



QDAT: A data set for Reciting the Quran

Hanaa Mohammed Osman¹, Ban Sharief Mustafa², Yusra Faisal³

^{1,2,3} College of Computer Science and Mathematics, Mosul University, Iraq

¹hanaosman@uomosul.edu.iq, ²banmustafa66@uomosul.edu.iq, ³yusrafaisalcs@uomosul.edu.iq

Abstract:

Dataset are considered as an important part of any audio research and an important resource for speech processing. Availability of dataset in speech processing field is important. The effort and time needed to build a complete good dataset are very long. The available public dataset in Arabic language are very little. This paper presents the "QDAT" dataset of audio Arabic speech files. The audio files are manually annotated by expert to show the correctness of the Reciting the Quran with Tajwid according to three rules of recitation of Quran. The dataset can be used for training and classification models based on machine learning and deep learning algorithms.

Keywords: audio dataset, Quran recitation, MFCC algorithm, machine & deep learning techniques

1-Introduction

Arabic is one of the six languages of the United Nations (UN), and it is one of the most widely spoken languages in the world. Among the statistics, it is the first language of around 400 million speakers, and it ranked fourth after Mandarin, Spanish and English in terms of the number of people who speak it as their native language (Abushariah, 2012). The Arabic language is distinguished as it is the language in which the Holy Qur'an was revealed. Corpus are required in different applications of speech and language processing. For example, Automatic Speech Recognition (ASR) systems use statistical models that are trained on corpus of related speech. Also, the development of ASR systems requires speech resources from a large number of speakers to obtain acceptable performance. Arabic speech corpus is less numerous compared to the linguistic resources devoted to other major spoken languages in the world. In recent years, the Language Data Consortium (LDC) published the first public Modern Standard Arabic (MSA) speech corpora that was designed for speech recognition experiments. The corpora, called West Point, contains speech data that was collected and processed by members of the Department of Foreign languages at the United States Military Academy at West Point (Selouani, 2010). A speech dataset is the basis component in speech processing research and in developing speech processing systems. An automatic speech(speaker) recognition system can be used successfully in real life only if it is developed using a real-life dataset. Without a good speech dataset, research on speech processing cannot be progressed. There are many datasets in special languages, like English, Japanese, Chinese, Spanish, German, etc. These datasets are rich in the number of speakers, amount of speech, variability of speakers and texts, transmission channels, and environments. But Arabic speech dataset are little bit in numbers and most of them are private. Therefore, there is a need for a publicly available comprehensive Arabic speech dataset. A rich and a publicly available dataset is an important and essential resource for research in the Arabic speech.

When a speech dataset needs to be developed, the following consideration may be taken into account: Scope of the dataset, Phonological distribution, Content, Gender, Number of speakers, speaking style, Environment, Recording materials, Partition into training and testing data sets (Alsulaiman, 2013).

In this paper a dataset (QDAT) has been built. The dataset is represented on more than 1500 audio files. The audio files are manually annotated by expert to show the correctness of the Reciting the Quran with Tajwid--- according to three rules of recitation of Quran.

2-Related work

Building an audio Arabic dataset for reciting Holy Quran is the main target in this research. Similar studies for building an audio dataset can be summarized as follows:

Lamiaa (2019) presents the Arabic Visual Speech Dataset (AVSD) for visual speech recognition. The dataset contains 1100 videos for 10 daily communication words collected from 22 speakers and recorded using smartphones' cameras in high-resolution and high-frame rate.

Yin May (2020) introduces an open source, multi-speaker, and multi-source text suite along with a comprehensive set of finite-state converter (FST) grammar rules for performing Burmese text normalization (Myanmar) language.

Abdel Rahim (2018) present ArSAS , an Arabic corpus of tweets annotated for the tasks of speech-act recognition and sentiment analysis. A large set of 21k Arabic tweets covering multiple topics were collected, prepared and annotated for six different classes of speech-act labels, such as expression, assertion, and question. In addition, the same set of tweets were also annotated with four classes of sentiment. This corpus promotes the research in both speech-act recognition and sentiment analysis tasks for Arabic language.

Iakushkin (2018) examines the methods to automatically build massive corpora of transcribed speech from open access sources in the internet, such as radio transcripts and subtitles to video clips. The study is focused on a method to build a speech corpus using the materials extracted from the You Tube video hosting. The study resulted in creating transcribed speech corpora in Russian containing 1000 hours of audio recordings. The quality of obtained data has been assessed by using a part of it to train a Russian-language automatic speech recognition system based on the Deep Speech architecture. Upon training, the system was tested on a data set consisting of audio recordings of Russian literature available on voxforge.com the best word error rate (WER) demonstrated by the system was 18%.

Afroz A., (2020) describes the main requirements for creating a well-structured data set from speech samples in non-native accents for training and testing of robust ASR systems. They present AccentDB, a dataset containing samples of 4 Indo-English dialects that brought together robust ASR systems. Also, they present an AccentDB, and the dataset contains samples of 4 Indo-English dialects collected by them, a sample set of 4 English languages, and an urban Indo-English dialect.

3- QDAT Construction Methodology

The methodology for the QDAT dataset is illustrated in Figure 1. The first step in creating the dataset is by specifying the design specifications of the dataset including the language and participant specification. The second step is the data collection step in which the environment is defined. The third step is the post-processing steps that are used to prepare the data set for testing. Finally, a dataset test step in which a speech scheme is proposed.

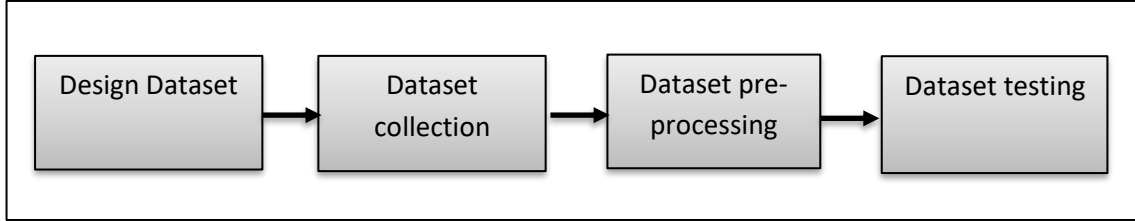


Figure. 1, Methodology of the QDAT dataset.

4-The proposed Model

To assess the utility of QDAT, a simple modern system was developed, trained, and tested with QDAT. The developed system shown in Figure 2 consists of two successive stages, feature extraction and feature recognition (or classification). For the feature extraction stage, the Mel-Frequency Cepstral coefficients MFCC technique was used. The machine learning classifier was used in the feature recognition stage.

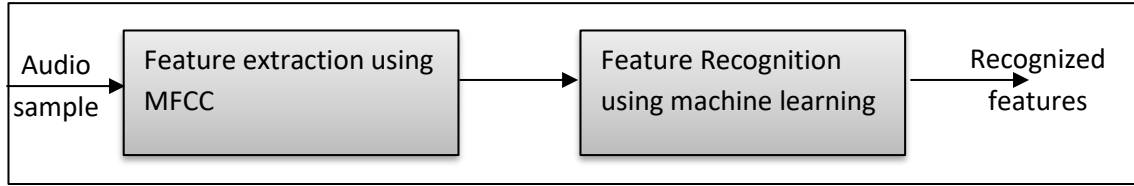


Figure. 2, The proposed Model.

5-MFCC Algorithm

The front-end features consist of the extraction of the MFCC. These coefficients were calculated from the bank amplitudes of the log filter m_i using a Discrete Cosine Transform as in the Equation. 1:

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \quad (1)$$

Where N is the number of the filter bank channels. The filters used were triangular and were equally spaced along the Mel-scale. The model uses MFCC, logarithmic energy, and delta coefficients that are the derivatives of the MFCC calculated through the use of regression analysis according to Equation. 2:

$$d_t = \frac{\sum_{\theta=1}^{\delta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\delta} \theta^2} \quad (2)$$

where d_t was the delta coefficient at time t computed by using the static coefficient $c_{t+\theta}$ to $c_{t-\theta}$ with a possible delay of δ . The same formula was applied to the delta coefficients to obtain acceleration (delta-delta) coefficients. To obtain a good performance of speech recognition system a derivative coefficient is added to static parameters [2].

In our ASR, the following parameters are used:

- 39-dimensional feature vector composed of 12 MFCCs and the normalized energy with their delta and acceleration coefficients;
- first-order digital filter with a transfer function $H(z) = 1 - Kz^{-1}$ was used for the pre-emphasis processing with $k=0.97$;
- 16 kHz sampling frequency;
- 25-millisecond Hamming window duration with a step size of 10 milliseconds;
- 22 as the length of cepstral liftering;
- 26 filter bank channels;
- delta coefficients are set to 3 in computation (Equation 2).

6- Scope of the dataset

In order to test the correctness of recitation of Quran according to rules of recitation, a dataset of recitation of Verse in the Quran is proposed. The Verse “(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ الْغُيُوبِ) [Surah Al-Ma'idah 109] “, contains three recitation rules:

- The separate stretching: is that the letter of the dowel is the last word, and the Hamza is the first word that follows it. The duration is four or five movements.
(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ الْغُيُوبِ)
- Tight Noon: Ghunnah must be shown in the aggravated Noon by two movements. This ruling is called an accentuated Ghunnah letter because the Ghunnah is a necessary attribute of letter Noon.
(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ الْغُيُوبِ)
- Hide: Language is the concealment, and as for idiomatically, it is pronouncing the consonant noun or Tanween in a state between showing and fading without emphasis, with the song remaining by two movements.
(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ الْغُيُوبِ)

These rules are checked manually by experts. A variety of recitation samples were obtained from different readers to form the dataset.

6-1 Content

The collected dataset includes over 1500 audio samples from over 150 different readers: 350 males and 1159 females. The reading is recorded using online recording (WhatsApp) in WAV files have a 11KHz sample rate, MONO channel and 16-bit resolution. QDAT dataset contains 1500 WAV files along with sound files stored on Excel CSV file format. The sound file contains links to the WAV files attached with other features: Age, Gender, and the correctness of the recitation of the three recitation rules and the final goal shows the correctness of the whole reading. The QDAT is available in Kaggle web site at the link: <https://www.kaggle.com/annealdahi/quran-recitation>.

6-2 Phonological distribution:

In table 1 explanation of the provisions of the Verse in the dataset are given:

Table 1: The provisions of the Verse

The provisions	Verse	Verse in symbols
Separate stretching of four movements	(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ) (الْعُيُوبِ)	ق _ _ ل _ _ ل _ _ ل _ _ ل _ _ م _ _ ل _ _ ن _ _ ن _ _ ع _ _ ي _ _ ي _ _ ب _ _ ك _ _ ع _ _ ن _ _ ت _ _ ع _ _ ل _ _ ل _ _ م _ _ ل _ _ ل _ _ غ _ _ ي _ _ ي _ _ ب _ _
Separate stretching of five movements	(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ) (الْعُيُوبِ)	ق _ _ ل _ _ ل _ _ ل _ _ ل _ _ ل _ _ م _ _ ل _ _ ن _ _ ن _ _ ع _ _ ي _ _ ي _ _ ب _ _ ك _ _ ع _ _ ن _ _ ت _ _ ع _ _ ل _ _ ل _ _ م _ _ ل _ _ ل _ _ غ _ _ ي _ _ ي _ _ ب _ _
Tight Noon	(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ) (الْعُيُوبِ)	ق _ _ ل _ _ ل _ _ ل _ _ ل _ _ ل _ _ م _ _ ل _ _ ن _ _ ن _ _ ع _ _ ي _ _ ي _ _ ب _ _ ك _ _ ع _ _ ن _ _ ت _ _ ع _ _ ل _ _ ل _ _ م _ _ ل _ _ ل _ _ غ _ _ ي _ _ ي _ _ ب _ _
Hide	(قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنْتَ عَلَّمُ) (الْعُيُوبِ)	ق _ _ ل _ _ ل _ _ ل _ _ ل _ _ ل _ _ م _ _ ل _ _ ن _ _ ن _ _ ع _ _ ي _ _ ي _ _ ب _ _ ك _ _ ع _ _ ن _ _ ت _ _ ع _ _ ل _ _ ل _ _ م _ _ ل _ _ ل _ _ غ _ _ ي _ _ ي _ _ ب _ _

6-3 Number of speakers

In table 2 the Age and Gender of speakers are explained.

Table 2: gender and ages of speakers

Age	Number of audio files	Gender	Number of audio files
<15	134	Female	1159
15<age<25	355	Male	350
25<age<35	252		
35<age<45	244		
45<age<55	312		
55<age<70	186		

6-4 Environment

The sounds were recorded in a noise-free environment, where the recording was performed either directly from the people themselves, or the recording was received via the internet (WhatsApp platform). The recording was repeated approximately 10 times for each person.

7- Experiments and Result

For using QDAT dataset, we experiment with two different classification algorithms using python libraries (Sklearn, Keras)¹. The first experiment runs with 10-fold cross validation using Gradient Boosting Classifier algorithm from Sklearn library. 50 of estimators for building the gradient boosting model are chosen. The average accuracy for training and testing data is:

<https://scikit-learn.org/stable/> , <https://keras.io/> ¹

Table 3: The average accuracy for training and testing data

Goal	AUC Score (Train)	Accuracy (Train)	AUC Score (Test)	Accuracy (Test)
S1	0.977188	0.9261	0.892396	0.8027
S2	0.991275	0.9436	0.918696	0.8821
S3	0.985809	0.9358	0.915496	0.839
Label (ALL)	0.972100	0.9037	0.907579	0.8209

The ratio of testing data takes 0.3 from all dataset. The second experiment use deep learning for building the classification model. The sequential model contains three Dense layers with Relu activation method. The last layer use Softmax as activation method. The average accuracies for validating and testing data are:

Table 4: The average accuracies for validating and testing data

Goal	Loss	Accuracy	Val-Accuracy	Val-Loss
S1	1.5838	0.6051	0.6281	0.7312
S2	3.1515	0.8045	0.8005	3.2163
S3	2.0358	0.6051	0.6190	0.7384
Label (All)	1.5838	0.6060	0.5646 ²	0.6885

8-Conclusion

In this research, a QDAT data set was constructed for the recitation of VERSE from the Holy Quran. They are audio files that contain correct and incorrect recitation, with a proportion of 40% of correct readings of the total readings.

To experiment with using the data, two models were built, the first depending on Gradient Boosting algorithm. The performance of the model was good compared to the second model based on deep learning and as shown in the figures (3,4).

These models are considered preliminary, their performance has not been improved and are only to illustrate the use of the QDAT dataset. The second part of QDAT dataset is currently being worked on, as it includes different verses for the same recitation rule. An intelligent model will be built to recognize the correct recitation of the Quran.

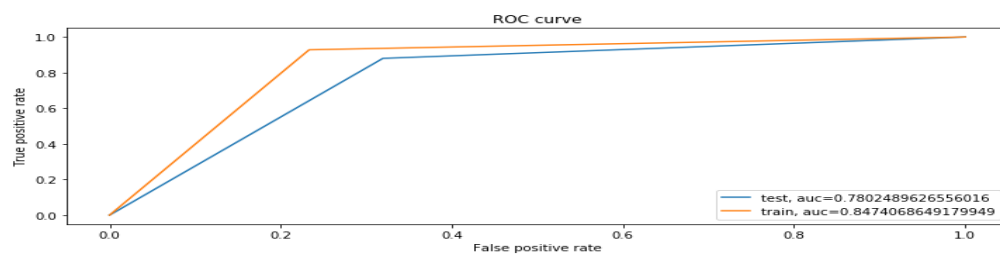


Figure. 3, ROC curve for Gradient Boosting model

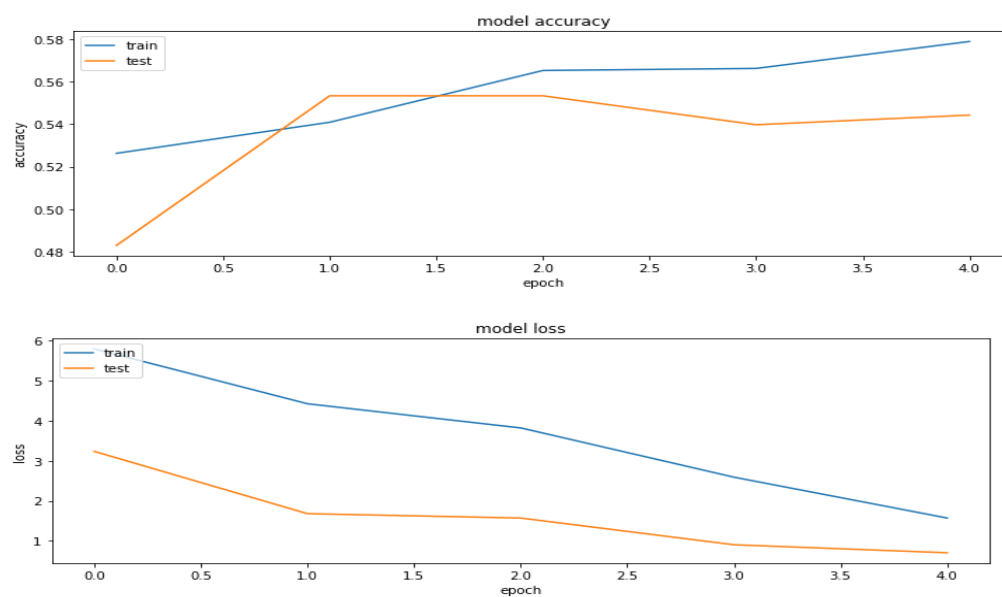


Figure. 4, Accuracy and Loss analysis for Deep Learning Model

References




- Abushariah, M. A. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2012). Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. *Int. Arab J. Inf. Technol.*, 9(1), 84-93.
- Ahamad, A., Anand, A., & Bhargava, P. (2020). AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition. *arXiv preprint arXiv:2005.07973*.
- Alsulaiman, M., Muhammad, G., Bencherif, M. A., Mahmood, A., & Ali, Z. (2013). KSU rich Arabic speech database. *Information (Japan)*, 16(6 B), 4231-4253..
- Elmadany, A., Mubarak, H., & Magdy, W. (2018). Arsas: An arabic speech-act and sentiment corpus of tweets. *OSACT*, 3, 20.
- Elrefaei, L. A., Alhassan, T. Q., & Omar, S. S. (2019). An Arabic Visual Dataset for Visual Speech Recognition. *Procedia Computer Science*, 163, 400-409.

Iakushkina, O., Fedoseev, G., & Shaleva, A. (2018). Building corpora of transcribed speech from open access sources. *Advisory committee*, 140.

Oo, Y. M., Wattanavekin, T., Li, C., De Silva, P., Sarin, S., Pipatsrisawat, K., ... & Gutkin, A. (2020, May). Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 6328-6339).

Selouani, S. A., & Boudraa, M. (2010). Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*, 35(2), 157-166..

Biodata

	<p>Hana Muhammad Osman ... A teacher at the University of Mosul ... Interested in Artificial Intelligence and its Applications in the service of the Holy Quran...</p>
	<p>Dr. Ban Sherif Mostafa ... Specialized in Artificial Intelligence technologies ... Interested in smart computer applications in the service of the Holy Quran and its Sciences...</p>
<p>Photo of author 3</p> 	<p>Dr. Yusra Faisal Al-Rahim ... Assistant Professor ... Specialized in Artificial Intelligence ... Interested in deep learning and speech processing ...</p>

QDAT: بيانات لقراءة القرآن الكريم

هناك محمد عصمان¹، د. بان شريف مصطفى²، د. يسرى فيصل الارحيم³

^{1,2,3} جامعة الموصل، كلية علوم الحاسوب والرياضيات

¹hanaosman@uomosul.edu.iq, ²banmustafa66@uomosul.edu.iq, ³yusrafaisalcs@uomosul.edu.iq

الخلاصة. تعتبر مجموعة البيانات جزءاً مهماً من أي بحث صوتي ومورد مهم لتطبيقات معالجة الكلام. ومن المهم توافر مجموعة البيانات للعمل في تطبيقات معالجة الكلام. تحتاج عملية بناء مجموعة بيانات جيدة ومتكاملة الى وقت طويل وجهد كبير. ان هناك ندرة في مجموعة البيانات والتي تخص القراءات القرآنية. تقدم هذه الورقة مجموعة بيانات "QDAT" لملفات تحوي القراءات القرآنية الصوتية. يتم تمييز الملفات الصوتية يدوياً بواسطة خبير لإظهار صحة تلاوة القرآن بالترتيل، وفقاً لثلاث قواعد لتلاوة القرآن. يمكن استخدام مجموعة البيانات في نماذج التدريب والتصنيف بناءً على التعلم الآلي وخوارزميات التعلم العميق.

الكلمات المفتاحية: مجموعة البيانات الصوتية ، تلاوة القرآن ، خوارزمية MFCC ، تقنيات التعلم الآلي والعميق