EasyChair Preprint

№ 11572

Enhancing Education Decision-Making with Deep Learning for Arabic Spoken Digit Recognition

Zineb Touati-Hamad and Mohamed Ridda Laouar

December 19, 2023

Enhancing Education Decision-Making with Deep Learning for Arabic Spoken Digit Recognition

Zineb TOUATI-HAMAD
Laboratory of mathematics, informatics and systems (LAMIS)
Echahid Cheikh Larbi Tebessi University
Tebessa, 12002, Algeria
zineb.touatihamad@univ-tebessa.dz

Mohamed Ridda LAOUAR
Laboratory of mathematics, informatics and systems (LAMIS)
Echahid Cheikh Larbi Tebessi University
Tebessa, 12002, Algeria
ridda.laouar@univ-tebessa.dz

***Abstract*—In the realm of education, it becomes imperative to gauge students' learning progress, enabling well-informed decisions and effective support. Recent years have witnessed the ascent of deep learning as a potent instrument for speech recognition. It provides a more exact and efficient examination of people receiving speech treatment. This research offers a deep learning model centered on convolutional neural networks (CNN) suited for the categorization of Arabic spoken digits spanning from 0 to 9. Our model is rigorously trained on a broad dataset including recordings of spoken Arabic numbers, covering authentic Arabic speakers of various ages and ability levels. The findings speak much about the capabilities of our CNN-based algorithm. It attains an outstanding accuracy rate in identifying and classifying Arabic spoken digits, claiming an overall accuracy of 96.10%. Furthermore, we dive into the larger ramifications of our results within the educational environment. This emphasizes the potential of our strategy to better the evaluation of adult learners' speech therapy and to create more effective support measures. This adaptable methodology finds relevance across many educational environments, making voice recognition technology in speech therapy for adult learners more accessible and productive.**

Keywords— Deep Learning, Convolutional Neural Networks, Speech Recognition, Adult Learning, Speech Therapy, Arabic Spoken Digit, Decision Making.

I. INTRODUCTION

In recent years, deep learning has emerged as a promising technology for speech recognition and has shown significant success in various languages [1]. However, the Arabic language presents unique challenges due to its complex phonological system, where subtle differences in pronunciation can alter the meaning of words [2]. Therefore, the application of deep learning in Arabic speech recognition is of particular interest in the field of education, where accurate assessment of language proficiency is essential for effective decision making and support for students.

The use of ASR technology in adult education has gained significant attention in recent years due to its potential to assess and support learners' progress in speech therapy.

However, the current ASR models based on neural networks, particularly MLPs, have limitations related to their learning style [3]. MLPs consider each neuron as independent and assign a different weight to each incoming signal, which can result in suboptimal recognition rates. CNNs, on the other hand, are designed to handle pattern recognition tasks and have demonstrated promising results in speech recognition.

Moreover, the proposed CNN-based model aims to recognize Arabic speech numbers from 0 to 9, which is a significant challenge due to the complex nature of the Arabic language. To address this challenge, we use the UCI Machine Learning Repository's Spoken Arabic Digit Dataset to train and evaluate our model. This dataset includes recordings of Arabic speech numbers pronounced by different speakers, which provides a diverse set of data for training and testing the proposed model.

The use of CNNs in ASR has shown remarkable results in recent studies, where it has achieved state-of-the-art performance in various speech recognition tasks [4]. The proposed model can contribute to improving the performance of ASR technology in adult learning by providing accurate assessments of learners' progress in Arabic spoken digit recognition as a part of their speech therapy. Ultimately, this can lead to more effective decision making and support for adult learners, enabling them to achieve their speech therapy goals.

The structure of this paper comprises several sections. The second section presents the background of the study. The third section outlines the research motivation. The fourth section presents the related works. The fifth section describes the methodology employed in the study. The sixth section details the experiments. Finally, the paper concludes with a conclusion and suggests avenues for future research.

II. BACKGROUND

A. Speech recognition

Speech recognition is a rapidly growing field in computer science, with a wide range of applications in various domains, including education. Automatic speech recognition (ASR) technology has the potential to provide valuable insights into student learning and support their progress in language learning [5]. In the context of education, ASR technology can be used to assess students' speaking and listening skills, providing feedback on pronunciation and identifying areas for improvement.

One of the main challenges in ASR is the variability of speech signals, which can be affected by factors such as speaker accent, background noise, and speaking rate. To address these challenges, deep learning techniques have shown promising results in speech recognition tasks. Convolutional neural networks (CNNs) have gained popularity in recent years, as they have been shown to be effective in handling complex patterns and achieving high accuracy in speech recognition tasks.

B. Decision support systems

Decision support systems (DSS) are computer-based tools that help people make better decisions [6]. They use data, models, and algorithms to analyze complex problems and provide recommendations to decision-makers. DSS can be used in a variety of fields, including business, healthcare, and government. They are especially useful for tackling problems that are too complex for humans to solve alone.

DSS have been around since the 1960s, but they have become much more powerful and sophisticated in recent years, thanks to advances in machine learning and big data analytics.

C. Deep learning

Deep learning (DL) is a subfield of machine learning that involves the use of neural networks with many layers [7]. It is inspired by the structure and function of the human brain and has been used to achieve breakthroughs in a wide range of applications, from image recognition to natural language processing. Convolutional neural networks (CNNs) are a type of deep learning architecture that has been particularly successful in image recognition tasks [8]. They work by using multiple layers of filters to detect features in images, and they have been used to achieve state-of-the-art performance on many benchmark datasets.

CNNs have become a popular choice for various pattern recognition tasks. CNNs can extract features from speech signals, such as mel-frequency cepstral coefficients (MFCCs), which can be used to train a model for speech recognition [9]. Moreover, CNNs have shown impressive results in speech recognition tasks, demonstrating their potential to improve the accuracy of ASR systems.

In the context of the Arabic language, there has been a growing interest in ASR technology, particularly in the field of Arabic speech recognition. However, the complex nature of the Arabic language, which includes various dialects and phonetic variations, poses a challenge for ASR systems. Therefore, the use of deep learning techniques, particularly CNNs, can contribute to improving the accuracy of ASR systems for Arabic speech recognition.

III. Research Motivation

The Arabic language is one of the most widely spoken languages in the world. It is a language of great cultural significance. However, Arabic is a morphologically rich and highly ambiguous language, with many colloquial dialects that differ from Modern Standard Arabic (MSA), used in formal communication [10]. This linguistic complexity presents significant challenges for Arabic Automatic Speech Recognition (AASR) systems, particularly for adult learners undergoing speech therapy. Recognizing spoken words accurately is critical for effective communication and treatment outcomes. However, the complexity of the language can pose significant challenges that impede the progress in therapy and hinder the development of speech and language skills in adult learners.

To address this challenge, we propose using Convolutional Neural Networks (CNNs) for Arabic speech number recognition, which has shown impressive performance in ASR applications. The use of CNNs is particularly effective at handling individual variations in the speech signal and improving the speaker invariance of the acoustic model [10].

Therefore, the main motivation for this research is to develop an accurate and reliable ASR system for Arabic spoken digit recognition that can be integrated into a decision support system to aid adult learners' speech therapy processes in the Arabic language. This will help to bridge the gap in the current Arabic education system, which lacks effective and accessible tools for adults to improve their speech and language skills.

IV. Related Works

Several research studies have been conducted using deep learning models for speech recognition in various languages, including Arabic. These studies have shown that deep learning models, such as Convolutional Neural Networks (CNNs), can significantly improve speech recognition accuracy compared to traditional methods. For example, in [10] the authors proposed a robust method for Arabic speech recognition using deep CNNs. Similarly, in [11] the authors provided an overview of deep learning-based Arabic speech recognition and highlighted the effectiveness of using CNNs in this field.

In the context of education, decision support systems have been developed to support student learning in various areas, such as academic performance prediction, student engagement, and personalized learning. The authors of [12] developed an intelligent decision support system to predict students' academic performance in Jordanian public universities.

Our contribution to this field is the development of a decision support system that focuses on adult learners' speech therapy progress in Arabic spoken digit recognition. The system employs deep learning models, particularly CNNs, for speech recognition and classification. The system's goal is to aid adult learners in improving their speech therapy progress by providing real-time feedback on their pronunciation of Arabic spoken digits from 0 to 9. The system is designed to be interactive and engaging, with visual and auditory feedback that motivates adults to practice and enhance their skills.

V. Proposed Model

In this study, a convolution neural network (CNN) approach was implemented and the architecture of the network was modified and improved by experimenting with various combinations of parameters (see Fig.1) .

The use of CNNs for processing spectrograms represented as 2D images was explored. Discriminatory features were extracted from the MFCC frequency in the selected dataset by applying a medium-sized CNN with varying numbers of convolutional layers and filters of different sizes. The number of convolutional layers was selected based on the complexity of the data, and after two or three layers, the gain in accuracy became stable while learning took a long time. The CNN operation with several filters was used to obtain a feature map, which was then passed through the activation function ReLU. The Max-pooling operation was used to obtain more significant features by selecting the maximum value of the other values of the map. This process was repeated with another convolutional layer and Max-pooling operation to obtain new features, which were then passed to two fully connected layers. A dropout layer was included to prevent overfitting, and the last layer contained the Softmax function, which provided the probability distribution on each class. The Adagrad stochastic gradient algorithm was used to optimize the network, and the categorical cross-entropy loss function was used since the study involved a multi-class classification problem (see Fig. 2) .

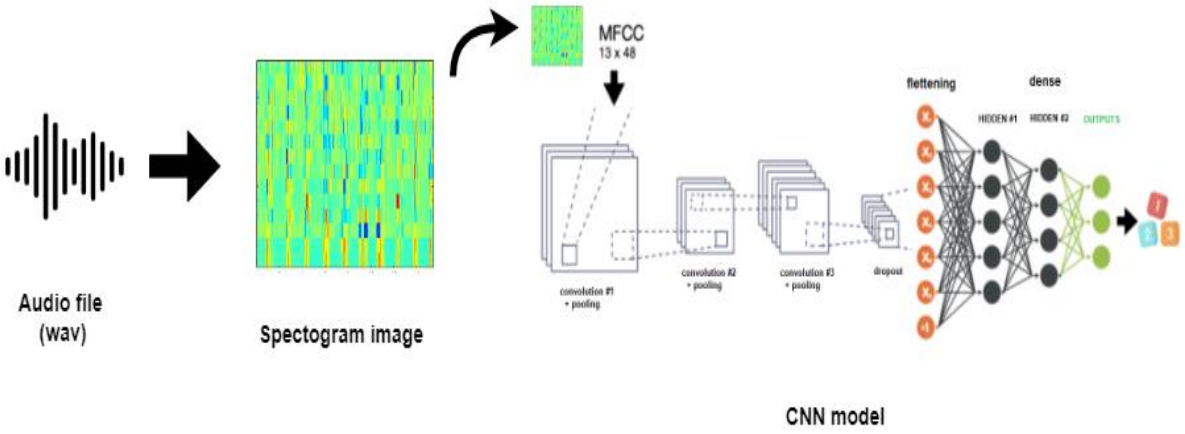

Fig. 1. Proposed architecture

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 256, 256, 16)	448
max_pooling2d_12 (MaxPooling2D)	(None, 128, 128, 16)	0
conv2d_13 (Conv2D)	(None, 128, 128, 32)	4640
max_pooling2d_13 (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_14 (Conv2D)	(None, 64, 64, 64)	18496
max_pooling2d_14 (MaxPooling2D)	(None, 32, 32, 64)	0
dropout_13 (Dropout)	(None, 32, 32, 64)	0
flatten_6 (Flatten)	(None, 65536)	0
dense_13 (Dense)	(None, 256)	16777472
dropout_14 (Dropout)	(None, 256)	0
dense_14 (Dense)	(None, 10)	2570

Total params: 16,803,626
Trainable params: 16,803,626
Non-trainable params: 0

Fig. 2. The CNN model

VI. EXPERIMENTS

A. Dataset

The Spoken Arabic Digit dataset is a collection of mel-frequency cepstrum coefficients (MFCCs) corresponding to spoken Arabic digits from 0 to 9.

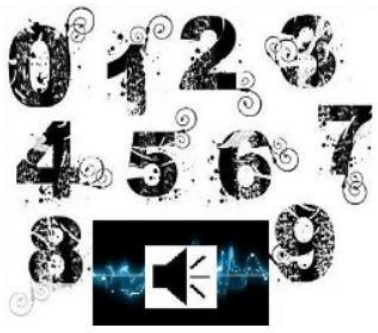

Fig. 3. SpokenArabicDigits Dataset Logo [13]

It includes recordings from 88 native Arabic speakers, consisting of 44 males and 44 females, between the ages of 18 and 40. Each speaker was asked to pronounce each digit ten times, resulting in a total of 880 recordings for each digit. The dataset provides a valuable set of data, with 75% allocated for training and 25% for testing, for speech recognition models, particularly for the recognition of Arabic numbers. The dataset is publicly available in the UCI Machine Learning Repository [13].

TABLE I. DATASET DESCRIPTION [13]

Data Set Characteristics:	Multivariate, Time-Series
Attribute Characteristics:	Real
Associated Tasks:	Classification
Number of Instances:	8800
Number of Attributes:	13
Missing Values?	No

B. Results and discussion

In this study, the proposed model was trained and evaluated using a 75% training dataset and a 25% test dataset, as previously mentioned. A series of tests were conducted to identify the optimal hyper-parameters for the model. These parameters, which include the network structure (such as the number of neurons and layers, activation functions), batch size, and number of iterations, play a significant role in the performance of the model during training. Once a suitable model was identified with minimum error rate and maximum accuracy, it was then tested on the independent test subset. The experiment involved 10-class classification on a dataset of 6600 training spectrograms and 2200 test spectrograms. The model was trained for a number of epochs ranging from 10 to 100. The results of the experiment are presented in the Fig.3.

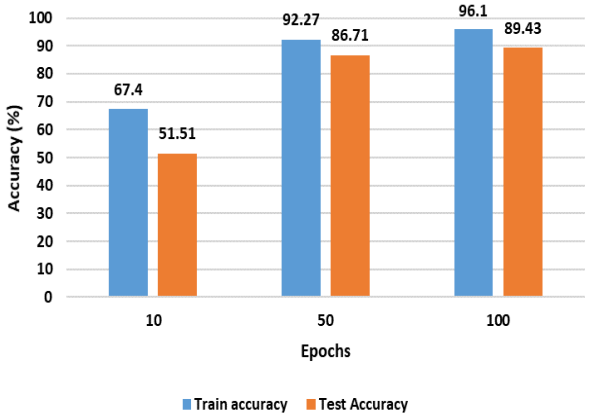

Fig. 4. Accuracy and loss results for different epochs

The results of the experiment indicate that the model begins to learn the classes from the first 10 epochs, and its accuracy improves as the number of epochs increases until it reaches an acceptable value of 96.10% for the learning dataset and 89.43% for the testing dataset. It is observed that the model begins to stabilize after reaching a certain threshold of epochs, and the increase in the number of epochs is not as significant as at the beginning. However, it is worth noting that while increasing the number of epochs improves the classification of Arabic digit, the execution time also increases with each iteration.

VII. Conclusion And Future Works

The ability of a computer to recognize human speech or voices poses a new challenge for modern scientific research, particularly as communication between humans and electronic devices has increased significantly. Consequently, researchers are striving to develop intelligent software capable of speech recognition without any human intervention.

In this study, we designed a system for automatic recognition of spoken Arabic digit using a CNN model to aid decision-making processes regarding adult learners' speech therapy progress.

We provided two subsets of data, one for learning and the other for testing, containing Arabic digit from 0 to 9 given by MFCC coefficients. The MFCC coefficient parameterization is widely used in this field to extract features and has also produced good results. Every recognition model undergoes a classification step, and we chose convolutional neural networks (CNN) in this study, which yielded satisfactory results after recognizing completely unknown speech samples. Therefore, CNNs proved to be a better technique for learning and identifying new data than other neural networks. Our research work serves as a starting point for launching future projects, such as continuous speech recognition, and also proposes ideas for improving our model, such as testing it on larger databases, adding additional training data to a database, and incorporating different languages.

Acknowledgment

The authors wish to acknowledge the support received from the Algerian General Directorate of Research (DGRSTD) and the Laboratory of Mathematics, Informatics and Systems (LAMIS) at the University of Larbi Tebessi in conducting this research

References

[1] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8599-8603.

[2] Z. Touati-Hamad, M. R. Laouar, I. Bendib, and S. Hakak, "Arabic Quran Verses Authentication Using Deep Learning and Word Embeddings," in International Arab Journal of Information Technology, vol. 19, no. 4, 2022, pp. 681-688.

[3] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4295-4299.

[4] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, 2014, pp. 1533-1545.

[5] J. Van Doremalen, L. Boves, J. Colpaert, C. Cucchiarini, and H. Strik, "Evaluating automatic speech recognition-based language learning systems: A case study," in Computer Assisted Language Learning, vol. 29, no. 4, 2016, pp. 833-851.

[6] S. Eom and E. Kim, "A survey of decision support system applications (1995–2001)," in Journal of the Operational Research Society, 2006, vol. 57, pp. 1264-1278.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," in Nature, vol. 521, no. 7553, 2015, pp. 436-444.

[8] J. Naranjo-Torres et al., "A review of convolutional neural network applied to fruit image processing," in Applied Sciences, vol. 10, no. 10, 2020, p. 3443.

[9] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," in IEEE Transactions on Information Forensics and Security, 2019, vol. 15, pp. 1616-1629.

[10] E. R. Abdelmaksoud, A. Hassen, N. Hassan, and M. Hesham, "Convolutional Neural Network for Arabic Speech Recognition," in The Egyptian Journal of Language Engineering, vol. 8, no. 1, 2021, pp. 27-38.

[11] R. Amari et al., "Deep Convolutional Neural Network for Arabic Speech Recognition," in International Conference on Computational Collective Intelligence, September 2022, pp. 120-134.

[12] Y. S. Alsalman, N. K. Abu Halemah, E. S. AlNagi, and W. Salameh, "Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance," in 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2019, pp. 104-109.

[13] M. Bedda and N. Hammami, "Spoken Arabic Digit," UCI Machine Learning Repository, 2010. [Online]. Available: https://doi.org/10.24432/C52C9Q.