Database for Arabic Speech Commands Recognition

Lina Tarek Benamer[1*], Osama A.S. Alkishriwo[2]

[1] l.benamer@uot.edu.ly, [2] o.alkishriwo@uot.edu.ly

[1] Department of Electrical and Electronics, College of Engineering, University of Tripoli, Libya
[2] Department of Electrical and Electronics, College of Engineering, University of Tripoli, Libya
*Corresponding author email: l.benamer@uot.edu.ly

ABSTRACT

Technology is all around us and it's changing rapidly, expanding Internet access has had huge impacts on everyday lives as people do everything on their phones and computers. The widespread growth in the use of digital computers, have an increasing need to be able to communicate with machines in a simpler manner. One of the main tasks that can simplify communication with machines is speech recognition. In this work, we introduce the Arabic speech commands database that contains six Arabic control order words and Arabic spoken digits. The created database is used to analyze and compare the recognition accuracy and performance of three recognition techniques which are, Wavelet Time Scattering feature extraction with Support Vector Machine (SVM) classifier, Wavelet Time Scattering feature extraction with Long Short-Term Memory (LSTM) classifier, and Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction with K-Nearest Neighbor (KNN) classifier. Finally, the experimental results show that the most accurate prediction of the database commands was 98.1250% given by Wavelet Time Scattering feature extraction and LSTM classifier and the fastest training time for the database was 144 minutes given by MFCC and KNN classifier.

Keywords: Speech Recognition - Arabic Speech Command Recognition - Wavelet Time Scattering - Support Vector Machine (SVM) - Long Short-Term Memory (LSTM) - Mel-Frequency Cepstrum Coefficients (MFCC) - K-Nearest Neighbor (KNN).

1 Introduction

Speech recognition is an automatic identification of speech by machine using some characteristics of the speaker's voice [1,2], and it is an important technique especially that the world is passing through the era of information and communication technology, where we use speech recognition in many application areas such as human–robot interaction, human–computer interaction, and telephone applications. Speech-based interaction [3,4] is performed using natural human voice and also has many difficulties such as noise, behaviour of humans, and accent of spoken words. Therefore, it is an open challenge for the researchers to develop speech recognition techniques that can recognize different words correctly.

Many speech recognition techniques exist and have been developed, several researchers did several research works related to this area. Truong et al. [5] presented a novel multi-speaker segmentation method that makes use of the wavelet analysis and support vector machine to separate various speech signals of the speakers through multi-dialog approach. Pawan and Raghunath [6] presented a text-independent speaker recognition technique with Fourier transform by means of MFCC and SVM. Mohamed and Ramachandran [7] described the improvement in property of the normalization by using HMM and artificial neural networks. M.A. Anusuya et al. [8] proposed a PCA-based Kannada speech recognition technique where discrete wavelet transforms are used for calculating wavelet coefficients. Current research in the computer science field focuses on deep learning for monitoring change in speech patterns, speech recognition and classification [9,10]. The authors in [11] present an expert neural network based on dynamic selection of classifiers for application in a speech signal pattern recognition system. M. Imtiaz et al. in [12] proposed an approach of speech recognition system based on isolated word structure using Mel-Frequency Cepstral Coefficients (MFCC's), Dynamic Time Warping (DTW) and K-Nearest Neighbour (KNN) techniques A combination of a hidden Markov model (HMM) and a deep long short-term memory (LSTM) network for speech recognition is given in [13]. An automatic language identification system using Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction with K-means clustering and Support Vector Machine (SVM) for classification is introduced in [14]. A comparative study of MFCC-KNN and LPC-KNN for Hijaiyyah letters pronunciation classification system is presented in [15].

As speech technology has developed, the number of individuals who would like to train and evaluate recognition models has grown rapidly, but the availability of datasets has not widened. The main contribution of this paper is to create a database of Arabic Speech Commands as an attempt to build a standard training and evaluation dataset for a class of simple speech recognition tasks. Its primary goal is to provide a way to build and test small models that detect when a single word is spoken, from a set of targeted words, with the fewest possible failures from background noise or unrelated speech. The database contains six Arabic control order words and Arabic spoken digits made with different users. Broadening access to databases will certainly encourage collaboration across groups and enables apples-for-apples comparisons between different approaches, helping the whole field move forward. The created database is used to study, implement, and compare the performance of three different speech recognition techniques.

2 Proposed Arabic Speech Commands Recognition System

The proposed speech recognition system is shown in Figure 1. The features of spoken Arabic commands that were created to contain Arabic control words and Arabic digits are extracted then classified using three machines and deep learning techniques. The classification is performed using Wavelet Time Scattering with a Support Vector Machine

(SVM) also with a Long Short-Term Memory (LSTM) network. In addition, an approach using machine learning to identify people based on features extracted from the recorded speech is also presented in this work. In this approach, the features are the Mel-Frequency Cepstrum Coefficients (MFCC), which used to train the K-Nearest Neighbor (KNN) classifier.

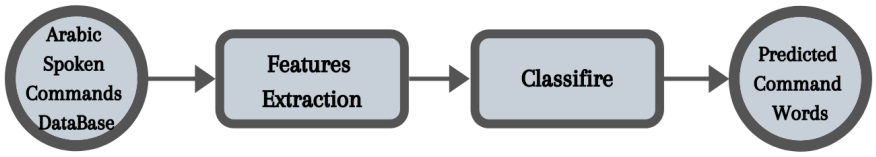

Figure 1: *Proposed Arabic speech commands recognition system*

2.1 Database

Data is very important to carry out tests, and when it comes to machine learning and deep learning tests we require having a lot of data to obtain the most accurate results. In this paper the database presented is created in Arabic language to support the researchers work in the field of Arabic speech recognition, carry out their tests using the Arabic commands database.

The created speech command database consists of 1600 recordings in Arabic of 6 control words: Add (إضافة), Back (رجوع), Cancel (إلغاء), Confirm (تأكيد), Continue (متابعة), Delete (حذف), and the digits 0 through 9 obtained from forty different speakers. The frequency of human voice ranges from 20Hz to 14,000Hz (typically from 300Hz to 4,000Hz). The frequency of a sound wave determines the human tone and pitch. In general, the frequencies, which have the most significant part of speech, lie between about 100Hz and 4,000Hz.

As a first step, the voice samples were collected together in one record from each volunteer reading the six words and the ten digits using the recording app on a mobile phone in a normal environment with as minimum background noise as possible. After that using Audacity software the collected data was divided so each audio file is for one specific command and converted to the WAV format so MATLAB could read and use the audio files. The presented database is available online at [16].

The database consists of 16 balanced classes with 100 recordings sampled at 48000Hz, managed to ensure the random division of the recordings into 80% training and 20% test sets as shown in Table 1.

Table 1: *Commands Database Labels and Count*

Name in English	Name in Arabic	Label	Total Count	Male Count	Female Count
Zero	Sefr "صفر"	0	100	59	41
One	Wahed "واحد"	1	100	37	63
Two	Ethnan "إثنان"	2	100	37	63
Three	Thalatha "ثلاثة"	3	100	41	59
Four	Arbaa "أربعة"	4	100	45	55
Five	Khamsa "خمسة"	5	100	40	60
Six	Seta "ستة"	6	100	40	60
Seven	Sabaa "سبعة"	7	100	44	56
Eight	Tamania "ثمانية"	8	100	40	60
Nine	Tesaa "تسعة"	9	100	43	43
Add	Edafa "إضافة"	A	100	34	66
Back	Rojou "رجوع"	B	100	36	64
Cancel	Elgha "إلغاء"	C	100	31	69
Delete	Hadef "حذف"	D	100	34	66
Confirm	Takeed "تأكيد"	E	100	33	67
Continue	Motaba "متابعة"	F	100	63	37

The recordings in the database are not of equal durations, and are not prohibitively large, so through reading the database files a histogram of the signal lengths was constructed as given in Figure 2. The histogram shows that the distribution of recording lengths is positively skewed. Classification uses a common signal length of 8192 samples, and it's considered as a conservative value that ensures that truncating longer recordings does not cut off speech content. Meaning that if the signal is greater than 8192 samples (1.024 seconds) in length, the recording is truncated to 8192 samples. And if the signal is less than 8192 samples in length, the signal is pre-padded and post-padded symmetrically with zeros out to a length of 8192 samples.

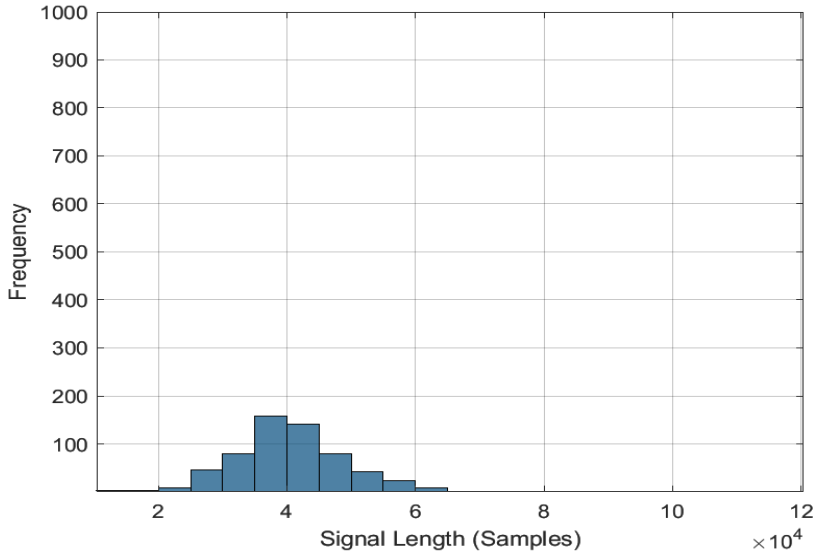

Figure 2: *Histogram of the signal lengths*

2.2 Feature Extraction Block

The process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing is called feature extraction. The characteristic of large data sets which has large number of variables, require a lot of computing resources to process. By extracting features from the input data the feature extraction increases the accuracy of learned models. This phase of the general framework reduces the dimensionality of data by removing redundant data. Of course, it increases training and inference speed while still accurately and completely describing the original data set.

2.2.1 Wavelet Time Scattering

Wavelet transform is a mathematical approach that is widely used for signal processing applications. It can be a decompose of special patterns hidden in the mass of data. Regarding the prediction issue through time series and neural networks, we need a modelling task. Neural networks as a general estimator in the estimation of extremely nonlinear systems have limited capability. Wavelet transform has the ability to simultaneously display functions and manifest their local characteristics in the time-frequency domain. The use of these characteristics facilitates the training of neural networks with accuracy to model extremely nonlinear signals. Wavelet techniques in general are effective tools for good data representations and feature extractions which can be used with most available classification algorithms. The wavelet scattering transform allows us to produce reliable and locally stable features to small deformations which can be used in conjunction with a deep neural network. Convolution, nonlinearity, and averaging, three successive main and required operations to produce a wavelet scattering transform of a time series input signal.

2.2.2 Mel-Frequency Cepstrum Coefficients (MFCC)

This is a method of extracting frequency information in speech signals and converting them into coefficients. Based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency the representation of the short-term power spectrum of a sound is made. Mel-frequency cepstral coefficients are a parametric representation of the speech signal that is commonly used in automatic speech recognition and they are calculated by applying a Mel-scale filter bank to the Fourier transform of a windowed signal. Subsequently, a DCT (discrete cosine transform) transforms the logarithmized spectrum into a cepstrum. The Mel filter banks consist of overlapping triangular filters with the cut off frequencies that can be determined by the centre frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. The logarithm has the effect of changing multiplication into addition. It converts the multiplication of the magnitude in the Fourier transform into addition. MFCCs simulate the properties of the human auditory system, that's one of the reasons they are widely applied in speech processing.

2.3 Classification Block

An algorithm that sorts data into labelled classes, or categories of information. And classification in general is the process of predicting the class of given data points. Classes are sometimes called targets or labels or categories. Classification predictive modelling is the task of approximating a mapping function from input variables to discrete output variables.

2.3.1 Support Vector Machine (SVM)

SVM is a binary classification algorithm that determines the decision boundary between feature vectors of two classes. It can be scaled well for multi-class classification and can be generalized well in high dimensional feature spaces. It is a supervised machine learning algorithm that is recognized as an easy-to-use and robust technique for classification, regression, and other learning tasks. It has a better classification performance on a small number of training samples using a technique called the kernel trick to transform data and then based on transformations, an optimal boundary between the possible outputs can found.

2.3.2 Long Short-Term Memory (LSTM)

An artificial recurrent neural network (RNN) architecture used in the field of deep learning, LSTMs are designed to overcome the vanishing gradient problem that restricts the memory capabilities of traditional RNNs, increases the chance of facing a gradient problem and losing information since too many time steps have been added. LSTMs are designed to allow the retention of information for longer periods compared to traditional RNNs. LSTMs continue learning over numerous time-steps and back propagate through time and layers because they can maintain a constant error. And they can also use gated cells to store

information outside the regular flow of the RNN. With these cells, the network can manipulate the information in many ways, including storing information in the cells and reading from them. The cells are individually capable of making decisions regarding the information and can execute these decisions by opening or closing the gates. Since there can be lags of unknown duration between important events in a time series, LSTM networks are well-suited to classifying, processing, and making predictions based on time series data.

2.3.3 K-Nearest Neighbor (KNN)

An unsupervised learning technique that is used for classifying objects based on closest training examples in the feature space. KNN is a type of lazy learning or instance-based learning where the function is only approximated locally and all computation is deferred until classification. The KNN is considered one of the simplest classification techniques when there is little or no prior knowledge about the distribution of the data. The followed up rule simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbors in the training set. In this method, each sample should be classified similarly to the surrounding samples. Therefore, if the classification of a sample is unknown, then this could be predicted by consideration of the classification of its nearest neighbor samples. In this classifier, we use Euclidean distance between the classes which can be further classified to know which class the data belongs to.

3 Results and Discussion

In order to obtain the recognition accuracy and analyse the performance of the Arabic speech commands database with the proposed techniques, experiments were carried out on the created database using MATLAB Audio Toolbox, Statistics and Machine Learning Toolbox, Deep Learning Toolbox and Wavelet Toolbox. The computed results were given using Intel Core i5 4250U at 1.3 GHz with 4GB RAM. The MATLAB used a single CPU with constant 0.0021988 learning rate of over 7500 iterations.

Table 2: *Comparison of different speech recognition techniques in terms of accuracy and computation*

Techniques	Training Time	Test Accuracy
Wavelet Time Scattering feature extraction with Support Vector Machine (SVM) classifier	261 min	96.2500%
Wavelet Time Scattering feature extraction with Long Short-Term Memory (LSTM) classifier	526 min	98.1250%
Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction with K-Nearest Neighbor (KNN) classifier	144 min	94.89%

As shown in Table 2, we can see that each technique had its pros and cons, which would make the technique that showed the most accurate prediction of the database commands the Wavelet Time Scattering feature extraction with Long Short-Term Memory (LSTM) classifier giving us 98.1250% test accuracy and the fastest training time for the database by the Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction with K-Nearest Neighbor (KNN) classifier given by 144 minutes of total training time. Based on these results we can use the technique that will serve the purpose of our application in which we would need the best prediction accuracy or the fastest training time.

4 Conclusions

This paper presents the analysis of Arabic speech recognition via Wavelet Time Scattering feature extraction with Support Vector Machine (SVM) classifier, Wavelet Time Scattering feature extraction with Long Short-Term Memory (LSTM) classifier, and Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction with K-Nearest Neighbor (KNN) classifier. The test experiments provided a recognition accuracy and training time performance of these techniques using the created Arabic speech commands database. Experiments with the three speech recognition techniques showed that the recognition accuracy for the Wavelet Time Scattering feature extraction with Long Short-Term Memory (LSTM) classifier was the best, although the difference between all of the techniques recognition accuracy was small. On the other hand, the best result in the training time performance was by Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction with K-Nearest Neighbor (KNN) classifier, and the difference between the training times was large. The Arabic Speech Commands dataset has shown to be useful for training and evaluating a variety of models.

As a future work, we recommend further investigation of Arabic speech recognition performance by different recognition techniques using this Arabic speech commands database to move forward with developing more speech control applications that support the Arabic language.

5 Acknowledgment

Massive thanks to everyone who volunteered with their recordings from my family, friends, and colleagues to create this database.

References

[1] C. S. Kumar and P. M. Rao, "Design of an automatic speaker recognition system using MFCC, vector quantization and LBG algorithm", *International Journal on Computer Science and Engineering*, vol. 3, no. 8, pp. 2942-2954, Aug. 2011.

[2] S. K. Gaikwad and B.A. Gawali, "A review on speech recognition technique," *International Journal of Computer Applications*, vol. 10, no. 3, pp. 16–24, Nov. 2010.

[3] Y. Lee and K. W. Hwang, "Selecting good speech features for recognition," *ETRI Journal*, vol. 18, no. 1, Apr. 1996.

[4] D. Y. Genoud, D. Ellis, and N.Morgan, "Combined speech and speaker recognition with speaker adapted connectionist models, *IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, , Dec. 1999.

[5] T. K. Truong, C. L. Chien, C. Shihuang, "Segmentation of specific speech signals from multi dialog environment using SVM and wavelet," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1307-1313, Aug. 2007.

[6] K. A. Pawan, S.H. Raghunath, "Fractional Fourier transform based features for speaker recognition using support vector machine," *Computers and Electrical Engineering*, vol. 39, no. 2, pp. 550-557, Feb. 2013.

[7] M. A. Ramachandran and K. N. Nair, "HMM/ANN hybrid model for continuous Malayalam speech recognition," *Procedia Eng*., vol. 30, pp. 616-622, 2012.

[8] M.A. Anusuya and S.K. Katti, "Mel frequency discrete wavelet coefficients for Kannada speech recognition using PCA*," In Proceedings of International Conference on Advances in Computer Science (ACEEE)*, pp. 225-227, 2010.

[9] A.B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143-19165, Feb. 2019.

[10] D. O'Shaughnessy, "Recognition and processing of speech signals using neural networks," *Circuits, Systems, and Signal Processing*, vol. 38, pp. 3454-3481, Mar. 2019.

[11] P. Rocha, W. Silva, and A. Barros, "Hierarchical expert neural network system for speech recognition*," Journal of Control, Automation and Electrical Systems*, vol. 30, pp. 347-359, Mar. 2019.

[12] M. A. Imtiaz and G. Raja, "Isolated word automatic speech recognition (ASR) system using MFCC, DTW & KNN*, "The 2016 Asia Pacific Conference on Multimedia and Broadcasting* , Bali, Indonesia, pp. 106-110, Nov. 2016.

[13] W. Ying, L. Zhang, and H. Deng, "Sichuan dialect speech recognition with deep LSTM network," *Frontiers of Computer Science*, vol. 14, pp. 378-387, Aug. 2019.

[14] V. K. Verma and N. Khanna, "Indian language identification using K-means clustering and support vector machine (SVM)," *2013 Students Conference on Engineering and Systems (SCES)*, Allahabad, India, Apr. 2013.

[15] Adiwijaya, M. Nur Aulia, M. S. Mubarok, W. U. Novia, and F. Nhita, "A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system," *2017 5th International Conference on \Information and Communication Technology (ICoIC7)*, Malacca City, Malaysia, May 2017.

[16] https://github.com/ltkbenamer/AR_Speech_Database.git