original	میں کافی عرصے سے یہ زبان سیکھ رہا ہوں	
	Inference	میں کافی یہ زبان رہا ہوں	35.14%
49	Original	دن مہینے سال لگتے ہیں	
	Inference	دن مہینے سال لگتے ہیں	0.00%
50	Original	واہ بہت خوب بہت اچھے	
	Inference	واہ بہت خوب بہت اچھے	0.00%
51	Original	آپ کی عمر کتنی ہے	
	Inference	آپ کی عمر کتنی ہے	0.00%
52	Original	میری عمر تیس سال ہے	
	Inference	میری عمر تیس سال ہے	0.00%
53	Original	مجھے اب جانا ہے اجازت چاہتا ہوں	
	Inference	مجھے اب جانا ہے اجازت چاہتا ہوں	0.00%
54	Original	جلدی واپس آؤں گا میرا انتظار مت کرنا	
	Inference	جلدی واپس آؤں گا میرا انتظار مت کرنا	0.00%
55	Original	یہ گھڑی آپ کو اچھی لگے گی	
	Inference	یہ گھڑی آپ کو لگے گی	16.00%
56	Original	مجھے گنتی سناؤ	
	Inference	مجھے گنتی سناؤ	0.00%
57	Original	ایک دو تین چار پانچ چھے سات آٹھ نو دس گیارہ بارہ	
	Inference	ایک دو تین چار پانچ چھے سات نو دس گیارہ بارہ	8.16%
58	Original	بیس تیس چالیس پچاس ساٹھ ستر اسی نوے سو	
	Inference	تیس چالیس پچاس ساٹھ اسی نوے سو	18.42%
59	Original	ہزار لاکھ کروڑ ارب کھرب	
	Inference	ہزار لاکھ کروڑ ارب	21.74%
60	Original	اللہ کا فضل اور شکر ہے	
	Inference	اللہ کا فضل اور شکر ہے	0.00%

61	Original	سالگرہ مبارک ہو مزے کرو	
	Inference	سالگرہ مبارک ہو مزے کرو	0.00%
62	Original	نیا سال مبارک ہو	
	Inference	نیا سال مبارک ہو	0.00%
63	Original	میں ایک دن یہاں جانا چاہتا ہوں چاہتی ہوں	
	Inference	میں ایک دن جانا چاہتا ہوں	37.50%
64	Original	سب کو میرا سلام کہنا	
	Inference	سب کو میرا سلام کہنا	0.00%
65	Original	اللہ رحم کرے ہم سب پر	
	Inference	اللہ رحم کرے ہم سب پر	0.00%
66	Original	شب بخیر اور سہانے خواب	
	Inference	شب بخیر	68.18%
67	Original	معذرت چاہتا ہوں یا چاہتی ہوں	
	Inference	معذرت چاہتا ہوں	46.43%
68	Original	معاف کرنا مجھے	
	Inference	معاف کرنا مجھے	0.00%
69	Original	کوئی بات نہیں معافی کی ضرورت نہیں	
	Inference	کوئی بات نہیں معافی کی ضرورت نہیں	0.00%
70	Original	کیا آپ اسے دوبارہ اور آہستہ کہہ سکتے ہیں	
	Inference	کیا آپ اسے دوبارہ اور کہہ	36.59%
71	Original	برائے مہربانی اپنا کام لکھیے	
	Inference	برائے مہربانی اپنا کام	21.43%
72	Original	میں سمجھا نہیں	
	Inference	میں سمجھا نہیں	0.00%
73	Original	مجھے نہیں معلوم	
	Inference	مجھے نہیں	40.00%
74	Original	مجھے اندازہ نہیں تھا	
	Inference	مجھے اندازہ نہیں تھا	0.00%
75	Original	اسے انگلش میں کیا بولتے ہیں	
	Inference	اسے میں کیا بولتے ہیں	22.22%
76	Original	یہ کیا ہے سب کچھ خراب ہے	
	Inference	یہ کیا ہے سب کچھ خراب ہے	0.00%
77	Original	پریشان نہیں ہوں سب ٹھیک ہو جائے گا	
	Inference	پریشان نہیں ہوں سب ٹھیک ہو جائے گا	0.00%
78	Original	مجھے اس کی مشق کرنی چاہیے	
	Inference	مجھے اس کی کرنی چاہیے	16.00%
79	Original	اچھا ہو یا برا ہے وہ معمولی سا	
	Inference	اچھا ہو یا برا ہے وہ	33.33%
80	Original	بڑا چھوٹا سب برابر ہے	
	Inference	بڑا سب برابر	42.86%
81	Original	مجھے گاڑی چاہیے آج اور ابھی	

	Inference	مجھے گاڑی چاہیے آج اور ابھی	0.00%
82	Original	آنے والا کل گزرے ہوئے کل سے بہتر ہے	
	Inference	آنے والا کل گزرے ہوئے کل سے بہتر ہے	0.00%
83	Original	ہاں یا نہ میں جواب دو	
	Inference	ہاں یا میں جواب دو	14.29%
84	Original	یہ لیجیے ہوگیا سارا کام	
	Inference	یہ ہوگیا سارا کام	26.09%
85	Original	کیا یہ آپ کو پسند آیا	
	Inference	کیا یہ آپ کو پسند آیا	0.00%
86	Original	مجھے تو بہت پسند آیا	
	Inference	مجھے تو بہت پسند آیا	0.00%
87	Original	مجھے بھوک اور پیاس لگ رہی ہے	
	Inference	مجھے اور پیاس لگ رہی ہے	17.86%
88	Original	میں کام کرتا ہوں صبح میں شام میں رات میں	
	Inference	میں کام کرتا ہوں صبح میں شام میں رات میں	0.00%
89	Original	یہاں بھی بارش ہو رہی ہے اور وہاں بھی	
	Inference	یہاں بھی ہو رہی ہے وہاں بھی	25.00%
90	Original	جلدی کرو یار	
	Inference	جلدی کرو یار	0.00%
91	Original	سچ میں تم پاگل ہو	
	Inference	سچ میں تم پاگل ہو	0.00%
92	Original	یہ ایک لڑکے اور لڑکی کی کہانی ہے	
	Inference	یہ ایک لڑکے اور لڑکی کی ہے	18.75%
93	Original	میں اور آپ اسے مل کر سنتے ہیں	
	Inference	میں اور آپ اسے مل کر ہیں	17.24%
94	Original	ابھی کیا وقت ہوا ہے	
	Inference	ابھی کیا وقت ہوا ہے	0.00%
95	Original	تمہارے منہ پر بارہ کیوں بجے ہیں	
	Inference	تمہارے منہ پر کیوں بجے ہیں	16.13%
96	Original	مجھے یہ دیجیے آدھے جملے ہو گے	
	Inference	مجھے یہ دیجیے آدھے جملے ہو گے	0.00%
97	Original	میں اپنے امی ابو سے پیار اور محبت کرتا ہوں	
	Inference	میں اپنے امی ابو سے پیار اور کرتا ہوں	11.90%
98	Original	یہ میرے دادا دادی نانا نانی ہیں	
	Inference	یہ میرے دادا دادی نانا نانی ہیں	0.00%
99	Original	تم سب میرے بہن بھائی ہو	
	Inference	تم سب میرے بہن ہو	26.09%
100	Original	یہ میرے چچا چاچی تایا تائی ہیں	
	Inference	یہ میرے چچا چاچی تایا تائی ہیں	0.00%
101	Original	یہ میرے خالہ خالو ماموں ممانی ہیں	
	Inference	یہ میرے خالہ ماموں ممانی ہیں	15.15%

102	Original	یہ بیوی اور شوہر کے آپس کا معاملہ ہے	
	Inference	یہ بیوی اور کے آپس کا معاملہ ہے	13.89%
103	Original	میرے ساس اور سسر بہت اچھے ہیں	
	Inference	میرے ساس اور بہت اچھے ہیں	13.79%
104	Original	تمہارا سسرال تو قریب ہی ہے	
	Inference	تمہارا سسرال تو قریب ہی ہے	0.00%
105	Original	بیٹا ہو یا بیٹی اولاد نیک ہونی چاہیے	
	Inference	بیٹا ہو یا بیٹی اولاد نیک ہونی چاہیے	0.00%
106	Original	والدین کی عزت اور خدمت کرو	
	Inference	والدین کی عزت اور	34.62%
107	Original	میری طبیعت خراب ہے	
	Inference	میری طبیعت خراب ہے	0.00%
108	Original	میرے لیے دعا کیجئے	
	Inference	میرے لیے دعا کیجئے	0.00%
109	Original	وہ بلڈ پریشر اور شوگر کا مریض ہے	
	Inference	وہ پریشر اور شوگر کا مریض ہے	12.50%
110	Original	مجھے ڈاکٹر کی ضرورت ہے	
	Inference	مجھے کی ضرورت ہے	27.27%
111	Original	مجھے انصاف چاہیے	
	Inference	مجھے انصاف چاہیے	0.00%
112	Original	سیاستدانوں نے ہمیں ٹرک کی بتی کے پیچھے لگایا ہے	
	Inference	نے ہمیں ٹرک کی کے پیچھے لگایا ہے	29.79%
113	Original	لائٹ چلی گئی	
	Inference	لائٹ چلی گئی	0.00%
114	Original	لائٹ آگئی	
	Inference	لائٹ آگئی	0.00%
115	Original	مجھے مختلف پھلوں کے نام بتاؤ	
	Inference	مجھے مختلف پھلوں کے نام بتاؤ	0.00%
116	Original	سیب انگور کیلا کنو آم انار چکوترا کھجور	
	Inference	سیب کیلا کنو آم انار	46.34%
117	Original	مجھے کھانے کھانا اور پانی پینا ہے	
	Inference	مجھے کھانا اور پانی پینا ہے	18.18%
118	Original	میرے پاس سب سواریاں ہیں	
	Inference	میرے پاس سب سواریاں ہیں	0.00%
119	Original	سائکل موٹر سائیکل کار بس ٹرک	
	Inference	سائکل موٹر سائیکل کار بس ٹرک	0.00%
120	Original	ہوائی جہاز بحری جہاز بیرون ملک سفر کے لئے استعمال ہوتے ہیں	
	Inference	ملک سفر کے لئے استعمال ہوتے ہیں	44.83%
121	Original	مجھے سبزیوں کے نام بھی یاد ہیں	
	Inference	مجھے کے نام بھی یاد ہیں	23.33%

122	Original	آلو گوبھی مٹر گاجر شلجم مولی کھیرا کدو ٹینڈے بھنڈی توری پالک	
	Inference	آلو مٹر گاجر مولی کھیرا کدو توری	46.67%
123	Original	جنگل میں جانور اور چرند پرند ہیں	
	Inference	جنگل میں جانور اور ہیں	31.25%
124	Original	بلی کتا شیر گدھا الو بندر لومڑی چڑیا طوطا مور	
	Inference	بلی کتا شیر الو بندر طوطا مور	35.56%
125	Original	باہر شدید گرمی ہے	
	Inference	باہر شدید گرمی ہے	0.00%
126	Original	آج تو کافی ٹھنڈ ہے	
	Inference	آج تو کافی ہے	27.78%
127	Original	پاکستان میں سب ھی موسم پائے جاتے ہیں	
	Inference	پاکستان میں سب ھی موسم پائے جاتے ہیں	0.00%
128	Original	موسم گرما سرما بہار خزاں	
	Inference	موسم گرما سرما بہار خزاں	0.00%
129	Original	طوفانی بارش برس رہی ہے	
	Inference	طوفانی بارش برس رہی ہے	0.00%
130	Original	برف باری بھی پڑ رہی ہے	
	Inference	برف باری بھی پڑ رہی ہے	0.00%
131	Original	سٹوڈنٹ موبائل پر فیس بک اور واٹس ایپ یوز کر رہے تھے	
	Inference	موبائل پر فیس بک اور کر رہے	45.10%
132	Original	میں فارغ وقت میں لیپ ٹاپ یوز کرتا ہوں وہ ٹی وی دیکھتا ہوں	
	Inference	میں فارغ وقت میں کرتا ہوں وہ ٹی وی ہوں	31.58%
133	Original	چائے پینے کا موڈ ہو رہا ہے	
	Inference	چائے پینے کا ہو رہا ہے	15.38%
134	Original	ادھر کا کھانا بہت مزے کا ہے	
	Inference	ادھر کا کھانا بہت مزے کا ہے	0.00%
135	Original	مسلمان مصیبت میں گھبرایا نہیں کرتے	
	Inference	مسلمان مصیبت میں نہیں کرتے	23.53%
136	Original	گریپ فروٹ جوڑوں کے درد میں مفید ہے	
	Inference	کے درد میں ہے	58.82%
137	Original	جسم سے فاسد مواد کا اخراج کرتا ہے	
	Inference	جسم سے مواد کا اخراج کرتا ہے	15.15%
138	Original	وزن کم کرنے میں مدد کرتا ہے	
	Inference	وزن کم کرنے میں مدد کرتا ہے	0.00%
139	Original	بے خوابی میں مبتلا لوگ اس کا استعمال ضرور کریں	
	Inference	بے خوابی میں لوگ اس کا ضرور کریں	30.43%
140	Original	کھانسی اور گلے کی خراش میں کھانا فائدہ مند ہے	
	Inference	کھانسی اور گلے کی میں کھانا فائدہ مند ہے	11.11%
141	Original	معدے کی تیزابیت دور کرتا ہے	

	Inference	معدے کی دور کرتا ہے	29.63%
142	Original	موٹاپے کو کم کرنے میں مدد گار ہے	
	Inference	موٹاپے کو کم کرنے میں ہے	25.00%
143	Original	چھاتی کے کینسر میں فائدہ مند ہے	
	Inference	چھاتی کے کینسر میں فائدہ مند ہے	0.00%
144	Original	لبلبے کے سرطان میں مفید ہے	
	Inference	کے میں مفید ہے	42.31%
145	Original	شریانوں کی رکاوٹ کو دور کرتا ہے	
	Inference	شریانوں کی رکاوٹ کو دور کرتا ہے	0.00%
146	Original	اسٹرابری دانتوں اور ہڈیوں کی کمزوری کو دور کرتی ہے	
	Inference	اسٹرابری دانتوں اور کی کمزوری کو دور کرتی ہے	12.00%
147	Original	سوجن میں کمی لاتی ہے	
	Inference	میں کمی لاتی ہے	20.00%
148	Original	نزلہ زکام کا خاتمہ کرتی ہے	
	Inference	زکام کا خاتمہ کرتی ہے	15.38%
149	Original	کینسر کے خلاف قوت مدافعت کو بڑھاتی ہے	
	Inference	کے خلاف قوت کو بڑھاتی	40.54%
150	Original	امراض چشم میں فائدہ مند ہے	
	Inference	امراض چشم میں فائدہ مند ہے	0.00%
151	Original	اس کے کھانے سے بلڈ پریشر نارمل رہتا ہے	
	Inference	اس کے کھانے سے بلڈ پریشر نارمل رہتا ہے	0.00%
152	Original	کولیسٹرول کی سطح کو نارمل رکھتی ہے	
	Inference	کی سطح کو نارمل رکھتی ہے	26.47%
153	Original	جھریاں پڑنے سے روکتی ہے	
	Inference	جھریاں پڑنے سے روکتی ہے	0.00%
154	Original	جوڑوں کے درد اور گنٹھیا میں مفید ہے	
	Inference	کے درد اور میں مفید ہے	34.29%
155	Original	پیاس کو کنڑول کرتی ہے	
	Inference	کو کنڑول کرتی ہے	19.05%
156	Original	گاؤں اور شہر ایک ہو رہے ہیں	
	Inference	گاؤں اور شہر ایک ہو رہے ہیں	0.00%
157	Original	سکول کالج یونیورسٹی دفتر بینک سب کھل گئے ہیں	
	Inference	سکول کالج یونیورسٹی دفتر بینک سب ہیں	18.18%
158	Original	لوگ مہنگائی سے تنگ آکر خودکشی کرنے پر مجبور ہیں	
	Inference	لوگ مہنگائی سے تنگ آکر خودکشی کرنے پر ہیں	12.77%
159	Original	ڈاکٹر اور مریض میں تلخ کلامی ہوئی	
	Inference	ڈاکٹر اور مریض میں ہوئی	30.30%
160	Original	مولوی مفتی قاری مسجد میں بیٹھے ہیں	
	Inference	مولوی قاری مسجد میں بیٹھے ہیں	14.71%
161	Original	اپنے اساتذہ کا احترام کریں	
	Inference	اپنے کا احترام کریں	26.92%

162	Original	پولیس غنڈہ گردی پر اتر آئی	
	Inference	پولیس پر اتر آئی	38.46%
163	Original	عدالت میں وکیل اور جج بیٹھے ہیں	
	Inference	عدالت میں وکیل اور بیٹھے ہیں	9.68%
164	Original	اللہ بہتر جانتا ہے	
	Inference	اللہ بہتر جانتا ہے	0.00%
165	Original	میرے سپروائزر بہت محنتی ہیں	
	Inference	میرے بہت محنتی ہیں	32.14%
166	Original	لڑکوں کے نام عبداللہ علی عمر ابوبکر عثمان ہیں	
	Inference	لڑکوں کے نام عبداللہ علی عمر ابوبکر عثمان ہیں	0.00%
167	Original	لڑکیوں کے نام عائشہ صبا مریم زینب ہیں	
	Inference	لڑکیوں کے نام عائشہ صبا مریم زینب ہیں	0.00%
168	Original	انشاءاللہ کامیابی تمہارے قدم چومے گی	
	Inference	انشاءاللہ کامیابی تمہارے قدم چومے گی	0.00%
169	Original	نماز ہم پر فرض ہے	
	Inference	نماز ہم پر فرض ہے	0.00%
170	Original	فجر ظہر عصر مغرب عشاء	
	Inference	فجر ظہر عصر مغرب عشاء	0.00%
171	Original	نزدیک کوئی قریبی ہسپتال ہے	
	Inference	نزدیک کوئی قریبی ہے	26.92%
172	Original	اکرم صاحب کا بیٹا انجینئر اور بیٹی ڈاکٹر ہے	
	Inference	اکرم صاحب کا بیٹا انجینئر اور بیٹی ڈاکٹر ہے	0.00%
173	Original	جو ہوتا ہے بہتر ہوتا ہے	
	Inference	جو ہوتا ہے بہتر ہوتا ہے	0.00%
174	Original	ہینڈ بیگ اور سوٹ کیس یاد سے اٹھا لینا	
	Inference	بیگ اور کیس یاد سے اٹھا لینا	21.62%
175	Original	ریموٹ نہیں مل رہا	
	Inference	ریموٹ نہیں مل رہا	0.00%
176	Original	میری جرابیں نہیں مل رہی	
	Inference	میری نہیں مل رہی	30.43%
177	Original	کمپیوٹر ہارڈ وئیر اور سافٹ وئیر سے چلتا ہے	
	Inference	کمپیوٹر ہارڈ وئیر اور سافٹ وئیر سے چلتا ہے	0.00%
178	Original	درخت اور پھول فضا کو صاف اور شفاف رکھتے ہیں	
	Inference	درخت اور پھول کو صاف رکھتے ہیں	30.23%
179	Original	شہری فیکٹری کے دھویں اور گردوغبار سے تنگ ہیں	
	Inference	شہری کے دھویں اور سے تنگ ہیں	36.36%
180	Original	کچھ عقل کرو	
	Inference	کچھ عقل کرو	0.00%
181	Original	مجھے مرغی کا گوشت اور انڈے پسند ہیں	
	Inference	مجھے کا گوشت اور پسند ہیں	27.78%
182	Original	عید کا چاند نظر آگیا ہے	

	Inference	عید کا نظر آگیا ہے	21.74%
183	Original	ہم اس بار بکرے اونٹ گائے کی قربانی کریں گے	
	Inference	ہم اس بار بکرے گائے کی قربانی کریں گے	11.90%
184	Original	بچے عیدی مانگتے ہیں	
	Inference	بچے عیدی مانگتے ہیں	0.00%
185	Original	عید مبارک خیر مبارک	
	Inference	عید مبارک خیر مبارک	0.00%
186	Original	سڑکیں پل اور انڈر پاس بن رہے ہیں	
	Inference	سڑکیں پل اور پاس بن رہے ہیں	15.63%
187	Original	کراچی لاہور فیصل آباد اور ملتان بڑے شہر ہیں	
	Inference	لاہور فیصل آباد اور ملتان بڑے شہر ہیں	11.63%
188	Original	پنجاب سندھ خیبر پختونخوا گلگت بلتستان پاکستان کے صوبے ہیں	
	Inference	پنجاب سندھ خیبر پختونخوا پاکستان کے	36.84%
189	Original	اللہ ہمارے ملک کو قائم و دائم رکھے	
	Inference	اللہ ہمارے ملک کو قائم و دائم رکھے	0.00%
190	Original	پاکستان زندہ باد	
	Inference	پاکستان زندہ باد	0.00%
191	Original	پاکستان پائندہ باد	
	Inference	پاکستان	61.11%
192	Original	اسلام آباد خوبصورت شہر ہے	
	Inference	اسلام آباد خوبصورت شہر ہے	0.00%
193	Original	اور پاکستان کا دارالحکومت ہے	
	Inference	اور پاکستان کا ہے	39.29%
194	Original	آنکھ کان ناک بال چہرے کا حصہ ہیں	
	Inference	آنکھ کان ناک بال چہرے کا حصہ ہیں	0.00%
195	Original	ٹانگیں بازو سر دھڑجسم کا حصہ ہیں	
	Inference	ٹانگیں بازو سر کا حصہ ہیں	21.88%
196	Original	سر درد کر رہا ہے	
	Inference	سر درد کر رہا ہے	0.00%
197	Original	تعاون کے لئے آپ کا شکریہ	
	Inference	تعاون کے لئے آپ کا	25.00%
198	Original	اللہ حافظ خدا حافظ اللہ نگہبان	
	Inference	اللہ حافظ خدا حافظ اللہ نگہبان	0.00%
199	Original	یہ آخری جملہ ہے آپ کا شکریہ	
	Inference	یہ آخری ہے آپ کا شکریہ	18.52%
200	Original	اللہ ایک ہے اور اس کا کوئی شریک نہیں	
	Inference	اللہ ایک ہے اور اس کا کوئی شریک نہیں	0.00%
Testing (Avg.)			**14.21%**

5.2 Test Results Comparison

We compared test results of our proposed method in term of WER with previously proposed methods for Urdu ASR. Some methods can only recognize words instead of sentences which are categorized separately. Comparison is presented in Table 5.4

Table 5.4: Test Results Comparison

Methods	WER	Category
Hidden Markov Model [22]	29.1%	Sentence Recognition
Hidden Markov Model [13]	21.8%	Words Recognition
Deep Neural Network [14]	21%	Sentence Recognition
Proposed	**9.75%**	**Sentence Recognition**

5.3 Training and Validation Results

Training and Validation step loss is calculated on each iteration of epochs. The minimum step loss logged after 350 epochs which is 1.29% for training set and 13.13% for validation set. Results are presented in Table 5.5.

Table 5.5: Training and Validation Step loss

Epochs	Step Loss	
	Training	**Validation**
100	2.91%	17.09%
200	1.81%	16.20%
350	1.29%	13.13%
500	1.33%	14.69%

5.4 Training and Validation Summary

For the calculations we utilized TensorFlow because preparing a gigantic profound neural system can be unpredictable and confounding. To make it clearer, troubleshootable and enhanced TensorFlow project incorporated a suite of representation devices called TensorBoard. You can utilize TensorBoard to imagine your TensorFlow diagram, plot quantitative measurements about the execution of your chart, and demonstrate extra information like pictures that go through it.

TensorBoard works by perusing TensorFlow occurrences records, which contain synopsis information that you can produce when running TensorFlow.

With the help of summary data which is collected on each iteration of epoch's graphical representation of step loss on each epoch iteration is logged. It includes Validation step loss, Training step loss and their combined results presented in Figure 5.1, 5.2 and 5.3 respectively.

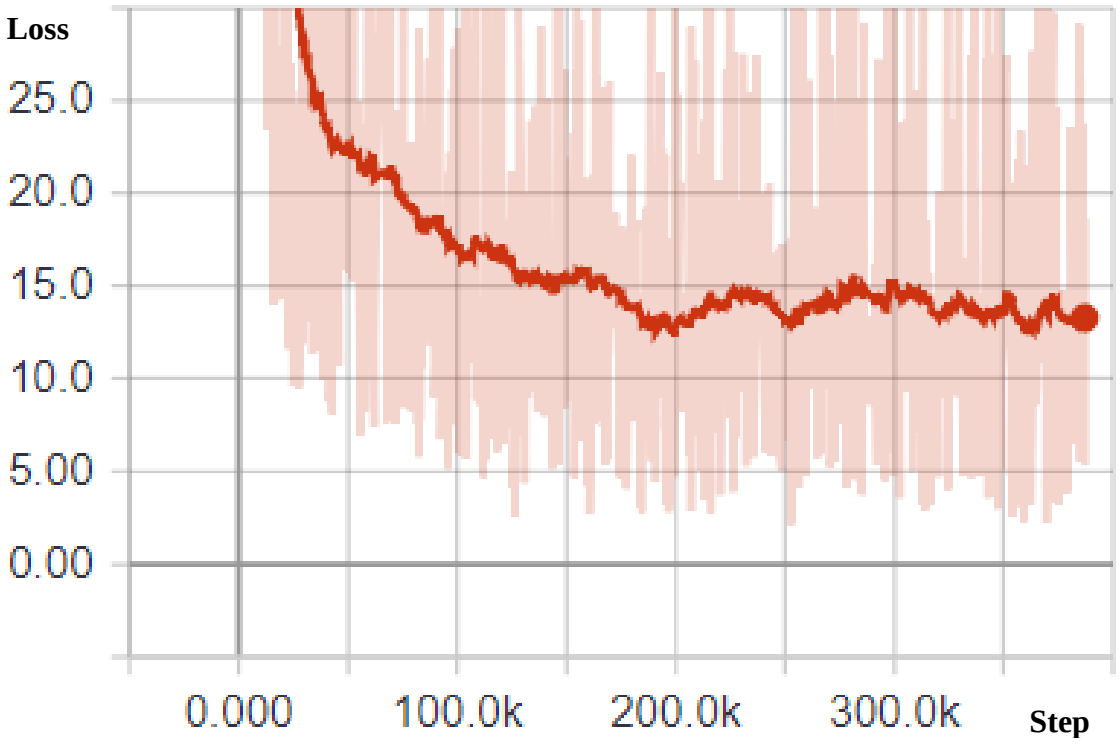

Figure 5.1: Validation Step Loss

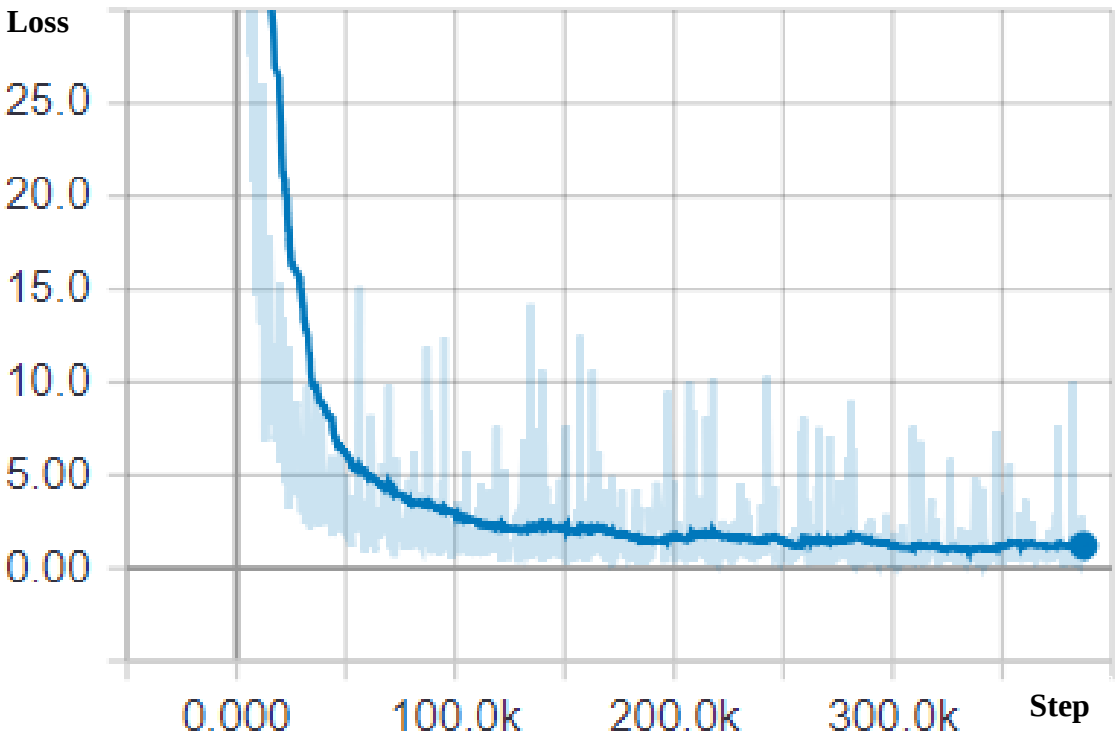

Figure 5.2: Training Step Loss

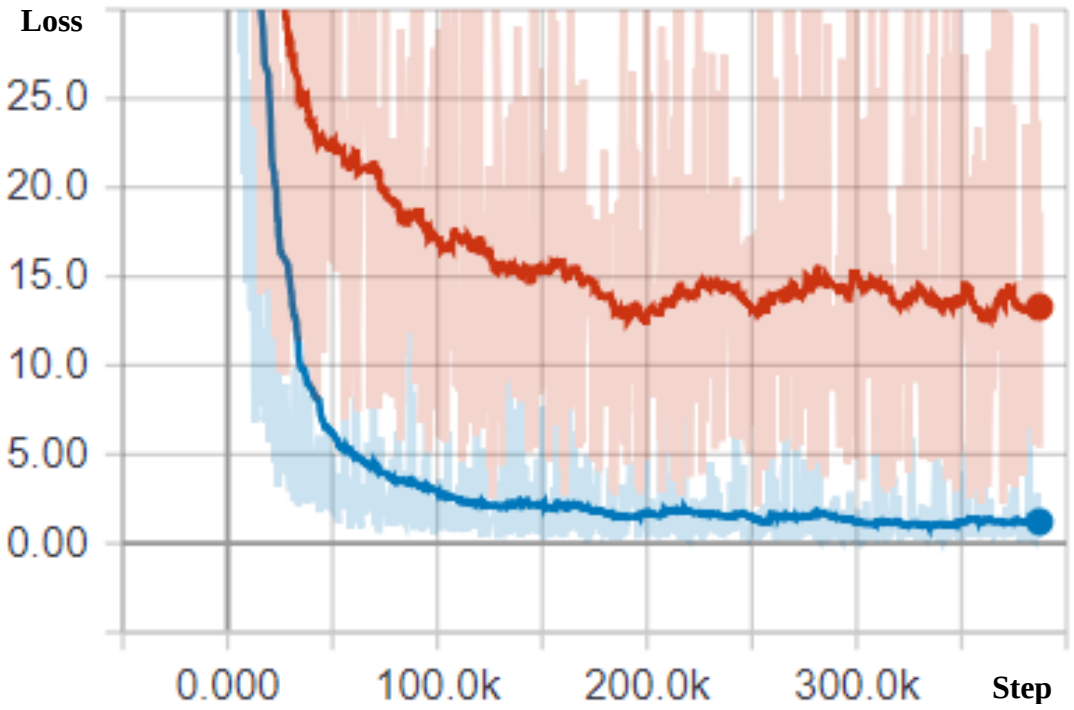

Figure 5.3: Training and Validation Step Loss

CHAPTER 6: CONCLUSION AND FUTURE WORK

In this paper we presented the development of ASR system for Urdu language using Deep Learning. We've used DeepSpeech which is an open source speech recognition framework implements Deep Learning techniques and supports almost all languages.

Results using Deep Learning are quite impressive as compare to other Urdu ASR systems which are available previously using other techniques like HMM and DNN. One of the major task we've done is the data collection in huge amount for common Urdu speech because data available for Urdu speech is very low in quantity and of single speaker or not open source.

We developed desktop and mobile application for data collection and preparation. Online storge is used to gather data from both application sources. We also developed utility application which makes the task of data arrangement and validation much easier. Language model and acoustic model for Urdu is developed which can be reused and retrained for inferences in future.

For training we used Google's cloud platform because Deep Learning requires huge computing power with large amount of data. Usage of GPU decreased the training time which is reduced from months and weeks to days. Also it makes the decording and inference of voice to text data faster.

We perform data testing for both speaker dependent voice data and speaker independent voice data. We got less than 10% WER in term of accuracy which is quite impressive as compare to other techniques used for ASR.

This ASR model could be retarined with more collected data and could be used in many daily life applications. Chat botes concept is emerging day by day, this model could be used in chat botes for the application which required Urdu language. This model could be used for Urdu to English or other languages translation.

Deep Learning requires huge amount of data to train any language acoustic model. We collected commonly used Urdu corpus which doesn’t cover the whole language; we are planning to collect more Urdu corpus to cover most of the language.

We are also planning to collaborate with Mozilla’s team and contribute in their project Common Voice to go live with the collected Urdu data and also our developed Urdu language and acoustic model. Additionally by adjusting hyper parameters and more data results could be achieved near to human level recognition.

AP PENDIX A: PARAMETERS FOR THE MODEL

--alphabet_config_path: path to the configuration file specifying the alphabet used by the network. See the comment in data/alphabet.txt for a description of the format.

(default: 'data/alphabet.txt')

--b1_stddev: standard deviation to use when initialising b1

(a number)

--b2_stddev: standard deviation to use when initialising b2

(a number)

--b3_stddev: standard deviation to use when initialising b3

(a number)

--b5_stddev: standard deviation to use when initialising b5

(a number)

--b6_stddev: standard deviation to use when initialising b6

(a number)

--beam_width: beam width used in the CTC decoder when building candidate transcriptions

(default: '1024')

(an integer)

--beta1: beta 1 parameter of Adam optimizer

(default: '0.9')

(a number)

--beta2: beta 2 parameter of Adam optimizer

(default: '0.999')

(a number)

--checkpoint_dir: directory in which checkpoints are stored - defaults to directory "deepspeech/checkpoints" within user's data home specified by the XDG Base Directory Specification

(default: '')

--checkpoint_secs: checkpoint saving interval in seconds

(default: '600')

(an integer)

--coord_host: coordination server host

(default: 'localhost')

--coord_port: coordination server port

(default: '2500')

(an integer)

--coord_retries: number of tries of workers connecting to training coordinator before failing

(default: '100')

(an integer)

--decoder_library_path: path to the libctc_decoder_with_kenlm.so library containing the decoder implementation.

(default: 'native_client/libctc_decoder_with_kenlm.so')

--default_stddev: default standard deviation to use when initialising weights and biases

(default: '0.046875')

(a number)

--dev_batch_size: number of elements in a validation batch

(default: '1')

(an integer)

--dev_files: comma separated list of files specifying the dataset used for validation. multiple files will get merged

(default: '')

--display_step: number of epochs we cycle through before displaying detailed progress - 0 means no progress display

(default: '0')

(an integer)

--dropout_rate: dropout rate for feedforward layers

(default: '0.05')

(a number)

--dropout_rate2: dropout rate for layer 2 - defaults to dropout_rate

(default: '-1.0')

(a number)

--dropout_rate3: dropout rate for layer 3 - defaults to dropout_rate

(default: '-1.0')

(a number)

--dropout_rate4: dropout rate for layer 4 - defaults to 0.0

(default: '0.0')

(a number)

--dropout_rate5: dropout rate for layer 5 - defaults to 0.0

(default: '0.0')

(a number)

--dropout_rate6: dropout rate for layer 6 - defaults to dropout_rate

(default: '-1.0')

(a number)

--[no]early_stop: enable early stopping mechanism over validation dataset. Make sure that dev FLAG is enabled for this to work

(default: 'true')

--earlystop_nsteps: number of steps to consider for early stopping. Loss is not stored in the checkpoint so when checkpoint is revived it starts the loss calculation from start at that point

(default: '4')

(an integer)

--epoch: target epoch to train - if negative, the absolute number of additional epochs will be trained

(default: '75')

(an integer)

--epsilon: epsilon parameter of Adam optimizer

(default: '1e-08')

(a number)

--estop_mean_thresh: mean threshold for loss to determine the condition if early stopping is required

(default: '0.5')

(a number)

--estop_std_thresh: standard deviation threshold for loss to determine the condition if early stopping is required

(default: '0.5')

(a number)

--export_dir: directory in which exported models are stored - if omitted, the model won't get exported

(default: '')

--export_version: version number of the exported model

(default: '1')

(an integer)

--[no]fulltrace: if full trace debug info should be generated during training

(default: 'false')

--h1_stddev: standard deviation to use when initialising h1

(a number)

--h2_stddev: standard deviation to use when initialising h2

(a number)

--h3_stddev: standard deviation to use when initialising h3

(a number)

--h5_stddev: standard deviation to use when initialising h5

(a number)

--h6_stddev: standard deviation to use when initialising h6

(a number)

--initialize_from_frozen_model: path to frozen model to initialize from. This behaves like a checkpoint, loading the weights from the frozen model and starting training with those weights. The optimizer

parameters aren't restored, so remember to adjust the learning rate.

(default: '')

--iters_per_worker: number of train or inference iterations per worker before results are sent back to coordinator

(default: '1')

(an integer)

--job_name: job name - one of localhost (default), worker, ps

(default: 'localhost')

--learning_rate: learning rate of Adam optimizer

(default: '0.001')

(a number)

--limit_dev: maximum number of elements to use from validation set- 0 means no limit

(default: '0')

(an integer)

--limit_test: maximum number of elements to use from test set- 0 means no limit

(default: '0')

(an integer)

--limit_train: maximum number of elements to use from train set - 0 means no limit

(default: '0')

(an integer)

--lm_binary_path: path to the language model binary file created with KenLM

(default: 'data/lm/lm.binary')

--lm_trie_path: path to the language model trie file created with native_client/generate_trie

(default: 'data/lm/trie')

--lm_weight: the alpha hyperparameter of the CTC decoder. Language Model weight.

(default: '1.75')

(a number)

--log_level: log level for console logs - 0: INFO, 1: WARN, 2: ERROR, 3: FATAL

(default: '1')

(an integer)

--[no]log_placement: whether to log device placement of the operators to the console

(default: 'false')

--[no]log_traffic: log cluster transaction and traffic information during debug logging

(default: 'false')

--max_to_keep: number of checkpoint files to keep - default value is 5

(default: '5')

(an integer)

--n_hidden: layer width to use when initialising layers

(default: '2048')

(an integer)

--one_shot_infer: one-shot inference mode: specify a wav file and the script will load the checkpoint and perform inference on it. Disables training, testing and exporting.

(default: '')

--ps_hosts: parameter servers - comma separated list of hostname:port pairs

(default: '')

--random_seed: default random seed that is used to initialize variables

(default: '4567')

(an integer)

--relu_clip: ReLU clipping value for non-recurrant layers

(default: '20.0')

(a number)

--[no]remove_export: whether to remove old exported models

(default: 'false')

--replicas: total number of replicas - if negative, its absolute value is multiplied by the number of workers

(default: '-1')

(an integer)

--replicas_to_agg: number of replicas to aggregate - if negative, its absolute value is multiplied by the number of workers

(default: '-1')

(an integer)

--report_count: number of phrases with lowest WER (best matching) to print out during a WER report

(default: '10')

(an integer)

--summary_dir: target directory for TensorBoard summaries - defaults to directory "deepspeech/summaries" within user's data home specified by the XDG Base Directory Specification

(default: '')

--summary_secs: interval in seconds for saving TensorBoard summaries - if 0, no summaries will be written

(default: '0')

(an integer)

--task_index: index of task within the job - worker with index 0 will be the chief

(default: '0')

(an integer)

--[no]test: whether to test the network

(default: 'true')

--test_batch_size: number of elements in a test batch

(default: '1')

(an integer)

--test_files: comma separated list of files specifying the dataset used for testing. multiple files will get merged

(default: '')

--[no]train: whether to train the network

(default: 'true')

--train_batch_size: number of elements in a training batch

(default: '1')

(an integer)

--train_files: comma separated list of files specifying the dataset used for training. multiple files will get merged

(default: '')

--[no]use_seq_length: have sequence_length in the exported graph (will make tfcompile unhappy)

(default: 'true')

--[no]use_warpctc: whether to use GPU bound Warp-CTC

(default: 'false')

--valid_word_count_weight: valid word insertion weight. This is used to lessen the word insertion penalty when the inserted word is part of the vocabulary.

(default: '1.0')

(a number)

--validation_step: number of epochs we cycle through before validating the model - a detailed progress report is dependent on "--display_step" - 0 means no validation steps

(default: '0')

(an integer)

--wer_log_pattern: pattern for machine readable global logging of WER progress; has to contain %%s, %%s and %%f for the set name, the date and the float respectively; example: "GLOBAL LOG:

logwer('12ade231', %%s, %%s, %%f)" would result in some entry like "GLOBAL LOG: logwer('12ade231', 'train', '2017-05-18T03:09:48-0700', 0.05)"; if omitted (default), there will be no logging

(default: '')

--word_count_weight: the beta hyperparameter of the CTC decoder. Word insertion weight (penalty).

(default: '1.0')

(a number)

--worker_hosts: workers - comma separated list of hostname:port pairs

(default: '')

tensorflow.python.platform.app:

-h,--[no]help: show this help

(default: 'false')

--[no]helpfull: show full help

(default: 'false')

--[no]helpshort: show this help

(default: 'false')

absl.flags:

--flagfile: Insert flag definitions from the given file into the command line.

(default: '')

--undefok: comma-separated list of flag names that it is okay to specify on the command line even if the program does not define a flag with that name. IMPORTANT: flags in this list that have arguments

MUST use the --flag=value format.

(default: '')

REFERENCES

[1] H. Xuedong, A. Alex and H. Hsiao-Wuen, “Spoken Language Processing: A guide to theory, algorithm, and system development,” 2001.

[2] D. Britz, “WILDML,” 17 September 2015. [Online]. Available: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/. [Accessed 21 March 2018].

[3] C. Olah, “Colah's blog,” 27 August 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 21 March 2018].

[4] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, Berlin: Springer-Verlag Berlin Heidelberg, 2012.

[5] S. Kostadinov, “Towards Data Science,” 2 December 2017. [Online]. Available: https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7. [Accessed 13 February 2018].

[6] “A Beginner’s Guide to Recurrent Networks and LSTMs,” [Online]. Available: https://deeplearning4j.org/lstm.html. [Accessed 13 March 2018].

[7] S. Hochreiter and S. Jürgen, “Long short-term memory,” *Neural Computation,* pp. 1735-1780, September 1997.

[8] A. Hannun, “Sequence Modeling with CTC,” *Distill,* 2017.

[9] A. Graves, S. Fernandez, F. Gomez and J. Schmidhuber, *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,* Pittsburgh, Pennsylvania: ICML '06 Proceedings of the 23rd international conference on Machine learning, 2006.

[10] D. Britz, “WILDML,” 3 January 2016. [Online]. Available: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/. [Accessed 12 March 2018].

[11] Y. Park, S. Patwardhan, K. Visweswariah and S. C. Gates, “An Empirical Analysis of Word Error Rate and Keyword Error Rate,” in *Ninth Annual Conference of the International Speech Communication Association*, Brisbane, 2008.

[12] I. S. MacKenzie and S. R. William, "A character-level error analysis technique for evaluating text entry methods," in *Proceedings of the second Nordic conference on Human-computer interaction*, Aarhus, 2002.

[13] Asadullah, Shaukat, A., Ali, H., & Akram, M.U. (2016). Automatic Urdu Speech Recognition using Hidden Markov Model. 2016 International Conference on Image, Vision and Computing (ICIVC), 135-139.

[14] Abbas Ali, Syed & Khan, Sallar & Perveen, Humaira & Muzzamil, Reham & Malik, Mahnoor & Khalid, Faiza. (2017). Urdu Language Translator using Deep Neural Network. Indian Journal of Science and Technology. 10. 1-7. 10.17485/ijst/2017/v10i40/120273.

[15] Hannun, Awni & Case, Carl & Casper, Jared & Catanzaro, Bryan & Diamos, Greg & Elsen, Erich & Prenger, Ryan & Satheesh, Sanjeev & Sengupta, Shubho & Coates, Adam & Y. Ng, Andrew. (2014). DeepSpeech: Scaling up end-to-end speech recognition.

[16] Ali, Hazrat & Khawaja, Muhammad &, Yahya & Ahmad, Nasir & Farooq, Omar. (2012). A Medium Vocabulary Urdu Isolated Words Balanced Corpus for Automatic Speech Recognition. Proceedings of 4th International Conference on Electronic Computer Technology, ICECT.

[17] Amodei, Dario & Ananthanarayanan, Sundaram & Anubhai, Rishita & Bai, Jingliang & Battenberg, Eric & Case, Carl & Casper, Jared & Catanzaro, Bryan & Cheng, Qiang & Chen, Guoliang & Chen, Jie & Chen, Jingdong & Chen, Zhijie & Chrzanowski, Mike & Coates, Adam & Diamos, Greg & Ding, Ke & Du, Niandong & Elsen, Erich & Zhu, Zhenyao. (2015). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.

[18] Battenberg, Eric & Chen, Jitong & Child, Rewon & Coates, Adam & Gaur Yi Li, Yashesh & Liu, Hairong & Satheesh, Sanjeev & Sriram, Anuroop & Zhu, Zhenyao. (2017). Exploring neural transducers for end-to-end speech recognition. 206-213. 10.1109/ASRU.2017.8268937.

[19] S Shaleva, Anna & Degtyarev, Alexander & S Sedova, Olga & Iakushkin, Oleg & Fedoseev, George. (2018). Russian-Language Speech Recognition System Based on Deepspeech.

[20] Abdel-Hamid, O., Rahman Mohamed, A., Jiang, H. & Penn, G. (2012). Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. ICASSP (p. /pp. 4277-4280), IEEE. ISBN: 978-1-4673-0046-9.

[21] Sainath, Tara & Vinyals, Oriol & Senior, Andrew & Sak, Hasim. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. 4580-4584. 10.1109/ICASSP.2015.7178838.

[22] Sarfraz, Huda & Hussain, Sarmad & Bokhari, Riffat & Raza, Agha Ali & Ullah, Inam & Sarfraz, Zahid & Pervez, Sophia & Mustafa, Asad & Javed, Iqra &

Parveen, Rahila. (2010). Large vocabulary continuous speech recognition for Urdu. 1. 10.1145/1943628.1943629.

[23] Agha Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah, Zahid Sarfraz, An ASR System for Spontaneous Urdu Speech, Oriental COCOSDA 2010 conference, Nov. 24-25, 2010, Katmandu, Nepal.

[24] Agha Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah, Zahid Sarfraz, Design and development of phonetically rich Urdu speech corpus, Proceedings of O-COCOSDA'09 and IEEE Xplore; O-COCOSDA'09, 10-13 Aug 2009, School of Information Science and Engineering of Xinjiang University, Urunqi, China (URL: http://o-cocosda2009.xju.edu.cn).

[25] Heafield, Kenneth. (2011). KenLM: Faster and smaller language model queries.

[26] D. Amodei, S. Ananthanarayanan, R. B. J. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen and J. Chen, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin," in *International Conference on Machine Learning*, New York City, 2016.

[27] W. Chan, N. Jaitly, Q. V. Le and O. Vinyals, "Listen, Attend and Spell," in *arXiv preprint arXiv:1508.01211*, 2015.

[28] C. C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina and N. Jaitly, "State-of-the-art speech recognition with sequence-to-sequence models," arXiv, 2017.

[29] A. Tjandra, S. Sakti and S. Nakamura, "Local Monotonic Attention Mechanism for End-to-End Speech and Language Processing," in In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, 2017.

[30] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao and N. Jaitly, "An analysis of "attention" in sequence-to-sequence models," in Interspeech, Stockholm, 2017.

[31] J. Hou, Z. Shiliang and D. Lirong, "Gaussian Prediction based Attention for Online End-to-End Speech Recognition," in Interspeech, Stockholm, 2017.

[32] P. M. H. R. S. Doetsch and Ney, "Inverted Alignments for End-to-End Automatic Speech Recognition," IEEE Journal of Selected Topics in Signal Processing , vol. 11, no. 8, pp. 1265-1273, 2017.

[33] S. Kim, T. Hori and S. Watanabe, "Joint CTC/attention decoding for end-to-end speech recognition," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017.

[34] S. Watanabe, "ESPnet," 2017. [Online]. Available: https://espnet.github.io/espnet/. [Accessed 3 January 2018].

[35] H. Soltau, H. Liao and H. Sak, "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition," in Interspeech, Stockholm, 2017.

[36] T. Zenkel, R. Sanabria, F. Metze, J. Niehues, M. Sperber, S. Stüker and A. Waibel, "Comparison of Decoding Strategies for CTC Acoustic Models," in Interspeech, Stockholm, 2017.

[37] O. Siohan, "CTC Training of Multi-Phone Acoustic Models for Speech Recognition," in Interspeech, Stockholm, 2017.

[38] R. Collobert, C. Puhrsch and G. Synnaeve, "Wav2letter: an end- to-end convnet-based speech recognition system," arXiv, 2016.

[39] Y. Zhou, C. Xiong and R. Socher, "Improving End-to-End Speech Recognition with Policy Learning," arXiv, 2017.

[40] Agha Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah, Zahid Sarfraz, An ASR System for Spontaneous Urdu Speech, Oriental COCOSDA 2010 conference, Nov. 24-25, 2010, Katmandu, Nepal.

[41] S. Watanabe, T. Hori, S. Kim, J. R. Hershey and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," IEE Journal, vol. 11, no. 8, pp. 1240-1253, 2017.

[42] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, Santa Barbara, CA, USA, 1997.