Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 24, S. 87-92, Nisan 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 24, pp. 87-92, April 2021

Research Article

A Turkish Broadcast News Speech Database for Investigation the Effect of Deep Neural Network and Long Short Term Memory Hyperparameters on Speech Recognition Based Systems

Serhat Ok[1*], Zekeriya Tüfekci[2],

[1*] ZİRAAT TEKNOLOJİ A.Ş, İstanbul, Turkey, (ORCID: 0000-0002-9764-2952)
[2] 2 Computer Engineering Department, Cukurova University, Adana, Turkey (ORCID: 0000-0001-7835-2741)

(2nd International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF)-10–12 March 2021)

(**DOI:** 10.31590/ejosat.900422)

ATIF/REFERENCE: Ok, S., & Tüfekçi, Z. (2021). A Turkish Broadcast News Speech Database for Investigation the Effect of Deep Neural Network and Long Short Term Memory Hyperparameters on Speech Recognition Based Systems. *European Journal of Science and Technology*, (24), 87-92.

Abstract

Speech recognition is the transformation of spoken words and sentences into text. There have been many studies on speech recognition in many countries recently. However, studies on speech recognition applications in our country are very few, one of the reasons is the lack of voice dataset. In this study, a Turkish speech database has been developed for Turkish speech recognition based systems. Sound recordings were obtained from news broadcasted by Turkish news tv channels at different times. The created data set was shared on the web in a way that everyone can access in order to set a precedent for other studies. Additionally, the effects of number of layers and number of cells hyperparameters of Long Short Term Memory (LSTM) and Deep Neural Network (DNN) models were investigated on the Turkish Broadcast News Speech Database.

Keywords: Speech Recognition, Deep Neural Networks, Long Short Term Memory, Turkish Speech Database.

Derin Sinir Ağları ve Uzun Kısa Süreli Bellek Hiperparametrelerinin Konuşma Tanıma Tabanlı Sistemler Üzerindeki Etkisinin İncelenmesi için Türkçe Yayın Haberleri Konuşma Veri Tabanı

Öz

Konuşma tanıma, söylenen kelime ve cümlelerin metne dönüştürülmesidir. Son zamanlarda birçok ülkede konuşma tanıma ile ilgili birçok çalışma yapılmıştır, fakat ülkemizde konuşma tanıma uygulamaları ile ilgili yapılan çalışmalar çok azdır, bunun nedenlerinden biri ses veri seti eksikliğidir. Bu çalışmada, Türkçe konuşma tanıma tabanlı sistemler için bir Türkçe konuşma veri tabanı geliştirilmiştir. Ses kayıtları Türkçe haber tv kanallarının farklı zamanlarda yayınladıkları haberlerden elde edilmiştir. Oluşturulan veri seti diğer çalışmalara da emsal teşkil etmesi açısından herkesin erişebileceği şekilde web ortamında paylaşılmıştır. Ek olarak, katman sayısı ve hücre sayısı hiper parametrelerinin Uzun Kısa Süreli Hafıza (LSTM) ve Derin Sinir Ağı (DNN) modelleri üzerindeki etkisi oluşturduğumuz Türkçe Yayın Haberleri Konuşma veri seti üzerinde incelendi ve karşılaştırıldı.

Anahtar Kelimeler: Konuşma Tanıma, Derin Sinir Ağları, Uzun Kısa Süreli Bellek, Türkçe Konuşma Veri Tabanı.

[*] Corresponding Author: serxadok47@gmail.com

1. Introduction

Speech is the fundamental communication tool between people. This communication flow not only has a transmission of sound, but also it is a common form of activity in the socialization process of people. There are many ways of communication exist in our lives to communicate with people, such as body language, sign language, textual language and speech. The most common types are; text-mail messaging services, telephones and face-to-face speech communications.

Speech recognition is the process in which a speech signal is converted to a sequence group of words by implementing algorithms in a computer program (Santosh K. Gaikwad, 2010). Speech Recognition-based systems and models have been used for seventy years (1950) as an effective research area in academic and commercial activities and increasing in popularity. It has various application areas such as speech recognition, voice interface systems, speaker identification, speech-to-text processing, and text-to-speech conversion. As time passed, the need for better control of complex machines emerged, and speech recognition systems began to play an important role in human-machine communication. (Patlar, F., 2009).

Speech recognition systems have three main steps (Yu and Deng, 2016). In first step, feature extraction process is performed on raw audio signals. Noise removal, signal conversion to feature domain are performed in feature extraction. The second step consists of acoustic and language models. Acoustic model structure gets extracted features as inputs and generates a model score for variable-length feature space, and language model estimates language model score for the words in training corpus. In third step, hypothesis search combines acoustic model score and language model score to generate final score and text transcription of the audio signal. The basic speech recognition steps are presented in Figure 1.

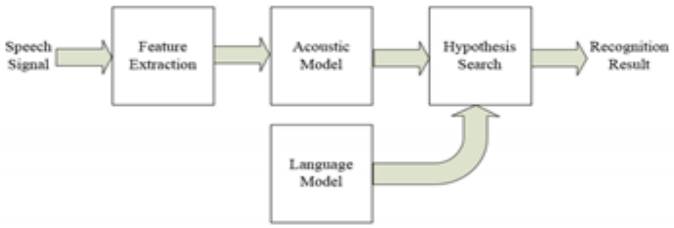

Figure 1. Basic Block Diagram of a Speech Recognition System

Nowadays, many studies have been done on speech recognition based systems. When we investigate the distribution of various applications on speech recognition studies according to languages, there are extensive studies and datasets in other languages, especially in; English, Japanese, Chinese and Arabic languages. (http://kaldi_asr.org/doc/examples.html). One of the most important reason for studies in these languages is that; they have large datasets, and most of them are improvable and accessible continuously. However, the number of speech database created for Turkish is extremely low. Speech databases are very important for researchers who tries to develop a Turkish speech recognition systems. Their number and diversity should be increased.

In the academic literature and other areas, many studies are presented for speech recognition based systems using deep learning architectures. The most used architectures are RNNs and its sub model Long-Short Term Memory (LSTM) but other architectures, such as Deep Neural Network (DNN), Deep Belief Networks (DBNs), and Deep Auto Encoder (DAE), are also used. (Tüfekci, Z., Dokuz, Y., 2020).

DNN is a neural network with multiple layers between the input and destination (output) layers. (Bengio, 2009). For a neural network system to be considered as a DNN, it is sufficient for the number of layers to be more than two. This value is the number of layers the model needs to complete the process. It can solve and analysis the linear and complicated non-linear relationships. The model moves through the between input and output layers calculating the probability of each output like a Feed Forward Neural Network (FFNN) structure.

RNN is one of the deep learning architectures which is efficient in processing sequential data inputs, like time series, or speech signals (Graves et al., 2013). RNNs process one input at a time and generate results at every time step. However, training RNNs is problematic because of the exploding or vanishing gradient problem at back-propagation process. To overcome this limitation, Long Short Term Memory (LSTMs) structure is proposed. (Hochreiter and Schmidhuber, 1997).

LSTM is a special kind of RNN architecture which is capable of forgetting previous inputs that is not useful for current output (Graves et al., 2013b). For defining usefulness of previous inputs, LSTMs unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. These gates carry several information between time steps and with the help of these gates exploding or vanishing gradient problem can be solved. LSTMs are also successfully applied in speech recognition based systems studies.

When the academic literature studies are analyzed, all of the deep learning systems have different architectures, frameworks, and configurations and optimization algorithms. Besides, all of these studies select hyperparameters for deep learning architecture based on their computation power. Therefore, real effect of deep learning architectures on speech recognition based systems performance is not easily observed from the literature.

Deep Learning Based Models have several hyperparameters that affect the performance of systems regardless of the area of application. These hyperparameters are values that change models's structure, and can help achieve better performance with the same dataset. In particular, batch size, number of layers, number of units (cells), and number of epochs are evaluated on accuracy of speech recognition system. The number of layers and the number of units are utilized as hyperparameters in this study.

We hope that this work will provide a further basis for further study in Turkish Speech Recognition and contribute to the accumulation of knowledge on this subject.

2. Material and Method

2.1. Data Creation Process and Speech Corpus

Çukurova University-Turkish Broadcast News Speech Database was created in Adana, Turkey by using the Turkish TV Broadcasts news channels published at different times. This dataset is a portion of structure of Turkish broadcast news dataset gathered and voice-text transcribed with the purpose of facilitating research in Turkish speech recognition based systems and their relevant applications such as; speech signal, speech recognition and speech retrieval. The dataset contains approximately 2 hours of Turkish Tv broadcasts news; a total of 1039 speech sentence files and corresponding text version of each speech file; a total of 1039 text sentence files, 8491 words (4101 different words), mono phoneme and three phoneme structures for all sentences.

Due to the lack of a Turkish speech dataset, this dataset was primarily created to be used in Turkish Speech Recognition Systems. Created dataset is shared on the web.

2.2. Deep Neural Network (DNN)

Deep Neural Network (DNN) is a neural network with multiple layers between the input and destination (output) layers. (Bengio, 2009). For a neural network system to be considered as a DNN, it is enough for the number of layers to be more than two.

The main purpose of DNN is to find an appropriate mathematical forming for relationships between input and output pairs. It moves through the between input and output layers calculating the probability of each output to solve a task. Also, it can solve the complicated non-linear relationships. The basic a DNN model's structure is presented in Figure 2.

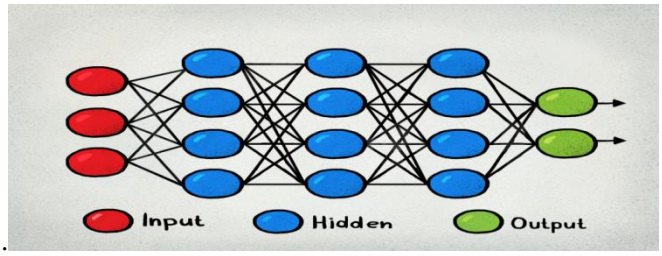

Figure 2. Example of a DNN structure

2.3. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a type of deep learning architectures which is capable of handling large sequential inputs (Graves et al., 2013). The main purpose of RNN is to extract outputs of current time step based on current input and previous inputs. This approximation has advantages for several tasks which needs information about previous inputs, such as speech recognition based systems, digital image processing on photo-video and natural language processing. The weights of input, hidden layers, and output do not change along the network.

RNN is the class of Artificial Neural Networks in which connections between nodes form a directed loop. This allows it to exhibit dynamic temporal behavior. Unlike DNN, RNN is able to use their input memory to process optional rows of inputs. This feature makes RNNs more usable method for speech recognition based systems. RRNs are not successful at learning dependencies in long sequences. Thus, LSTMs were developed (Zuo, Shuai, Wang, Liu, Wang, & Wang, 2016).

2.4. Long-Short Term Memory (LSTM)

LSTMs are clearly designed to avoid the problem of long-term dependence, thus can learn about long-term dependence. They are widely used today because they work very well in a wide variety of problems. An LSTM network contains LSTM units and can remember long or short time periods. The key to this capability is that it does not use any activation functionality in its repeated components. Thus, the stored value is not changed recursively and the slope is not lost when trained by back propagation over time.

LSTM introduced a cell state *c(t)* and four gates whose names are input gate *i(t)*, output gate *o(t)*, forget gate *f(t)* and input modulation gate *g(t)*. To calculate a value between 0 and 1 these LSTM gates perform logistic function. Replication is applied with this value to partially allow or deny information to enter and exit from memory. As a sample, an "input" port controls how much a new value flows on memory struct. A "forget" gate checks the extent to which value stay in memory. An "output" gate checks how much that value in memory is used to calculate the output activation of the block.

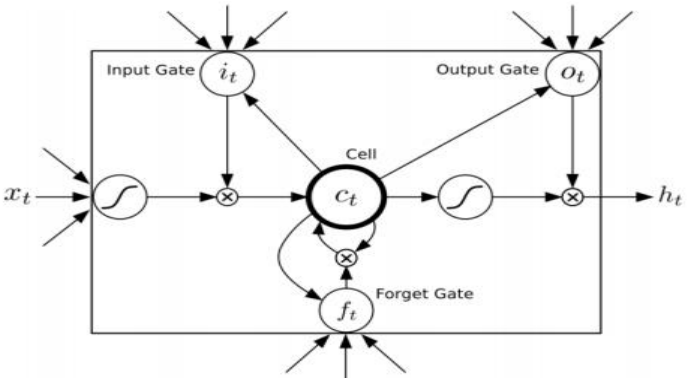

Figure 3. LSTM structure with gates

Figure 3. presents a basic LSTMs network with gates. x, c, h and o are input, cell, hidden state and output vectors. i, f and o present input, forget and output gates. Equations (1) to (5) show how to calculate each vector. The weight matrices present the weights of denoted two parts connections, i.e. W_{xi} denotes the weight of input to input gate connection.

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

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \sigma(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}ht\text{-}1 + W_{co}c_t + b_o) \quad (4)$$

$$h_t = o_t \sigma(c_t) \quad (5)$$

2.5. DNN and LSTM Hyperparameters

LSTM and DNN models have several hyperparameters that affect the performance of deep learning systems regardless of the area of application. These hyperparameters are values that change these models structure, and can help achieve better

performance with the same dataset. The number of layers and the number of units were used as hyperparameters in this study

• **Number of Layers:** This hyperparameter controls the number of layers of which deep learning systems will be built. When the number of layers are increased, the deep learning could better handle variations in the feature space but also complexity of the structure increases. (Tüfekci, Z., Dokuz, Y., 2020). Increasing the number of layers generally increases success while at the same time increasing the computation time proportionally. Therefore, the number of layers are special hyperparameters that need to be considered when building the model. (Hızlısoy, S., 2020). In this study, the number of layers were chosen from 1 to 6 on LSTM and DNN models.

• **Number of Units:** This hyperparameter controls the number of cells that will be constructed on models. When the number of cells are increased, the more backward dependencies could be handled by the deep learning system, but also it increases complexity. (Tüfekci, Z., Dokuz, Y., 2020). Increasing the number of cells generally increases success at the same time increasing the computation time proportionally. Therefore, the number of units are special hyperparameters that need to be considered when building the model. (Hızlısoy, S., 2020). In this study, the number of units were chosen from 50 to 300 on LSTM and DNN models.

3. Results and Discussion

In this section, experimental evaluation of LSTM and DNN hyperparameters on Cukurova University – Turkish Broadcast News Speech corpus is presented. The hyperparameters of number of layers, number of units and batch size are used for evaluation. Each hyperparameter is evaluated while other hyperparameters remain constant.

When examining the effect of the number of layers on LSTM and DNN models; number of units and batch size are set to 200, 32, respectively and number of layers were selected from 1 to 6 for both models.

When examining the effect of the number of units on LSTM and DNN models; number of layers for LSTM and DNN were set as 5 and 4, respectively, according to the their most successful rates in our setup. Number of units were selected as 50, 100, 150, 200, 250 and 300, and Label Error Rate (LER) and Test Cost were calculated for both models. Also, out of 1039 sentences created in this study, 839 were used for training, 200 were used for testing.

3.1. Effect of Number of Layers on LSTM

Table 1 and Figure 4 present the effect of number of layers on speech recognition performance for LSTM model. As the number of layers increase, LSTM gives better result for both LER and Test cost. However, after 4 layers, the increase of number of layers decreases the performance of the LSTM.

Table 1. Effect of Number of Layers on LSTM Model

Number of Layers	Test Cost	Test LER
1	58.116	49.40%
2	49.725	42.25%
3	44.438	38.30%
4	**41.985**	**37.40%**
5	52.174	44.10%
6	55.206	47.90%

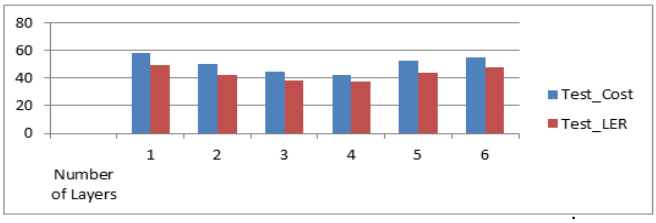

Figure 4. Effect of Number of Layers on LSTM Model

3.2. Effect of Number of Layers on DNN

In this experiment, the effect of number of layers on speech recognition performance were evaluated on DNN model. The results are presented in Table 2 and Figure 5. When the number of layers increases, DNN becomes more accurate for both LER and Test cost. However, after 5 layers, the increase of number of layers decreases the performance of the DNN. The best performance was observed at 5 layers for speech recognition task.

Table 2. Effect of Number of Layers on DNN Model

Number of Layers	Test Cost	Test LER
1	72.524	62.35%
2	63.682	55.45%
3	56.741	50.90%
4	54.954	49.30%
5	**53.035**	**48.25%**
6	60.782	53.45%

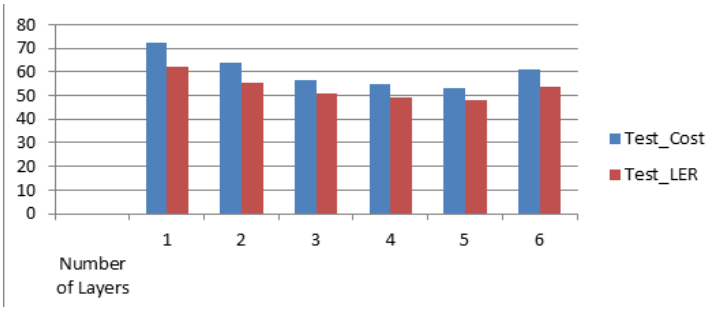

Figure 5. Effect of Number of Layers on DNN Model

3.3. Effect of Number of Units on LSTM

In this experiment, effect of number of units on speech recognition performance were evaluated on LSTM model. The results are presented in Table 3 and Figure 6. When the number of units increases, the accuracy of the LSTM increases too for both LER and test cost. However, after 250 units, the accuracy keeps constant and does not increase. The best performances are observed for the number of units of 250 and 300.

Table 3. Effect of Number of Units on LSTM Model

Number of Units	Test Cost	Test LER
50	64.233	53.75%
100	53.415	45.30%
150	47.744	41.60%
200	41.985	37.40%
250	37.248	**34.25%**
300	**36.312**	**34.25%**

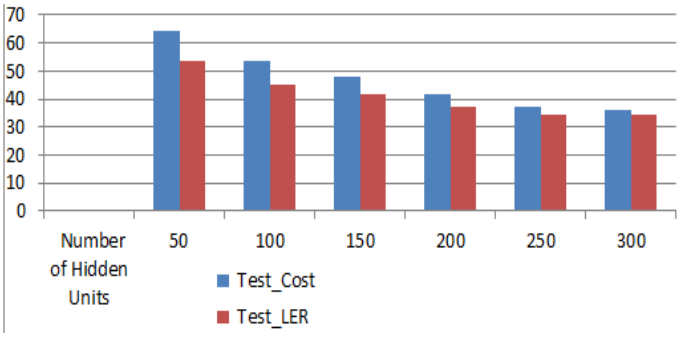

Figure 6. Effect of Number of Units on LSTM Model

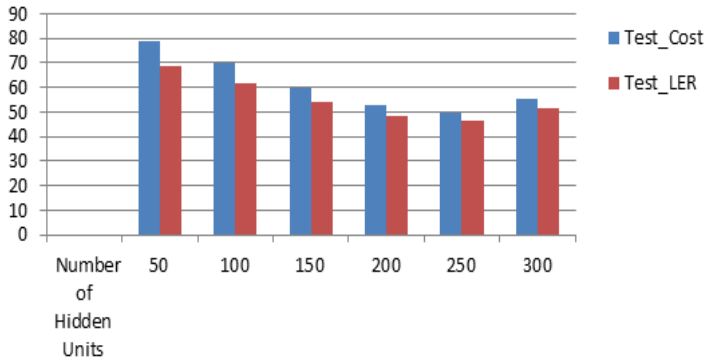

Figure 7. Effect of Number of Units on DNN Model

3.4. Effect of Number of Units on DNN

In this experiment, effect of number of units on speech recognition performance were evaluated on DNN model. The results are presented in Table 4 and Figure 7. When the number of units increases, the accuracy of the DNN increases too for both LER and test cost. However, after 250 units, the accuracy decreases. The best performances were observed for the number of units of 200 and 250**.**

Table 4. Effect of Number of Units on DNN Model

Number of Units	Test Cost	Test LER
50	78.524	68.35%
100	69.682	61.45%
150	59.741	53.90%
200	53.035	48.25%
250	**49.854**	**46.70%**
300	55.603	51.45%

4. Conclusions and Recommendations

One of the main purpose of this study is to Create a speech recognition database for Turkish language and to share it in the web environment as a source for other speech recognition based studies. For this purpose, approximately 2 hours of Turkish Broadcasts News Speech Dataset from a total of 1039 sound sentence files with their corresponding text transcripts were created. Also 8491 words (4101 different words), mono-phoneme and three-phoneme structures were created from all sentences.

Additionally, we investigated and compared the effect of the number of layers and number of cells for RNNs (LSTMs) and DNN hyperparameters on Turkish Broadcasting News Speech Dataset that we created. When the number of layers were increased for LSTM, the performance increases until 3 and 4 layers, after 3 and 4 layers, the performance gets worse in our setup. When the number of layers were increased for DNN, the performance increases until 4 and 5 layers, after 4 and 5 layers, the performance gets worse in this study. In both models, the success rate decreases after a certain layer due to the situation of memorization. When the number of units were increased with fixed number of layers for LSTM, the performance of the system gets better and later on the recognition rate saturates. When the number of units were increased with fixed number of layers for DNN, the performance of the system gets better and later on recognition rate decreases.

For our installation, the best performance was observed at the rate of % 34.25 LER in the LSTM model, while the best performance was observed as % 46.70 LER in the DNN model. LSTM produced more successful results than DNN. In previous studies on the effect of speech recognition on models, it was expected that LSTM would give better results and therefore LSTM would also give more successful results in this study.

Experimental results show that each parameter has its specific values for the selected number of training instances to provide lower error rates and better speech recognition performance. It is shown in this study that before selecting appropriate values for each LSTM and DNN parameters, there should be several experiments performed on the speech corpus to find the most eligible value for each parameter. It is observed that all hyperparameters that we applied have effect on the performance of LSTMs and DNN for speech recognition.

For future work, larger dataset and better modelling strategies and hyperparameters can be used to obtain better performance. Currently, our models have less dataset for training

and testing than those of the studies that obtain speech recognition based models performances. Secondly the number of speech databases created for Turkish is very few. Speech databases are extremely important for researchers who are trying to develop a Turkish Speech Recognition System. Their number and diversity should be increased. As a result, this study set a precedent for creating a database for studies in Turkish speech recognition. The created dataset can be accessed from the following web site;

http://cusesveri.com/

References

Bengio, Y., 2009. "Learning Deep Architectures for AI" (PDF). Foundations and Trends in Machine Learning. 1–127.

Gaikwad, S., Gawali, B. W., & Yannawar, P. 2010. A review on Speech Recognition Technique. , pp. 16-24

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.

Graves, A., Jaitly, N., & Mohamed, A. R. (2013b, December). Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding (pp. 273-278). IEEE.

Hizlisoy, S., 2020. Music Emotion Recognition Using Convolutional Long Short Memory Deep Neural Networks.

Patlar, F., 2009. A Continuous Speech Recognition System For Turkish Language Based On Triphone Model.

Sepp Hochreiter; Jürgen Schmidhuber (1997). "LSTM can Solve Hard Long Time Lag Problems". Advances in Neural Information Processing Systems 9. Advances in Neural Information Processing Systems. Wikidata Q77698282.

Tüfekci, Z., and Dokuz, Y., 2020. Investigation of the Effect of LSTM Hyperparameters on Speech Recognition Performance , European Journal of Science and Technology: p. 165.

Yu, D., & Deng, L. (2016). Automatic Speech Recognition: A Deep Learning Approach. Springer

Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B. (2016). Learning Contextual Dependence with Convolutional Hierarchical Recurrent Neural Networks. IEEE Transactions on Image Processing, 25, 2983-2996.

http://kaldi asr.org/doc/examples.html