ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Halil İbrahim YALMAN

COMPARISON OF RNN-CTC, LSTM-CTC AND GRU-CTC MODELS AND PARAMETERS ON A NEW TURKISH AUDIOBOOK DATASET

DEPARTMENT OF COMPUTER ENGINEERING

ADANA-2022

ABSTRACT

MSc THESIS

COMPARISON OF RNN-CTC, LSTM-CTC AND GRU-CTC MODELS AND PARAMETERS ON A NEW TURKISH AUDIOBOOK DATASET

Halil İbrahim YALMAN

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
Year: 2022, Pages: 57
Jury : Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
: Prof. Dr. Umut ORHAN
: Asst. Prof. Dr. Gökay DİŞKEN

Speech is very important in human communication. Speech recognition systems work to convert sounds and text. Devices that use speech recognition systems make daily life easier. Although there are many studies on Turkish speech recognition systems, the lack of data sets is obvious.

In this thesis, the original Turkish Audiobook Dataset was developed and neural network models were examined. An original data set obtained from audiobook recordings was prepared. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), Connectionist Temporal Classification (CTC) models were examined and compared on the obtained data set.

Keywords: Long Short Term Memory, Recurrent Neural Network, Gated Recurrent Units, Connectionist temporal classification, Turkish Audiobook Dataset

ÖZ

YÜKSEK LİSANS TEZİ

YENİ BİR TÜRKÇE SESLİ KİTAP VERİ SETİ ÜZERİNDE RNN-CTC, LSTM-CTC VE GRU-CTC MODELLERİNİN VE PARAMETRELERİNİN KARŞILAŞTIRILMASI

Halil İbrahim YALMAN

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Doç. Dr. Zekeriya TÜFEKÇİ
Yıl: 2022, Sayfa: 57
Juri : Doç. Dr. Zekeriya TÜFEKÇİ
: Prof. Dr. Umut ORHAN
: Dr. Öğr. Üyesi Gökay DİŞKEN

İnsan iletişiminde konuşma çok önemlidir. Konuşma tanıma sistemleri, sesleri ve metinleri dönüştürmek için çalışır. Konuşma tanıma sistemlerini kullanan cihazlar günlük hayatı kolaylaştırmaktadır. Türkçe konuşma tanıma sistemleri ile ilgili çok sayıda çalışma olmasına rağmen veri setlerinin eksikliği aşikardır.

Bu tezde özgün Türkçe Sesli Kitap Veri Kümesi geliştirilmiş ve sinir ağı modelleri incelenmiştir. Sesli kitap kayıtlarından elde edilen özgün bir veri seti hazırlanmıştır. Elde edilen veri seti üzerinde Evrişimsel Sinir Ağları (ESA), Yinelemeli Sinir Ağı (YSA), Uzun Kısa Süreli Bellek (UKSB), Geçitli Tekrarlayan Birimler (GTB), Bağlantıcı Zaman Sınıflandırma (BZS) modelleri incelenmiş ve karşılaştırılmıştır.

Anahtar Kelimeler: Uzun Kısa Süreli Bellek, Basit Tekrarlayan Ağlar, Kapılı Tekrarlayan Hücreler, Bağlantıcı Zamansal Sınıflandırma, Türkçe Sesli Kitap Veri seti

EXTENDED ABSTRACT

The most basic means of communication between people is speech. Writing is one of the most important tools in communication, along with speech. Speech recognition can be defined as finding the text equivalent of information in sound waves. Speech recognition allows us to verbally give a written command to a machine. Speech recognition has gained even more importance with technological developments. Speech recognition is a natural way to communicate and command with machines.

The effort of technology to facilitate the life of humanity has also paved the way for advances in the development of speech recognition systems. With the increase in the processing capacity of computers in recent years, speech recognition, which has become one of the topics that has increased in popularity, has made serious progress, especially with the parallel use of video cards.

In the previous studies, there are different sizes of audiobook dataset studies in different languages. English, German, Dutch, Russian French Spanish, Polish, Italian, Portuguese etc. There are studies of different sizes in languages. Although there is no dataset created from the Turkish audio book, there are studies to create a dataset created from other sources.

There are dataset studies in Turkish language created from different sources. However, these studies are limited in number. Due to this deficiency, to create a dataset in our study, 15,000 sentences of audio data and text data were prepared from the books voiced by 42 people, 21 females and 21 males, and 19 hours, 4 minutes and 14 seconds of audio data were prepared. The dataset with 15000 samples is named dataset-4. The dataset with 7500 samples is named dataset-3. The dataset consisting of 3750 samples is named dataset-2. The dataset consisting of 3000 samples is named dataset-2.

In our experiments, 97% of each data set was partitioned to be used for training and 3% for testing. Feature extraction was performed with the Mel

Spectrogram model of the dataset audio files. CNN+LSTM/CTC, CNN+RNN/CTC, CNN+GRU/CTC, CNN+BIRNN/CTC, CNN+BILSTM/CTC and CNN+BIGRU/CTC models were used in Neural Networks.

In this study, the comparison of models with equal parameters, the effect of increasing the node on the performance of the models and the effect of increasing the data size on the performance of the models were investigated.

As a result of the comparison of the models with equal parameters, it was determined that the LSTM model gave more successful results than the other models in unidirectional and bidirectional models.

In the study of the effect of node increase on the performance of the models, the number of layers was kept the same, the number of nodes was doubled and an increase in success rates was observed. Node increase had a positive contribution to the performance in all models.

In the study, in which the effect of the increase in data size on the performance of the models was examined, CNN+LSTM/CTC, CNN+RNN/CTC, CNN+GRU/CTC CNN+BILSTM/CTC, CNN+BIRNN/CTC and CNN+BIGRU/CTC models were used in the 5 hour, 10 hour and 20 hour data sets and the success rates were compared. CNN+LSTM/CTC model was found to be more successful than CNN+ RNN/CTC and CNN+GRU/CTC models. It was also observed that the CNN+BILSTM/CTC model was more successful than the CNN+BIRNN/CTC and CNN+BIGRU/CTC models. Another conclusion that can be drawn from this study is that bidirectional models give more successful results than unidirectional models. The GRU unidirectional and bidirectional model drew attention with its strikingly low operating time, as well as giving close results to the LSTM model. It can be said that the most successful model is LSTM, but the GRU model is also performing with its runtime and success rate.

The importance of the size of the data set in future studies shows itself seriously. A data set for speech recognition was created using Turkish audiobook recordings for use in future studies.

GENİŞLETİLMİŞ ÖZET

İnsanlar arasındaki en temel iletişim aracı konuşmadır. Yazı, konuşma ile iletişimde en önemli araçlar arasındadır. Konuşma tanıma ses dalgalarındaki bilgilerin metin karşılığını bulma olarak tanımlanabilir. Konuşma tanıma bir makineye yazılı olarak verdiğimiz komutu sesli olarak da vermemizi sağlar. Konuşma tanıma konusu teknolojik gelişmelerle birlikte daha da önem kazanmıştır. Konuşma tanıma makinelerle iletişim ve komut vermede doğal bir yol olarak karşımıza çıkmaktadır.

Teknolojinin insanlığın hayatını kolaylaştırma çabası konuşma tanıma sistemlerini gelişmesinde de ilerlemelerin önünü açmıştır. Son yıllardaki bilgisayarların işlem kapasitesinin artması ile popülaritesi artan konulardan biri haline gelen konuşma tanıma, özellikle ekran kartlarının paralel olarak kullanılması ile ciddi gelişme kaydetmiştir.

Daha önceki çalışmalarda farklı dillerde farklı boyutlarda sesli kitap veri seti çalışmaları bulunmaktadır. İngilizce, Almanca, Felemenkçe, Rusça Fransızca İspanyolca, Lehçe, İtalyanca, Portekizce vb. Dillerde farklı boyutlarda çalışmalar mevcuttur. Türkçe sesli kitaptan oluşturulmuş bir veri seti bulunmamakla birlikte, başka kaynaklardan oluşturulmuş bir veri seti oluşturmaya yönelik çalışmalar mevcuttur.

Türkçe dilinde farklı kaynaklardan oluşturulan veri seti çalışmaları vardır. Ancak bu çalışmalar kısıtlı sayıdadır. Bu eksiklikten dolayı çalışmamızda veri seti oluşturmak amacıyla 21 kadın, 21 erkek olmak üzere 42 kişinin seslendirdiği kitaplardan 15.000 cümle ses verisi ile metin verisi hazırlanarak 19 saat 4 dakika ve 14 saniye ses verisi hazırlanmıştır. 15000 örnekli veri seti veriseti-4 olarak adlandırılmıştır. 7500 örnekli veri seti veriseti-3 olarak adlandırılmıştır. 3750 örnekten oluşan veri seti veriseti-2 olarak adlandırılmıştır. 3000 örnekten oluşan veri seti veriseti-2 olarak adlandırılmıştır.

Deneylerimizde, her bir veri setinin %97'si eğitim ve %3'ü test için kullanılmak üzere bölümleme yapılmıştır. Veri seti ses dosyalarının Mel Spectrogram modeli ile özellik çıkarımı gerçekleştirilmiştir. Sinir Ağlarında CNN+LSTM/CTC, CNN+RNN/CTC, CNN+GRU/CTC, CNN+BIRNN/CTC, CNN+BILSTM/CTC ve CNN+BIGRU/CTC modelleri kullanılmıştır.

Bu çalışmada Eşit parametreli modellerin karşılaştırılması, düğüm artışının modellerin performansına etkisi ve veri boyutunun artışının modellerin performansına etkisi araştırılmıştır.

Eşit parametreli modellerin karşılaştırılması sonucunda tek yönlü modellerde ve çift yönlü modellerde LSTM modelinin diğer modellere göre daha başarılı sonuçlar verdiği saptanmıştır.

Düğüm artışının modellerin performansına etkisi çalışmasında katman sayıları aynı tutularak düğüm sayıları iki katına çıkarılmış ve başarı oranlarında artış gözlenmiştir. Düğüm artışının bütün modellerde başarıma olumlu katkısı olmuştur.

Veri boyutundaki artışın modellerin performansına etkisinin incelendiği çalışmada 5 saat, 10 saat ve 20 saatlik veri setlerinde CNN+LSTM/CTC, CNN+RNN/CTC, CNN+GRU/CTC CNN+BILSTM/CTC, CNN+BIRNN/CTC ve CNN+BIGRU/CTC akustik modelleri kullanıldı ve başarı oranları karşılaştırıldı. CNN+LSTM/CTC modeli CNN+ RNN/CTC ve CNN+GRU/CTC modellerinden daha başarılı bulunmuştur. Ayrıca CNN+BILSTM/CTC modeli de CNN+BIRNN/CTC ve CNN+BIGRU/CTC modellerinden daha başarılı olduğu gözlemlendi. Bu çalışmadan çıkarılabilecek bir diğer sonuç, çift yönlü modellerin tek yönlü modellere göre daha başarılı sonuçlar verdiğidir. GRU tek yönlü ve çift yönlü modeli, LSTM modeline yakın sonuçlar vermesinin yanı sıra, dikkat çekici derecede düşük çalışma süresi ile dikkat çekti. En başarılı modelin LSTM olduğu söylenebilir ancak GRU modeli de çalışma süresi ve başarı oranı ile performans sergiliyor.

Gelecekte yapılacak çalışmalarda veri setinin büyüklüğünün önemi ciddi vaziyette kendini göstermektedir. Gelecek çalışmalarda kullanılması amacıyla

Türkçe sesli kitap kayıtları kullanılarak konuşma tanıma için bir veri seti oluşturulmuştur.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Assoc. Prof. Dr. Zekeriya TÜFEKCI for his endless support, guidance and motivation in the successful completion of this thesis.

I would like to thank members of thesis jury Prof. Dr. Umut ORHAN and Assoc. Dr. Gökay DİŞKEN for their constructive suggestions and corrections.

I would like to thank my dear wife Seda YALMAN, my daughter İpek Nevra YALMAN, my mother Hamide YALMAN, my father Mamo YALMAN and other family members for their limitless support.

CONTENTS **PAGES**

LIST OF TABLES **PAGES**

LIST OF FIGURES **PAGES**

ABBREVIATIONS

BIGRU : Bidirectional Gated Recurrent Unit
BILSTM : Bidirectional Long Short Term Memory
BIRNN : Bidirectional Recurrent Neural Network
CNN : Convolutional Neural Network
CTC : Connectionist Temporal Classification
GELU : Gaussian Error Linear Unit
GRU : Gated Recurrent Unit
HMM : Hidden Markov Model
LER : Label Error Rate
LPC : Linear Predictive Analysis
LPCC : Linear predictive cepstral coefficients
LSTM : Long Short Term Memory
MFCC : Mel-frequency cepstral coefficients
PLP : Perceptual Linear Predictive
RASTA : Relative Spectra Filtering of Log Domain Coefficients
RELU : Rectified Linear Unit
RNN : Recurrent Neural Network
SRN : Simple Recurrent Network

1. INTRODUCTION

The place of speech in human communication is indisputably important. However, in communication, writing is also very important along with speaking. Since machines used in daily life work with commands, it is necessary to convert speech into text.

Speech recognition systems are needed in all areas of life. For example, it can be used more comfortably while commanding a phone, driving an autonomous vehicle, controlling smart watches and wristbands, and in many other areas in the coming years. Because technological developments are evolving towards this point.

In electronic systems used, working only with commands is more troublesome than working with voice. For this reason, systems that can be commanded by voice are needed. This need has led to the development of speech recognition systems. Although it is known that the first studies were carried out at Bell Laboratory in the 1960s, the "Shoebox" developed by IBM in 1961 has an important place in this field.

The Markov Model was used in speech recognition studies in the 1970s, and the Hidden Markov Model (HMM) was used in the 1980s. In recent years, developments in mobile technology and autonomous systems and the ability to perform parallel and powerful operations in computer systems have accelerated the developments in this field.

In the 2010s, with the use of Neural Networks in speech recognition and the increase in the processing capacity of computers, the studies accelerated. CNN, RNN, LSTM and GRU Neural Network models are used for speech recognition.

Convolutional Neural Network (CNN) is a special model for image processing presented in the literature by LeCun et al. (1989). CNN is used as both feature extraction and acoustic models in speech recognition systems. Abdulhamid et al. (2014) proposed the model in which he used the CNN model as an acoustic

model. It has been shown that the CNN Model gives 6% to 10% better results than the Deep Neural Network model.

Rumelhart et al. (1986) proposed the Recurrent Neural Network (RNN) structure to process sequential data. Graves et al. (2013) also proposed an acoustic model of speech recognition using the deep RNN model. RNN model forms the basis of the LSTM and GRU models. LSTM and GRU models are RNN models. Bidirectional Recurrent Neural Network (BIRNN), a bidirectional version of the RNN model, was introduced by Schuster and Paliwal (1997).

Hochreiter and Schmidhuber (1997) presented an RNN model, the Long Short Term Memory (LSTM) model, to eliminate some of the shortcomings of the RNN model. Sak et al. (2014) proposed to use LSTM model for speech recognition. They showed that the LSTM model produces approximately 50% less error rate than the RNN model.

Cho et al. (2014) proposed Gated Recurrent Unit model by making simplifications in LSTM model. GRU model is an RNN model that has fewer gates than LSTM model and gives similar results compared to LSTM. Shewalkar (2018) showed that the LSTM model achieves better results than the GRU model. It achieved this result with the same number of nodes and longer running time of the LSTM model compared to the GRU model.

The bidirectional models of the RNN, LSTM, and GRU models are called Bidirectional Recurrent Neural Network (BIRNN), Bidirectional Long Short Term Memory (BILSTM), and Bidirectional Gated Recurrent Unit (BIGRU), respectively. Zeyer et al. (2017) showed that the BILSTM model gives approximately 8% better results than a feedforward neural network. Arisoy et al. (2015) explained that the BIRNN model gives 0.2% better results than the Unidirectional RNN.

Audio files and text files are sequence data and it is a problem for models to find the probability that which letter corresponds to which sound at which moment between these two files. Alex Graves et al. (2006) proposed the Connectionist Temporal Classification (CTC) model to overcome this problem. The CTC model is

used as a classifier to transform the outputs of the acoustic model into a conditional probability distribution over label sequence.

There are many dataset studies in many languages used in the world. It is necessary to create a Turkish dataset due to the lack of diversity and size of studies conducted in Turkish language.

There are datasets created from audiobooks in many languages. However, there is no dataset study created from audiobooks for Turkish. Some of the studies conducted in languages other than Turkish using audio books are given below.

One of the dataset studies is the dataset created by tagging the books downloaded from the Librivox site in 8 different languages within the scope of the Multilingual LibriSpeech project. This dataset contains approximately 45000 hours of English, 2000 hours of German, 1600 hours of Dutch, 1100 hours of French, 950 hours of Spanish, 260 hours of Italian, 170 hours of Portuguese, 110 hours of Polish. (Pratap et al.,2020)

Russian language audiobook dataset is the dataset created by tagging the books downloaded from the Librivox site, within the scope of the Russian LibriSpeech project. This dataset contains 98 hours of Russian voice data. (Bakhturina et al., 2021)

The Turkish Speech dataset, developed by Boğaziçi University, contains 194 hours of audio data obtained from television and radio programs. (Arisoy et al., 2009)

Another Turkish Speech dataset is the Middle East Technical University Turkish Microphone Speech dataset developed by the Middle East Technical University. This dataset contains 5.6 hours of audio data and text. (Salor et al., 2006)

Since some of the dataset in Turkish language have not been published, they cannot be a source for another researcher. Another problem is that the size of the studies is far from being sufficient.

The main feature extraction techniques of speech recognition are Linear predictive analysis (LPC), Linear predictive cepstral coefficients (LPCC), perceptual

linear predictive coefficients (PLP), Mel-frequency cepstral coefficients (MFCC), Power spectral analysis, Mel scale cepstral analysis, Relative spectra filtering of log domain coefficients (RASTA), First order derivative and Mel spectrogram. (Shrawankar and Thakare, 2013; Hwang et al.,2021)

In the second section, the details of the material and method we used are given. In the third section, experimental setup and results are presented. In the fourth section, conclusion and future works are given.

2. BASIC CONCEPT

2.1. Speech Recognition

Speech recognition is the process of converting speech data to text data. A dataset is required to perform speech recognition. In Figure 2.1, after the sound information in the dataset is pre-processed, the sound information is given to the acoustic model. The part called pre-processing is feature extraction. Acoustic neural network models are trained on training data. It is tested with test data and its performance is measured. If you have a language model, it is included in the process.

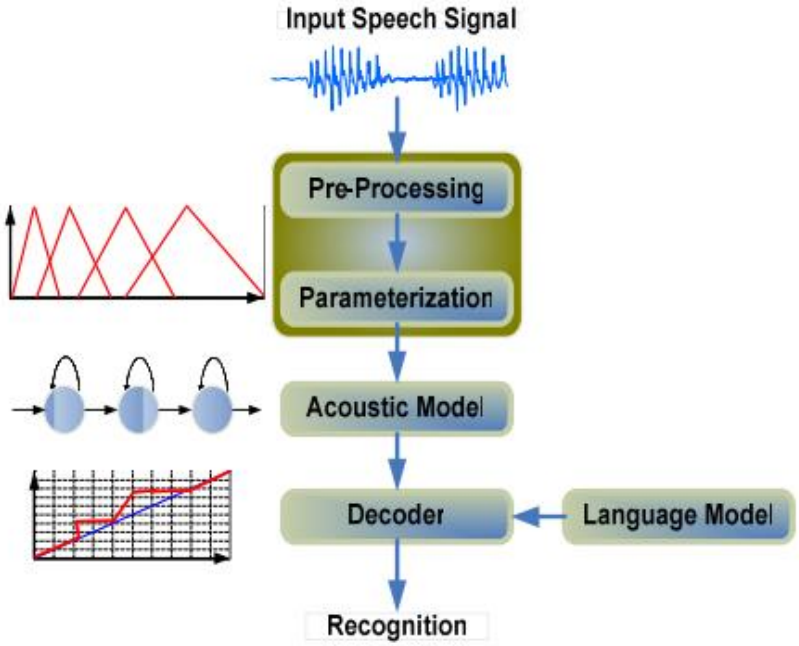

Figure 2.1. Block Diagram of a Speech Recognition System. (Mporas et al., 2007)

In recent years, developments in parallel processing, graphics card power and computer power have contributed to speech recognition. Therefore, developments in speech recognition have increased in parallel. In addition, the development of intelligent systems has increased the need for speech recognition.

As shown in Figure 2.1, the first step of the speech recognition process is to convert the raw audio data into matrix form by feature extraction. This process is explained in the next section.

2.2. Feature Extraction

In speech recognition, many different techniques are used for feature extraction of voice data. The most frequently used features techniques are the MFCC and Mel Spectrogram. Mel spectrogram was used in this study. The block diagram of Mel spectrogram feature extraction is presented in Figure 2.2.(Tak et al., 2017)

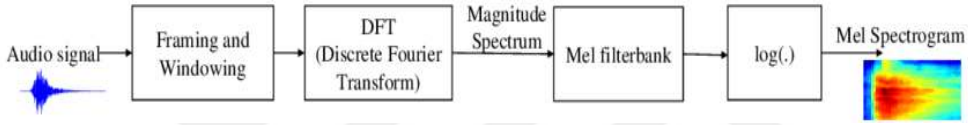

Figure 2.2. Block diagram of Mel Spectrogram feature extraction (Tak et al., 2017)

The first step in mel spectrogram feature extraction process is framing, which splits audio file into segments at specified milliseconds (ms). Since the frequency content of the speech signal does not change in the 20-30 millisecond interval, a value between 20 ms and 30 ms is used for the frame size. In this study, 20 ms frame size was used. Since the sampling frequency is 44100 Hz, 20 ms corresponds to 882 samples.

As shown in Figure 2.3, Frame Size and Step Size are determined for the framing process. The section with the specified Frame Size is selected. It is shifted in Step Size to determine a new frame on the signal. The process is repeated by selecting the Frame Size section again. For example, if a Frame Size of 20 ms and a Step Size of 10 ms are specified, the first frame will be framed between 0-20 ms and

the second frame between 10-30 ms.

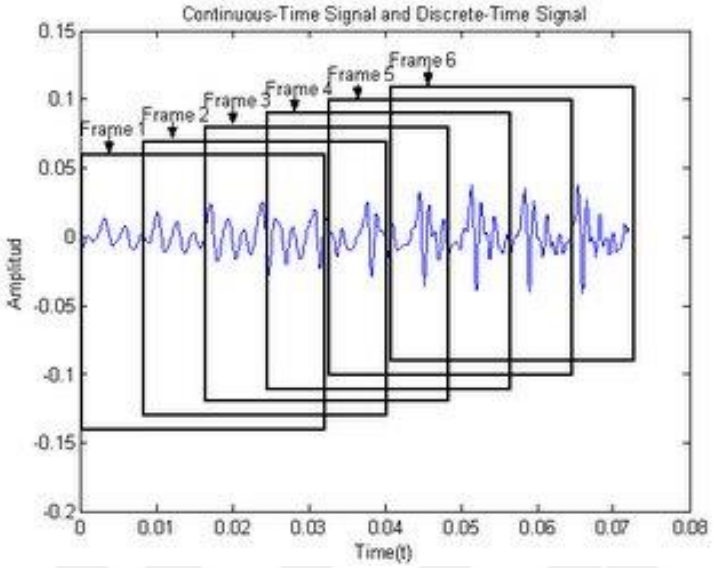

Figure 2.3. Framing Process

The second step is the windowing process. Various windows are used such as Hamming, Hanning, Bartlett, Blackman, Kaiser. The windowing process is performed in order not to avoid the data due to spectral leakage at the start and end points of the divided frame segments. By multiplying the split frame and one of the window types, the beginning and end of the frame are smoothed. Hanning Window shown in Figure 2.4 is used in this study.

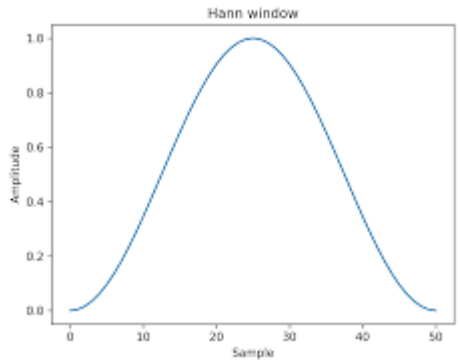

Figure 2.4. Hanning Window

In the third step, the Fourier transform is applied to frames of speech signal. The Fourier transform transforms a signal from the time domain to the frequency domain. Then magnitude of each frequency bin is computed to obtain magnitude spectrum.

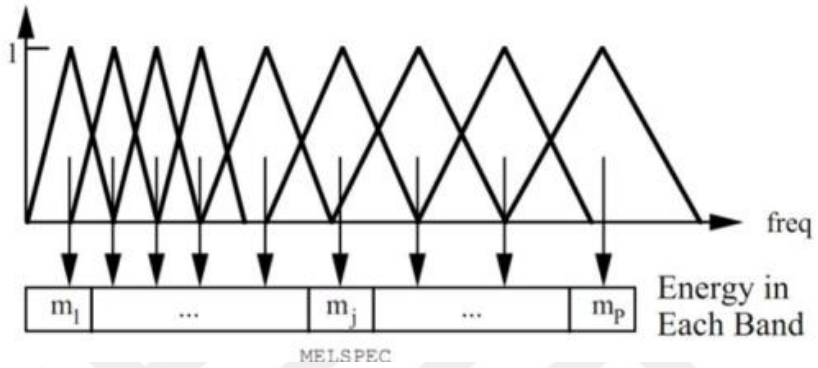

Figure 2.5. Mel-Scale Filter bank

The Mel Scale shown in Figure 2.5 is used in the fourth step. The Mel scale indicates the frequencies perceived by the human hearing organs. The human ear is more sensitive to low frequencies than high frequencies. Therefore, the center frequencies of the Mel scale filter bank used are adjusted according to the Mel scale. The frequency bandwidth increases in direct proportion to the frequency. The magnitudes in the filter bank are calculated by the inner product of the magnitude spectrum with these filter banks. This process results in Mel Spectrogram features.

2.3. Deep Learning

Deep learning is a sub-branch of machine learning that uses a set of algorithms that attempt to model high-level abstractions in data using a deep structure with multiple processing layers, thanks to multiple linear or non-linear structures. (Goodfellow et al., 2016). The most basic privilege of deep learning in machine learning is the use of hidden layers. Deep learning is implemented using neural network models with hidden layers.

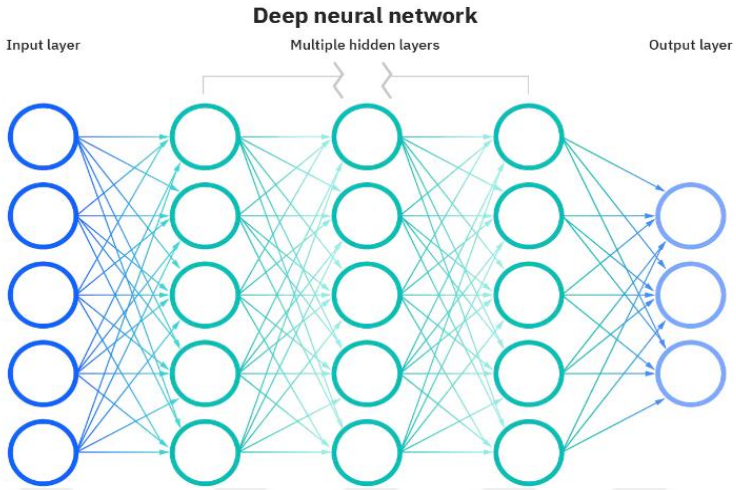

Figure 2.6. Structure of Deep Neural Network

As seen in Figure 2.6, the word "Deep" in Neural Network denotes hidden layers. In other words, it can be said that the deeper a neural network is, the more hidden layers it has.

Convolutional Neural Network and Recurrent Neural Network were used in this study. Three subtypes of Recurrent Neural Network, Simple Neural Network, Long Short Term Memory and Gated Recurrent Units were used.

2.3.1. Convolutional Neural Network

Convolutional networks, also known as convolutional neural networks or CNNs, are a specialized kind of neural network for processing data that has a known, grid-like topology. (Goodfellow et al. 2016)

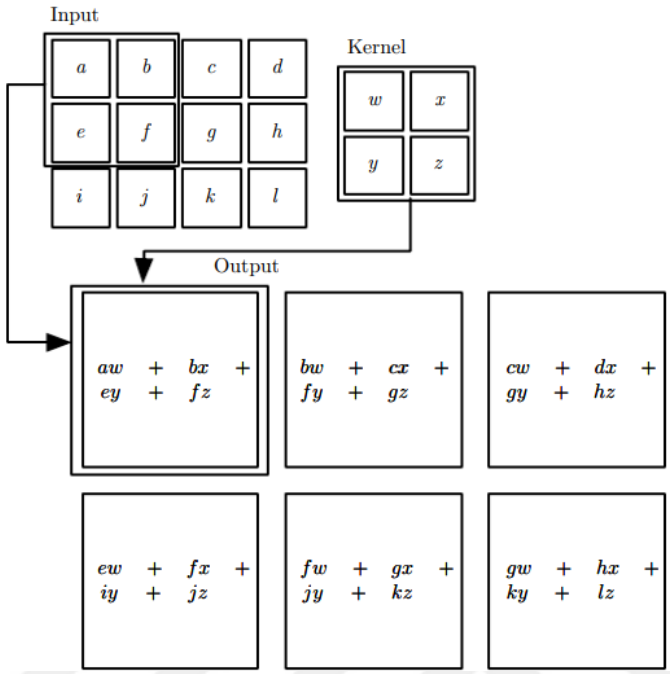

Figure 2.7. An example of 2-D convolution (Goodfellow et al. 2016)

As seen in Figure 2.7, the kernel is shifted at each step and multiplied by the kernel size of the input matrix. While doing this, we specify the number of steps to be shifted with the Stride parameter. As seen here, 8x3 matrix is reduced to 6x2 matrix. If we want the matrix to stay the same size, the process of adding a certain number to its sides is called padding. Abdulhamid et al. (2014) revealed in their study that it reduces the noise in the voice, reveals the clean parts and increases the performance. CNN can be used both as an acoustic model and as a feature extraction process in speech recognition processes.

2.3.2. Recurrent Neural Network(RNN)

Recurrent neural networks or RNNs (Rumelhart et al., 1986) are a family of neural networks for processing sequential data. (Goodfellow et al. 2016)

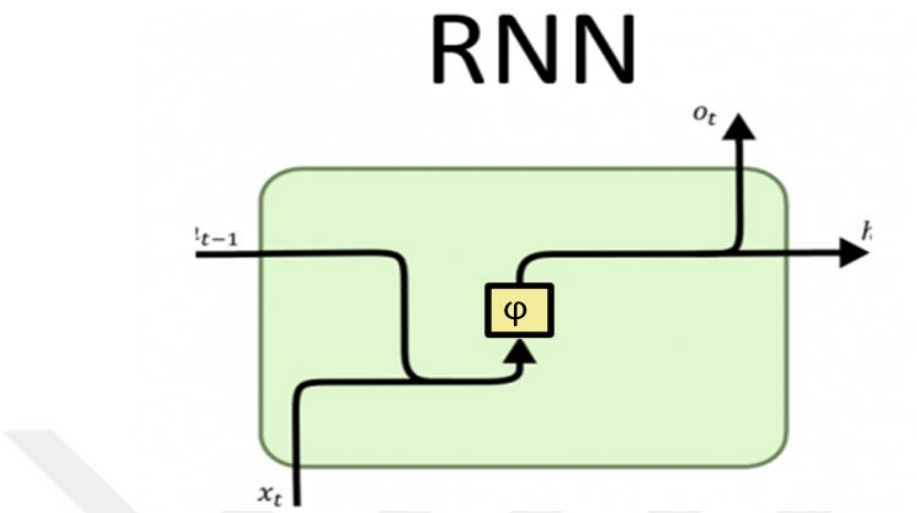

Figure 2.8. Structure of RNN

As shown in Figure 2.8, RNN is a neural network model in which only the current state is not important and the previous time state is also included. It is used in Speech Recognition, Handwriting Recognition, Natural language processing etc. In this section, Simple Recurrent Network (SRN), Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) models of RNN types are examined.

2.3.2.1. Simple Recurrent Network

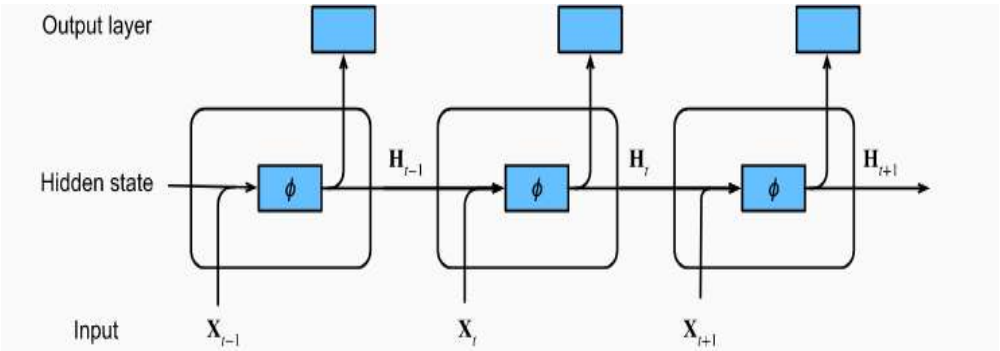

Figure 2.9. Structure of SRN

It works with a structure in which data from previous time hidden layer affects the current time output. As seen in Figure 2.9, the information in the hidden state is transferred sequentially from each time to the next time. While the previous time and current input information affect output information, they also influence the future time. SRN output equation is as follows:

$$h_t = \tanh (W_{xh} x_t + W_{hh} h_{t-1} + b_h) \quad (2.1)$$

$$y_t = \tanh (W_{hy} h_t + b_y) \quad (2.2)$$

In the Equation (2.1) and Equation (2.2), the current time input is shown as x_t, the hidden layer information from the previous time is h_{t-1} and the current time hidden layer information is ht. h_t is also transferred to the next time. W_{xh} , W_{hh} , and W_{hy} are input, hidden and output weight matrices, respectively. b_h and b_y are input and output bias vectors, respectively. The activation function tanh is Hyperbolic Tangent Function. Hyperbolic Tangent Function is explained in section 2.4.1.

Since SRN model is known as RNN model, RNN nomenclature will be used in this study.

2.3.2.2. Long Short Term Memory(LSTM)

Long Short Term Memory (LSTM) is a subtype of RNN architecture.

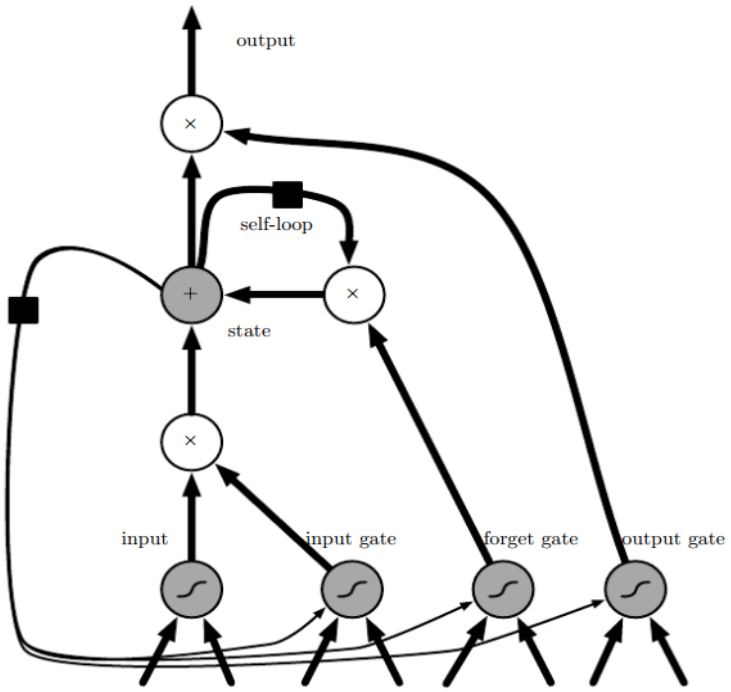

Figure 2.10. Structure of LSTM

The block diagram of LSTM is given in Figure 2.10. LSTM operates with 4 control gates. Among LSTM gates, forget gate decides which information will be forgotten, input gate decides which information to add, and output gate decides which information to be transferred to the other state. The weight matrices decide how much each door will affect. The LSTM equations describing the operation steps of this system are as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2.3)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (2.4)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \quad (2.5)$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \quad (2.6)$$

$$h_t = o_t tanh(c_t) \quad (2.7)$$

In LSTM Equation (2.3) to Equation (2.7), t represents the current time, t-1 represents the previous time. In Equation (2.3), input gate (i_t) is formulated. W_{xi}, W_{hi}, W_{ci} represent the weight matrices of current input, previous time hidden state, and the previous c_t, respectively. b_i is input gate bias vector. Forget gate (f_t) is formulated in Equation (2.4). W_{xf}, W_{hf}, W_{cf} represent the weight matrices of the current time input, previous time hidden state, and the previous c_t, respectively. bf is the forget gate bias vector. In Equation (2.5), cell state (c_t) is formulated. W_{xc}, W_{hc} represent the weight matrices of the current input and previous time hidden state, respectively. b_f is cell state bias vector. In Equation (2.6), Output gate (o_t) is formulated. W_{xo}, W_{ho}, W_{co} represent the weight matrices of the current input, the previous time hidden state, and the previous c_t, respectively. b_o is output gate bias vector. In Equation (2.7), the current output is obtained by multiplying the Hyperbolic Tangent Function of the cell state from the other equations with the output. σ represents Sigmoid Function. Sigmoid Function is explained in section 2.4.1.

2.3.2.3. Gated Recurrent Unit (GRU)

Gated Recurrent Units (GRU) is a subtype of RNN, which is a simplified version of LSTM system. When GRU structure is examined, the difference of the method from LSTM is that a single gate unit controls both the forget factor and the update decision of the state unit at the same time. (Goodfellow et al., 2016)

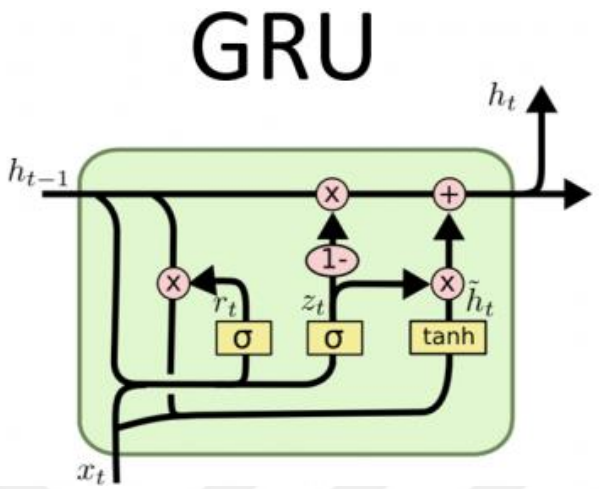

Figure 2.11. Structure of GRU

As shown in Figure 2.11, GRU structure consists of update unit (z_t), reset unit (r_t), and the candidate hidden state ($\tilde{h}_t$), which are used as forget and state units. GRU structure is formulated from Equation (2.8) to Equation (2.11) below.

$$r_t = \sigma\left(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}\right) \tag{2.8}$$

$$z_t = \sigma\left(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}\right) \tag{2.9}$$

$$\tilde{h}_t = \tanh\left(W_{in}x_t + b_{in} + r_t * \left(W_{hn}h_{(t-1)} + b_{hn}\right)\right) \tag{2.10}$$

$$h_t = (1 - z_t) * \tilde{h}_t + z_t * h_{(t-1)} \tag{2.11}$$

In Equations (2.8 - 12), the current input x_t, h_{t-1} represents the hidden layer in the previous time state. W_{ir}, W_{hr}, W_{iz}, W_{hz}, W_{in}, W_{hn} denote the weight matrix. b_{ir}, b_{hr}, b_{iz}, b_{hz}, b_{in}, b_{hn} are bias vectors.

2.3.2.4. Bidirectional Recurrent Neural Network

Bidirectional Recurrent Neural Network (BIRNN) is a type of RNN in which the current information is decided by the information received from both the previous and the next time node at the time t.

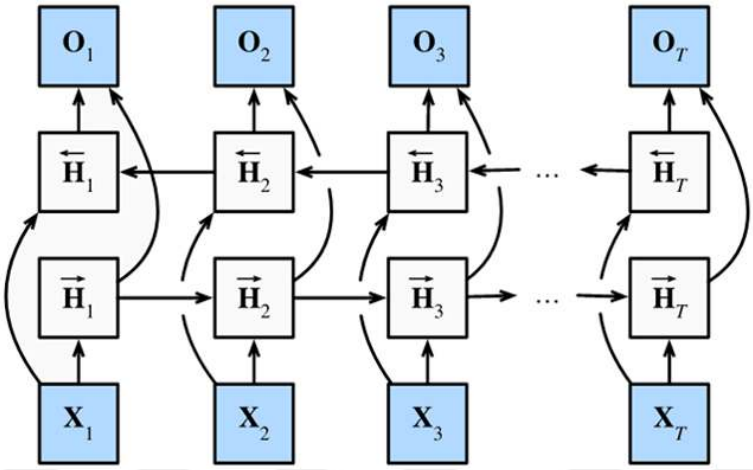

Figure 2.12. Structure of Bidirectional RNN(BIRNN)

As seen in Figure 2.12, while RNN model is only processed with the information from the previous time node, in bidirectional RNN version it processes both the previous and the next time node. Equation (3.12) to Equation (3.14) are equations which describe the BIRNN. (Zhang et al., 2021):

$$\overrightarrow{H}_t = \tanh(X_t W_{xh}^{(f)} + \overrightarrow{H}_{t-1} W_{hh}^{(f)} + b_h^{(f)}) \tag{2.12}$$

$$\overleftarrow{H}_t = \tanh(X_t W_{xh}^{(b)} + \overleftarrow{H}_{t+1} W_{hh}^{(b)} + b_h^{(b)}) \tag{2.13}$$

$$O_t = \tanh(H_t W_{hq} + b_q) \tag{2.14}$$

Forward hidden state and backward hidden state are shown in Equation (2.12). W denotes the weight matrix. b denotes bias vector. Arrows to the right ($\rightarrow$) indicate forwards, and arrows to the left ($\leftarrow$) indicate backwards. The matrix H_t is obtained by concatenating $\vec{H}_t$ and $\overleftarrow{H}_t$ .Bidirectional models have different parameters in matrix operations compared to unidirectional models because forward and backward structures are combined.

2.3.2.5. Bidirectional Long Short Term Memory

While a single layer provides information transfer in one direction in the LSTM model, the model in which forward and backward LSTM layers are concatenated is called Bidirectional Long Short-Term Memory (BILSTM). As seen in Figure 2.13, in a dataset, both the information from the previous time and the information from the next time affect the current information.

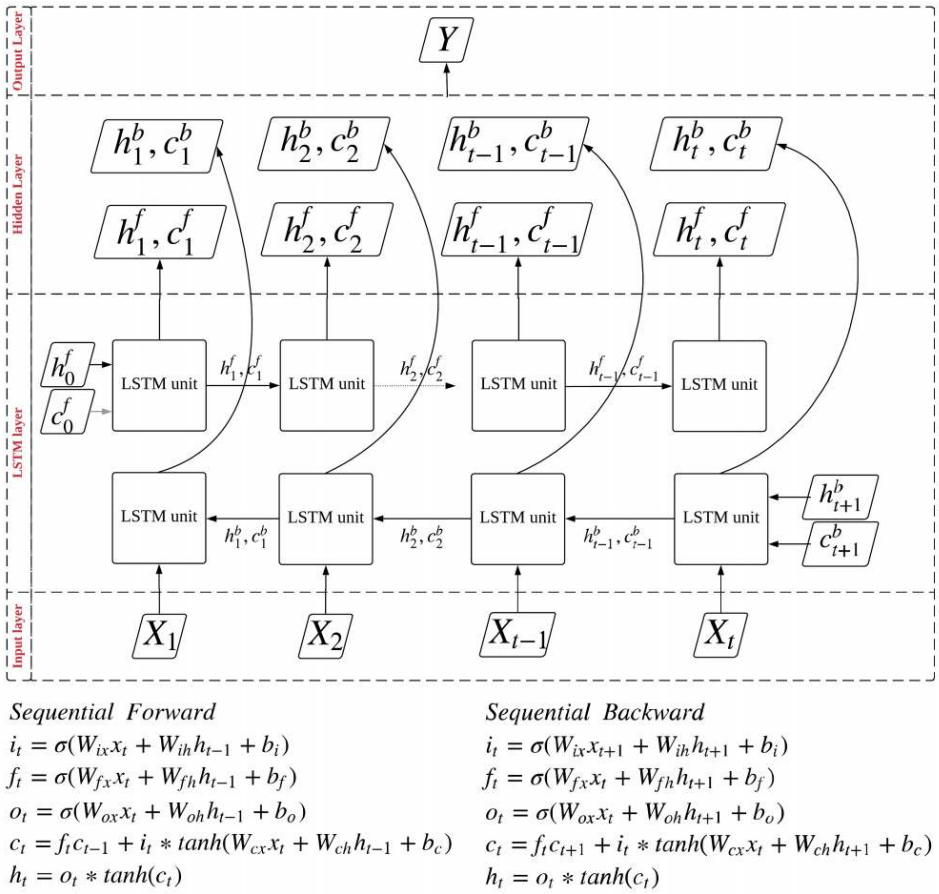

Figure 2.13. Structure and Formula of BILSTM (Hu and Zhang, 2018)

Bidirectional Long Short Term Memory (BILSTM), which is the system in which LSTM model works bidirectionally, is expressed by the formula as in Figure 2.13. Input gate, forget gate, output gate and cell state are the same as in LSTM model. Not only the previous state, but also the next state affects the current state. W symbols all represent weights and b symbols represent biases.

2.3.2.6. Bidirectional Gated Recurrent Unit

As shown in Figure 3.11, Bidirectional Gated Recurrent Unit (BIGRU) model is bidirectional of GRU model. It is GRU model in which the current state is affected by both the previous state and the next state.

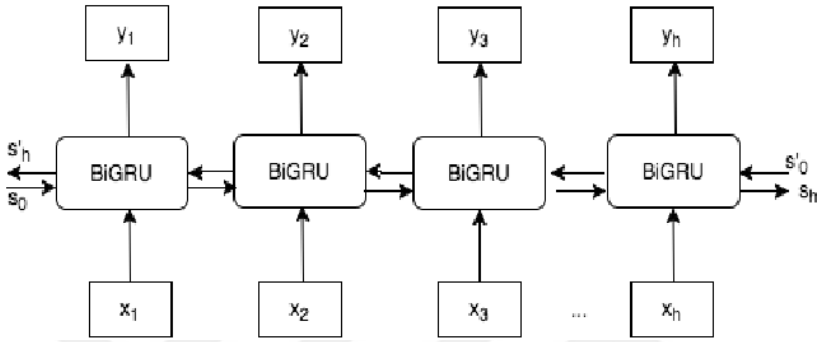

Figure 2.14. Structure of BIGRU

BIGRU model was formulated from Equation (2.15) to Equation (2.23) (Bhuvaneswari et al., 2019):

Right Direction:

$$\overrightarrow{z_t}^{(i)} = \partial\left(\vec{W}_{(i)}^{(z)} x_t^i + \vec{U}_{(i)}^{(r)} h_{t-1}^{(i)}\right) \tag{2.15}$$

$$\overrightarrow{r_t}^{(i)} = \partial\left(\vec{W}_{(i)}^{(r)} x_t^i + \vec{U}_{(i)}^{(r)} h_{t-1}^{(i)}\right) \tag{2.16}$$

$$\vec{\tilde{h}}_t^{(i)} = \tanh\left(\vec{W}_{(i)} x_t + r_t^{\circ} \vec{U}_{(i)} h_{t-1}\right) \tag{2.17}$$

$$\overrightarrow{h_t}^{(i)} = z_t^{(i)\circ} h_{t-1}^{(i)} + \left(1 - z_t^{(i)}\right)^{\circ} \tilde{h}_t^{(i)} \tag{2.18}$$

Left Direction:

$$\overleftarrow{z_t}^{(i)} = \partial\left(\bar{W}_{(i)}^{(z)} x_t^i + \bar{U}_{(i)}^{(r)} h_{t-1}^{(i)}\right) \tag{2.19}$$

$$\overleftarrow{r_t}^{(i)} = \partial\left(\bar{W}_{(i)}^{(r)} x_t^i + \bar{U}_{(i)}^{(r)} h_{t-1}^{(i)}\right) \tag{2.20}$$

$$\bar{h}_t^{(i)} = \tanh\left(\bar{W}_{(i)} x_t + r_t \circ \bar{U}_{(i)} h_{t-1}\right) \tag{2.21}$$

$$\overleftarrow{h_t}^{(i)} = z_t^{(i)} \circ h_{t-1}^{(i)} + \left(1 - z_t^{(i)}\right) \circ h_t^{(i)} \tag{2.22}$$

Output:

$$y_t = \text{softmax}(U\left[\overrightarrow{h_t}^{(top)}, \overleftarrow{h_t}^{(top)}\right] + a) \tag{2.23}$$

In Equations, update unit (z_t), reset unit (r_t), and candidate hidden state ($\tilde{h}_t$) is formulated and current input x_t, h_{t-1} represents hidden layer in previous time state. The arrow sign above the symbols indicates the right and left directions. W_{ir}, W_{hr}, W_{iz}, W_{hz}, W_{in}, W_{hn} denote the weight matrix and b_{ir}, b_{hr}, b_{iz}, b_{hz}, b_{in}, b_{hn} denote bias vector. Right and left direction matrices are concatenated in the output. Softmax represents Softmax function. Softmax function is explained in section 2.4.1.

2.3.3. Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) is a layer used for classification in Recurrent Neural Networks. CTC is a model that classifies the label data equivalents of sound waves. Synchronizing phoneme times and corresponding labels is a problem in speech recognition. CTC eliminates this problem.

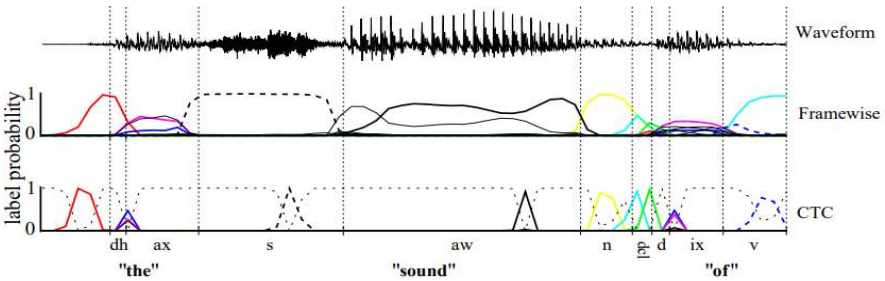

Figure 2.15. Framewise and CTC networks classifying a speech signal. (Graves et al., 2006)

As can be seen in Figure 2.15, when outputting the “the sound of” label, the probability of the letters chosen among all the letters of the possible alphabet and the space character is evaluated. As a result of this evaluation, output is taken according to CTC probability result.

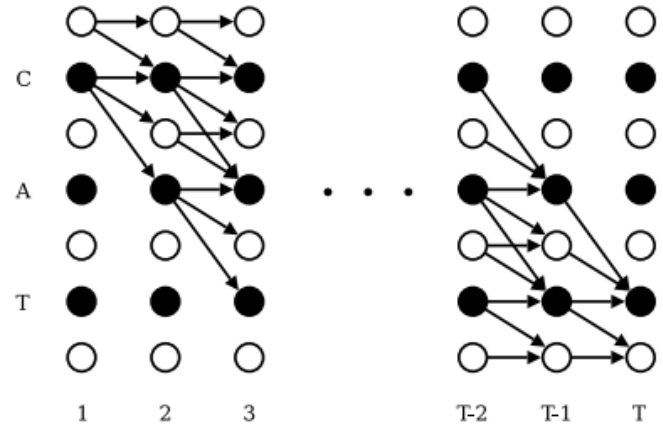

Figure 2.16. Examples of Word “cat” with CTC Model (Graves et al., 2006)

$$p(\mathbf{l} \mid \mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi \mid \mathbf{x}) \tag{2.24}$$

As shown in Figure 2.16 and Equation 2.24, the CTC model predicts a label because of the sum of the highest possible probabilities (Graves et al., 2006). The labels with the highest probability corresponding to the phonemes in each sound data are combined. Duplicates and spaces are deleted from these labels. As shown in Figure 2.16, the "cc-aaaa---t" label "cat" is the most likely result. As a result of these operations, the CTC model gives results.

2.4. Other Parameters

2.4.1. Activation Functions

In each node or node set in Neural Network structures, the activation function decides output of the node or node set. There are many different activation functions used.

- **Rectified linear unit (ReLU):** It is an activation function that provides a nonlinear transformation. As seen in Figure 2.17 and Equation 2.25, the output is zero for negative input values. The output is same as input for the positive input values.

$$\text{ReLU}(x) = \max(x, 0) \tag{2.25}$$

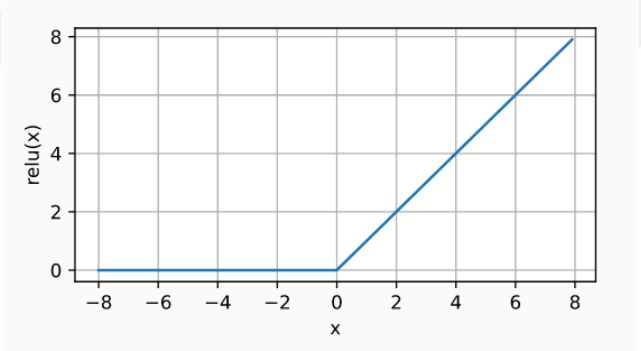

Figure 2.17. ReLU Activation Function

- **Sigmoid:** It is a function that converts the entered value to a value between 0 and 1 as seen in Figure 2.18 and Equation 2.26.
-

$$sigmoid(x) = \frac{1}{1+\exp(-x)} \tag{2.26}$$

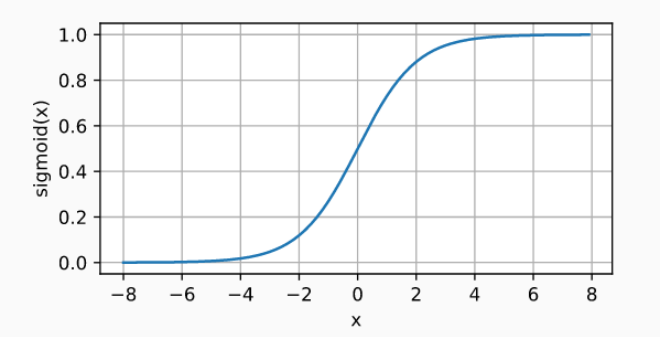

Figure 2.18. Sigmoid Activation Function

- **Tanh:** It is a function that converts the entered value to a value between -1 and 1 in a system like the sigmoid function as shown in Figure 2.19 and Equation 2.27.

$$\tanh(x) = \frac{1-\exp(-2x)}{1+\exp(-2x)} \tag{2.27}$$

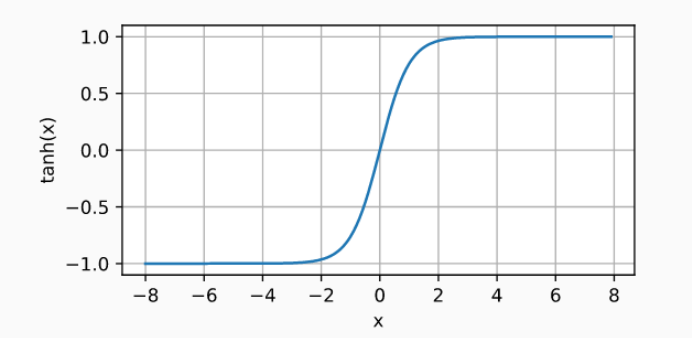

Figure 2.19. tanh Activation Function

- **Gaussian Error Linear Units (GELU):** Φ(x) is the Cumulative Distribution Function for Gaussian Distribution. (Hendrycks et al. 2017). As seen in Figure 2.20 and Equation 2.27, it is a function that

determines the output according to the percentile of the value in the input.

$$GELU(x) = x * \Phi(x) \tag{2.27}$$

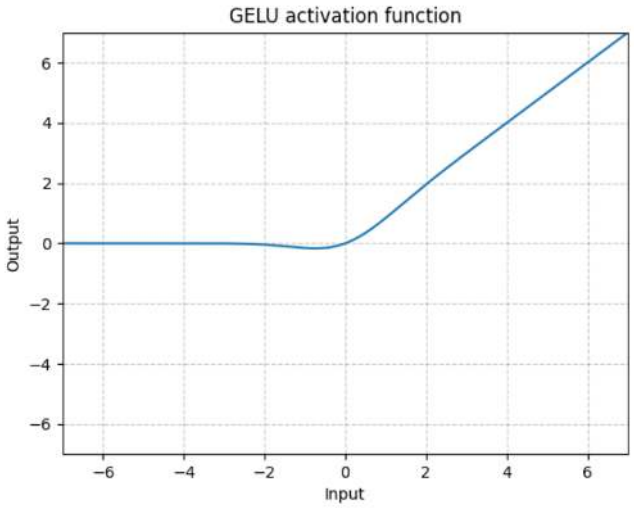

Figure 2.20. GELU Activation Function

2.4.2. Overfitting and Dropout

If any Neural Network model memorizes the data instead of understanding the relationship between the information given during training, the model is overfitting. As seen in Figure 2.21, models must learn the relationship between values for a good fit. However, in overfitting, it tries to memorize every point. This indicates that the model is not trained.

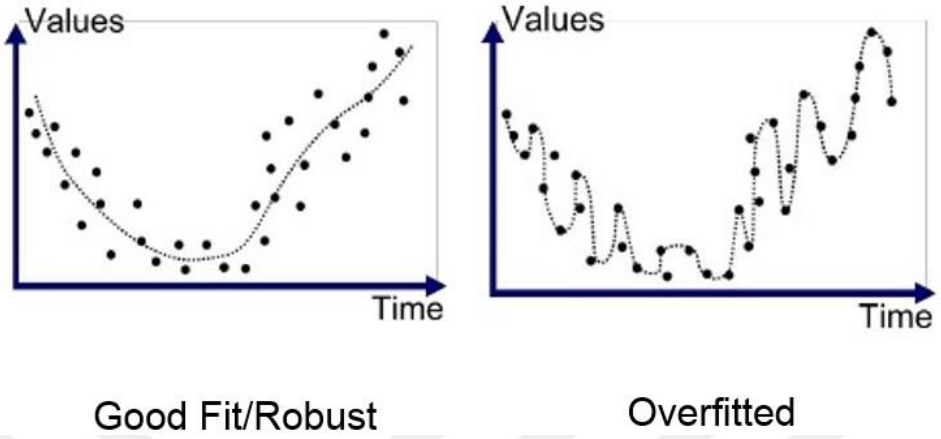

Figure 2.21. Overfitting

Dropout is a regularization method developed to prevent overfitting by disconnecting some nodes. (Srivastava et al., 2014) Disconnection is done randomly at the rate you specify. As shown in Figure 2.22, by disconnecting some nodes, the dropout model is applied and overfitting is prevented.

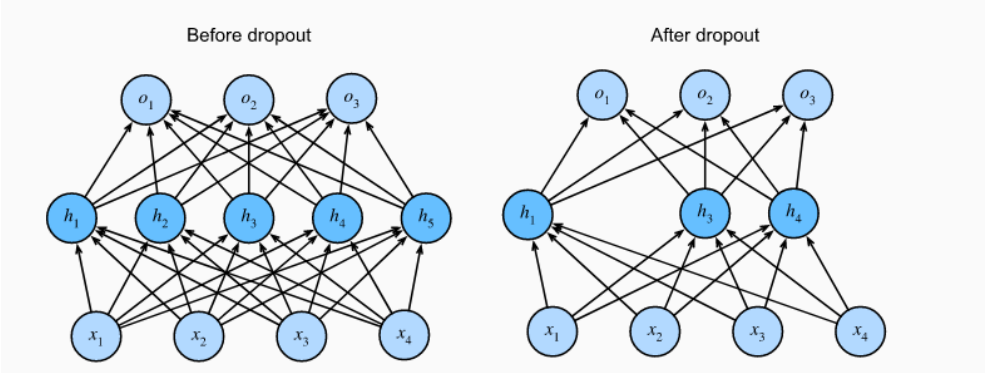

Figure 2.22. Dropout

2.4.3. Gradient descent and Learning Rate

Gradient descent is an optimization technique that allows us to find the most optimal parameter according to the cost calculation made while learning the model.

Considering that the optimal point is the base, as shown in Figure 2.23, the Gradient Pattern provides the process of approaching the base according to the starting point.

Learning Rate is the parameter that determines the descent rate while reaching the optimal point in the Gradient Descent process. If the learning rate is chosen too low, the process of reaching the optimal point will be prolonged. If set too high, it may miss the optimum point.

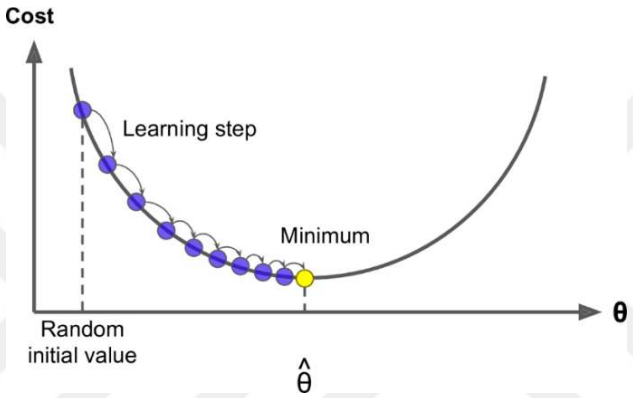

Figure 2.23. Learning Rate

2.4.4. Normalization

The process of reducing data to a value between 0 and 1 is called normalization. This process applied to the layer in the model is called layer normalization (Ba et al., 2016).

2.4.5. Label Error Rate

Label Error Rate (LER) is used to find the error rate of the character string outputted from CTC. The difference between target and prediction is calculated by Edit Distance. The LER value is calculated with the formula given below.

$$LER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2.28}$$

In Equation 2.28, LER is a measure of similarity between two strings with the number of letters inserted (I), the number of letters deleted (D), the number of letters replaced (S), and the number of correctly known letters (C) to convert between two expressions. N represents the size of the reference expression between the two arrays.

3. EXPERIMENTAL SETUP

In this section, information about dataset, models and model parameters is given.

3.1. Data Creation Process and Speech Corpus

In the original dataset, there are 15000 speech files in total, one sentence in each speech file. In addition, there is a word-level label file for each speech file in the dataset. The speech signals in the dataset were created from the speeches of 42 speakers, twenty-one females and twenty-one males. There are 1500 speech files for one male and one female speaker in the dataset. For each of the other 40 speakers, there are 300 speech files in the dataset. The maximum duration of speech files is 10 seconds. The speech signal was recorded in Microsoft wav format. The sampling frequency is 44100 Hz.

The dataset created using all audio data is called Data-Set-4 and contains approximately 20 hours of audio data.

A dataset of 9 hours, 21 minutes and 41 seconds, called Data-Set-3, consisting of 7500 audio files created by using half of all audio dataset, was created.

A dataset called Data-Set-2 was created, consisting of 3750 audio files with a duration of 4 hours, 50 minutes and 51 seconds, which were created by using a quarter of all audio dataset.

A dataset called Data-Set-1, consisting of 3000 sound files, was created. The purpose of creating the Data-Set-1 file is to compare models with approximately equal parameters.

In Data-Set-2, it is examined the contribution of doubling the layer to the result. Data-Set-2, Data-set-3 and Dataset-4 were created to examine the effect of training data size on speech recognition performance. Experiments were performed using these four datasets.

In all experiments in this study, Laptop with Amd Ryzen 5600H processor, 16 Gb ram, and Rtx 3050 (4 Gb) graphics card was used. In the experiments, the Python programming language and the PyTorch library were used.

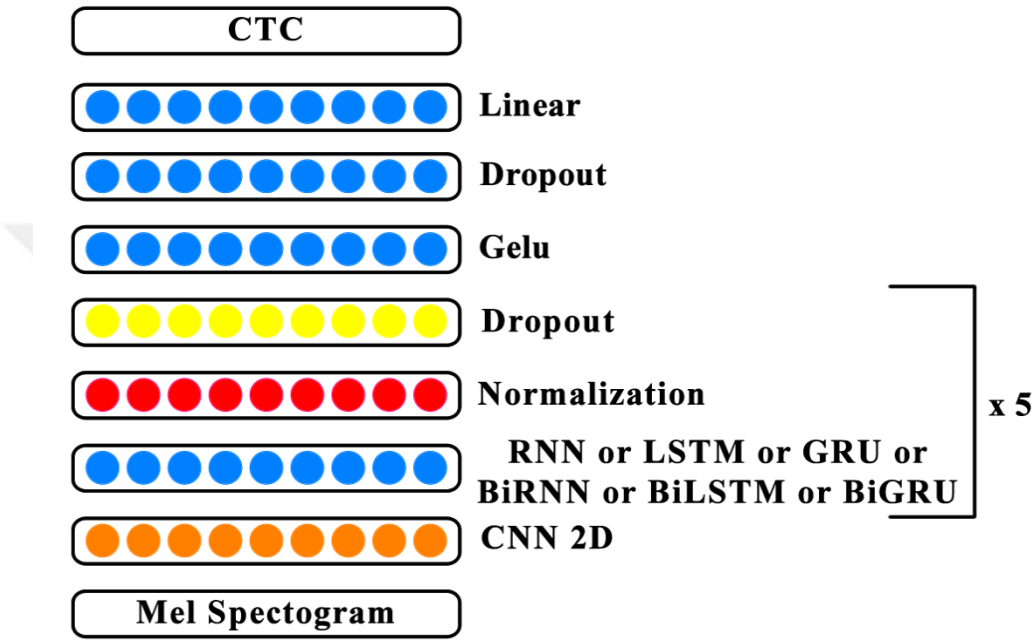

Figure 3.1. Experimental Models

6 different Models used in the experiments are shown in Figure 3.1. Mel spectrogram was applied to Sound data in all of these models. CNN, Gelu, Dropout, Linear and CTC layer are the same in all models. There are changes in the section with RNN, LSTM, GRU, BIRNN, BILSTM and BIGRU models. In the model used, the following 3 layers are repeated 5 times:

- RNN, LSTM, GRU, BIRNN, BILSTM and BIGRU layer
- Normalization layer
- Dropout layer

3.2. Hyperparameters

There are many hyperparameters that affect the performance of the speech recognition system. In this study, the same hyperparameters were used in all experiments. These parameters are explained below:

- **Feature Extraction:** Speech signal is separated into frames using Hanning window of 20 milliseconds (with 50% overlap) every 10 milliseconds using torch audio library and 128 Mel Spectrogram features are extracted from each frame.
- **Number of Layer:** A single 2D-CNN layer is used. 32 filters are used for 2D-CNN. The kernel size of each filter is (3,3), and the stride size is (2,2). Five layers are used for RNN, LSTM, GRU, BIRNN, BILSTM and BIGRU models. After each layer, normalization and dropout were done. As shown in Figure 4.1, each model has Gelu, Droput and Linear layers, and CTC is used as the last layer.
- **Number of Units:** The number of nodes differs in each experiment.
- **Number of Epoch:** The number of epochs is 300 for each experiment
- **Learning Rate:** Learning rate was set as 0.005
- **Number of Train and Test Dataset:** 97% of the dataset was used as training set and 3% as test set.
- **Batch Size:** It is the size of the piece where you divide the dataset in case you cannot process the dataset in one piece while training or testing. If you have a computer with a high processing power, there is the possibility of adjusting the batch size as you want. You can even process all the data at once. The batch size is set to 4 due to the computer I am using.
- **Dropout:** 0.1 dropout is applied at the end of each layer.

- **Activation Function:** Tanh, SoftMax and sigmoid activation functions are used in the structure of the models. Gelu activation function is used as layer.

3.3. Experimental Setup

4 different experiments were conducted on 4 datasets. In the first experiment, models with equal parameters were compared. In the second experiment, the effect of augmentation the number of nodes was examined. In the third experiment, the effect of dataset size in unidirectional models was examined. In the fourth experiment, the effect of dataset size in bidirectional models was examined.

3.3.1. Comparison of RNN, LSTM and GRU models with equal parameters

The comparison of RNN, LSTM and GRU models with equal parameters were done over Data-Set-1. There are 3000 samples in this dataset. RNN, LSTM, GRU, BIRNN, BILSTM and BIGRU models were compared. The total number of parameters is given in Table 3.1.

Table 3.1. Parameter Counts of Models

Unidirectional Models	RNN	4560247
	LSTM	4547939
	GRU	4552349
Bidirectional Models	BIRNN	11190499
	BILSTM	11179363
	BIGRU	11167331

By varying the number of nodes in each layer of the RNN, LSTM, and GRU models, we approximately equalized the total number of parameters for all models. When we approximately equalize the total number of parameters of the models, we found that the number of nodes in each layer of the RNN, LSTM, and GRU models should be 578, 256, and 307, respectively.

Similarly, by varying the number of nodes in each layer of the BIRNN, BILSTM, and BIGRU models, we approximately equalized the total number of parameters for all models. When we approximately equalize the total number of parameters of the models, we found that the number of nodes in each layer of the BIRNN, BILSTM, and BIGRU models should be 562, 256, and 304, respectively.

3.3.2. Comparison of the effect of node augmentation in RNN, LSTM and GRU models

The Comparison of the effect of node augmentation in RNN, LSTM and GRU models was made over Data-Set-2. There are 3750 samples in this dataset. The aim of this experiment is to observe the performance increase when we double the number of nodes. Experiments were made with RNN, LSTM, GRU, BIRNN, BILSTM, BIGRU models.

Table 3.2. Parameter Counts for Unidirectional Models

256 Nodes	RNN	1197923
	LSTM	4547939
	GRU	3431267
512 Nodes	RNN	3702115
	LSTM	13941091
	GRU	10528099

The Parameter counts of unidirectional models are shown in Table 3.2. By increasing the number of hidden layer nodes from 256 to 512, the number of parameters increased by approximately 200%.

Table 3.3. Parameter Counts of Bidirectional Models

128 Nodes	BIRNN	996707
	BILSTM	3855203
	BIGRU	2902371
256 Nodes	BIRNN	2906467
	BILSTM	11179363
	BIGRU	8421731

Parameter counts of Bidirectional models are given in Table 3.3. When the number of hidden layer nodes was increased from 128 to 256, there was a parameter increase of approximately 200%. Forward and Backward layers are concatenated while calculating the hidden layer node counts.

3.3.3. Comparison of the effect of dataset size on RNN, LSTM and GRU, models

Although the number of layers and node numbers of RNN, LSTM and GRU models, which are the compared Neural Network models, are the same, the total number of parameters of each model is different from each other because the internal structure of each model is different. Total Parameter Counts of acoustic models are shown in Table 3.4. As can be seen from Table 3.4, the total number of parameters of the RNN model is the least. The model with the highest total number of parameters is LSTM.

Table 3.4. Parameter Counts of Unidirectional Models

Models	**Parameter**
RNN	1197923
LSTM	4547939
GRU	3431267

3.3.4. Comparison of the effect of dataset size on BIRNN, BILSTM and BIGRU models

Although the number of layers and the number of nodes of the compared Neural Network models BIRNN, BILSTM and BIGRU are the same, the total number of parameters of each model is different from each other due to the structural

differences of the models. The total parameter counts of the models are shown in Table 3.5. As can be seen from Table 3.5, the BIRNN model has the lowest total number of parameters and the model with the highest total parameter number is BILSTM.

Table 3.5. Parameter Counts of Bidirectional Models

Models	**Parameter**
BIRNN	2906467
BILSTM	11179363
BIGRU	8421731

4. RESULT AND DISCUSSION

4.1. Comparison of the performances of RNN, LSTM and GRU models with equal total number of parameters

Figures 4.1 and 4.2 show the change in loss by epoch for Data-Set-1. As can be seen from the three figures, it is seen that the loss regularly decreases as the epoch value increases for the RNN, LSTM and GRU acoustic models. It is seen that the RNN model has a higher loss function than the LSTM and GRU models.

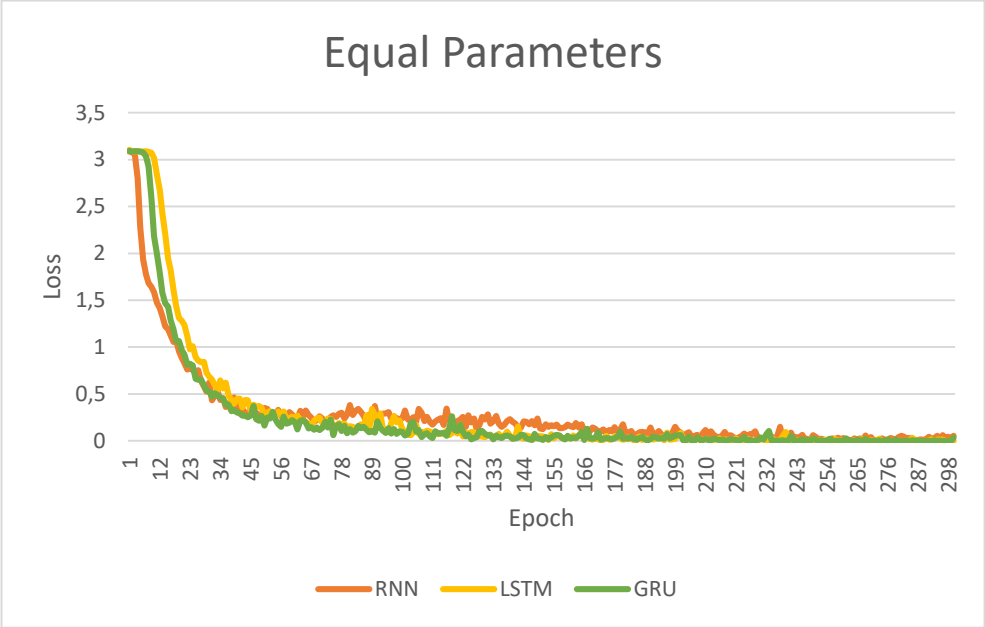

Figure 4.1. Unidirectional Models Loss Function with Equal Parameters

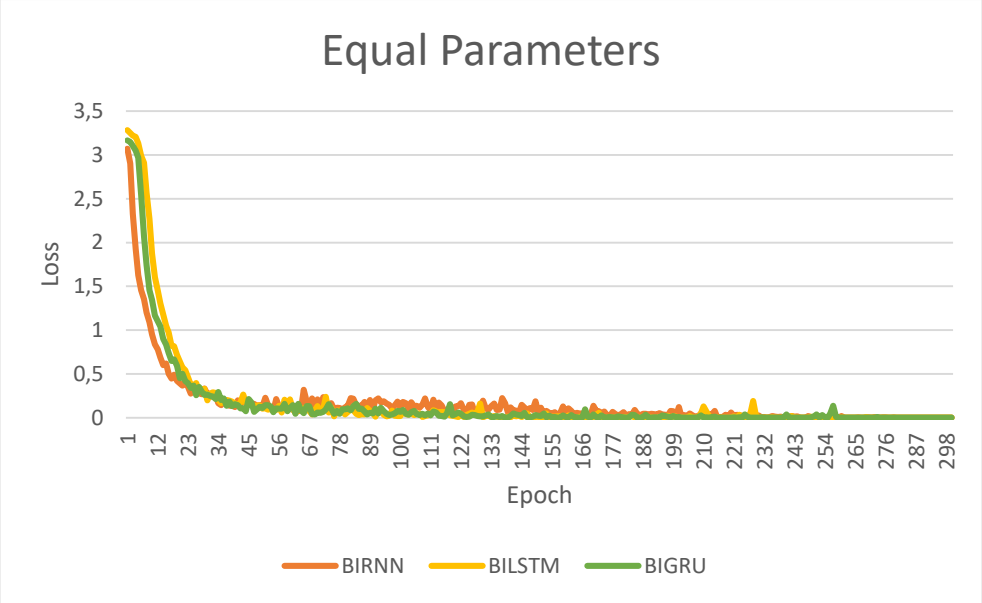

Figure 4.2. Bidirectional Models Loss Function with Equal Parameters

The LER and run time values of all models for the Data-Set-1 are shown in Table 4.1. As seen in the related tables, it is seen that RNN gives the highest LER value and LSTM gives the lowest LER value in both unidirectional and bidirectional models. From this we can conclude that LSTM arguably gives the best LER value. However, it can be seen from the Tables that the LER values of BIGRU are very close to the LER values of BILSTM for each dataset.

As can be seen from Table 4.1, the LSTM has the shortest run time and the GRU has the longest run time for each dataset. We can conclude that the unidirectional and bidirectional versions of the LSTM model have the lowest LER values and the lowest training time compared to the other models, thus making it suitable for speech recognition.

Since the number of parameters is equal, the number of hidden layer nodes of the RNN model is 125% more than the LSTM model. The hidden layer nodes of the GRU model are 20% more than the LSTM model. Therefore, the LSTM model

has the lowest running time and despite the low node count, the lowest LER number indicates the success of the LSTM model.

Since the number of parameters is equal, the number of hidden layer nodes of the BIRNN model is 120% more than the BILSTM model. The number of hidden layer nodes of the BIGRU model is 18% more than the BILSTM model. However, the running time of the BILSTM model is the lowest. In addition, despite the low number of nodes, the low number of LERs is the success of the BILSTM model. The LER values of the BIGRU and BILSTM model with the same number of parameters are approximately the same, and the advantage of the BILSTM model is the low operating time.

Table 4.1. Comparison of models with equal parameters

Data-Set-1		
MODELS	LER	RUNNING TIME (Minute)
RNN	0.4123	209
LSTM	**0.3670**	**201**
GRU	0.3864	215
BIRNN	0.3266	369
BILSTM	**0.2834**	**307**
BIGRU	0.2846	427

4.2. Comparison of the effect of the number of nodes on the result in RNN, LSTM and GRU models

The variation of the loss value of unidirectional models with 256 nodes is given in Figure 4.3. While LSTM and GRU models give similar results, RNN model has a higher Loss value than other models. The variation of the loss value of unidirectional models with 512 nodes is given in Figure 4.4. The LSTM and GRU acoustic models gave similar results here. Although RNN model gave a higher Loss value up to 200 epochs than other models, it gave similar Loss values above 200 epochs with other models.

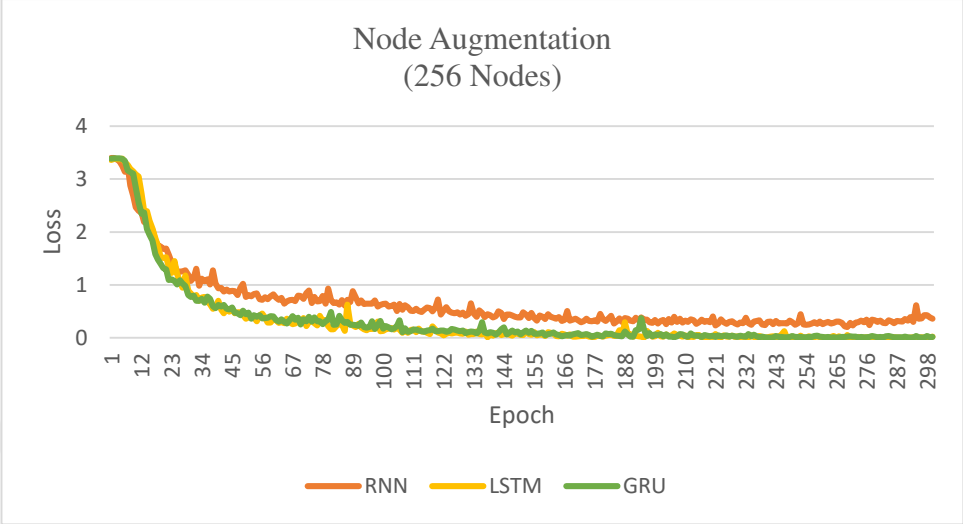

Figure 4.3. Loss function for the models with 256 nodes on Dataset-2 (unidirectional)

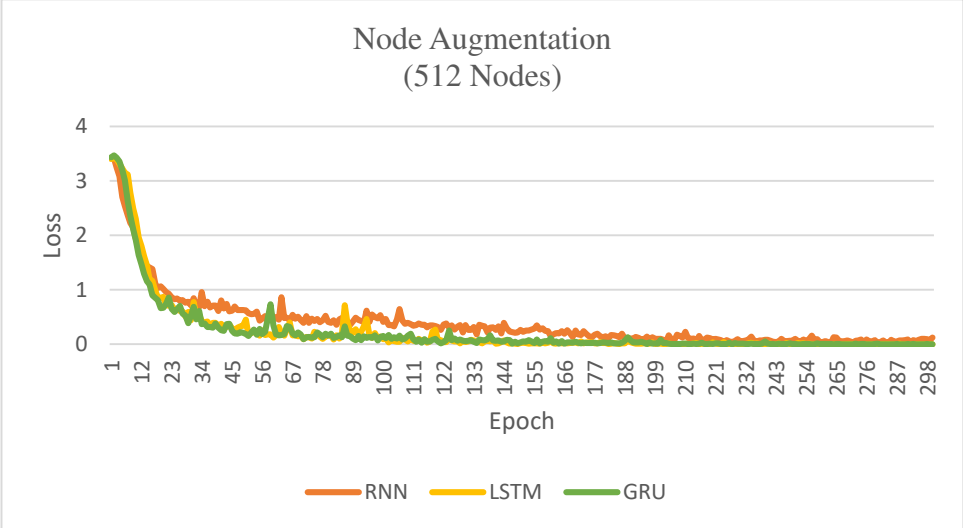

Figure 4.4. Loss function for the models with 512 nodes on Dataset-2 (unidirectional)

Figure 4.5 shows the variation of losses for the bidirectional models with 256 nodes in each hidden layer. Figure 4.6 shows the variation of the losses for the bidirectional models with 512 nodes in each hidden layer. While BILSTM and

BIGRU models give similar results for both 256 and 512 nodes, BIRNN model has higher loss value at 256 and 512 nodes than the BILSTM and BIGRU models. BIRNN model showed similar results with the BILSTM and BIGRU models with 512 nodes compared to 256 nodes in the hidden layer.

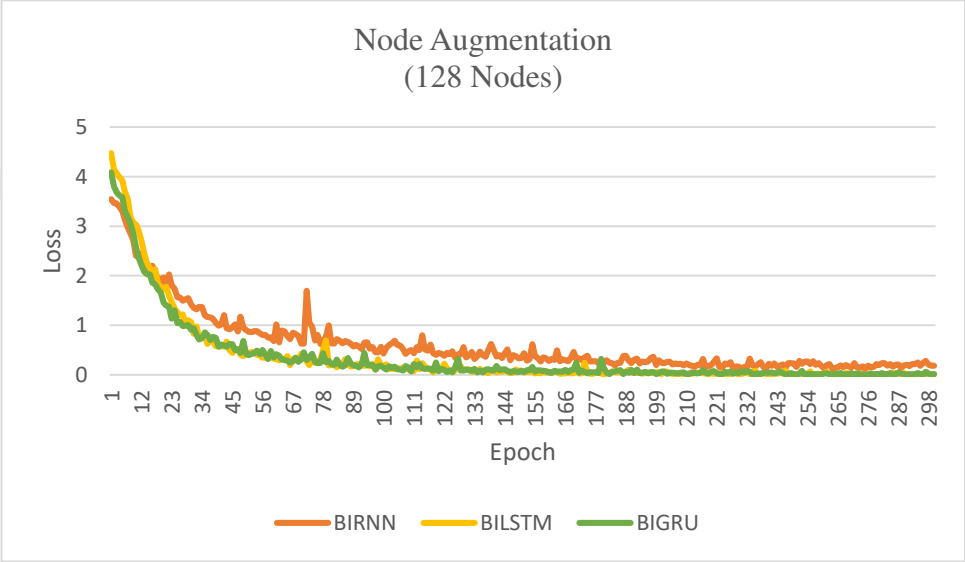

Figure 4.5. Loss function for the models with 128 nodes on Dataset-2 (Bidirectional)

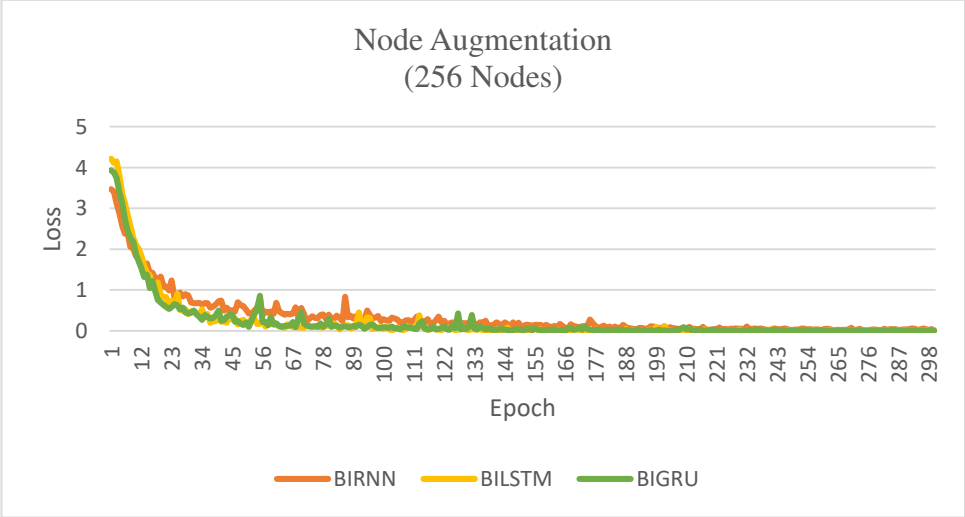

Figure 4.6. Loss function for the models with 256 nodes on Dataset-2 (Bidirectional)

LER, running time and total loss values of all models for Data-Set-2 are shown in Table 4.2. As can be seen from the related tables, in each dataset, it is seen that the RNN acoustic model gives the highest LER value in the use of both 256 nodes and 512 nodes, while LSTM acoustic model gives the lowest LER value in the use of both 256 nodes and 512 nodes. From these results we can see that LSTM arguably gives the best LER value. However, it is noteworthy from the Tables that the LER values of GRU are very close to LSTM and obtain these results with a lower running time. GRU acoustic model In 256 node usage, GRU model can be preferred with 1% LER difference and approximately 20% less running time compared to LSTM Model. The same is true for 512 nodes.

When we compare the running times of the models, it can be seen from Table 4.2 that LSTM model has the shortest running time for both 256 and 512 nodes, and RNN model has the longest. It is obvious that LSTM and BILSTM models have the lowest LER values and the highest running time compared to other models. This is due to the inherently higher parameter handling. If only the LER is very important, it seems more advantageous to use LSTM and the bidirectional version as the acoustic model. However, based on the running time, the performance of GRU model is also important.

Table 4.2. Comparison of the Effect of Node Augmentation

Data-Set-2						
	256 NODES			512 NODES		
MODELS	LER	RUNNING TIME (Minute)	TOTAL LOSS	LER	RUNNING TIME (Minute)	TOTAL LOSS
RNN	0.4346	**185**	194,6976	0.4176	**237**	119,2079
LSTM	**0.3608**	267	**102,0223**	**0.3284**	503	71,9418
GRU	0.3701	226	103,0146	0.3398	367	**68,0445**
BIRNN	0.3642	**303**	186,3182	0.3336	356	101,4044
BILSTM	**0.3015**	321	108,2406	**0.2586**	378	68,6523
BIGRU	0.3117	306	**103,9647**	0.2669	**346**	**66,9942**

4.3. Comparison of the effect of dataset size on the result in RNN, LSTM and GRU models

Figures 4.7, 4.8 and 4.9 show the loss value verses epoch value for Dataset-1, Dataset-2 and Dataset-3, respectively. As can be seen from the three figures, it is seen that the loss regularly decreases as the epoch value increases for the LSTM and GRU acoustic models. It is seen that the loss function does not decrease with the same stability for RNN.

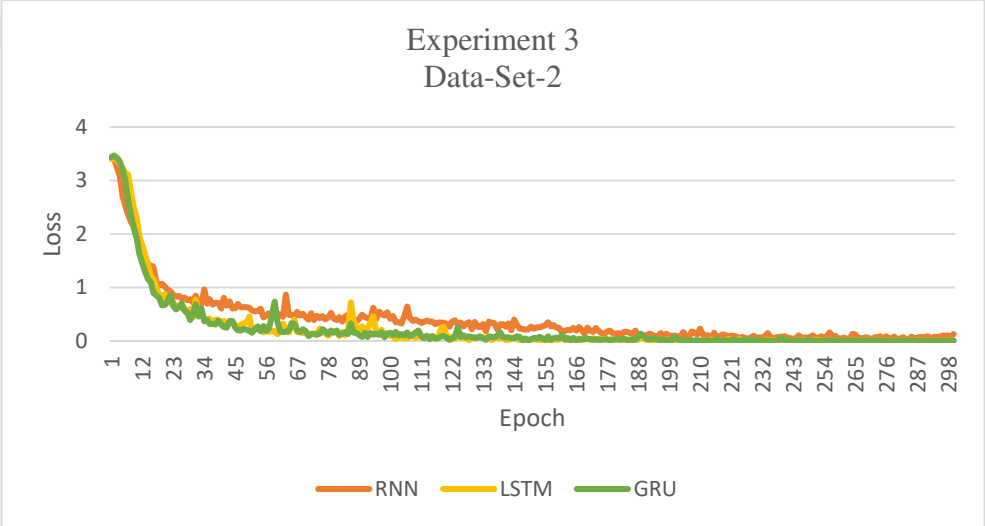

Figure 4.7. Data-Set-2 Loss Function (Unidirectional)

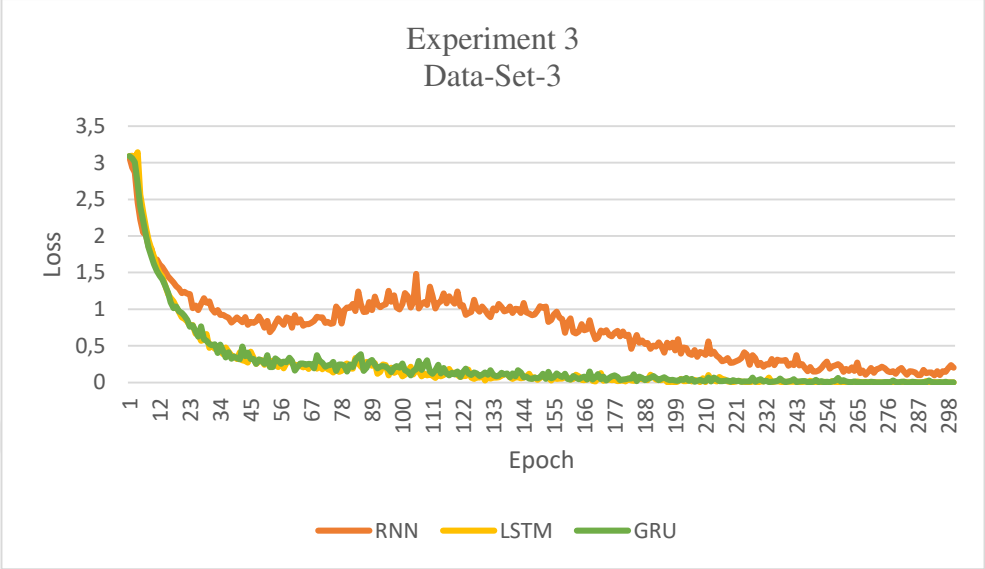

Figure 4.8. Data-Set-3 Loss Function (Unidirectional)

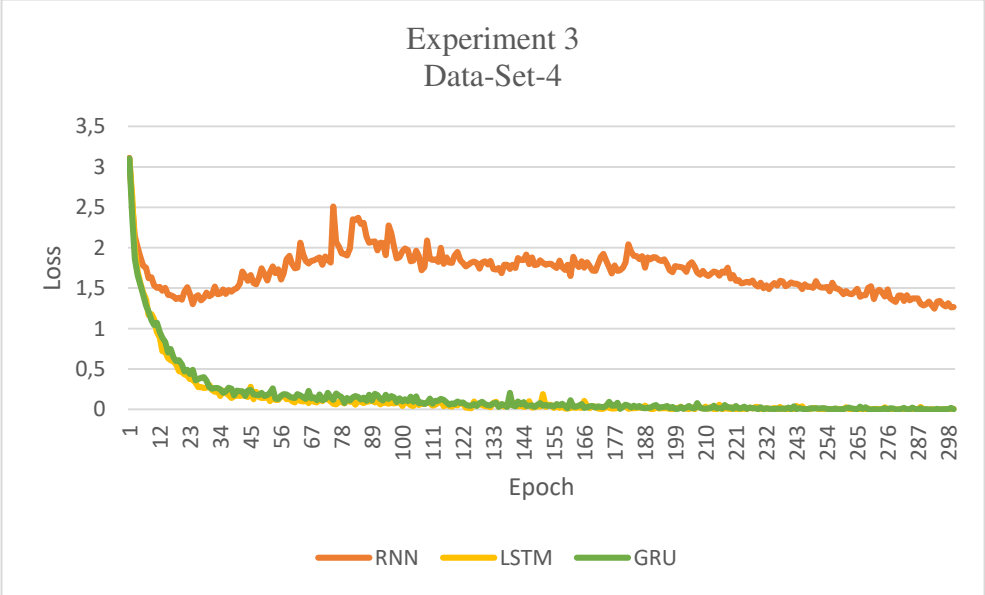

Figure 4.9. Data-Set-4 Loss Function (Unidirectional)

Table 4.3 shows the LER and running time values of all models for Data-Set-2, Data-Set-3 and Data-Set-4. As can be seen from the related tables, it is seen that RNN gives the highest LER value and LSTM gives the lowest LER value in each dataset. From this we can conclude that LSTM gives arguably the best LER value for all datasets. However, it can be seen from the Tables that the LER values of the GRU are very close to the LER values of the LSTM for each dataset.

When we compare the running times of the models, Table 4.10, Table 4.11 and Table 4.12 show that RNN has the shortest running time and LSTM has the longest run time for each dataset.

Since the LER values of the RNN model are very high compared to other models, we can conclude that the RNN model is not suitable for speech recognition.

When we compare the LER values of the LSTM and GRU acoustic models, Table 4.10, Table 4.11, and Table 4.12 show that LSTM has an LER value of approximately 3% lower for each dataset compared to GRU. Therefore, if LER is very important to us, it is more advantageous to use LSTM as an acoustic model.

When the running time is compared, it is seen in Table 4.3 that the run time of the LSTM model is approximately 30% longer than the GRU model for each dataset. Also, as can be seen from Table 3.4, the number of parameters of the LSTM is about 32% more than the number of parameters of the GRU. Although LSTM gives better LER results, GRU can be trained faster and has fewer parameters than LSTM. For this reason, although GRU gives 3% worse results than LSTM, GRU can be preferred instead of LSTM as an acoustic model.

Another result seen in Table 4.3 is that the LER values of all models decrease as the dataset size increases.

Table 4.3. Comparison of the Effect of Dataset Size

	MODELS	LER	RUNNING TIME (Minute)
Data-Set-2 (Approximately 5 Hours)	CNN-RNN	0.4176	**237**
	CNN-LSTM	**0.3284**	503
	CNN-GRU	0.3398	367
Data-Set-3 (Approximately 10 Hours)	CNN-RNN	0.3592	**479**
	CNN-LSTM	**0.2825**	1024
	CNN-GRU	0.2920	734
Data-Set-4 (Approximately 20 Hours)	CNN-RNN	0.4561	1993
	CNN-LSTM	**0.2329**	2022
	CNN-GRU	0.2485	**1476**

4.4. Comparison of the effect of dataset size on the result in BIRNN, BILSTM and BIGRU models

Figures 4.10, 4.11 and 4.12 show the change of loss values of BIRNN, BILSTM and BIGRU models according to epoch value for Data-Set-2, Data-Set-3, and Data-Set-4, respectively. As can be seen from the three figures, as the epoch value increases for BILSTM and BIGRU acoustic models, it is seen that the loss decreases regularly. For BIRNN, it is seen that the loss value does not decrease with the same stability. Especially as the size of the dataset grows, BIRNN acoustic model gives worse loss value than BILSTM and BIGRU models.

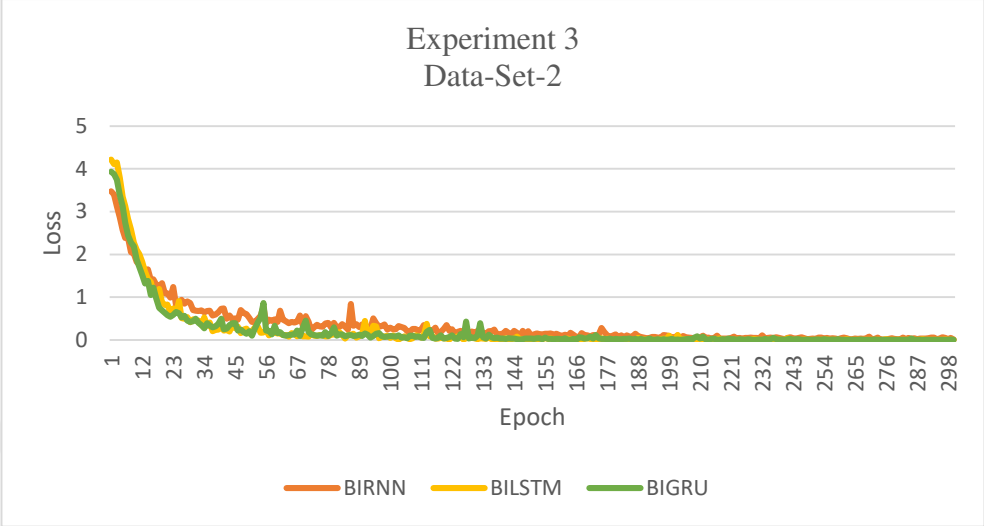

Figure 4.10. Data-Set-2 Loss Function (Bidirectional)

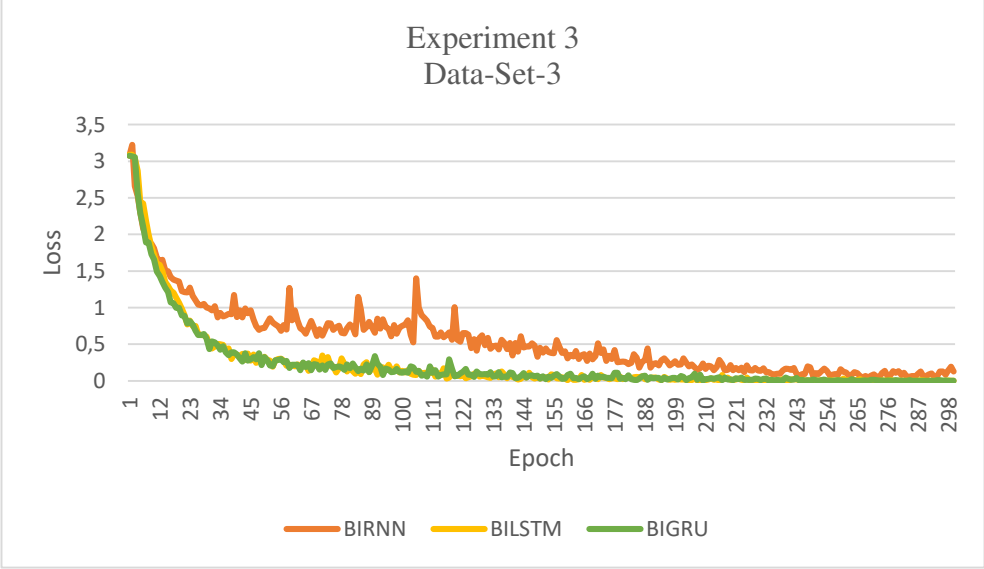

Figure 4.11. Data-Set-3 Loss Function (Bidirectional)

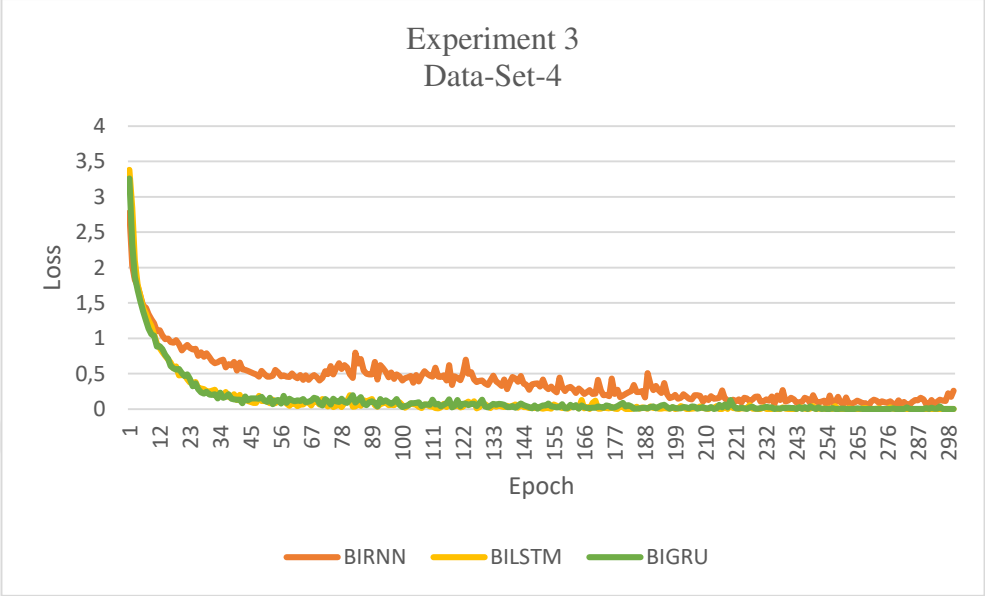

Figure 4.12. Data-Set-4 Loss Function (Bidirectional)

As seen in Table 4.4, it shows the LER, runtime and total loss values of all models for Data-Set-2, Data-Set-3 and Data-Set-4. As seen in the related table, it is seen that the BIRNN model gives the highest LER value and the BILSTM model gives the lowest LER value in each dataset. From this, we can conclude that the BILSTM acoustic model is the most successful model for all datasets and gives the lowest LER value. However, it can be seen from the Tables that the LER values of BIGRU are very close to the LER values of LSTM for each dataset.

As it can be seen from Table 4.4 when we compare the training times of the models, BIGRU has the lowest training time for the dataset of approximately 5 hours. BIRNN appears to have the shortest training time for datasets of about 10 hours and 20 hours. It is seen that BILSTM has the longest training time for each dataset.

Since the LER values of the BIRNN model are very high compared to other models, it can be concluded that the BIRNN acoustic model is not suitable for speech recognition. When we compare the LER values of BILSTM and BIGRU acoustic models, it is seen from Table 4.4 that BILSTM has approximately 3% lower LER

value for each dataset than BIGRU. Therefore, if the LER value is very important, using BILSTM as an acoustic model seems to be more advantageous than other acoustic models.

When the running time is compared, it is seen from Table 4.4 that the run time of the BILSTM model is approximately 9% longer than the BIGRU model for each dataset. Also, as can be seen from Table 3.5, the number of parameters of BILSTM is approximately 33% more than the number of parameters of BIGRU. The use of BIGRU as an acoustic model is more advantageous than BILSTM, considering the running time and the number of parameters. Although BILSTM gives better LER results, BIGRU can be trained faster and has fewer parameters than BILSTM. For this reason, BIGRU can be preferred instead of BILSTM as an acoustic model, although BIGRU gives 3% worse results than BILSTM.

Another conclusion can be drawn from Table 4.4 is that the LER values of all acoustic models decrease as the dataset size increases.

Table 4.4. Data-Set-2 Model Comparison (Bidirectional)

	MODELS	LER	RUNNING TIME (Minute)
Data-Set-2 (Approximately 5 Hours)	CNN-BIRNN	0.3336	356
	CNN-BILSTM	**0.2586**	378
	CNN-BIGRU	0.2669	**346**
Data-Set-3 (Approximately 10 Hours)	CNN-BIRNN	0.2935	**644**
	CNN-BILSTM	**0.2109**	761
	CNN-BIGRU	0.2282	710
Data-Set-4 (Approximately 20 Hours)	CNN-BIRNN	0.2509	**1274**
	CNN-BILSTM	**0.1797**	1594
	CNN-BIGRU	0.1968	1467

5. CONCLUSION AND FUTURE WORK

In this thesis, experimental studies were carried out under 4 different headings. Both unidirectional and bidirectional models of 3 different acoustic models were compared from different aspects.

In RNN, LSTM and GRU models with equal parameters comparison study, RNN, LSTM and GRU models with the same parameter counts on Data-Set-1 and their bidirectional versions (BIRNN, BILSTM, BIGRU) were compared. LSTM and BILSTM models gave the most successful result with the lowest LER values at the lowest running times.

In RNN, LSTM, GRU, BIRNN, BILSTM and BIGRU models node augmentation comparison study, it was observed that the label error rate decreases when the number of nodes in each layer is increased. Increasing the number of nodes by a certain amount increases the performance. However, it has been observed that increasing the number of nodes also increases the running time.

In dataset size effect comparison studies, three datasets, 5, 10 and 20 hours, were prepared to examine the effect of the size of the dataset on the speech recognition.

When dataset size effect comparison studies were examined, it was observed that the LER value for the relevant acoustic model decreased as the size of the dataset used for each acoustic model increased. This result shows us the importance of using large datasets for training.

As seen in dataset size effect comparison studies, the unidirectional and bidirectional versions of the LSTM have approximately 32% more parameter counts than the unidirectional and bidirectional versions of the GRU. However, it is seen that there is not much difference between the LER value of the GRU and LSTM models. Therefore, only when the LER value is not important, the GRU model can be preferred because of running time advantage.

If we compare RNN, LSTM and GRU models with their bidirectional versions, each model's bidirectional version appears to have 145% more parameters than the unidirectional version. However, in general, more successful results are obtained in a shorter running time. In other words, BIRNN, BILSTM and BIGRU models achieve more successful results than RNN, LSTM and GRU models.

In all experiments on all datasets, it has been shown that LSTM model gives the highest speech recognition rate, while RNN gives a very low recognition rate compared to the other two models. Therefore, there is no advantage of using RNN as an acoustic model.

In future studies, datasets with higher dimensions can be prepared. Because the dataset size directly affects the performance in Speech Recognition systems. Dataset diversity can also be increased. It is planned to use different models, use more parameters in the models and try higher layered models.

REFERENCES

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., and Yu, D., 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10), 1533-1545.

Arisoy, E., Sethy, A., Ramabhadran, B., & Chen, S. (2015, April). Bidirectional recurrent neural network language models for automatic speech recognition. In 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5421-5425). IEEE.

Arisoy, E., Can, D., Parlak, S., Sak, H., & Saraçlar, M. (2009). Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(5), 874-883.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bakhturina, E., Lavrukhin, V., & Ginsburg, B. (2021). NeMo Toolbox for Speech Dataset Construction. *arXiv preprint arXiv:2104.04896*.

Bhuvaneswari, A., Thomas, J. T. J., & Kesavan, P. (2019). Embedded Bi-directional GRU and LSTMLearning Models to Predict Disasterson Twitter Data. *Procedia Computer Science*, *165*, 511-516.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv*:1409.1259.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *In Proceedings of the 23rd international conference on Machine learning* (pp. 369-376).

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. *In 2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv*:1606.08415.

Hu A, Zhang K. Using Bidirectional Long Short-Term Memory Method for the Height of F2 Peak Forecasting from Ionosonde Measurements in the Australian Region. *Remote Sensing*. 2018; 10(10):1658. https://doi.org/10.3390/rs10101658

Hwang, Y., Cho, H., Yang, H., Won, D. O., Oh, I., & Lee, S. W. (2020). Mel-spectrogram augmentation for sequence to sequence voice conversion. *arXiv preprint arXiv*:2001.01401.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Mporas, I., Ganchev, T., Siafarikas, M., Fakotakis, N.: 'Comparison of speech features on the speech recognition task', *J. Comput. Sci.*, 2007, 3, (8), pp. 608–616

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv*:2012.03411.

Ravanelli, M., Brakel, P., Omologo, M., & Bengio, Y. (2018). Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2), 92-102.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv*:1402.1128.

Salor, O., Qiloglu, T., & Demirekler, M. (2006). METU Turkish Microphone Speech Corpus. *In 2006 IEEE 14th Signal Processing and Communications Applications.*

Shewalkar, A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, *9*(4), 235-245.

Shrawankar, U., & Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: *A comparative study. arXiv preprint arXiv*:1305.1145.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Tak, R. N., Agrawal, D. M., & Patil, H. A. (2017, December). Novel phase encoded mel filterbank energies for environmental sound classification. *In International Conference on Pattern Recognition and Machine Intelligence* (pp. 317-325). Springer, Cham.

Y. LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *in Neural Computation*, vol. 1, no. 4, pp. 541-551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

Zeyer, A., Doetsch, P., Voigtlaender, P., Schlüter, R., & Ney, H. (2017, March). A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. *In 2017 IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 2462-2466). IEEE.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv*:2106.11342.

CURRICULUM VITAE

Halil İbrahim YALMAN. Primary Education is at Hatice Büyükbeşe Primary School and Gazi Secondary School, and high school education is at Mehmet Akif Ersoy Anatolian Technical High School. He completed his undergraduate education in Kocaeli University Kocaeli / TURKEY Computer Education (TEF) department between 2007-2011 and Istanbul University Istanbul / TURKEY Computer Engineering between 2013-2015. He has been working in the "Ministry of National Education" (MEB) since 2012.