Harla Journals

Harla Journal of Engineering and Technology

Journal home page: https://journals.ddu.edu.et/index.php/HJET

Harla journal of Engineering and Technology

Investigation of deep learning techniques in speech recognition for under-resourced languages: the case of Afaan Oromo

[1, 2, *] Yadeta Gonfa, [1]Kula Kekeba

[1]*Department of Software Engineering, College of Mechanical and Electrical Engineering, Addis Ababa Science and Technology University, Addis Ababa, Addis Ababa, Ethiopia*

[2]*Department of Information Technology, School of Technology and Informatics, Ambo University, Ambo, Ethiopia*

yadeta4@gmail.com, *kuulaa@gmail.com*

Abstract

Human-machine interactions are increasing in day-to-day human activities. Automatic Speech Recognition (ASR) is one of the hot research areas to invent the machine that can understand human languages to give responses. Many researchers show the possibility of developing a speech recognition system for assisting human beings in communicating with their machines like computers. ASR work started in the mid of 19th century, and several improvements were presented by implementing various tools and techniques. Several works of literature show that the deep learning approach is currently state-of-the-art in speech recognition. Still, there needs to be more research on learning approaches for under-resourced languages. However, the need for large datasets to implement deep learning approaches is challenging for under-resourced languages. Exploring a deep learning approach for Ethiopian languages, in general, and Afaan Oromo, in particular, should have been emphasized. Therefore, investigating deep learning techniques in speech recognition for Afaan Oromo is the main objective of this study. The experiment was conducted on 2953 utterances of total datasets, and the Convolutional Neural Network (CNN) model was used. The datasets were partitioned into training, validating, and testing datasets. The best test accuracy of 51.27% was obtained when batch size, number of epochs, and learning rate were set to 32, 40, and 0.001, respectively. This result is incredible when compared with the result obtained using the Hidden Markov Model (HMM). Therefore, we have a conclusion on the possibility of investigating deep learning techniques in speech recognition for Afaan Oromo by implementing the CNN model. Further work could be experimented with by using other deep learning algorithms and techniques to improve the accuracy.

Keywords: Deep learning, convolutional neural network, speech recognition, under-resourced languages, Afaan Oromo

1. INTRODUCTION

Deep learning is one of the approaches from a class of machine learning that uses multiple layers of computational nodes to progressively extract higher-level and lower-level features from the raw

---

Corresponding author: Yadeta Gonfa, yadeta4@gmail.com,

DOI:

 Manuscript received October, 2022; revised December, 2022; accepted December, 2022.

input. Starting from the 20th century, this class of machine learning has developed strongly and has been incorporated into many types of research [1]. For this reason, recently the areas in which deep learning has been merged are started from information processing to artificial intelligence [2]; besides, deep learning has emerged as a powerful promising technique in fields ranging from computer vision to natural language processing and even recommendation systems [3]. Deep learning is usually implemented using neural networks and deep neural networks that are becoming a fundamental component of high-performance speech recognition systems by using a large dataset of training [4]. Because of the performance that deep learning-based systems a Deep Neural Network (DNN) have lately exceeded that of their Gaussian mixture model (GMM) counterparts in acoustic modeling.

Many questions are often raised when deep learning is targeted to use in speech recognition for under-resourced languages. Some questions are: (i) why is deep learning in speech recognition, (ii), and how it can be implemented for under-resourced languages because for training a huge amount of data is required in deep learning.

Before the investigation of deep learning-based approaches, the Hidden Markov Model (HMM) is the dominating approach in the work of speech recognition for several years [1], [5]. In addition, works of the literature showed that the research in ASR is becoming significantly advanced using deep learning approaches by producing the-state-of-the-art results when compared to conventional approaches. Mostly, the three deep learning architectures utilized in the field of speech recognition are Deep Belief Networks (DBN), (CNN), and Recurrent Neural Networks (RNN) [2], [6]. Nowadays, because of its several advantages like accuracy improvement, using a deep learning-based approach for the work of ASR is recommended. Therefore, it is also possible to implement the deep learning approaches for under-resourced languages by using different techniques such as data augmentation, multi-lingual and cross-lingual approaches, and transfer learning [7].

Speech is one of the most natural interfaces of human communication. Naturally, human beings are knowledgeable about how to speak and make communication through speech. Typically, speech recognition technology allows computers to take the spoken audio forms of a natural language as an input, interpret it, and generate a text form as an output [4], [8]. Nowadays, speech recognition is occupying our lives in several ways: it is built into our phones, our game consoles, our smartwatches, and even our homes. This justifies that a speech recognition system is useful in several aspects of life activities, such as the telephony system, the command control system, and the dictation system [9].

Several tasks were performed to develop a speech recognizer system with the help of spoken/natural language processing; like creating acoustic modeling, language modeling, and having a lexicon of pronunciations [10]. Different statistical models were applied for the completion of those tasks and in the work of ASR, N-gram language models were typically used in the past and acoustic models are built using GMM [11].

Speech recognition has been around for decades using a deep learning approach to overcome the problems of using statistical models. Deep learning is finally able to make the speech recognition with enough accuracy to be useful. For example, the neural language models were built instead of using n-gram language models to feed them into a speech recognition system to restore things produced. By looking at the pronunciation models, figure out to do comparing pronunciation for a new sequence of characters that have never been seen before using a neural network.

As mentioned above, deep learning has emerged as a powerful favorable technique in several fields, ranging from computer vision to natural language processing and even to recommendation systems. Speech recognition is also one field that requires the application of deep learning models to achieve good accuracy with a reduction in error rates.

It is clear that a large amount of dataset for training is required when deep learning is used, and this is also good for the very rich (resourced) languages. Under-resourced languages were challenged to use the deep learning-based approaches because of the scarcity of large datasets; but using different techniques like transfer learning, multi-lingual approach, and the like, it is possible to implement the work of speech recognition in such languages. Deep learning models in speech recognition were not more emphasized yet for the under-resourced languages like Afaan Oromo. Therefore, investigating a deep learning technique by utilizing the CNN model in speech recognition for Afaan Oromo was taken as the focus of this paper.

2. REVIEW OF THE LITERATURE

2.1. Overview of Afaan Oromo

Ethiopia has more than 80 different languages. Afaan Oromo is one of several Ethiopian languages that have a speaker over 34 million people in the Horn of Africa [12]. Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written and the language is free from the problem of homonymy [13]. This means there are no words written in different ways and pronounced in the same way in Afaan Oromo. This can be taken as one good feature of that language (Afaan Oromo). Also, researchers cannot be challenged for finding solutions to homonymy's problem. The language has 33 alphabets (5 vowels, 21 consonants, and 7 doubled letters called "*qubee dachaa*") where Latin letters are used for the writing system. Afaan Oromo has 59 phones. The language has also a special symbol for glottal sound [13], but it is challenged to have a standard International Phonetics Alphabets (IPA) representation for the glottal sound which is called "*hudhaa*".

2.2. Related Works

A work of ASR for Afaan Oromo is started by using isolated words [14]. The researcher used HMM model and an open-source speech recognition toolkit Sphinx4. A researcher prepared 50 Afaan Oromo words as a corpus by consulting the domain experts; then these words were read by 20 persons. In total, 1000 utterances of Afaan Oromo isolated words were obtained for the experiment. In his study, when 66.67% of the data was used for training, the remaining 33.33% of the data was used for testing purposes. In the end, the word-level accuracy achieved by the researcher's work was 82.83% and 81.081% for the context-dependent phoneme-based model and context-independent word-based, respectively.

Another study was conducted by using continuous speech, to develop a continuous speaker-independent speech recognizer system for Afaan Oromo [15]. The study aims to explore the possibility of developing a continuous Afaan Oromo speech recognition system by using the already available tools and techniques; the researcher used the HMM model and the sphinx system (Sphinx train for training and Sphinx4 for decoding). In the study [15], 70 Afaan Oromo long words, phrases, and simple sentences were selected and read by 30 people who are different by their age and gender to prepare a corpus consisting of 2100 utterances. For the experiment, 66.67% of the data is used for training and the remaining 33.33% of the data is used for testing purposes. The performance level achieved was 68.514% of word accuracy with sentence accuracy of 28% and 89.459% of word accuracy with the sentence accuracy of 42%, in the context-dependent and context-independent model, respectively.

An additional study was conducted to explore the possibility of developing a speech recognition system for Afaan Oromo using broadcast news [13]. The researcher has used a piece of broadcasting news of the language to prepare a speech corpus. For the study, a Hidden Markov Model (HMM) and HTK tools were used. Accordingly, 2953 utterances/sentences with lengths of about 6 hours of speech corpus are prepared for the experiment. Out of the total speech datasets, 2653 utterances were used for training and the remaining 300 utterances were used to test the developed system. Hence, the WER's best performance was 91.46% and 89.84%, for context-independent and context-dependent, respectively. However, the obtained result is not as interesting because the WER achieved was high.

Syllable-based speech recognition was developed for Afaan Oromo by using Hidden Markov Model (HMM) and HTK tool [16]. In this study [16], the researcher used a speech corpus of more than 4 hours long in total (i.e., 4 hours of speech for training and 40 minutes of speech for testing). The speech was collected from 63 speakers who are different in gender (39 males and 24 females). To build the speech recognizer, a bigram language model, phone-based, and syllable-based alternative pronunciation dictionary was also built. As a result, the correctly recognized word was 39.55%, 47.21, 55.35, and 43.96% by using mono-phone, tri-phone, tied state tri-phone, and syllable-based recognition models respectively. The main finding of the study indicates that the performance obtained for syllable-based Afaan Oromo speech recognition is highly increasing as the frequency of syllables increases.

Another developed speech recognition is using a hybrid of the Hidden Markov Model the and Artificial Neural Network (HMM/ANN) model [17]. The Center for Spoken Language Unit (CSLU) hybrid toolkit was used along with other tools used for recording and labeling the speech corpus. In this study [17], the experiment was conducted on the recognition of limited vocabulary by using a total of a hundred Afaan Oromo words, and the words are organized to form sentences to make recording easy. The recorded speech was labeled manually for the experimental process. After training and testing the recognizer system with the labeled speech, the word accuracy of 98.11% was achieved as the best result.

Recently, a Deep Neural Network (DNN) based automatic speech recognition was developed for four Ethiopian languages namely Afaan Oromo, Amharic, Tigrigna, and Wolaytta [12]. An experiment was conducted by having a training speech corpus of about 20 to 29 hours and about 1 hour of speech was used for evaluation. The researcher developed the lexical and language models for two languages namely Amharic and Tigrigna. But, for Afaan Oromo and Wolaytta, the trained lexicon models were used for decoding. In the end, researchers have achieved relative WER reductions that range from 15.1% to 31.45% for the four Ethiopian languages by using DNN-based acoustic models.

A multilingual speech recognition system was developed based on an end-to-end approach for under-resourced Ethiopian languages namely Amharic, Afaan Oromo, Tigrigna, and Wolaytta [18]. The end-to-end (E2E) framework is more attractive for less-resourced languages. Because the E2E approach maps a sequence of input features into a sequence of graphemes or words. The researchers address the problem of training data scarcity by using a multilingual approach. Accordingly, the experiment was conducted for Afaan Oromo by using a training speech which have a length of 22.8 hours. The result shows that the use of multilingual data leads to E2E ASR performance improvement over the use of monolingual data.

2.3. Deep Learning Tools and Techniques in ASR

Several works of the literature show that deep learning becoming a popular topic with high research interest due to its extensive application generally in natural language processing and speech recognition [19]. There are many algorithms namely Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM), and Transformers that deep learning provides for the development of a speech recognition system. From these several algorithms, we have implemented by using the CNN algorithm for this study.

The CNN model can be utilized in speech recognition, because, this is a type of separated deep architecture in which every model contains a convolutional layer and a merging layer that are stacked on top of each other [20]. A CNN is a multi-layer neural network that consists of two different types of layers that alternate: the convolution layer and the pooling layer. The convolution layer is used to extract features. It consists of several feature maps made up of neurons. Each neuron in the convolution layer processes data only for its receptive field which features a limited range and not the whole input. Currently, CNNs have been effectively applied in many research areas like computer vision and speech recognition tasks [21]. Traditional CNNs for speech recognition consists of several convolutional layers and pooling layers, followed by several fully connected layers for acoustic modeling.

3. METHODOLOGY

The first phase in developing a speech recognition system is preparing the corpus required for the work. Therefore, this section gives the descriptions of the methodology used for data collection, audio pre-processing methods, and preparing the speech datasets methods.

3.1. Data Collection Methods

It is obvious that the acoustic model is one of the components of speech recognition; speech is one and the primary input for the recognizer system. Hence, for modeling acoustic, we need audio/speech data.

Table 1. Audio data collected from several sources

Source of speech	Length (in Hour)
OBN	02:20:14
FBC	01:05:11
OBS	01:03:07
FIB	01:07:11
VOA Afaan Oromo	01:08:09
BBC Afaan Oromo	01:03:08
Total	**07:47:00**

We have collected the speech data by downloading it from the websites of Oromia Broadcasting Network (OBN), Fana Broadcasting Corporate (FBC), Finfinnee Integrated Broadcasting (FIB), Oromia Broadcasting Service (OBS), BBC Afaan Oromo, and VOA Afaan Oromo. The reason why we select these broadcasting media is only they have the open domain Afaan Oromo speech data. The collected speech data is about 07:47 hours in length before implementing the data pre-processing techniques as depicted in Table 1.

3.2. Audio Pre-processing

The collected data needs to be pre-processed because all audio corpus gathered could not be used in this work as it is. Accordingly, we conducted pre-processing on the collected audio data of Afaan Oromo news to construct the speech data that is a suitable format. Also, to make the audio files in the supported file formats, we have used an open-source software called Audacity for preparing the audio file in sentence-level and saving the audio files separately in .wav file format. We open the downloaded audio files with audacity. The task of pre-processing audio (speech) segmentation takes a more place in setting the speech.

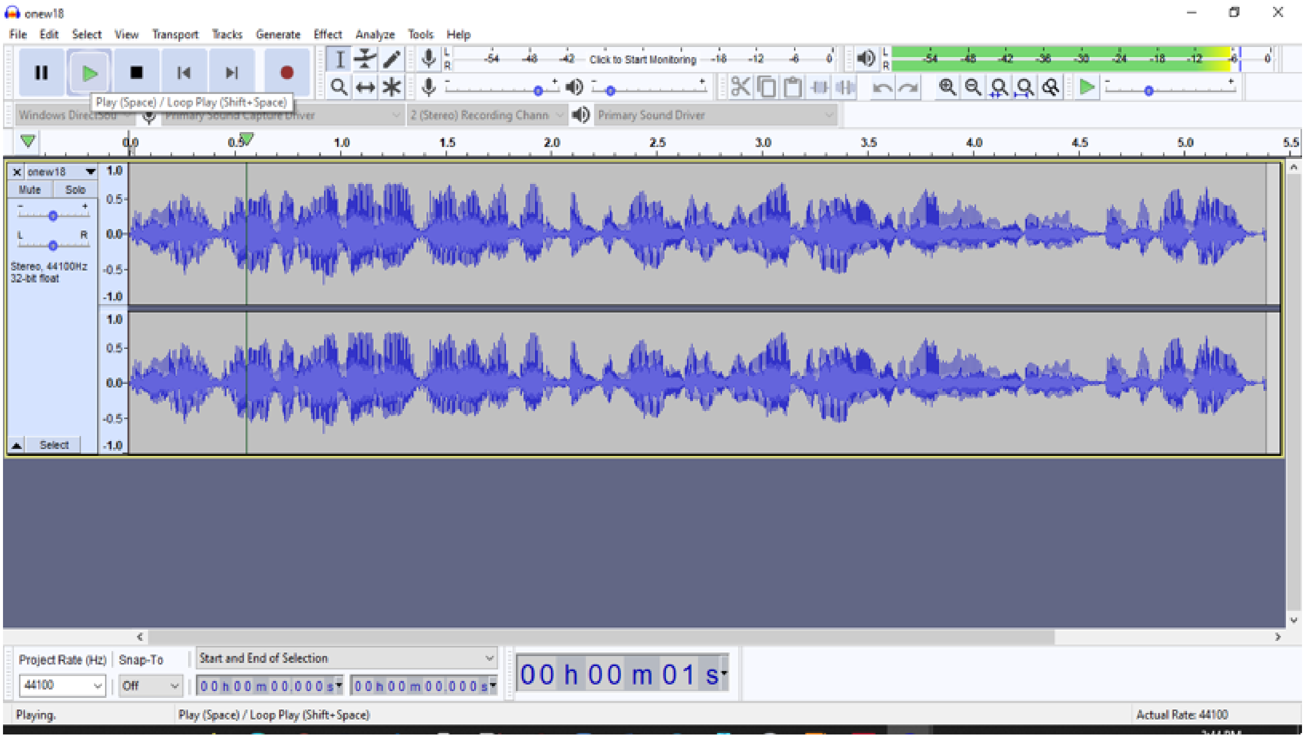

Fig. 1: Audio Segmentation by Audacity

To the best of our knowledge, there is no preferred tool to segment the audio data of Afaan Oromo automatically. Therefore, audio (speech) segmentation is done manually by listening to the audio file as shown in Fig 1. During segmentation, telephone reports, interviews, and other non-speech-like background music were skipped. Accordingly, we have constructed 2953 total speech utterances that have a length of about 6 hours. The utterances of prepared speech corpus have different lengths in second as described in the following Table 2. In general, the prepared speech corpus has a length of 1 second for the minim length and 26 seconds for the maximum length.

Table 2. The pre-processed audio data

Length (in Seconds)	# of utterances
1 – 5	1142
6 – 10	1279
11 – 15	448
16 – 20	72
21 – 26	12
Total	**2953**

3.3. The designed architecture

The overall architecture for this work showed in Fig 2. After completing the work of pre-processing, speech in a waveform was prepared. Because having such speech which is an input for speech recognition work is the initial task. In the work, first, we have collected audio data of broadcast news that have a length of more than 7 hours. Next, we have performed the task of pre-processing.

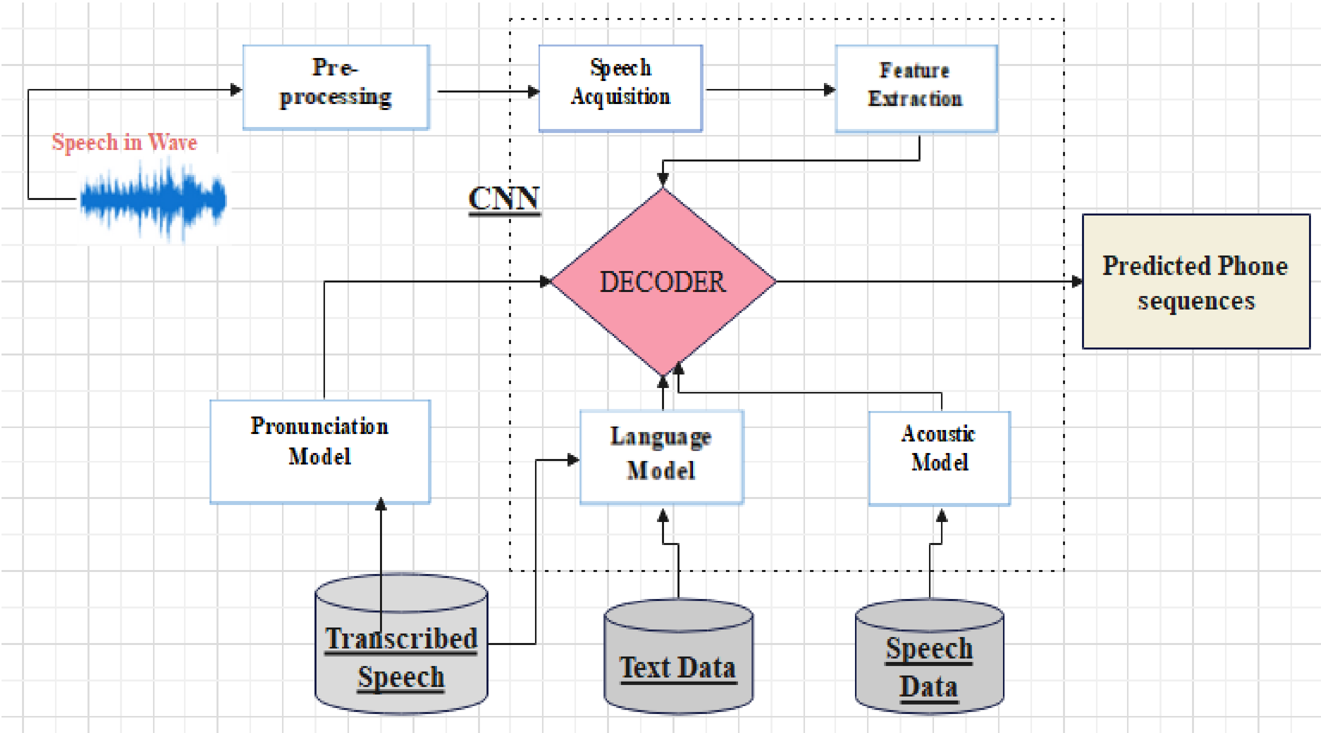

Fig. 2. Conceptual Framework of ASR Development for Afaan Oromo

We have used a CNN model to explore the investigation of deep learning techniques for the Afaan Oromo speech recognition system.

4. EXPERIMENT AND DISCUSSION

4.1. Splitting the Datasets

For the experiment, we have split our total dataset into three as 80%, 10%, and 10% of datasets used for training, validating, and testing, respectively. For this purpose, the most useful and robust library for machine learning in Python called Scikit-learn (Sklearn) was used. Hence, we have set the training, test validation, and testing dataset.

4.2. Result and Discussion

We have performed several experiments by changing the parameters like batch size, the number of epochs, and the learning rate. To begin with, the experiment was conducted by changing the batch size and using a learning rate=0.0001, epoch=40 shown in Table 3. Therefore, we have obtained different results on the test accuracy of 46.08% was achieved as best when the batch size is 32.

Table 3. Result Experiment I by changing Batch size.

Batch Size	Epochs	Learning Rate	% of Test Accuracy
1	40	0.0001	38.57
16	40	0.0001	39.93
32	**40**	**0.0001**	**46.08**
64	40	0.0001	41.98
128	40	0.0001	38.91

Now by taking the batch size of 32 and the number of epochs = 40, we experimented by changing the learning rate. The test accuracy rate is 51.27% at a 0.001 learning rate. The learning rate can be between 0.0 & 1.0, i.e. 0.0 stands for low learning rate and 1.0 indicates a high learning rate as justified in Table 4. At a low learning rate, the training is more reliable.

Table 4: Result Experiment II by changing learning rate.

Batch Size	Epochs	Learning Rate	% of Test Accuracy
32	40	0.0	6.83
32	40	0.00001	40.61
32	40	0.0001	45.39
32	**40**	**0.001**	**51.27**
32	40	0.01	44.37
32	40	0.1	35.15
32	40	1.0	39.93

As Table 5 shows, another experiment is conducted based on changing the epoch number and keeping the batch size and learning rate are 32 and 0.001, respectively that performing the best accuracy rate of testing as 51.27%.

Table 5: Result Experiment III by changing the number of epochs.

Batch Size	Epochs	Learning Rate	% of Test Accuracy
32	11	0.001	37.88
32	20	0.001	41.30
32	30	0.001	45.05
32	**40**	**0.001**	**51.27**
32	50	0.001	44.37
32	80	0.001	40.61
32	100	0.001	39.25

As shown in the above Tables 2, 3, and 4, the highest performance obtained was 51.27% of accuracy. CNN model can achieve the presented accuracy by using 32, 40, and 0.001 for batch size, the number of epochs, and learning rate, respectively.

4.3. Comparing results with previous work

From the ASR works of Afaan Oromo, we have taken the study conducted [13] by using the HMM model with a total dataset of 2953 utterances. The study is conducted by pure HMM and but the obtained result was not promising. Because, the word error rate was very high (WER of 91.46% and 89.84%, for context-independent and context-dependent, respectively). However, in this study, we have used the same dataset by only using the deep learning approach of the CNN algorithm. The result shows that it is better than the previous work.

5. CONCLUSION AND FUTURE WORK

5.1. Conclusion

From the result of different experiments, it is understandable as there is a possibility of error rate reduction by using the CNN model of deep learning for Afaan Oromo. Accordingly, from several experiments, 51.27% of test accuracy was obtained as the best result using 40, 32, and 0.001 parameters for the number of epochs, batch size, and learning rate. When this result is compared with the previous work conducted on the same language and datasets by pure HMM that achieved an 89.84% word error rate, it is optimistic about implementing the deep learning application in speech recognition for Afaan Oromo.

5.2. Acknowledgments

Foremost, we would like to thank Dire Dawa University Institute of Technology who preparing for its first International Conference on Computing, Informatics, and Engineering Conference. Again, we thank Addis Ababa Science and Technology University (AASTU). So, my thanks also to Ambo University. I hope that I am waiting for a more constructive comment while presenting.

REFERENCES

[1] A. Kumar, S. Verma, and H. Mangla, "A Survey of Deep Learning Techniques in Speech Recognition," *Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018*, pp. 179–185, 2018, DOI: 10.1109/ICACCCN.2018.8748399.

[2] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, no. February, pp. 19143–19165, 2019, DOI: 10.1109/ACCESS.2019.2896880.

[3] L. Deng, "Deep learning: From speech recognition to language and multimodal processing," *APSIPA Trans. Signal Inf. Process.*, vol. 5, no. 2016, pp. 1–15, 2016, DOI: 10.1017/atsip.2015.22.

[4] P. Vanajakshi and M. Mathivanan, "A detailed survey on large vocabulary continuous speech recognition techniques," *2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017*, 2017, DOI: 10.1109/ICCCI.2017.8117755.

[5] S. Singhal and R. K. Dubey, "Automatic speech recognition for connected words using DTW/HMM for English/Hindi languages," *Int. Conf. Commun. Control Intell. Syst. CCIS 2015*, pp. 199–203, 2016, DOI: 10.1109/CCIntelS.2015.7437908.

[6] Z. Zhang, J. Geiger, J. Pohjalainen, A. E. D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally

robust speech recognition: An overview of recent developments," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 1–16, 2018, DOI: 10.1145/3178115.

[7] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 2613–2617, 2019, DOI: 10.21437/Interspeech.2019-2680.

[8] X. Lu, S. Li, and M. Fujimoto, "Automatic speech recognition," *SpringerBriefs Comput. Sci.*, pp. 21–38, 2020, DOI: 10.1007/978-981-15-0595-9_2.

[9] H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems: Recent Advances and New Frontiers," vol. 19, no. 2, pp. 25–35, 2017.

[10] G. Thimmaraja Yadava and H. S. Jayanna, "Creating language and acoustic models using Kaldi to build an automatic speech recognition system for Kannada language," *RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc.*, vol. 2018-Janua, pp. 161–165, 2017, DOI: 10.1109/RTEICT.2017.8256578.

[11] K. Jeeva Priya, S. S. Sree, T. Navya, and D. Gupta, "Implementation of Phonetic Level Speech Recognition in Kannada Using HTK," *Proc. 2018 IEEE Int. Conf. Commun. Signal Process. ICCSP 2018*, pp. 82–85, 2018, DOI: 10.1109/ICCSP.2018.8524192.

[12] S. T. Abate, M. Yifiru Tachbelie, and T. Schultz, "Deep Neural Networks Based Automatic Speech Recognition for Four Ethiopian Languages," pp. 8274–8278, 2020, DOI: 10.1109/icassp40776.2020.9053883.

[13] Y. G. Gutu, "A Large Vocabulary, Speaker-Independent, Continuous Speech Recognition System for Afaan Oromo: Using Broadcast News Speech. Masters Thesis: Addis Ababa University, Addis Ababa, Ethiopia," 2016.

[14] A. Demissie, "A Speech Recognition System for Afaan Oromo Using isolated Words. MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia," 2009.

[15] K. G. Micho, "A Continuous, Speaker Independent Speech Recognizer For Afaan Oromoo. Masters Thesis, Addis Ababa University. Addis Ababa, Ethiopia," 2010.

[16] A. Dessalegn, "Syllable Based Speaker Independent Continous Speech Recognition for Afan Oromo," 2020.

[17] T. Kebebew, "Speech Recognition for Afan Oromo Using Hybrid Hidden Markov Models and Artificial Neural Networks. Addis Ababa University School: Masters Thesis. Addis Ababa, Ethiopia.," no. June 2010.

[18] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "End-to-End Multilingual Automatic Speech Recognition for Less-Resourced Languages : The Case of Four Ethiopian Languages. CSL, University of Bremen, Germany," *IEEE*, pp. 7013–7017, 2021.

[19] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," *ACM Int. Conf. Proceeding Ser.*, pp. 142–147, 2016, DOI: 10.1145/3015166.3015210.

[20] V. Delić *et al.*, "Speech technology progress based on new machine learning paradigm," *Comput. Intell. Neurosci.*, vol. 2019, 2019, DOI: 10.1155/2019/4368036.

[21] A. Vilardi, "Very Deep Convolutional Neural Networks for face identification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2263–2276, 2016, DOI: 10.1109/TASLP.2016.2602884.