Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230064

Translation from spoken Arabic digits to sign language based on deep learning approach

Mothanna Almahmood[1], Sayel Abualigah[1], Rehab M. Duwairi[1], Laith Abualigah[2,3,4,5,6,9], Raed Abu Zitar[7], Anas Ratib Alsoud[3] and Sathishkumar V. E.[8]

[1]Computer Information System Department, Jordan University of Science and Technology, Jordan.
[2]Computer Science Department, Prince Hussein Bin Abdullah Faculty for Information Technology, Al al-Bayt University, Mafraq 25113, Jordan.
[3]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan.
[4]Faculty of Information Technology, Middle East University, Amman 11831, Jordan.
[5]Applied science research center, Applied science private university, Amman 11931, Jordan.
[6]School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang 11800, Malaysia.
[7]Sorbonne Center of Artificial Intelligence, Sorbonne University-Abu Dhabi, Abu Dhabi, United Arab Emirates
[8]Department of Software Engineering, Jeonbuk National University, Jeonju, Republic of Korea

[9]aligah.2020@gmail.com and sathish@jbnu.ac.kr

Abstract. Deaf-and-dumb humans make up about 5% of the world's population, and they need special care by providing them alternative methods that help them to communicate with the outside world, whereas the sense of hearing is the main element of human communications, which is indispensable. From the standpoint of introducing helpful applications that help deaf-and-dumb population, the idea of this research aimed used deep learning techniques to create a model based on the principle of converting Arabic spoken digits to sign language images, through a study of two different datasets that were freely taken from open-source websites. The first one contains audio records of Arabic spoken digits that was used to train on-dimensional CNN model to generate a text translation of any Arabic spoken digit record. The second one contains sign language images of Arabic digits, where used to build IF-THEN rules system that can generate the sign language image as a translation of given Arabic digit text. The whole idea conducted through using both systems in one prediction model that can generate the sign language image of any giving spoken Arabic digits' record, where it had accurate results with 86.85% accuracy value and 0.5039 loss value. The goal of this research is to add a new technology based on deep learning, in order to help this group of people with a simple idea that opens the researchers' minds to produce a model of all Arabic spoken speech, which in turn can be a complete technology that helps deaf-and-dumb humans' to easily communicate with the outside world.

DOI: 10.54254/2755-2721/30/20230064

Keywords: one-dimensional CNN, sign language, spoken Arabic digits; deep learning.

1. Introduction

These years, data science has proved its magic in several areas that fill human needs, and the most important of these areas is Speech Recognition, which has had many applications benefited the population all over the world. Arabic is one of the world's most spoken languages, while it is spoken by more than 467 million people globally. Although English is the first language in the world, Arabic is the most widely spoken language and ranks fourth among the languages spoken online [1]. Arabic has 28 letters, unlike other languages, it is the most abundant in terms of vocabulary and in comparison, to English, Ibn Manzoor's dictionary contains 80,000 Arabic words, while Samuel Johnson's dictionary contains 42,000 words in English [2]. Arabic language is divided into two parts, the first is the dialect language, and the second type is the standard modern language, which I have used in this research. Sign language is the official language of the deaf-and-dumb humans, and it's considered as a complex language, because each word can be represented by specific hand sign. A Literature review was carried out on the pathophysiology including genetics, clinical presentation, etiology, diagnosis and various management, using internet Google, search PubMed, and found that hearing loss is a common disorder and can be conductive, sensorineural or mixed types, also It can be congenital or acquired, where in pediatric population more than 50% of deafness is genetic in origin, and the patients may present as Deaf, mute or hard of hearing [3]. With the development of deep learning, the researchers have come up with many applications using different techniques that helped the community especially in the healthcare sector. The idea of this research came from spotting the light and the deaf-and-dumb humans' needs, since they need something to help them to communicate with the population, because they cannot make sure that people around them knows how to communicate by sign language. It is known that sign language is not a standard language, thus this research has come up the idea of making people communicate with the deaf-and-dumb humans' without need to use the sign language, specifically for the Arabic digits. It is worth to mention that deep learning contributed significantly to its development, and this work give a chance to implement a model that could be an efficient step of increasing the healthcare scores in the Arabian world. The remaining parts of the paper are structured as follows: section 2 shows relevant research papers that were collected, section 3 describes the applied research methodology, section 4 presents the experiments and results, and section 5 presents the conclusion of the research.

2. Related work

The field of deep learning is advancing so quickly that it is worth noticing that neural-based techniques that work today may be replaced by new ones in the near future. thus,computational power required was prohibitive in the past, and data available was not enough to train such complex systems, thus ideas were abandoned. In a study on designing an application to assist the Arab deaf community [4], the researchers have made a mobile application based on cloud computing, where they undertook complex manipulation of the Arabic text. The results showed that this application can help the deaf to communicate with anyone who was on the move, and without the need for multiple devices to do so. Also, in another study on designing an application V2M to assist the Arab Deaf community [5], the researchers have made a mobile application based on Cyberstalk frequency slope coefficients, in addition to using the hidden Markov model kit (HTK). They recognized the words of the deaf and converted it into a form that could be recognized by the deaf community. With the integration of the application with the 3D avatar, the results showed that this app can help the deaf to communicate with anyone. The research was not limited to this extent only, as the previously available sources regarding speech, texts, and sign language are widely available. In the study [6], the researchers have conducted a new system for translating the Arabic text to Arabic sign language based on Java. The systems work by concatenating the Arabic words to a sign from the Arabic sign language dictionary, which gives a video clip the reflects the right translation. Such language-oriented applications require a large amount of data, which in turn will help to give

DOI: 10.54254/2755-2721/30/20230064

powerful results in translation from one language to another. In the study [7], the researchers have conducted a system called V2S that translates the language through speech and image processing techniques, where the approach first used a speech pattern based on a set of spectral parameters. The ultimate goal was to store the set of spectral parameters in a database, where the system designed to perform the speech recognition process by matching the set of input parameters with the previously stored to display the sign language at the end in the form of a formatted video. In the study [8], the researchers have conducted the SSLIS system to translate the English speech into a sign language video in streaming mode, where the speech recognition model was proposed through using the Sphinx 3.5 for translation. The sign language syntax was not based on specific criteria, but the signed English manual was followed as a parallel to English language. The results showed that SSLIS system can help deaf people to fill the gap between deaf and non-deaf populations. Recently, the modern techniques of deep learning have become the most popular methods for spoken languages analysis, where it gives the ability to minimize error rate for optimization problems. In the study [9], the researchers have conducted Arabic digit's speech classification model through using the Recurrent Neural Network. Where the model takes the best speech signal representation by the feature extraction techniques of Mel-Frequency Cepstrum Coefficients. Also, the output features from the speech of digit are go into Long Short-Term Memory cells of the network. The results showed that the model has obtained 69% accuracy level in classifying the spoken Arabic digits.

3. The proposed method

In my study framework, two datasets were used in training the model, that consist of two parts; One-dimensional convolutional neural network to build spoken Arabic digits' classification model, and IF-THEN rules to concatenate the resulted text of the first part with its unique sign language image. Finally, the whole model Arabic Spoken Digit to Sign Language Image (ASD-SLI) accepts any record of an Arabic spoken digit and converts it to a sign language image. The diagram in Fig. 1 shows the framework of this research.

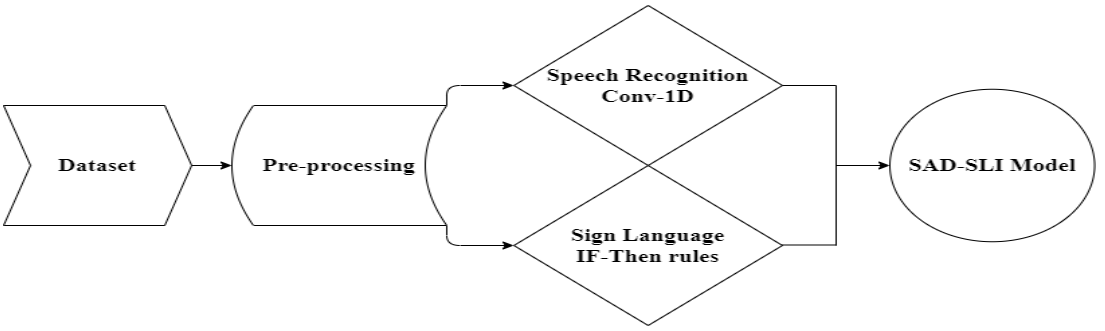


Figure 1. The steps of the research methodology.

3.1. Dataset

Two different datasets are experimented with in this study. The first one is "The Arabic Speech Corpus for Isolated Words"[10] taken from University of Stirling, which is a free source complete dataset that contains 9992 utterances of 20 words spoken by 50 native male Arabic speakers. It has been recorded with a 44100 Hz sampling rate and 16-bit resolution. Since the interest of this work is the Arabic spoken digits, the dataset has been filtered out, in which only the needed records have been taken, where the final form is 5000 utterances of 10 spoken digits by 50 native male Arabic speakers. This provides enough samples to train a CNN with higher accuracy and with less computational resources and time. A 70/30 data split is used for training and validating the model. The second dataset was taken from open source website [11].

DOI: 10.54254/2755-2721/30/20230064

3.2. Pre-processing

Python environments provide us with a useful library called "librosa" [12], that helps in importing a wav sounds dataset, in which it can be used for analytics later on. The sampling rate of the audio signals has a direct impact on the dimensionality of the input sample and eventually on the computational cost of model [13]. The sampling rate of 16000 Hz may be considered as the best trade-off between the quality of the input records and the computational cost of the used approach [14]. The sampling rate of the used dataset is 44100 Hz, which looks too high, hence it is going to be resampled to 16000 Hz. The aim of this model is to classify the recorded audio according to the classes of each Arabic digit, thus the Encoder labeling technique [15] is going to be used to introduce the needed 10 classes. Each image of the second dataset is going to be labeled, according to the Arabic digit that the image indicates to.

3.3. SAD-SLI model

One-dimensional convolutional neural network is corresponding to regular neural network, but it is characterized by its raw input data, which keeps out from the need of using the manual features [16]. The convolutional layers learn the best representation of the input data, in which it become able to be processed later. According to the theorem of local connectivity [17], the neurons in a layer are connected to a small sector of the last layer, where this connectivity is called the responsive field. The inputs of the one-dimensional convolutional neural network are arrays that reflects the waveform of the recorded audios, where the network designed to learn some parameters, to link the inputs to the classification based on hierarchical feature extraction. Pooling layers are conducted between the convolutional layers, to rise the area that is covered by the next responsive fields. Then, the output of the last convolutional layer is flattened, and used as an input of many stacked fully connected layers. The proposed methodology of this work aims to be a compact one-dimensional convolutional neural network architecture with a reduced number of parameters. The number of parameters of a one-dimensional convolutional neural network is straight connected to the computational strive to train a network as well as to the need of a huge amount of data in the training process.

Table 1. The parameters of one-dimensional CNN model.

Parameter	Description	Model parameters
Batch size	Number of datasets given to the network per iteration.	32
Epochs	Number of times the dataset is being complete to the network.	100
Loss	Calculated value after each iteration to define the error, which is calculated by the loss function.	Categorical cross-entropy
Optimizer	A technique that reduces the output error of the loss function.	Adam
Activation Function	A function used to choose whether the neuron should fire the data or not.	Softmax
Learning rate	A factor that changes the weights of the function, in order to end up with high accuracy.	0.001
Patience	Number of epochs to wait before early stop if no progress on the validation set.	10

The parameters in Table 1 are going to be used to tune the training in the model. The early stopping technique [18] is going to be used to stop the training process after ten epochs without decreasing the loss value and save the minimum loss value epoch. Also, checkpoint technique [19] is going to be used to save the best model based on the maximum accuracy value saved by the early stopping. Finally, the model is going to be compiled and fitted with the same train and test dataset. Next step is going to be concatenate each label of the SAD-SLI model outputs with its appropriate label of sign language images,

DOI: 10.54254/2755-2721/30/20230064

though using IF-THEN rules, in which the final result achieves the goal of this work to translate spoken Arabic digits to sign language images.

4. Experimental results

This research project was carried out using Dell computer with 3.20GHz CPU Intel Core (TM) i7-8700, and memory of 16GB. The operating system is windows 10, and the software that has been used is Python 3.7. The model has been fitted with the training and testing sets, and Fig. 2 shows the obtained results.

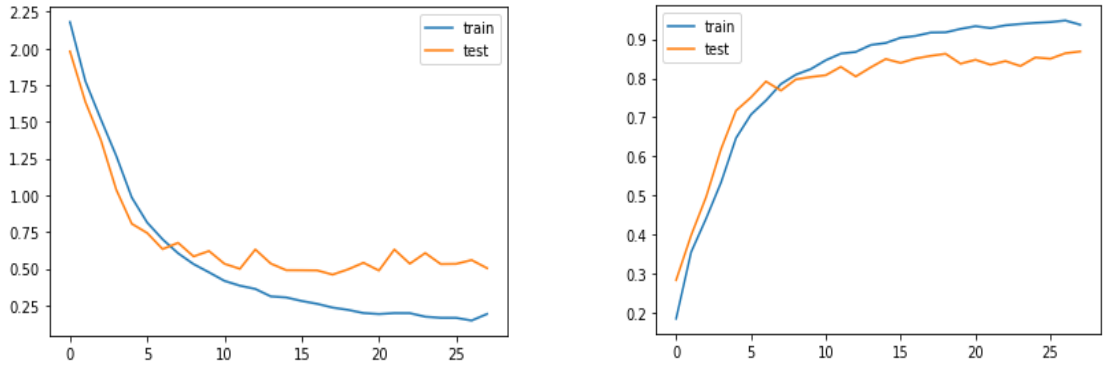
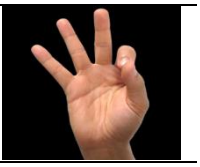


Figure 2. The results of accuracy and loss for train and test in the one-dimensional CNN model.

The saved model was early stopped at the 28th Epoch, since that the accuracy value was improved from 0.8638 to 0.8685 rate. A high level of accuracy of 78.47% was obtained with a 0.5039 value of loss, where the model was trained concurrently on the training dataset. The results showed that the model performance was enhanced due to the given dataset, and the model was able to improve the translation performance on the training dataset of SAD-SLI model.

Table 2. The results of performing samples in the SAD-SLI model.

Num. Spoken Digit recorded voice	1 Khamsa	2 Settah	3 Saba'ah	4 Thmanyah	5 Tesa'ah
Sign Language Image					

After saving the model, five test experiments for the spoken Arabic digits; khamsa, settah, saba'ah, thmanyah and tesa'ah were conducted, in order to check the efficiency of the saved model. As indicated in Table 2, the output images that were concatenated by IF-THEN rules were awesome. Thus, the SAD-SLI model can be considered as an efficient technology that can be used to translate the spoken Arabic digits to sign language.

5. Conclusion

Deaf-and-dumb humans need special care by providing them alternative methods that help them to communicate with the outside world. The applied one-dimensional CNN model with IF-THEN rules of concatenations on a 5000 utterances dataset of 10 digits spoken by 50 native male Arabic speakers, and 10 images of sign language digits, has been improved its results of 86.85% accuracy level and 0.5039

DOI: 10.54254/2755-2721/30/20230064

loss level. These results have reached the goal of the research; thus, it has been considered that SAD-SLI model was perfectly efficient in translating the spoken Arabic digits to sign language. The modern technologies of deep learning have opened the way to build different models, by discovering new mathematical equations that come with a high level of accuracy and performance. Also, it is worth mentioning that the healthcare sector had many applications conducted by deep learning, which benefited the people with their several situations. The high accurate SAD-SLI model in this work is a new efficient method that can help the deaf-and-dumb humans, where they can communicate with the community without any issues and exempt other people from using sign language to communicate with them. In the future work, we are going to use larger datasets that contain huge amount of spoken Arabic digits and alphabets, with their sign language images, in order to develop a model that can translate both Arabic digits and Arabic alphabets to sign language.

References

[1] T. Tinsley and K. Board, “Languages for The Future,” Br. Counc., 2013.
[2] “Arabic - Wikipedia” [Online]. Available: https://en.wikipedia.org/wiki/Arabic/ [Accessed: 12-May-2020].
[3] W. Tin, Z. Lin, - Swe, and N. K. Mya, “Deaf mute or Deaf,” Asian J. Med. Biol. Res., vol. 3, no. 1, pp. 10–19, 2017.
[4] M. M. El-Gayyar, A. S. Ibrahim, and M. E. Wahed, “Translation from Arabic speech to Arabic Sign Language based on cloud computing,” Egypt. Informatics J., vol. 17, no. 3, pp. 295–303, 2016.
[5] K. Yousaf et al., “A Novel Technique for Speech Recognition and Visualization Based Mobile Application to Support Two-Way Communication between Deaf-Mute and Normal Peoples,” Wirel. Commun. Mob. Comput., vol. 2018, pp. 1–12, 2018.
[6] L., H., & M., S., A., Automatic translation of Arabic text-to-Arabic sign language. Universal Access in the Information Society, 2019, 18(4), 939–951.
[7] D., P., Hermawan Nugroho Centre for Intelligent Signal and Imaging Research Universiti Teknologi PETRONAS, Bandar Sri Iskandar Perak. 2015, 1–5.
[8] K. El-darymli, O. O. Khalifa, H. Enemosah, K. Lumpur, and K. K. El-darymli, “The citation of this paper is as follows: Khalid El-Darymli , Othman O . Khalifa, and Hassan Enemosah , " Speech to Sign Language Speech to Sign Language Interpreter System (SSLIS),” no. May, 2006.
[9] A. S. Mahfoudh Ba Wazir and J. Huang Chuah, “Spoken Arabic Digits Recognition Using Deep Learning,” 2019 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2019 - Proc., no. June, pp. 339–344, 2019.
[10] “The Arabic Speech Corpus for Isolated Words” [Online]. Available: http://www.cs.stir.ac.uk/~lss/arabic/ [Accessed: 12-May-2020].
[11] “Numbers in sign language” [Online]. Available: http://easyenglishforme.blogspot.com/2013/03/numbers-in-sign-language.html. [Accessed: 12-May-2020].
[12] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” Proc. 14th Python Sci. Conf., no. Scipy, pp. 18–24, 2015.
[13] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, “End-to-end environmental sound classification using a 1D convolutional neural network,” Expert Syst. Appl., vol. 136, no. September, pp. 252–263, 2019.
[14] A. Kipnis, A. J. Goldsmith, and Y. C. Eldar, “Optimal trade-off between sampling rate and quantization precision in Sigma-Delta A/D conversion,” 2015 Int. Conf. Sampl. Theory Appl. SampTA 2015, pp. 627–631, 2015.
[15] “Label Encoder” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html [Accessed: 12-May-2020].
[16] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D Convolutional Neural Networks and Applications: A Survey,” pp. 1–20, 2019.

DOI: 10.54254/2755-2721/30/20230064

[17] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An Introductory Review of Deep Learning for Prediction Models with Big Data," Front. Artif. Intell., vol. 3, no. February, pp. 1–23, 2020.
[18] G. Montavon et al., "Neural Networks: Tricks of the Trade," Springer Lect. Notes Comput. Sci., no. MAY 2000, p. 432, 2012.
[19] H. Chen, S. Lundberg, and S.-I. Lee, "Checkpoint Ensembles: Ensemble Methods from a Single Training Process," no. October 2017, 2017.