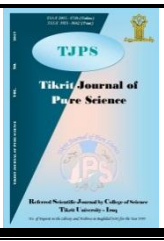
IASJ المجلات الاكاديمية العراقية

Tikrit Journal of Pure Science

ISSN: 1813 – 1662 (Print) --- E-ISSN: 2415 – 1726 (Online)

Journal Homepage: http://tjps.tu.edu.iq/index.php/j



Speaker Age and Gender Estimation Based on Deep Learning Bidirectional Long-Short Term Memory (BiLSTM)

Aalaa Ahmed Mohammed[1], Yusra Faisal Al-Irhayim[2]

[1] *College of Engineering, University of Tikrit, Tikrit, Iraq*

[2] *Dept. of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq*

DOI: http://dx.doi.org/10.25130/tjps.26.2021.068

ARTICLE INFO.

Article history:

-Received: 1 / 12 / 2020

-Accepted: 15 / 6 / 2021

-Available online: / / 2021

Keywords: speaker age estimation, speaker gender estimation, MFCC, BiLSTM

Corresponding Author:

Name: Aalaa Ahmed Mohammed

E-mail: aalaa.alrashidy@tu.edu.iq

Tel:

ABSTRACT

Estimating the age and gender of the speaker has gained great importance in recent years due to its necessity in various commercial, medical and forensic applications. This work estimates the speakers gender and ages in small range of years where every ten years has been divided into two subcategories for a span of years extending from teens to sixties. A system of speaker age and gender estimation uses Mel Frequency Cepstrum Coefficient (MFCC) as a features extraction method, and Bidirectional Long-Short Term Memory (BiLSTM) as a classification method. Two models of two deep neural networks were building, one for speaker age estimation, and the other for speaker gender estimation. The experimental results show that the deep neural network model of age estimation achieves 94.008 % as accuracy rate, while the deep neural network model of gender estimation achieves 90.816% as accuracy rate.

I. Introduction

The importance of estimating the age and gender of the speaker comes from many applications in our life that is need it, including effective shopping and advertising strategies for goods, as they are used in Client Relationship Management (CRM) systems that depend on the susceptibility of gender influence. As well as enhancing computer and human interaction systems, especially the chat systems used by customer services in various companies. In addition to advertisements that can be customized based on the age and gender of the person on the phone. In the field of forensic medicine, the number of suspects can be reduced if there is evidence of the age and gender of the speaker through his voice during a phone call, for example [1]. Speaker age estimation is a useful tool in various applications, especially those that require age knowledge to give permission for its use [2]. Some other uses of this system can be statistics that need to learn about age and gender information for a specific population [3].

The process of distinguishing the gender of the speaker does not have significant difficulties, because there are clear differences between male and female voices in puberty due to the physiological differences and physical characteristics of the vocal cords in terms of length and degree of tension, which leads to the production of different sound signals in frequency and other characteristics [4]. While the process of estimating the age of the speaker is difficult from different points of view. Firstly, There is usually a difference between the age of the perceived speaker, i.e. that which is perceived by the listener, and the real age of the speaker. Secondly, it is difficult to develop a robust system for estimating age because it requires a coded database with a wide and balanced age range. Thirdly, Sounds patterns are affected by several factors such as weight, height, and emotional conditions [5].

Speaker age and gender estimation is affected by many factors related to speech speed, articulation, and the speaker's health and psychological conditions [6]. So the efficient feature extraction plays an important role in achieving high-fidelity results in the estimation process, as well as the classifier that used to detect the gender and to estimate the age class [7]. In this work, system uses Mel Frequency Cepstrum Coefficient (MFCC) (introduced by Davis and Mermelstein in 1980), Delta (which is the first derivative of MFCC), and Delta-delta (which is the second derivative of MFCC) as a features extraction method, and deep learning classifier that is Bidirectional Long-Short Term Memory (BiLSTM).

This paper is arranged as follows: Section 2 contains the related works. Section 3 introduces the methodology of the system. In the section 4 practical experiments were presented. The results and discussion are illustrated in section 5. The paper finishes with conclusions in section 6.

II. Related Works

Hye-Jin Kim, Kyungsuk Bae, and Ho-Sub Yoon in 2007 describe a method to recognize the age and gender for a human speech. They use the Mel Frequency Cepstral coefficients (MFCCs) as features extraction method, a Gaussian Mixture Model (GMM) technique is applied for discovering the age, gender. The data were classified into three categories, which are the children, the adult male, and the adult female. The experimental results show that the gender classification accuracy was 94.9 %, and the age classification accuracy was 94.6 % [8]. In 2011, Heba Idris proposed an algorithm that classifies the speaker's age into one of two categories: the young and the senior based on their speech signal. The characteristic values of the variance matrix formed from the (one-dimensional) speech signal data after rearranging it into a number of square binary matrices were adopted as a fundamental factor in the process of distinguishing between the two classes. After implementing the algorithm on 50 people of both genders (male/female), it was found that it succeeded in classifying the ages of the speakers by 80% as accuracy rate [4]. In 2015, Fatima K. Fayeq examined automatic identification of gender and age through speech where she used the first four basic frequencies (F1, F2, F3, F4) and the features of the twelve MFCCs for the purpose of extracting traits, then she made the classification using the support vector machine classifier to distinguish gender, while to distinguish age she used the Nearest Neighbor Classifier (K-NN), two gender recognition models were designed, the first consisting of two classes (adult male and adult female) achieving a defining accuracy of 96%, and the second consisting of three categories (adult male, adult female and children) achieving a recognition accuracy of 94%. Fatima showed in her work that F1 and F4 are more related to age recognition than F2 and F3, so they were chosen for the age recognition model, and the accuracy of age recognition with noise reached 75.3% and with noise reduction it reached 81.44% [20]. Qawaqneh Z., Abu Mallouh A. and Buket D. in 2017 present a new approach for speaker age and gender classification that utilizes Deep Neural Networks (DNNs) as both feature extractor and classifier. They compared their work with the Mean Super Vector, GMM Base, MLLR Super Vector, SVM Base, TPP Super Vector, and the fused system of all these systems. Seven categories were classified, which are children, young females, young males, middle-aged females, middle-aged males, elderly females, and elderly males. The results showed that the proposed model was the best among other systems by achieving 57.21% overall classification accuracy [9]. Leo Kristopher Piel in 2018 analyses the performance of various kinds of neural networks on children's age and gender identification depending on speech and then compares the results to a selected i-vector baseline system. After comparing the results show that the model reached 4.6 percentage points higher accuracy on age identification, while about gender identification the model reached 2.6 percentage points higher accuracy on [10].

III. Methodology

A. Dataset

An Arabic text containing a noble Quranic verse was used through the Arabic data set (QDAT) specially prepared for academic and research purposes, such as training and classification models based on machine learning and deep learning algorithms, which are available at the following link: (https://www.kaggle.com/annealdahi/quran-recitation). The dataset contains more than 1500 voice clips recorded by people of both genders (males and females) and of different age groups. The audio clips included recording a part of verse in Surah Al-Ma'idah 109: (قَالُوا لَا عِلْمَ لَنَا إِنَّكَ أَنتَ عَلَّامُ الْغُيُوبِ) that means (They will say we have no knowledge it is Thou Who knowest in full all that is hidden). It is recorded using (WhatsApp) online in (WAV) files with a sample rate of (11kHz), mono channel, and (16-bit) resolution. The dataset included audio recordings of more than 150 persons whose age groups and genders were mentioned in an Excel CSV file attached to the recorded audio file folder [22]. The data set has been reformatted to suit the requirements of the proposed system by renaming the age categories of speakers in the attached (.csv) file, where the age categories of the speakers in the data set used are divided into 12 age categories and the table No.1 shows these categories and the number of audio clips for each category.

Table 1: The number of speakers in different age categories

No.	Category Name	Age/Years	No. of clips for each Category
1	Early-teens	10-14	101
2	Late-teens	15-19	77
3	Early-twenties	20-24	209
4	Late-twenties	25-29	187
5	Early-thirties	30-34	151
6	Late-thirties	35-39	90
7	Early-forties	40-44	194
8	Late-forties	45-49	106
9	Early-fifties	50-54	162
10	Late-fifties	55-59	115
11	Early-sixties	60-64	86
12	Late-sixties	65-69	30

The data used in this work was divided into two groups, the first is the training audio clips, which are 1214 clips, or approximately 80% of the total number of the data set used, and the second is the test audio

clips, which are 294 clips, or approximately 20% of the total number.

B. Pre-processing

Two stages of preprocessing are implemented on the input audio data before extracting their features, silence removal and normalization.

1. Silence removal

The process of removing silence periods of the audio clip is one of the important steps in the development of any powerful and efficient sound processing system. It means separating the acoustic region in the speech signal from the silent part. It is necessary because that most of the characteristics of the speaker's voice are present in the phonemic part of the speech signal. As well as, extracting the audio part only from speech signaling and processing it greatly reduces the complexity of calculations [11].

2. Normalization

It is the process of re-measuring the data so that all values are within the range [-1,1]. It is an important and necessary process when the time series data contain input values of different scales, and it includes re-distributing the values so that the average of the observed values is 0 and the standard deviation is 1, and can be represented mathematically, by the following equation [21]:

$$x_{new} = \frac{(x - Mean(x))}{MAX(ABS(x - Mean))} \quad \ldots(1)$$

Where (x) represents the sound signal, and ($Mean$) represents the average obtained by dividing the sum of signals by their number.

C. Features Extraction

Three stages of features extraction are performed to input audio data after preprocessing, MFCCs, Delta and Delta-delta.

1. Mel Frequency Cepstral coefficients (MFCCs)

The extraction of the characteristics of the speech signal is the basic step of any speech or speaker recognition system. It is responsible for extracting important information that expresses the most prominent features of the speech signal after divided it into frames. The characteristics of the audio signal are extracted for each frame. One of the most important features extraction algorithms is a mathematical representation of speech data that reflects the characteristics of the human auditory system. It is used here as a first step for typical applications to distinguish speakers age and gender because it reduces noise effects and perfectly represents sounds when the source features are fixed [12]. The Mel scale is defined as a perceptual measure of frequencies whose distance between each other is equal. MFCCs are parameters that together form a Mel-frequency cepstrum in which the frequency bands are evenly spaced on the Mel scale. MFCC produces 13 features for each frame. The response of MFCCs is more similar to the human audio system than conventional methods that use linearly spaced frequency bands [13]. The MFCC is derived through the stages illustrated in figure (1).

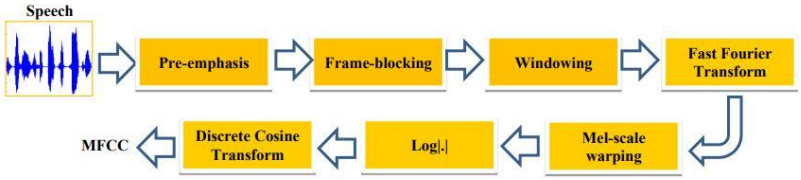


Fig. 1: MFCC block diagram [14]

- **Pre-emphasis**: This step deals with passing the signal through a filter that emphasizes the higher frequencies and excludes the lower frequencies. This means that pre-emphasis process will increase the signal energy at higher frequencies, as shown in equation (2) [7].

$$Y[n] = X[n] - 0.95\, X[n\text{-}1] \quad \ldots(2)$$

Where the number 0.95 represents the filter factor, which assumes that any sample originates from 95% of the previous sample.

- **Frame-blocking**: Is the process of dividing the speech signal into small frames with a length ranging from 20 to 40 milliseconds. The audio signal is divided into frames composed of N numbers of samples, the adjacent frames are separated by M number of samples, where M <N, and the typical values that used are M = 128 and N = 256 [7].
- **Hamming window**: This step aims to reduce the interruption of the speech signal before and after each frame, and is implemented through the two equations (3) and (4) below [7]:

$$Y(n) = X(n) \times W(n) \quad \ldots(3)$$

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \ldots(4)$$

Where N represents the number of samples in each frame, Y (n) is the output signal, X (n) is the input signal, and finally W (n) is the Hamming window.

- **Discrete Fourier Transform (DFT)**: estimating the power spectrum of the short frames of the signal by converting each frame of samples from the time domain to the frequency domain through the following equation [26]:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \quad \ldots(5)$$

- **Mel Filter Bank**: The range of frequencies in the spectrum resulting from the intermittent Fourier transform (DFT) is very wide, so the tonal scale is applied to the energy spectra by using overlapping triangular windows as in Figure (2), and summing the energy in each filter as in the equation below [7]:

$$F(Mel) = [2595 * log_{10}[1+f]\,700] \quad \ldots(6)$$

Where f represents frequency in HZ.

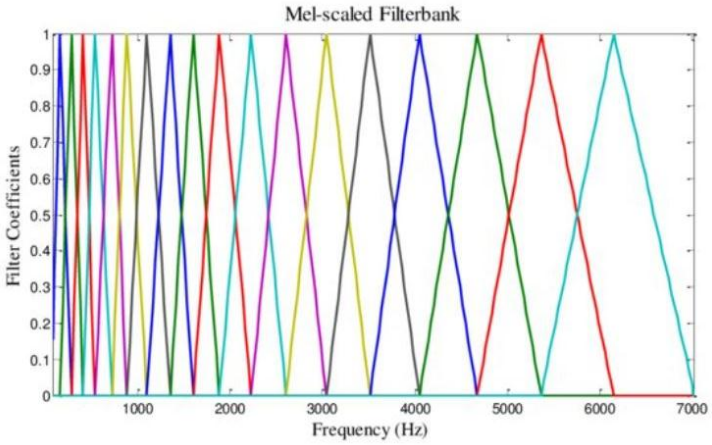


Fig. 2: Mel filter banks [14]

• **Discrete Cosine Transform**: In this step, the log Mel spectrum resulting from the previous step is converted to the time domain. The result of the conversion is called the Mel Frequency Cepstrum Coefficient. In this way, the input speech is converted into sound vectors [7]. Figure 3 shows MFCC coefficients.

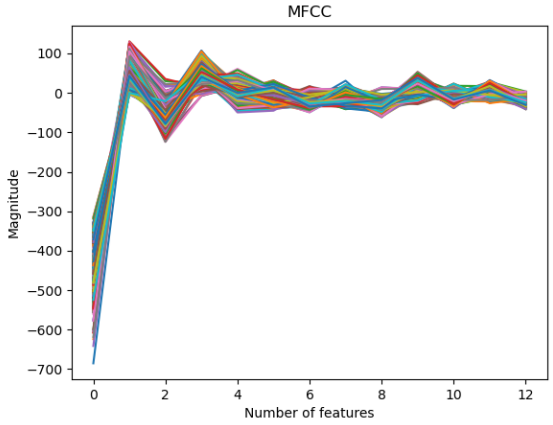


Fig. 3: MFCC coefficients

Delta

The characteristics extracted from the MFCC stage are a spectral representation of a short-term, semi-stable signal that describes the shape of the instantaneous spectral envelope of the speech signal. Since the speech signal is time-variable, so it needs a more accurate representation that describes the signal in general during its change over time. This is done by taking the first derivative (Delta) for the extracted features in the first stage by determining the difference in the extracted features between two different times, and this stage also produces 13 features for each frame [27]. Figure 4 shows Delta coefficients.

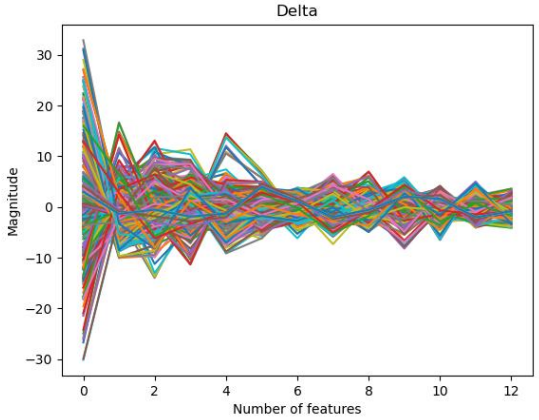


Fig. 4: Delta coefficients

2. Delta-delta

It is an implementation of the second derivative (Delta-Delta) of the extracted features and is used to represent the signal in a longer time context. This stage also produces 13 features for each frame. Both Delta coefficients and Delta-delta coefficients Calculated using the equation 7 below [27]. Figure 5 shows Delta-delta coefficients.

$$d(t) = \frac{c(t+1)-c(t-1)}{2} \quad \ldots.(7)$$

In the above equation, (t) is represented the delta coefficient from the t frame and is calculated in terms of the constant coefficients $(t + 1)$ and $(c - 1)$, and the delta-delta coefficients are calculated in the same way but they are computed from delta, not from the constant coefficients.

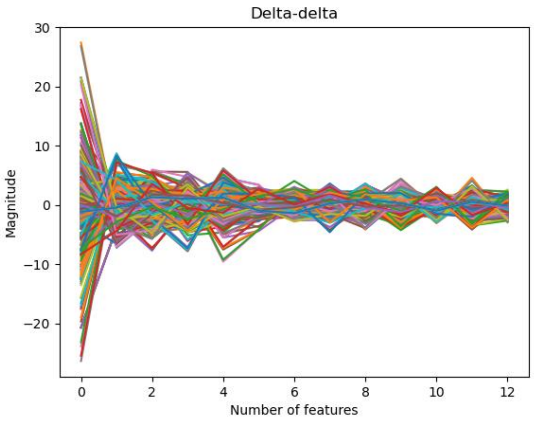


Fig. 5: Delta-delta coefficients

D. Classification

To implement the classification process and estimate the age and gender of the speaker, two models of deep neural networks are created, one for estimating the speaker age and the other for estimating the speaker gender. The classification was done using the bidirectional long-short (BiLSTM) term memory algorithm.

1. LSTM

The long-short term memory (LSTM) network is a type of Recurrent Neural Network (RNNs) that was introduced by Hochreiter and Schmidhuber in (1997). It is widely used in research areas concerned with serial data such as text, audio and video. However, Recurrent Neural Networks (RNNs) that consist of sigma cells or tanh cells are unable to learn information about the input data when the input gap is large. However, Long-Short Term Memory (LSTM) can deal with the dependency problem. It works well by introducing gate functions into the cell structure so LSTM becomes the focus of deep learning [15]. The LSTM differs from other types of RNNs in terms of the structure of the repeating pattern (internal structure). In RNNs there is a single neural network layer and in the LSTM there are four layers called gates interacting in a very special way, which is the forgetting gate, the input gate, the cell gate, and the output gate, as shown in Figure (6). Two states are transferred to each cell, which are the cell state and the hidden state. The cell state is the key to the LSTM, which is represented by the horizontal line at the top of the figure (6) and it passes straight through all the repetitive forms and the information is transmitted through it. The LSTM has the ability to delete and add information to the state of the cell through the three sigma gates that make the passage of information optional because the output of this gate is either the number 0 or the number 1, so the value 0 means no data passes and the value 1 allows passage all information. The equations (8), (9), (10), (11), (12) and (13) represent how LSTM works [24].

$$f_t = \sigma (W_f x_t + w_f h_{t-1} + b_f) \ldots(8)$$

$$i_t = \sigma (W_i x_t + w_i h_{t-1} + b_i) \ldots(9)$$

$$\check{C}_t = \tanh (W_c x_t + w_c h_{t-1} + b_i) \ldots(10)$$

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t \ldots(11)$$

$$o_t = \sigma (W_o x_t + w_o h_{t-1} + b_o) \ldots(12)$$

$$h_t = o_t * \tanh (C_t) \ldots(13)$$

Where x_t is the input at time t, h_t the output at time t, σ is the sigmoid function, W_f is the weight matrices, b_f is the bias of the forget gate, C_t is cell state, f_t is a vector with values ranging from 0 to 1, corresponding to each number in the cell state, Ct−1, o_t is the output of the sigmoid gate.

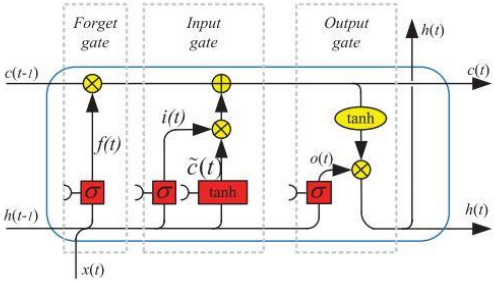


Fig. 6: LSTM block diagram [15]

2. BiLSTM

Bidirectional LSTMs are an extension classification method of the traditional LSTMs that can improve the model's performance in series classification problems. In cases where all the time steps of the input sequence are available, Bidirectional LSTMs train two networks instead of one. The first one works on the correct input sequence and the second works on the inverted copy of the input sequence. This training can provide additional network context and lead to faster and more complete learning of the problem [17]. Figure (9) illustrates a Bidirectional LSTMs network diagram.

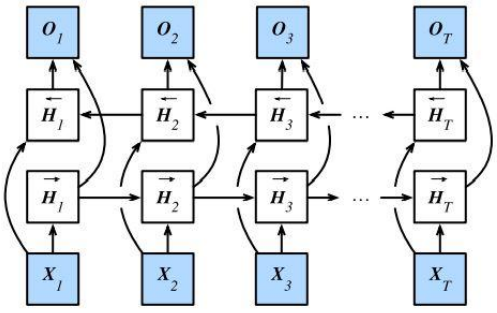


Fig. 9: Bidirectional LSTM block diagram [25]

The use of the bidirectional network will lead to the operation of the inputs in two ways, one from the past to the future and the other from the future to the past, in order to retain information from the past and the future as shown in Figure (8). This is done by adding two hidden cases the forward hidden sequence output $\overrightarrow{ht}$ is first computed, then the result of the back hidden sequence $\overleftarrow{ht}$ is computed; finally they are combined to generate the final product y_t. As shown in the following equations [28]:

$$\vec{h}_t = H(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}) \ldots..(14)$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}} \ldots..(15)$$

$$y_t = W_{\vec{h}y} \vec{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \ldots..(16)$$

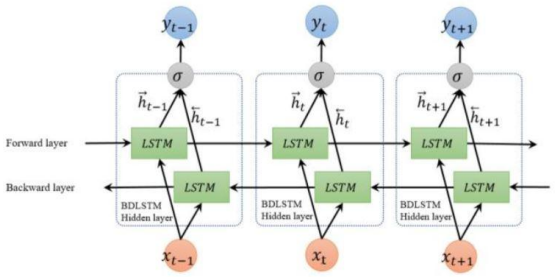


Fig. 8: Bidirectional LSTM [17]

E. The proposed system

A system of two deep neural network models is built, one for speaker age estimation, and the other for speaker gender estimation. In the proposed system, the two deep neural networks were trained with 25 training epochs and the system is built using Python 3.7 as a programming tool.

a. Deep neural network for speaker age estimation

The deep neural network used to estimate the speaker age as shown in Figure (7) consists of an input layer, an output layer, and six hidden layers with different number of units to obtain the best results in the prediction process. The input layer contains nodes with the number of network inputs and these inputs are three-dimensional data obtained from the process of features extracted by MFCC, Delta, and Delta-Delta as first dimension, second dimension and third dimension, respectively, in addition to the Labels data on which the network will be trained. These three-dimensional data are combined with each other to form network input.

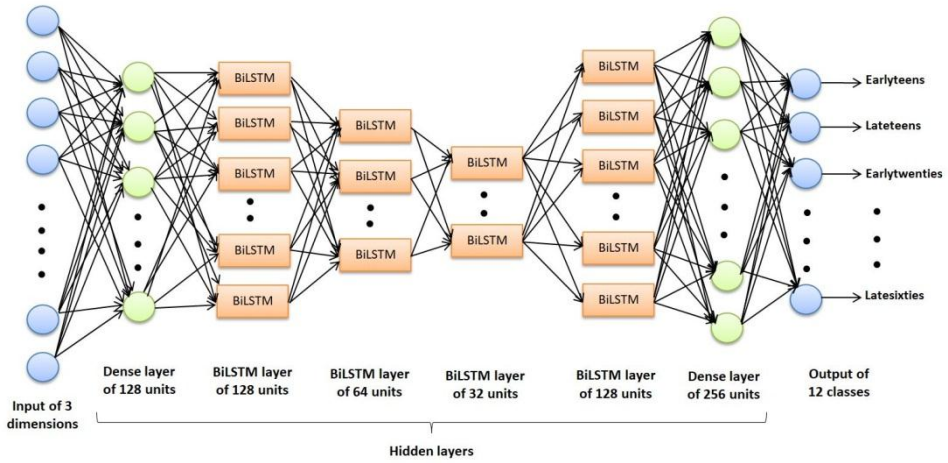


Fig. 7: Deep neural network for speaker age estimation using BiLSTM

The seven hidden have four BiLSTM layers, and two Dense (fully connected) layers that are very necessary in building the network as BiLSTM layers cannot be used only.

The output layer contains twelve nodes according to the number of age classes to be estimated for the speaker. This deep neural network model is described as a sequential model as it consists of sequential blocks and the network work is done by entering the input layer data into the first hidden layer (Dense layer) that consisting of 128 nodes, and the output of the first hidden layer is used as the input for the second layer (BiLSTM layer) that consisting of 128 nodes, and so on. finally comes the (Dense) output layer consisting of twelve nodes.

b. Deep neural network for speaker gender estimation

The deep neural network used to estimate the speaker gender as shown in Figure (8) consists of an input layer, an output layer, and five hidden layers. The input layer is similar to the input layer in the age estimation network in that it contains nodes with the number of input data with three dimensions in addition to the labels data. The five hidden layers have three BiLSTM layers and two Dense (fully connected) layers. The output layer contains two nodes that represent the network output, which is the gender of the speaker, either male or female. This deep neural network works in serial model similar to how the age estimation network works.

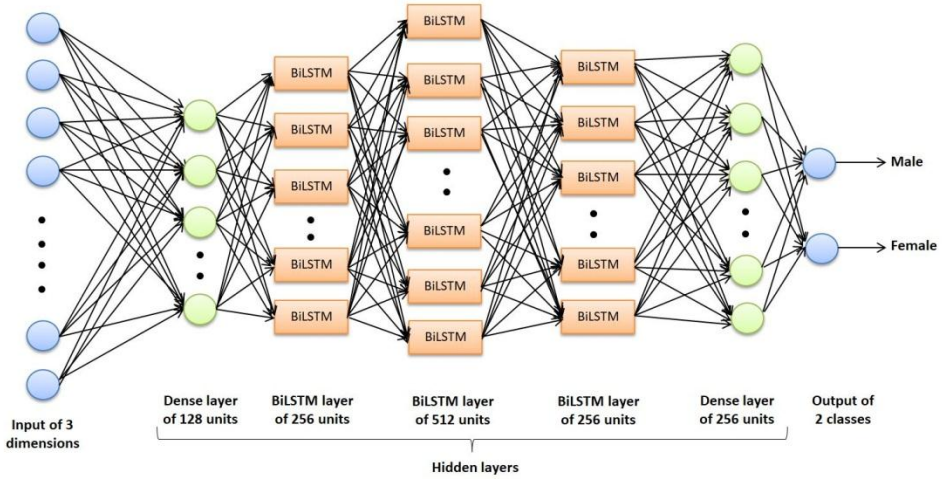


Fig. 8: Deep neural network for speaker gender estimation using BiLSTM

IV. Practical experiments

After building the two deep neural networks for age and gender estimation, the two networks are trained using the training dataset and the final weights are saved for use in the test phase using the test dataset. Some different experiments were conducted in terms of the length of the audio clip, as well as the batch size of the audio features for each training epoch of the network. Three experiments were conducted to test the effect of the audio clip length on the accuracy of the results of the two networks of estimation, which are four seconds, seven seconds and ten seconds, and four experiments were conducted to test the effect of batch size on the accuracy of the results of the two networks that are 25 features, 50 features, 75 features and 100 features. These experiments were carried out for each network separately.

The efficiency of the two networks of the proposed system was evaluated by using the Accuracy Scale, which is one of the metrics used to measure the efficiency of the deep artificial neural network. The accuracy ratios for each network were calculated by the following equations:

$$\text{Accuracy of age network} = \frac{\text{Number of correctly estimated speaker age clips}}{\text{The total number of clips}} \times 100 \quad (16)$$

$$\text{Accuracy of gender network} = \frac{\text{Number of correctly estimated speaker gender clips}}{\text{The total number of clips}} \times 100 \quad (17)$$

V. Results and Discussion

The results showed that the best accuracy was obtained in the case that the audio clip was 10 seconds in length and the size of the batch of audio features for each training epoch was 50 attributes for both networks. The accuracy rate for the age estimation network reached 94.008%, while the accuracy for the gender estimation network reached 90.816%, as the following two tables show:

Table 2: The accuracy results of the age estimation network

Batch Size	Length of audio clips in seconds		
	4	7	10
25	84.474	86.972	93.747
50	86.196	89.656	**94.008**
75	86.505	88.093	93.982
100	86.145	87.051	93.145

Table 3: The accuracy results of the gender estimation network

Batch Size	Length of audio clips in seconds		
	4	7	10
25	86.98	87.075	89.286
50	86.898	89.136	**90.816**
75	86.306	88.136	90.273
100	85.878	88.266	90.192

When the network training process begins, the accuracy ratios were low, then as the training progresses and the weights are adjusted for several training epochs, the accuracy rate increases gradually as shown in figure 9.

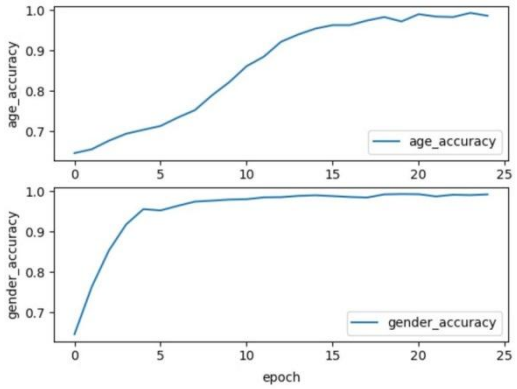


Fig. 9: The models accuracy ratios

VI. Conclusions

It is well known in the deep learning community that no general conclusions can be made about the effects of the parameters because the behavior often varies from dataset to dataset and model to model.

Therefore, the conclusions reached are for the functioning of our system only. In this work, two deep neural network models of estimating the age and the gender of the speaker have been built. The experimental results show that the best accuracy rate was obtained if the audio clip was of 10 seconds in length and the batch size of audio features for each training epoch was 50 features for both networks. Thus, it can be said that the length of the audio clip has a role in extracting more and stronger features to distinguish the voice of each speaker, so the accuracy rate in estimating the age and gender of the speaker was higher, but the length of the audio clip should not be very long because it will lead to an increase in the calculations and complexity as it takes longer time to prepare and train data. As for the size of the batch of features entering the network in each training epoch, we note that the batch size should not be large, as each training epoch is automatically divided into several steps, and in the case that the size of the batch of data is large, the number of steps per epoch will be small and the network will not be able By training on the required results with high accuracy, and at the same time the batch size should not be so small that there is a waste of time and a waste of resources.

References

[1] Alkhawaldeh, *R. S.* (2019). DGR: Gender recognition of human speech using one-dimensional conventional neural network. *Hindawi. Scientific Programming*, **(2019)7213717**:1-12.

[2] Sedaaghi, M. H. (2009). A comparative study of gender and age classification in speech signals. *Iranian Journal of Electrical & Electronic Engineering*, **(5) 1**:1-12.

[3] Erokyar H. (2014). Age and gender recognition for speech applications based on support vector machine. M.Sc. thesis, University of South Florida, USA: 69 pp.

[4] Younis H. A. (2011). Speaker age detection using eign values. M.Sc. thesis, University of Mosul, Mosul, Iraq: 101 pp.

[5] Bahari M. H., Hamme H. V. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. Workshop on Biometric Measurements and Systems for Security and Medical Application (BIOMS), Milan, Italy: p. 27-32.

[6] Přibil J., Přibilová A., Matouš J. (2017). GMM-based speaker age and gender classification in Czeck and Slovak. *Journal of Electrical Engineering*, **68(1)**: 3-12.

[7] Muda L., Begam M., Elamvazuthi I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing*, **2(3)**: 138-143.

[8] Kim H., Bae K., Yoon H. (2007). Age and gender classification for a home-robot service. 16th IEEE International Conference on Robot & Human Interactive Communication, 26 – 29 Aug, 2007, Korea. Jeju: p. 122-126.

[9] Faek F. K. (2015). Objective gender and age recognition from speech sentences. *ARO-The Scientific Journal of Koya University*, **III(2)** :06.

[10] Qawaqneh Z., Abu Mallouh A., Buket D. (2017). DNN-based models for speaker age and gender classification. The 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC), Porto, Portugal: p. 106-111.

[11] Piel, L. K. (2018). Speech-based identification of children's gender and age with neural networks. M.Sc. thesis, Tallinn University of Technology, Tallinn, Estonia: 85 pp.

[12] Osman H. M., Mustafa B. S., and Faisal Y. (2021). QDAT: A data set for reciting the quran. *International Journal on Islamic Applications Computer Science And Technology*, **9(1)**: 1-9.

[13] Hong Z. (2017). Speaker gender recognition system. M.Sc. thesis, University of Oulu, Oulu, Finland: 54 pp.

[14] Rehmam B., Halim Z., Abbas Gh., and Muhammad T. (2015). Artificial neural network-based speech recognition using DWT analysis applied on isolated words from oriental language. *Malaysian Journal of Computer Science*, **28(3)**: 242-262.

[15] Ranjan, R. and Thakur, A. (2019). Analysis of feature extraction techniques for speech recognition system. *International Journal of Innovative Technology and Exploring Engineering*, **(8)7C2**: 197-200.

[16] Muda L., Begam M., Elmavazuthi I. (2010). Voice recognition algorithms using mel frequency cepstral coefficients (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing*, **2(3)**: 138-143.

[17] Yusnita M. A., Paulraj M. P., Yaacob S., Yusuf R., Shahriman A. B. (2013). Analysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English. *International Journal of Automotive and Mechanical Engineering (IJAME)*, **7**: 1033-1073.

[18] Kulkarni A. G., Qureshi M. F., and Jha M. (2014). Discrete fourier transform: approach to signal processing. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, **3(10)**: 12341- 12348.

[19] Iliadi K. (2016). Bio-inspired voice recognition for speaker identification. Ph.D. thesis, University of Southampton, Southampton, United Kingdom: 203 pp.

[20] Yu Y., Si X., Hu C., Zhang J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, **31(7)**: 1235-1270.

[21] Shrestha A., Mahmood A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, **7**: 53040-53065.
[22] Apaydin H., et. al. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water Journal*, **12(1500)**: 1-18.
[23] Zhang A. , Lipton Z. C. , Li M. , Smola A. J. (2020). Dive into Deep Learning. E-book available from: https://d2l.ai.
[24] Basaldella M., Antolli E., Serra G., Tasso C. (2018) Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction. In: Serra G., Tasso C. (eds) Digital Libraries and Multimedia Archives. IRCDL 2018. Communications in Computer and Information Science, **806**, Springer, Cham.

تقدير عمر وجنس المتحدث استناداً الى التعلم العميق طريقة الذاكرة طويلة–قصيرة المدى ثنائية الاتجاه

الاء احمد محمد احمد[1] ، يسرى فيصل الارحيم[2]

[1]كلية الهندسة ، جامعة تكريت ، تكريت ، العراق

[2]قسم علوم الحاسوب ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

الملخص

لقد اكتسب تقدير عمر وجنس المتحدث أهمية كبيرة في السنوات الأخيرة لضرورته في مختلف التطبيقات التجارية والطبية والطب الشرعي. يقدر هذا العمل جنس المتحدثين وأعمارهم في نطاق صغير من السنوات حيث تم تقسيم كل عشر سنوات إلى فئتين فرعيتين لفترة سنوات تمتد من عمر المراهقة إلى الستينيات. يستخدم هذا النظام معاملات درجة النغم (MFCC) كطريقة لاستخراج صفات الصوت، وخوارزمية التعلم العميق الذاكرة طويلة–قصيرة المدى ثنائية الاتجاه (BiLSTM) كطريقة تصنيف. وتم بناء نموذجين لشبكتين عصبيتين عميقتين، واحدة لتقدير عمر المتحدث والثانية لتقدير جنس المتحدث. وأظهرت النتائج التجريبية أن النموذج المقترح لتقدير عمر المتحدث يحقق 94.008٪ كمعدل دقة و النموذج المقترح لتقدير جنس المتحدث يحقق 90.816٪ كمعدل دقة للشبكة العصبية العميقة.