Dari Speech Classification Using Deep Convolutional Neural Network

Mursal Dawodi
School of Engineering and Science
University of the Ryukyu
Okinawa, Japan

Jawid Ahamd Baktash
Faculty of Computer Science
Kabul University
Kabul, Afghanistan

Tomohisa Wada
School of Engineering and Science
University of the Ryukyu
Okinawa, Japan

Najwa Alam
Faculty of Computer Science
Kabul University
Kabul, Afghanistan

Mohammad Zarif Joya
Faculty of Computer Science
Kabul University
Kabul, Afghanistan

***Abstract*— Recently, speech recognition is one of the most advanced research topics in the world. Many recent research papers have proven the power of deep neural networks in speech recognition systems. The main purpose of this paper is to identify isolated words in Dari speech using deep learning algorithms. This research is one of the new studies in Dari speech recognition and focuses on one-word speech recognition. This collection uses our audio files as a database because there were no Dari language databases on the market at that time. In this paper, the Convolutional Neural Network (CNN) is implemented to detect automatically isolated words in Dari. Besides, it uses Mel frequency coefficients (MFCC) to learn the representation of features during training. This model achieved 88.2% in the test set. The results show that the model can predict samples of words seen during training with high accuracy. However, it is somewhat trying to generalize terms outside the scope of training data and very noisy examples.**

Keywords—speech classification, Dari, convolutional neural network, deep neural network, speech recognition

I. INTRODUCTION

Anyone using technology devices is usually faced with speech recognition systems such as automatic dictation, human-computer dialogue, speech-to-speech translation, system control, etc. Researchers have already evaluated many alternative algorithms to enhance this technology and discover more precise algorithms. In the past few years, deep learning, a specific subfield of machine learning has been a subject of intense media hype and achieved impressive results in speech recognition like its outstanding outcomes in alternative natural language processing themes, image processing, and other research articles [11, 12, 13].

Although there are several studies in speech recognition technology for numerous human languages, there is still a gap for Dari Language. This study develops an automatic speech recognition system in the context of Dari speeches using MFCC for feature extraction and deep CNN for prediction and classification tasks. Moreover, there is not any dataset of Dari language for this context. Therefore, in the early stages of this analysis, we prepared a collection of 2753 audio records.

II. RELATED WORKS

Much research has been done on the implementation of deep neural networks in speech recognition programs. As an example, A. Graves et. Al. [1] represented the efficiency of combining deep, bidirectional Long Short-term Memory (LSTM) RNNs with end-to-end training and weight noise on the TIMIT phoneme recognition benchmark that reduced test set error rate to 17.7%. Similarly, [3] showed that CNN outperforms in speech recognition context and reduces the error-rate. Some of the studies focused on improving accuracy for large vocabulary speech recognition while some other researchers focused on leveraging the power of deep learning in context-based speech recognition. G. E. Dahl et. al. [2] proposed a novel context-dependent pre-trained deep neural networks model. Their investigated model had higher sentence accuracy and lower relative error compared to the former proposed conventional context-dependent Gaussian mixture model.

In recent years some natural processing language researchers started research on Persian NLP. They used different methods and developed several models. One of the best projects is Navisa [8] Persian continuous speech recognition system. [7] proposed a model for isolating Persian spoken words that were evaluated on the first interactive robot with Persian commands in 2008. Another research [4] employed MFCC and a multilayer perceptron feed-forward artificial neural network to distinguish vowel and consonant characters in Persian speeches. They used the Persian Consonant-Vowel Combination dataset that consists of 20 sets of audio samples from 10 speakers and implemented MFCC on every partitioned sound sample. Farsi LTS (letter to the sound system) [5] is a system that translates Persian letters and words to speak. The researchers used neural networks with rule-based and MLP layers. Subsequently, they obtained 61% to 87% conversion accuracy. Likewise, S. Malekzadeh [6] utilized MLP deep learning algorithms to detect Persian phonemes by artificial intelligence and improve voice signal processing.

III. DATA PREPARATION

This research formed a new data set called Isolated Dari Speech Database, the first Dari-speaking audio data set, because we could not find any product on the Internet, as mentioned in the first section. The dataset contains 2753 one-second long utterances of 20 short words, by several people including both male and female. The files are classified into 20 classes each represents a common word or command in Dari. The categories are Gol, Salaam, Qand, Jaan, Nek, Dast, Paa, Bebin, Boro, Beeyaa, Beradar, Khaahar, Haazer, Bad, Bogoo, Boot, Cheshm, khoob, Boo, and Goft. Fig. 1 lists all classes within our dataset and Fig. 2 demonstrations the distribution of each class. In addition to preparing the audio sets, data augmentation was used to reduce overfitting,

978-1-7281-9615-2/20/$31.00 ©2020 IEEE

artificially synthesizing new training data and increase the size of training sets. We structure our data in a csv file where name of audios is listed with their specified label. Fig. 3 shows the metadata for each record of the employed corpus. Similarly, Fig.4 demonstrates an example of depicting words within the corpus as the raw waveform. This study used supervise learning techniques and employed labeled dataset to train the models.

1	گل	Gol
2	سلام	Salaam
3	قند	Qand
4	جان	Jaan
5	نیک	Nek
6	دست	Dast
7	پا	Paa
8	ببین	Bebin
9	برو	Boro
10	بیا	Beeyaa
11	برادر	Beradar
12	خواهر	Khaahar
13	حاضر	Haazer
14	بد	Bad
15	بگو	Bogoo
16	بوت	Boot
17	چشم	Cheshm
18	بو	Boo
19	گفت	Goft
20	خوب	khoob

Fig. 1. Classes within dataset

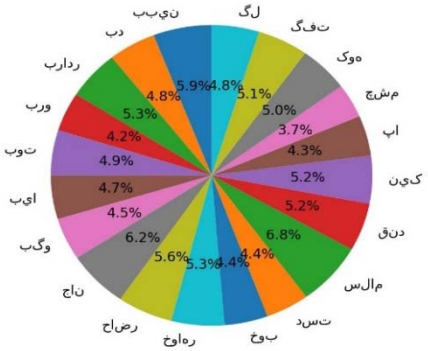

Fig. 2. Class distribution

	slice_file_name	fold	class_name
0	bad1.wav	1	بد
1	bad10.wav	1	بد
2	bad11.wav	1	بد
3	bad12.wav	1	بد
4	bad13.wav	1	بد
5	bad14.wav	1	بد
6	bad15.wav	1	بد
7	bad16.wav	1	بد
8	bad17.wav	1	بد
9	bad18.wav	1	بد

Fig. 3. Metadata example for each prepared Dari sound in the dataset

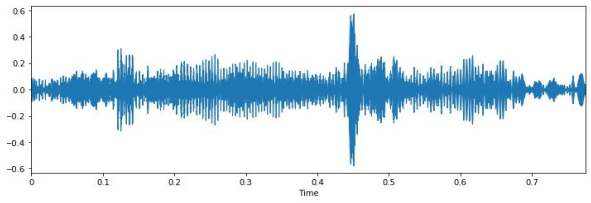

Fig. 4. A sample depicting the word "Jaan" as the raw waveform

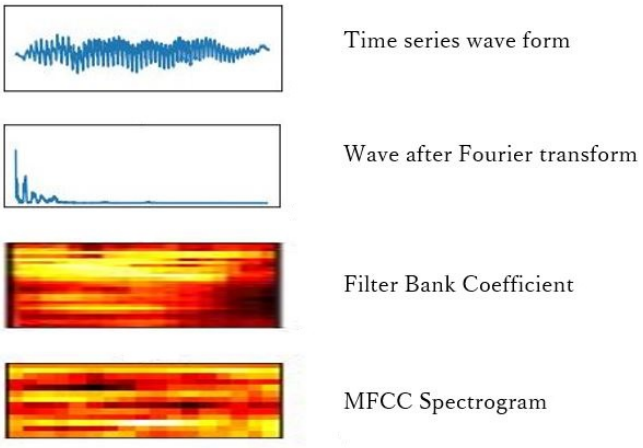

Fig. 5. Time series, Fourier transform, Filter bank coefficient and MFCC of an audio in Paa class

IV. Data Preprocessing

We have pre-processed the audio data before using them. Hence, in the first step, we normalized the sounds to range from -1 to +1. We changed the audio signals to mono because all sounds are considered on one channel only. To extract features from the raw waveform, we used spectroscopy, log-Mel filter banks, and Mel-Frequency implantation coefficients (MFCCs) from stereotyped samples to convert the raw waveform to a time-frequency domain [9] (see Fig. 5). These features are considered as the inputs for neural nets [10]. Moreover, we transformed the data to numeric vectors and subsequently split the dataset into two separate sets of training and validation 80% and 20% respectively.

V. Deep Neural Network Models

Our model is based on CNN to classify utterances and predict associated text. The CNN model consisted of four blocks containing two-dimensional convolutional layers. The following layer is a two-dimensional max pooling layer with a 50% dropout and three dense layers. This model utilizes the ReLU activation function. The number of features, frames, and channels are the input shape of the first layer. The entire layers have a kernel size of 3 throughout the network. The model has 16 convolutional filters in the first layer that is doubled in the proceeding layers. The output layer is a dense layer with an output size of 20 and Softmax activation function as exposed in Fig. 6. The total amount of parameters in the network is roughly 45.892 thousand.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 39, 173, 16)	80
max_pooling2d_1 (MaxPooling2	(None, 19, 86, 16)	0
dropout_1 (Dropout)	(None, 19, 86, 16)	0
conv2d_2 (Conv2D)	(None, 18, 85, 32)	2080
max_pooling2d_2 (MaxPooling2	(None, 9, 42, 32)	0
dropout_2 (Dropout)	(None, 9, 42, 32)	0
conv2d_3 (Conv2D)	(None, 8, 41, 64)	8256
max_pooling2d_3 (MaxPooling2	(None, 4, 20, 64)	0
dropout_3 (Dropout)	(None, 4, 20, 64)	0
conv2d_4 (Conv2D)	(None, 3, 19, 128)	32896
max_pooling2d_4 (MaxPooling2	(None, 1, 9, 128)	0
dropout_4 (Dropout)	(None, 1, 9, 128)	0
global_average_pooling2d_1 (	(None, 128)	0
dense_1 (Dense)	(None, 20)	2580

Total params: 45,892
Trainable params: 45,892
Non-trainable params: 0

Fig. 6. CNN model architecture summarization

VI. Model Evaluation and Discussion

The model was evaluated based on both training and test set accuracies including 80% and 20% of the samples. This model obtained overall accuracy of 88% and 0.45 validation loss compared as shown in Fig. 7.

For this study, we evaluate the best performance of the models by decreasing and increasing the number of dense layers, pooling layers, dropouts, and kernel size. Finally, the best options were selected for this research as described in previous sections. However, the impressive accuracy result, 88% on testing data, for CNN model was achieved through resampling the audio sample rate to 8000, changing dimensionality of the model to two and increasing the kernel size both in conventional and pooling layers.

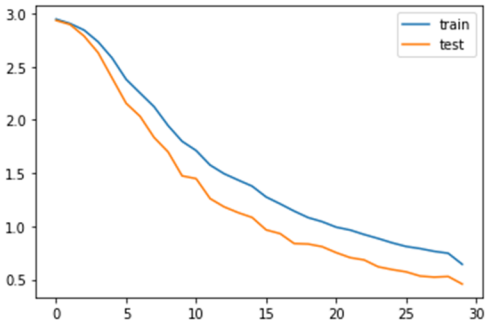

Fig. 7. CNN model: training and evaluation loss

VII. Conclusion and Future Work

In this paper, the effectiveness of the MFCC feature extraction technique and the deep CNN model along with sensible training techniques for Dari one-word Dari speech recognition is demonstrated. This is a preliminary study of Dari natural language processing, and more research is needed to deal with very noisy examples and unsupervised Dari words. In addition, we used the limited vocabulary of Dari words. However, the implementation of more continuous and accurate models requires a very large corpus. Providing a collection of information enriched by the Dari speeches and creating more effective and accurate models are the future work of this research.

References

[1] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30–42, Jan. 2012, doi: 10.1109/TASL.2011.2134090.

[3] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 22, no. 10, pp. 1533–1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.

[4] S. Malekzadeh, M. H. Gholizadeh, and S. N. Razavi, "Persian Vowel recognition with MFCC and ANN on PCVC speech dataset," arXiv preprint arXiv:1812.06953, 2018.

[5] M. Namnabat and M. M. Homayounpour, "A Letter to Sound System for Farsi Language Using Neural Networks," in 2006 8th international Conference on Signal Processing, Nov. 2006, vol. 1, doi: 10.1109/ICOSP.2006.345518.

[6] S. Malekzadeh, "Phoneme-Based Persian Speech Recognition," arXiv:1901.04699 [cs, eess], Jan. 2019, doi: 10.13140/RG.2.2.32856.96007.

[7] H. Hasanabadi, A. Rowhanimanesh, H. T. Yazdi, and N. Sharif, "A Simple and Robust Persian Speech Recognition System and Its Application to Robotics," in 2008 International Conference on Advanced Computer Theory and Engineering, Dec. 2008, pp. 239–245, doi: 10.1109/ICACTE.2008.125.

[8] H. Sameti, H. Veisi, M. Bahrani, B. Babaali, and K. Hosseinzadeh, "Nevisa, a Persian Continuous Speech Recognition System," in Advances in Computer Science and Engineering, Berlin, Heidelberg, 2009, pp. 485–492, doi: 10.1007/978-3-540-89985-3_60.

[9] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello Edge: Keyword Spotting on Microcontrollers," ArXiv, 2017.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[11] D. Amodei et al., “Deep speech 2: end-to-end speech recognition in English and mandarin,” in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, New York, NY, USA, Jun. 2016, pp. 173–182, Accessed: Aug. 22, 2020. [Online].

[12] T. Sainath and C. Parada, “Convolutional neural networks for small-footprint keyword spotting,” 2015.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014.

[14] M. Dawodi, T. Wada, and J. Baktash, “An Intelligent Recommender System Supporting Decision-Making on Academic Major,” Information : an international Interdisciplinary journal, vol. 22, no. 3, pp. 241–254, May 2019.

[15] M. Dawodi, T. Wada, and J. A. Baktash, “Applicability of ICT, Data Mining and Machine Learning to Reduce Maternal Mortality and Morbidity: Case Study Afghanistan,” p. 13.

[16] M. Mohammadi, M. Dawodi, W. Tomohisa, and N. Ahmadi, “Comparative study of supervised learning algorithms for student performance prediction,” in 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Feb. 2019, pp. 124–127, doi: 10.1109/ICAIIC.2019.8669085.

[17] M. Dawodi, J. A. Baktash, and T. Wada, “Data-Mining Opportunities in E-Government: Agriculture Sector of Afghanistan,” in 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Oct. 2019, pp. 0477–0481, doi: 10.1109/IEMCON.2019.8936193.

[18] M. Dawodi, M. H. Hedayati, J. A. Baktash, and A. L. Erfan, “Facebook MySQL Performance vs MySQL Performance,” in 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Oct. 2019, pp. 0103–0109, doi: 10.1109/IEMCON.2019.8936259.