

Towards developing impairments arabic speech dataset using deep learning

Sura Ramzi Shareef, Yusra Faisal Al-Irhayim

Department of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq

Article Info

Article history:

Received Sep 10, 2021

Revised Jan 11, 2022

Accepted Jan 19, 2022

Keywords:

Arabic speech classification

long short-term memory

Impairment's children

Mel-frequency cepstral-coefficients

Speech sound error

ABSTRACT

The effective and efficient recognition of speech sounds errors for impaired children is important if a defective phonological process is early detected and corrected. This study deals with the topic of classification of speech sound errors in Arabic impairments children when Arabic letters and numbers are incorrectly pronounced. For 18 standard Arabic isolated numerals and characters, we created an impaired children speech recognition system. We utilized the Mel frequency cepstral coefficients throughout the feature extraction step. then deep long short-term memory network recognition phase. We used the developed model with the developed dataset and the classification accuracy was 97.99% and lose 0.18%, additionally, the results have been compared and yielded extremely intriguing results with previously existing recognition rates models.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sura Ramzi Shareef

Department of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul
Mosul, Iraq

Email: sura.ramzishareef@uomosul.edu.iq

1. INTRODUCTION

Deep learning has recently developed greatly in several fields of machine learning such as language recognition, natural language processing and machine learning [1]. For more than 22 countries, Arabic is the official language. Arabic is one of over 313 million worldwide speakers' most common native language [2], [3]. it is written and consists of 28 letters, From the right to the left. The letter shapes are altered inside a word depending on its position. Any Arabic word might have several connotations. Languages are communications systems that allow speakers to more efficiently employ well-learned words. The features of a voice sound are based on human speech. Naturally speaking, the audience may get a great amount of information within a few minutes. Speech is the major communication technique between people, it is necessary to comprehend, learn to read or write. Speech technology now enables robots to react quickly and correctly by using the voices of people instead of keyboards [2], [4].

The process of making voice sounds is articulation. The motion of mouth and joints (lips, tongue, teeth, mandible, soft palate, hard palate and the back of the mouth's roof) makes spoken sounds. Each articulator configuration produces various sounds. Initially, airflow is generated through the larynx and respiratory system. The sounds can vibrate or not depending on whether the sound generated is voiced (e.g., I vowel) or not (e.g., [s] fricative consonant). The way the articulators move for a particular sound is varied from speaker to speaker and for each phoneme, in a language, the pronunciation results in distinct pronunciation.

No standard way to pronounce a phoneme properly in a particular language is available. However, human beings have to distinguish properly sounds, syllables and words. Due to several physiological variables, the right pronunciation might result in numerous joint issues. Articulation disorders consist of

sounds, syllables and phrases that make it hard to grasp what is being said or need listeners to concentrate attentively on how words sound more than meaning. In fast sound creation in regular speech, articulation problems may become more evident. Joint issues develop for various causes from physical impairment, such as neurological issues, cleft palate or hearing loss, to other oral difficulties, such as dental or language-related challenges. Disorders of speech may vary from a moderate to almost incomprehensible language. Although joint issues impact both children and adults, young children typically face these difficulties when they learning their language [5].

In 1950, the technique of speech recognition began. It was then enhanced by the digital identification system in 1952 to isolate fundamental words. Speech recognition technology allows people to isolate words via a machine [6]. Speech recognition recognizes specified words according to the feature effectiveness of the voice signals obtained. The aim is to identify the variety of features in the words utilized. "The automatic voice recognition system has been able to analyze individual words to conduct pattern recognition templates for the last sixty and seventy years [7]. Automatic speech recognition (ASR) is employed to recognize the words used to authenticate the identification of an individual. Moreover, utilizes computer hardware and software to recognize and process the speech of people. In many sectors of our life, ASR is frequently used in communication, education, healthcare, and protection [7], [8].

In 1980, the neural artificial networks of deep neural networks transformed the direction in which speech was recognized. Then deep learning is a newer area in the machine learning field such as handwriting, speech recognition language processing and computer vision. Therefore the field of machine learning (ML), a deep neural network tries to learn from multiple layers concurrently by extracting certain features and information [9], [10]. A high-quality dataset of voice disorders can help address rising speech problems in and outside the Arab area. In recent years, the number of individuals with voice disorder has grown considerably, with around 17.9 million people in the United States alone experiencing vocal difficulties [11]. 15% of the total King Abdulaziz University Hospital visitors in Saudi Arabia have been reportedly complaining of voice problems [12]. The problems created by voice problems in a teacher are much higher than in non-teachers and studies have shown that voice abnormalities are 57.7% in the United States and 28.8% in non-teachers throughout a lifetime [13], [14]. Roughly 33% of instructors in the Riyadh region of Saudi Arabia suffer from voice problems [12]. spasmodic dysphonia is nevertheless a vocal issue due to unintentional movements of the laryngeal muscles. We have a large number of voice disorders (about 760 cases annually) among people of diverse occupational and etiological backgrounds at our voice centre at the Communicatory and Swallowing Disorders unit of King Abdulaziz university hospital.

The reason for this research is that an exact and computationally effective method must be developed to classify speech sound mistakes in impairments children, to enable speakers and language pathologists to diagnose speech sound disorder (SSDs). In this research, we emphasise speech voice errors among Arabian children who are impaired to speak Arabic. We developed an Arabic speech dataset from the impairment children which have been collected from 38 children and collected 380 samples. Whereas, the developed dataset was including Arabic pronounced letters and numbers. Correspondingly, we identify the articulation disorder categories from a given speech segment by employing a deep learning technique based on long short-term memory to classify the pronounced numbers and letters. Specifically, the model will be able to recognize incorrectly pronounced letters or numbers.

2. METHOD

The proposed system for impairment arabic speech recognition would generally consists of three main stages are data collection stage, feature extraction stage, and recognition/classification stage as shown in Figure 1. Whereas, each phase will be explained in the followed sub section. Moreover, we have used python in order to build the system.



Figure 1. Methodological flowchart

2.1. Phase 1: data collection

To contribute to the field of automatic detection and classification of speech disorders for children speaking Arabically we produced children's Arabic language error data set for pupils who are disabled. The

speech records have been collected from impairments students from schools for children with special needs in Mosul-Iraq. The dataset contains 770 speech trials from 38 scholars while the student's age rate is between 7 to 11 years old. The specialist advised the youngsters to pronounce three letters and numerals carefully. For each letter and each number, each kid was instructed to repeat the pronunciation 10 times. The number used in the dataset were from 0 to 9 and letters are "Aleef", "Baa", "Taa", "Thaa", "Geem", "Haa", "Khaa", and "Daal".

2.2. Phase 2: feature extraction

We utilized MFCC to represent the speech signals as one of the most common acoustic characteristics [15]. Five stages are included in the feature extraction process: enhancement of speech signal, signal slicing, windowing, fast fourier transformation (FFT), and mel spaced filter bank coefficient calculation. We begin by improving the voice signal by reducing the low frequency in the speech signals. A high pass filter of format $H(z) = 1 - 0,97z^{-1}$, whereas z represents the speech signal, passes the speech signal. The improved voice signal is then divided into frames in 30 milliseconds with 10 milliseconds of overlap. Frame size is the best effect for speech-recognition applications between 20-40 milliseconds with a $50 \pm \%$ gap between consecutive frames [16].

This technique permits FFT to be applied over small time frames, so that the frequency contour of the speech stream may be well approximated. We use the Hamming window function $h(\hat{z}) = 0.54 - 0.46 \cos \frac{2\pi\hat{z}}{n-1}$ in each framework, where z is a frame, and n is the window length in the FFT, to decrease spectral leakage effects. In each window frame, we then execute a 512-point FFT to calculate its frequency and power range. We calculate the coefficients of the filter bank from each frame's power spectrum. then we utilized a set of 40 triangular filters to determine the bank filter coefficients. We increase the power range of each filter to 40 bank energies and combine the outputs. We use discrete cosine transformation (DCT) on them after log each of the 40 filter bank energies. 40 cepstral coefficients may be produced using the DCT method. Mel-frequency cepstral-coefficients (MFCC) features selected the first 13 cepstral coefficients. We concatenate the 13 sets of MFCC characteristics extracted from all speech signals in a single vector to represent the features of a speech signal. However, as voice signals include a variety of frames, feature vectors of the same size for speech signals are problematic for obtaining. To address the challenge, for each of the speech signals we calculate a fixed-size super vector [17]. Using the MFCC features, we derive a GMM and then stack the means to produce a super vector of 25 sizes. The supervisor is used to train the proposed classification model.

2.3. Phase 3: recognition/classification long short-term memory

Generally, it might be quite difficult to train such a conventional RNN network since mistakes calculated in the back-propagation method are compounded by each other at each time, resulting in gradient issues. Thus, the gradients readily go away or explode by repeated multiplications throughout the training [18]. To address this problem an alternate architecture is developed based on long-term memory (LSTM) [19]. A memory block consisting of a memory cell ct with three sigmoid portals forms the architecture of LSTM.

$$f(a) \triangleq \sigma(a) = \frac{1}{1 + \exp(-a)} \quad (1)$$

Includes it as input gate, a gate to as output, and gate forget as f_t , as shown in Figure 2. The activation functions of $g(\cdot)$ and $h(\cdot)$ are $\tanh(\cdot)$. The input x_t and output y_t LSTM memory block is implanted and could be vectored as:

$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \\ f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) \\ g_t &= \tanh(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) \\ m_t &= o_t \odot \tanh(c_t) \\ y_t &= s(W_{ym}m_t + b_y) \end{aligned} \quad (2)$$

where \odot represents the product of an element, the weight matrices of W_{ix} , W_{fx} , W_{cx} , and W_{ox} , from the input-to-input gate, forget gate, cell and output gate, correspondingly. Thus, the diagonal weight matrices W_{ic} , W_{fc} , and W_{oc} for peepholes are the cell-output c_{t-1} diagonal connections that are represented in dashed

lines. In the previous step, m_{t-1} weight matrices for the memory block output vector are shown in red lines by W_{im} , W_{fm} , W_{cm} , and W_{om} .

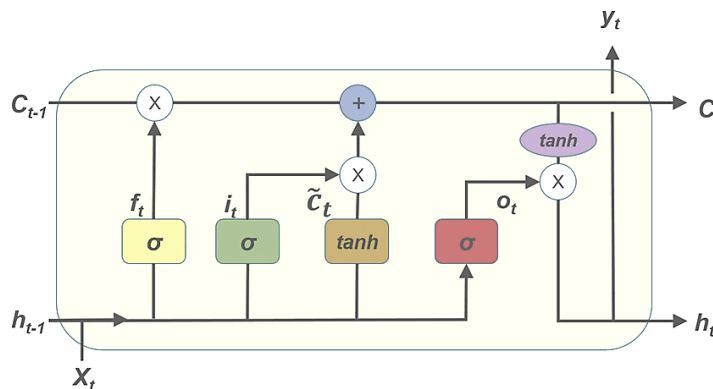


Figure 2. Long short-term memory diagram [20]

2.4. Experiment setup

To evaluate the proposed method, we use the developed dataset for impairment children which is described in section 2.1 and 2.2. The dataset comprises 770 samples of speech, we have assigned 616 samples as the training set for the LSTM classifier and 154 samples of speech are for testing and validation reserved, which is described in section 2.3. Moreover, we set the parameters of the deep LSTM classifier including which was one hidden layer with 32 units or neurons, we used ‘adam’ as optimizer and tanh as activation function. To measure the performance of the proposed model in terms of the classification accuracy and loss of the classification results. We compare the performance of the proposed model with recent related work commonly used for speech recognition.

3. RESULT AND DISCUSSION

The classification results for impaired children speech are summarized in Table 1. The table shows that the accuracy of deep learning based on LSTM classified in isolated Arabic numbers and letters was 97.99%, and loss was 0.18%. The Table 1 shows the comparative results of our developed model with our developed dataset compared to other researches in speech recognition in terms of using isolated Arabic words, letters, or numbers. With the respect to other developed models, our proposed model that has been developed using a deep learning technique based on LSTM gives more recognition accuracy than others models.

Table 1. Evaluation of our proposed model compared to other models

Models	Accuracy	Dataset type
[21]	97.8%	10 Isolated Arabic Words
[22]	94.39% - 94.56%	40 Arabic words
[23]	Used Error rate 0.68%	11 standard Arabic isolated words
[24]	71.75%	3 Arabic isolated words
[25]	58.4% - 76.7%	29 isolated Arabic Letters
Our developed model	97.99%	Impairments children Dataset contains 0-9 and 8 Arabic Letters

As shown in the Table 1 researchers, A novel methodology for the identification of speakers has been proposed in [21]. The first approach is the combination of linear predictive coding (LPC) and skewness equation. The first method is based. A mixture of linear predictive coding (LPC), discrete transform wavelet (DWT) and cpestrum analysis was used in the second weighted linear prediction cepstral coefficients (WLPCC). Their outcome in term of accuracy were 97.8% which is very near to our result but less with 0.19%. Moreover, [22] created a novel approach based on a combination of multiple extractions and classifying features, isolated Arabic speech identification. The system uses a voting mechanism to merge the method outputs. While the mean calculation time is 1.56 seconds and the accuracy of the system increased

to 94.56%. Although in the paper, Boussaid and Hassine [23], system was built for 11 common Arabic isolated words with a speech recognition system. During the extraction stage several methods have been employed, including cepstral coefficients for the mel frequency (MLF), linear perceptual prediction, linear perceptual prediction and time derivatives in their initial order. To decrease the feature dimension, the primary component analysis was employed. The recognition stage is based on two learning algorithms, the Levenberg–Marquardt "Trainlm" and the scaled conjugate gradient "Trainscg," which are the basis of the feed backward neural propagation network. With dimensions 26 and Trainscg learning algorithm, the best results were obtained. When utilizing the corporation of 5, 10 and 20 speakers, respectively, they achieved a 0.05, 0.21 and 0.68% test error rate. While, Hammami *et al.* [24], addresses the topic of the classifying of speech sound errors in native Arabic children, where Arabic words including the letter r (pronounced as/ra/) are improperly spoken. they determine if there is a sound mistake in the speech when a letter appears at the start, middle or end. To categorize the words said, they describe the speaker signal using the functions of mel frequency cepstral coefficients (MFCC). They test the performance of their method used a real-world dataset of native Arabic speech children who record their voice. For Arabic words with letters r at the beginning, middle and end of words, A suggested technique obtains a classification accuracy of 71.75%, 77.20%, and an average of 74.06%. These findings are higher than those achieved using the Hidden Markov model. Finally, Khudeyer *et al.* [25], explore the combination of multi-machine learning methods for the recognition of Arabic isolated letters recognition with imperfect and dimensional variables. There is no such work in this regard, to the best of our knowledge. They integrated various machine classifications to recognize isolated Arabic written characters in multifaceted formats. The method is based on the combination, with a majority vote, of three machine classifier(s) (k-nearest neighbors (kNN), support vector machine (SVM), and generalized neural regression network (GRNN)) and for final decision it return to the structural similarity index (SSIM). Experimental findings reveal that System 1, System 2, Systems 3, and System 4 have 63.8%, 64.5% and 58.4%, and 76.7%, respectively.

4. CONCLUSION

The experimental results show that by using the MFCC feature extraction technique with deep learning the results are higher as compared to other proposed models been developed by other researchers. The recognition accuracy is higher by using LSTM as a deep learning model instead of using machine learning techniques. The recognition accuracy might be different if we used other extraction techniques or a combination of two techniques as well as other deep learning models such as CNN.




REFERENCES

- [1] S. R. Shareef and Y. F. Irhayim, "A Review: Isolated Arabic Words Recognition Using Artificial Intelligent Techniques," *Journal of Physics: Conference Series*, vol. 1897, no. 1, p. 12026, May 2021, doi: 10.1088/1742-6596/1897/1/012026.
- [2] M. M. El Choubassi, H. E. El Khoury, C. E. J. Alagha, J. A. Skaf, and M. A. Al-Alaoui, "Arabic speech recognition using recurrent neural networks," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2003*, 2003, pp. 543–547, doi: 10.1109/ISSPIT.2003.1341178.
- [3] L. Baghai-Ravary and S. W. Beet, "Introduction," *SpringerBriefs in Speech Technology*, pp. 1–6, 2013, doi: 10.1007/978-1-4614-4574-6_1.
- [4] S. Alharbi *et al.*, "Automatic Speech Recognition: Systematic Literature Review," *IEEE Access*, vol. 9, pp. 131858–131876, 2021, doi: 10.1109/ACCESS.2021.3112535.
- [5] L. M. Bedore and E. D. Peña, "Assessment of bilingual children for identification of language impairment: Current findings and implications for practice," *International Journal of Bilingual Education and Bilingualism*, vol. 11, no. 1, pp. 1–29, Jan. 2008, doi: 10.2167/beb392.0.
- [6] L. Schillingmann, J. Ernst, V. Keite, B. Wrede, A. S. Meyer, and E. Belke, "AlignTool: The automatic temporal alignment of spoken utterances in German, Dutch, and British English for psycholinguistic purposes," *Behavior Research Methods*, vol. 50, no. 2, pp. 466–489, Jan. 2018, doi: 10.3758/s13428-017-1002-7.
- [7] J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe, and K. J. Ballard, "Automated speech analysis tools for children's speech production: A systematic literature review," *International Journal of Speech-Language Pathology*, vol. 20, no. 6, pp. 583–598, Jul. 2018, doi: 10.1080/17549507.2018.1477991.
- [8] L. Verde, G. De Pietro, and G. Sannino, "Voice Disorder Identification by Using Machine Learning Techniques," *IEEE Access*, vol. 6, pp. 16246–16255, 2018, doi: 10.1109/ACCESS.2018.2816338.
- [9] C. Rana, "International Journal of Computer Science and Mobile Computing A Review: Speech Recognition with Deep Learning Methods," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 5, pp. 1017–1024, 2015, Accessed: Jan. 20, 2022. [Online]. Available: www.ijcsmc.com.
- [10] V. N. Gudivada and C. R. Rao, "Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications," *Handbook of Statistics*, vol. 38, pp. 197–228, 2018.
- [11] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," *Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, May 2014, doi: 10.1002/lary.24740.
- [12] T. A. Mesallam *et al.*, "Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/8783751.




- [13] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions," *Journal of Voice*, vol. 31, no. 1, pp. 3–15, Jan. 2017, doi: 10.1016/j.jvoice.2016.01.014.
- [14] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of Voice Disorders in Teachers and the General Population," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 2, pp. 281–293, Apr. 2004, doi: 10.1044/1092-4388(2004)023).
- [15] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task," *October*, vol. 1, no. 3, pp. 191–194, 2005, doi: 10.1.1.75.8303.
- [16] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," Dec. 2010, doi: 10.1109/ICSPCS.2010.5709770.
- [17] C. H. You, H. Li, and K. A. Lee, "A GMM-supervector approach to language recognition with adaptive relevance factor," in *European Signal Processing Conference*, 2010, pp. 1993–1997.
- [18] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, Dec. 2014, [Online]. Available: <http://arxiv.org/abs/1312.6026>.
- [19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [20] A. A. Ismail, T. Wood, and H. C. Bravo, "Improving Long-Horizon Forecasts with Expectation-Biased LSTM Networks," Apr. 2018, [Online]. Available: <http://arxiv.org/abs/1804.06776>.
- [21] Y. Faisal and A. M. Khalaf, "Speech Recognition of Isolated Arabic words via using Wavelet Transformation and Fuzzy Neural Network," *Computer Engineering and Intelligent Systems*, vol. 7, no. 3, pp. 21–31, 2016, Accessed: Jan. 20, 2022. [Online]. Available: <https://iiste.org/Journals/index.php/CEIS/article/view/29405>.
- [22] A. Kourd and K. Kourd, "Arabic Isolated Word Speaker Dependent Recognition System," *British Journal of Mathematics & Computer Science*, vol. 14, no. 1, pp. 1–15, Jan. 2016, doi: 10.9734/bjmcs/2016/23034.
- [23] L. Boussaid and M. Hassine, "Arabic isolated word recognition system using hybrid feature extraction techniques and neural network," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 29–37, Nov. 2018, doi: 10.1007/s10772-017-9480-7.
- [24] N. Hammami, I. A. Lawal, M. Bedda, and N. Farah, "Recognition of Arabic speech sound error in children," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 705–711, Sep. 2020, doi: 10.1007/s10772-020-09746-3.
- [25] R. S. Khudeyer, M. Alabbas, and M. Radif, "Multi-Font Arabic Isolated Character Recognition Using Combining Machine Learning Classifiers," *Journal of Southwest Jiaotong University*, vol. 55, no. 1, 2020, doi: 10.35741/issn.0258-2724.55.1.12.

BIOGRAPHIES OF AUTHORS



Sura Ramzi Shareef    received the BSc degree in 1996, the M.S in 2003, in 2018 accepted as PhD student at the college of computer science and mathematics, from University of Mosul, Iraq. Currently a PhD student in research field of computer sciences and a teacher in the computer Engineering Department college of engineering. Much of my research particularly concerned with computer science, intelligent techniques concerned with the technique's automatic speech recognition. She can be contacted at email: sura.ramzishareef@uomosul.edu.iq



Yusra Faisal Al-Irhaiym    received the B.Sc. degree in mathematics from the University of Mosul, Iraq, the M.Sc. degree in computer science from the University of Mosul, Iraq, and the Ph.D. degree in artificial intelligence from The University of Mosul, Iraq. She has supervised more than 8 masters and 3 Ph.D. students. She has authored or coauthored more than 30 publications. Her research interests include speech recognition, machine learning, and intelligent systems. She can be contacted at email: yusrafaisalc@uomosul.edu.iq