

## Gender detection in children's speech utterances for human-robot interaction

Ameer Abdul-Baqi Badr<sup>1,2</sup>, Alia Karim Abdul-Hassan<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Technology, Baghdad, Iraq

<sup>2</sup>College of Managerial and Financial Sciences, Imam Ja'afar Al-Sadiq University, Salahaddin, Iraq

### Article Info

#### Article history:

Received Jul 27, 2021

Revised May 17, 2022

Accepted Jun 12, 2022

#### Keywords:

Children's gender identification

Formant average feature

Formant dispersion feature

Formant position feature

Logistic regression

### ABSTRACT

The human voice speech essentially includes paralinguistic information used in many real-time applications. Detecting the children's gender is considered a challenging task compared to the adult's gender. In this study, a system for human-robot interaction (HRI) is proposed to detect the gender in children's speech utterances without depending on the text. The robot's perception includes three phases: Feature's extraction phase where four formants are measured at each glottal pulse and then a median is calculated across these measurements. After that, three types of features are measured which are formant average (AF), formant dispersion (DF), and formant position (PF). Feature's standardization phase where the measured feature dimensions are standardized using the z-score method. The semantic understanding phase is where the children's gender is detected accurately using the logistic regression classifier. At the same time, the action of the robot is specified via a speech response using the text to speech (TTS) technique. Experiments are conducted on the Carnegie Mellon University (CMU) Kids dataset to measure the suggested system's performance. In the suggested system, the overall accuracy is 98%. The results show a relatively clear improvement in terms of accuracy of up to 13% compared to related works that utilized the CMU Kids dataset.

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



### Corresponding Author:

Ameer Abdul-Baqi Badr

Department of Computer Science, University of Technology

Baghdad, Iraq

Email: cs.19.53@grad.uotechnology.edu.iq

## 1. INTRODUCTION

The human being's sound consists of using the vocal cords to talk, laugh, cry, shout, and sing. Since it is the essential source of a sound, the human vocal cords play a major role in the conversation [1]. In addition to the content of a speech, a listener can understand several characteristics of a speech such as identity, accent, gender, emotional state, and age range [2]. Automatic recognition of this kind of speech characteristics can guide human-robot interaction (HRI) systems for adapting to various requirements of users in an automatic way [3], [4].

Identifying a child speaker's gender is a more difficult task than an adult speaker, it is not easy to recognize the gender of a speaking child. Their acoustic-phonetic properties have no significant differences due to the under-developed vocal tract as well as thin vocal folds in female and male children. Also, children have a tendency to mispronounce a few phonemes that are difficult for them to pronounce. Thus, that difficult phoneme is replaced by those children with another simpler phoneme. These replacement patterns are referred to as phonological processes [5]. Children are enthusiastic adopters of new technology. New technology can also give exciting chances to enhance children's lives, such as for educational and therapeutic

purposes. In telecommunication and speech therapy, determining the gender of a child speaker based on his or her speech is critical [5], [6]. Children's speech has distinct linguistic and acoustic characteristics from that of adults. Children's speech, for example, has a higher pitch, and perceptually important features like formants present at higher frequencies. As a result, bandwidth reduction has a greater impact on speech recognition accuracy in children's speech than in adults' speech [7].

Many factors influence the performance of gender identification systems. The input speech content can be text-independent or text-based. The use of a suitable classifier is another factor. On the other hand, selecting the feature set to use in the model is an essential part of the speaker's gender identification systems [8], [9]. A lot of the variability between the voices of children and adults is because of the difference in the vocal tract's length and vocal folds' mass. Concerning a certain vowel, such differences result in a considerable difference in formant frequencies and glottal-pulse rate (GPR) (i.e., the rate of closing and opening of vocal folds). The formant frequencies shift towards lower frequencies as the vocal tract lengthens. Between the ages of four and twelve (i.e., puberty), the length of a child's vocal tract increases steadily, while the formant frequencies decrease [10].

This study essentially relies on the formant features and expects their significance to accurately identify the gender of the children. The primary contributions of the present study can be highlighted and summarized: i) combining three formant-based features which are formant average ( $A_F$ ), formant dispersion ( $D_F$ ), and formant position ( $P_F$ ) to produce robust features dimensions and ii) taking advantage of the logistic regression classifier strength to detect the children's gender with the least possible latency.

The rest of this study is organized in the following way: the related works of the suggested system are presented in section 2. The dataset used is presented in section 3. Section 4 deals with the proposed method. The results of simulations and experiments are shown in section 5. Finally, section 6 sets out the study conclusions and future works.

## 2. RELATED WORKS

Given the difficulty of detecting the gender of children, the related works are rather few compared to the detection of the gender of adults. Safavi *et al.* [7] presented an approach to identify gender from children's speech. They identified the spectrum regions which contain significant children's gender information through carrying out their system experiments across 21-frequency sub-bands. They used gaussian mixture model-support vector machine (GMM-SVM) and gaussian mixture model-universal background model (GMM-UBM) in their system. Their experiments have been conducted on the OGI Kids corpus; their best result was a 79.18% identification rate. Ramteke *et al.* [5] presented an attempt for exploring the features effective in differentiating the gender from the speech of children. Also, they examined various spectral features' combinations, like mel-frequency cepstral coefficients (MFCCs) as well as its 1st and 2<sup>nd</sup> derivate, linear predictive cepstral coefficients (LPCCs), formants, jitter, shimmer and prosodic features like pitch and its statistical variations. Their features were evaluated using non-linear classifiers, which are random forest (RF), artificial neural networks (ANNs), and deep neural networks (DNNs). Their experiments have been conducted on the Carnegie Mellon University (CMU) Kids corpus; their best accuracy was about 84.79 % when an RF classifier has been used. The children-based HRI related works are also few compared to adult-based HRI. Kruijff-Korbayova *et al.* [6] presented a conversational system for child-robot interaction. Their system involves components for interpretation, recognition, and generation of speech, user modeling, and dialogue management. Also, they conducted 3 game-like activities that a child can undertake: Quiz, in which a robot and a child ask each other some multiple-choice questions, and provide assessment feedback. In addition to imitation, where either the robot or child presents some simple arm poses which the other attempts to imitate and memorize. The third game-like activity is dance, where a series of dance movements are learned by the child via the robot. Sandygulova *et al.* [11] presented a robotic system for gathering 3-D body metrics and using them for effectively estimating the gender and age of previously-unseen children in real-world situations. They evaluated the performance of the systems on 428 children volunteers; they showed that even a small number of biometrics might achieve excellent gender and age estimation results in perceptually challenging environments.

## 3. THE DATASET USED

CMU Kids corpus is the database applied in the presented study; it contains sentences in the English language that are read aloud by children of both genders [12]. Originally, the database has been developed for creating a training set of children's speeches for SPHINX-II automatic speech recognizer within LISTEN project at Carnegie Mellon University (CMU). There is a total of 818 audio records. The ages of the children ranged between (6 and 11) years. In this study, the same number for male and female speakers was chosen,

which is 19. In total, the number of utterances that have been used in this work was 250 for males, and 250 for females.

#### 4. PROPOSED METHOD

As shown in Figure 1, the proposed robot's perception consists of three main stages: features extraction, features standardization, and semantic understanding (i.e., speaker gender detection). Initially, formant features are extracted from each children's utterance, followed by features scaling to fall within a smaller range using standardization technique (i.e., z-score). Finally, a logistic regression classifier is used to detect the children's gender. The proposed robot's action is specified via a speech response using the text to speech (TTS) technique.

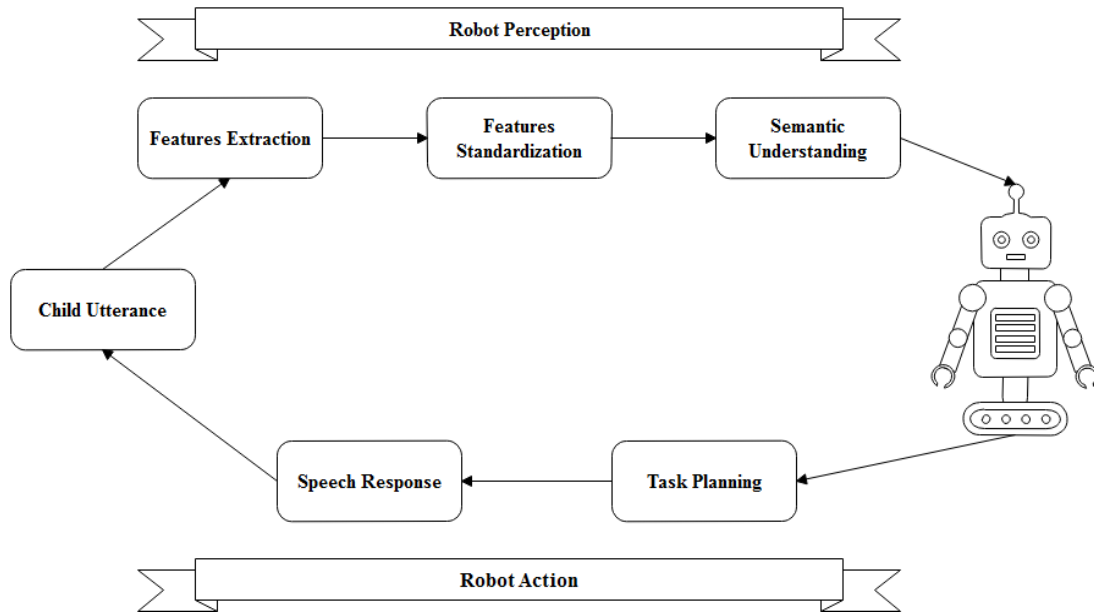


Figure 1. The general simulation framework of the proposed system

##### 4.1. Features extraction based on formants

With different vocal tract configurations which correspond to various resonances, the formant's frequencies change. It is possible to see the difference in formant frequencies between adult females and males. The formant frequencies' values decrease as the vocal tract length increases. Both male and female adults have higher formant frequencies compared to children [5], [13], [14]. Formants were only measured at the glottal pulse to make the measurement easier along with the whole utterance. As a result, instead of measuring individual vowels, samples are taken from a wide range of vocal tract configurations. Furthermore, because the sound source of fricatives is turbulence in the mouth rather than vocal fold vibration, this method sampled only voiced speech and avoided fricatives, which artificially reduce apparent vocal tract length [15]. The formants extraction is done using the Praat software [16]. As shown in Figure 2, four formants (i.e., F1, F2, F3, and F4) are measured at each glottal pulse and then a median is calculated across these measurements. After that, three types of features are measured which are formant average ( $A_F$ ) as given in (1) [17], formant dispersion ( $D_F$ ) as given (2) [18], and formant position ( $P_F$ ) as given in (3) [15].

$$A_F = \frac{\sum_{i=1}^n F_i}{n} \quad (1)$$

$$D_F = \frac{\sum_{i=1}^n (F_{i+1} - F_i)}{n-1} \quad (2)$$

$$P_F = \frac{\sum_{i=1}^n \tilde{F}_i}{n} \quad (3)$$

Where  $F_i$  is the  $i_{th}$  formant,  $\tilde{F}_i$  is the standardized  $i_{th}$  formant, and  $n$  is the number of formants measured.

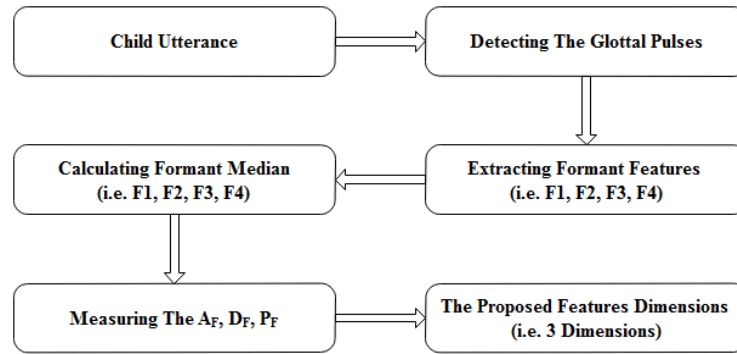


Figure 2. The proposed features extraction method

#### 4.2. Features standardization based on the z-score method

The expression of features in smaller units will result in a wider range for these features and thus will tend to give such features a greater effect. In the z-score method, the values for each feature are standardized based on their mean and standard deviation as given in (4) [19], [20]. Therefore, due to the great usefulness of the standardization process in machine learning methods, the 3-dimensional features, extracted from the previous stage, will be standardized using the z-score method.

$$z = \frac{x_i - \mu}{\sigma} \quad (4)$$

Where,  $x_i$  is the feature vectors,  $\mu$ ,  $\sigma$  are the mean and standard deviation, respectively, of the  $x_i$ .

#### 4.3. Semantic understanding based on logistic regression classifier

Logistic regression is considered very accurate and reliable among most statistical methods. In this model, the dependent variables are predicted by the independent variables. A binary format is used for the dependent variable, while independent variables might be evaluated on an ordinal, nominal, or ratio scale. Despite logistic regression being based on various assumptions as to the relation of dependent-independent variables, logistic regression is considered a special case of a linear regression model. Logistic regression conditional distribution is a Bernoulli distribution instead of Gaussian distribution because the dependent variable has a binary variable form [21]–[23]. The relationship between the occurrence and its dependency on several variables in logistic regression analysis might be expressed via (5) [21].

$$p = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n)}} \quad (5)$$

Where  $p$  is the occurrence probability,  $b_0$  represents the model's intercept,  $b_i$  ( $i=0, 1, 2, \dots, n$ ) represents the slope coefficients related to logistic regression model, while  $x_i$  ( $i=0, 1, 2, \dots, n$ ) are the independent variables.

There are several strong points when using the logistic regression classifier, i.e., it is easy to implement, very fast at classifying unknown data, performs well when the input data is linearly separable, and can be regularized to avoid overfitting. Therefore, to take advantage of the logistic regression classifier strength (i.e., fast), the logistic regression classifier is trained on the previous stage output (i.e., 3-dimensional standardized features) to detect the children's gender accurately.

#### 4.4. Speech response based on TTS method

TTS synthesis can be defined as one of the systems by which normal language text is converted into speech. There are many differences between machine speech production and human, however, the increase in the capability of machine learning paradigms for simulating human speech production mechanisms will result in a more natural and accurate TTS [13], [24]. In this study, the pytt3 library [25] has been used for TTS synthesis as a robot's speech response.

### 5. EXPERIMENTAL RESULTS AND DISCUSSION

Different experiments are conducted to find the optimal configuration of the proposed gender detection system parameters. The average latency time is 0.4465 seconds. In the first experiment, the performance evaluation of the proposed gender detection system in terms of precision, recall, F-score, and

accuracy is conducted. This experiment also shows the influence of data training size on the proposed system. Table 1 shows the results of this experiment on the CMU Kids dataset. As shown in Table 1, the effectiveness of the suggested system is evaluated using four measures: precision, recall, F-score, and accuracy. All these measures are applied to different sizes of training data; they are 80%, 66%, and 50%. The F-score of the proposed system for male children is almost equal to the F-score of the female children for all sizes of training data and that because of using a balanced number of utterances. On the other hand, the overall accuracy of the proposed system is affected slightly by the size of training data, as it decreased from 98% when the training size is 80% to 97.20% when the training size becomes 50%. That demonstrates the robustness of the proposed feature dimensions for the proposed gender detection system. The second experiment shows the impact of each feature on the performance in terms of accuracy. In addition, it shows the impact of the combined features on the performance of the proposed system in terms of accuracy. Table 2 shows the results of this experiment on the CMU Kids dataset (i.e. test size is 20%).

Table 2 shows the impact of each of the features on the performance of the proposed system in terms of accuracy. The table demonstrates that the prediction errors from these different features are complementary. This will make the predictions from these features to be combined to further improve the results as shown in the table. Finally, a comparison of the proposed system with related works is presented in the third experiment. It utilizes the children's gender detection issue in terms of accuracy. Table 3 shows the results of this experiment. Table 3 demonstrates the clear superiority of the proposed children's gender detection system. In spite of the minimum number of feature dimensions, the proposed system taking advantage of using features fusion based on formants, and standardization technique. In addition to using the logistic regression as a classifier.

Table 1. The performance evaluation of the suggested system on CMU Kids dataset

Train Size	Test Size	Male Children (%)			Female Children (%)			Accuracy (%)
		Precision	Recall	F-Score	Precision	Recall	F-Score	
80%	20%	100	95.35	97.62	96.61	100	98.28	98
66%	33%	98.88	96.70	97.78	96.05	98.65	97.33	97.57
50%	50%	99.19	95.35	97.23	95.24	99.17	97.17	97.20

Table 2. The feature's impact on the proposed system performance in terms of accuracy on the CMU Kids dataset

Feature Type	Accuracy (%)
Formant average ( $A_F$ )	63
Formant dispersion ( $D_F$ )	56
Formant position ( $P_F$ )	44
The combined features	<b>98</b>

Table 3. Accuracy-based comparative study between the proposed system and related works

Authors	Database used	Number of feature dimensions	Accuracy (%)
Safavi <i>et al.</i> [7]	The OGI Kids Corpus	57	79.18
Ramteke <i>et al.</i> [5]	The CMU Kids Corpus	68	84.79
The proposed system	The CMU Kids Corpus	<b>3</b>	<b>98</b>

## 6. CONCLUSION AND FUTURE WORKS

An automatic system to detect gender in children's speech utterances without depending on the text is proposed in this study. Three formants-based features (i.e.  $A_F$ ,  $D_F$ ,  $P_F$ ) are combined to further improve system performance. Then, the 3-dimensional features are standardized using the z-score method. Finally, the proposed system detects the children's gender using the logistic regression classifier. The experimental results clearly show the effectiveness of the proposed system with an accuracy of 98% using the CMU Kids dataset. For future work, the same proposed feature vectors can be used for an adult's gender detection to see their ability in cross-language condition.

## REFERENCES




- [1] P. Kumar, P. Baheti, R. Kumar Jha, P. Sarmah, and A. Professor, "Voice gender detection using gaussian mixture model," *Journal of Network Communications and Emerging Technologies (JNCET)*, vol. 8, 2018.
- [2] E. Yucesoy and V. V. Nabiyev, "Gender identification of a speaker using MFCC and GMM," in *2013 8th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov. 2013, pp. 626–629., doi: 10.1109/ELECO.2013.6713922.
- [3] M. Li, C.-S. Jung, and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," in *Interspeech 2010*, Sep. 2010, pp. 2826–2829., doi: 10.21437/Interspeech.2010-747.

*Gender detection in children's speech utterances for human-robot interaction (Ameer Abdul-Baqi Badr)*




- [4] A. A. Badr and A. K. Abdul-Hassan, "Estimating age in short utterances based on multi-class classification approach," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1713–1729, 2021, doi: 10.32604/cmc.2021.016732.
- [5] P. Bhaskar Ramteke, A. A. Dixit, S. Supanekar, N. V. Dharwadkar, and S. G. Koolagudi, "Gender identification from children's speech," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Aug. 2018, pp. 1–6, doi: 10.1109/IC3.2018.8530666.
- [6] I. Kruijff-Korbayova *et al.*, "Spoken language processing in a conversational system for child-robot interaction," in *14<sup>th</sup> Annual Conference of the International Speech Communication Association*, Lyon, France, 2012, pp. 2440–2444.
- [7] S. Safavi, P. Jančovič, M. Russell, and M. Carey, "Identification of gender from children's speech by computers and humans," in *Interspeech 2013*, Aug. 2013, pp. 2440–2444, doi: 10.21437/Interspeech.2013-567.
- [8] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Applied Acoustics*, vol. 98, pp. 52–61, Nov. 2015, doi: 10.1016/j.apacoust.2015.04.013.
- [9] G. Alipoor and E. Samadi, "Robust gender identification using EMD-based cepstral features," *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 7, no. 1, pp. 71–81, Jun. 2018, doi: 10.17576/apjtm-2018-0701-06.
- [10] D. R. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3177–3186, Nov. 2005, doi: 10.1121/1.2047107.
- [11] A. Sandygulova, M. Dragone, and G. M. P. O'Hare, "Real-time adaptive child-robot interaction: age and gender determination of children based on 3D body metrics," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Aug. 2014, pp. 826–831, doi: 10.1109/ROMAN.2014.6926355.
- [12] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids speech corpus LDC97S63," *Linguistic Data Consortium*, 1997.
- [13] A. Badr and A. Abdul-Hassan, "A review on voice-based interface for human-robot interaction," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 16, no. 2, pp. 1–12, Dec. 2020, doi: 10.37917/ijeee.16.2.10.
- [14] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, 2021, doi: 10.1016/j.ins.2021.02.016.
- [15] D. A. Puts, C. L. Apicella, and R. A. Cárdenas, "Masculine voices signal men's threat potential in forager and industrial societies," *Proc. of the Royal Society B: Biological Sciences*, vol. 279, no. 1728, pp. 601–609, 2012, doi: 10.1098/rspb.2011.0829.
- [16] V. van Heuven and P. Boersma, "Speak and unSpeak with PRAAT," *Glott International*, vol. 5, no. 9–10, pp. 341–347, 2001.
- [17] K. Pisanski and D. Rendall, "The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2201–2212, Apr. 2011, doi: 10.1121/1.3552866.
- [18] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *The Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, Aug. 1997, doi: 10.1121/1.421048.
- [19] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan Kaufmann, 2012, doi: 10.5860/CHOICE.49-3305.
- [20] K. Cho, T. Yoon, S. Park, and D. Park, "Using standardization for fair data evaluation," in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, Feb. 2018, pp. 585–589, doi: 10.23919/ICACT.2018.8323842.
- [21] P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size," *CATENA*, vol. 145, pp. 164–179, Oct. 2016, doi: 10.1016/j.catena.2016.06.004.
- [22] A. Jacob, "Modelling speech emotion recognition using logistic regression and decision trees," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 897–905, Dec. 2017, doi: 10.1007/s10772-017-9457-6.
- [23] F. Salehi, E. Abbasi, and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.
- [24] V. Delić *et al.*, "Speech technology progress based on new machine learning paradigm," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–19, Jun. 2019, doi: 10.1155/2019/4368036.
- [25] V. Mittal, "FALCON (personal assistant)," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 6, pp. 404–408, Jun. 2020, doi: 10.22214/ijraset.2020.6061.

## BIOGRAPHIES OF AUTHORS



**Ameer Abdul-Baqi Badr**    received the B.Sc. degree and the M.Sc. degree in System Software from Computer Science Department, University of Technology, Baghdad, Iraq, in 2014 and 2018 respectively, and the Ph.D. degree in artificial intelligence (AI) from Computer Science Department, University of Technology, Baghdad, Iraq, in 2021. Currently, he is a lecturer at Imam Ja'afar Al-Sadiq University, Salahaddin, Iraq. He has authored or co-authored more than 10 refereed journal and conference papers. His research interests include AI, machine learning, speech processing, speech enhancement, speech recognition, speaker recognition and verification, and, voice-based HRI. He can be contacted at email: amir.abdulbaqi@sadiq.edu.iq.



**Alia Karim Abdul-Hassan**    received the B.Sc. degree, the M.Sc. degree and the Ph.D. degree from Computer Science Department, University of Technology, Baghdad, Iraq, in 1993, 1999 and 2004 respectively. She is working as a Dean of Computer Science Department, University of Technology, since Feb 2019 till now. She was supervised on more than 40 M.Sc. and Ph.D. thesis in Computer Science since 2007. Her research interests include soft computing, green computing, artificial intelligence (AI), data mining, software engineering, electronic management, and computer security. She can be contacted at email: 110018@uotechnology.edu.iq.