Enhancement and modification of automatic speaker verification by utilizing hidden Markov model

Imad Burhan Kadhim, Ali Najdet Nasret, Zuhair Shakor Mahmood
Electronic Department, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq

Article Info

Article history:

Received Nov 1, 2021
Revised May 17, 2022
Accepted Jun 20, 2022

Keywords:

French data set
Generalized method of moments
Text dependent
Viterbi algorithm

ABSTRACT

The purpose of this study is to discuss the design and implementation of autonomous surface vehicle (ASV) systems. There's a lot riding on the advancement and improvement of ASV applications, especially given the benefits they provide over other biometric approaches. Modern speaker recognition systems rely on statistical models like hidden Markov model (HMM), support vector machine (SVM), artificial neural networks (ANN), generalized method of moments (GMM), and combined models to identify speakers. Using a French dataset, this study investigates the effectiveness of prompted text speaker verification. At a context-free, single mixed mono phony level, this study has been constructing a continuous speech system based on HMM. After that, suitable voice data is used to build the client and world models. In order to verify speakers, the text-dependent speaker verification system uses sentence HMM that have been concatenated for the key text. Normalized log-likelihood is determined from client model forced by Viterbi algorithm and world model, in the verification step as the difference between the log- likelihood. At long last, a method for figuring out the verification results is revealed.

Corresponding Author:

Ali Najdet Nasret
Electronic Department, Kirkuk Technical Institute, Northern Technical University
Kirkuk, Iraq
Email: alinajdet@ntu.edu.iq

1. INTRODUCTION

In today's world, there are numerous uses for methods for automatically recognizing individuals with security, enterprise-safe electronic banking and o financial activities; law enforcemen, health , counter-terrorism measures retail sales and social services are some examples of physiological or behavioral features found in the field of homeland security. Technology based on biometrics [1] are already being used by the majority of these organizations. Probability distributions of random variables, states, which can take on values from a set, can be predicted using the Markov chain model [2], [3]. Words, tags, and symbols, such as those used to depict the weather, can all be included in these collections. A Markov chain assumes that if we wish to forecast the future, the current state is the most important factor. Prior to the existing situation, no one can predict what would happen in the future. An authentication system that uses biometrics must compare a previously registered biometric sample with a newly acquired biometric sample [4], [5]. A biometric sample is collected, analyzed, and saved for comparison later on during registration. A person's claimed identity may be verified or authenticated using recognition in either identification or identification mode (the system determines the best match from the full enrolled population) are examples of this [6], [7]. Since the 1970s, techniques of speaker recognition have been deemed "classic" in the field of biometrics [8], [9]. The acoustic characteristics of each speaker are extracted by speaker identification systems from the spoken stream [10]-

[12]. The following characteristics are reflected in: i) Anatomy: the geometrical and size forms of the voice lips, velum, lungs, teeth, tongue and lips cords [13]; ii) learned behavioral patterns: the manner in which one speaks [14]; Learning: the manner in which one speaks [9].

Speech signal processing include both verification and identification in speaker recognition. An individual is verified as the person they claim to be by using speaker verification. A speaker identification system determines whether or not a speaker belongs to a given individual or group [15]-[17]. In speaker verification, a claim of identification is made by a person . If the phrase is recognized using text dependent recognition, the system will recognize it regardless of its phonetic content, whereas using text independent recognition, the phrase must be recognized regardless of whether it is prompted visually or vocally [18], [19].

This article makes advantage of a previously developed continuous speech recognizer to construct an autonomous surface vehicle (ASV) system that relies on text dependence, high quality speech and cooperative speakers. To verify a speaker's identity, the process usually involves a few simple steps: the claimant speaks a phrase into a microphone, which is picked up by the system, which decides whether to reject or accept or wither to report a lack of confidence or reject the claim and ask for more speech before making a decision [7], [20].

2. PROPOSED METHOD IN ASV SYSTEM DEPENDENT ON TEXT

This study focuses on the text-dependent system constructed on top of the prior continuous speech recognition (CSR) system. Next section describe depth on the french CSR data set. It utilizes context-free, single-mixture hidden Markov model (HMM) at the level of monophone to generate the final product [8], [21]. In order to develop an automatic speech verification system for a certain speaker (client), there are five stages to take:

a. Firstly training a speaker independent automatic speech recognizer (SIASR), all of the database's accessible phrases (training set) are employed; client sentences are omitted from the training pool. Training the global model.
b. To train a speaker dependent automatic speech recognizer (SDASR), all the sentences acquired from the client are employed. This enrolling operation is carried out for each new client.
c. Test phrases are all aligned using the Viterbi forced alignment process, and two acoustics scores are produced for each phrase (observation):
   - SD-ASR logP (O|λSD) Log-likelihood calculated.
   - SI-ASR log P (O|λSI) Log-likelihood calculated.

   where O is the observation sequence of the sentence, we get λSI and λSD which are speaker-specific HMMs.
d. Normalization of the acoustic score-the normalized score is calculated as shown in for the robustness of recognition:

$$L(O) = \log P\ (O|\lambda_{SD}) - \log P(O|\lambda_{sr}) \tag{1}$$

e. Test phrase set performance assessment includes determining the false rejection rate (FRR) and false acceptance rate (FAR) for a specified threshold T range .

The Figure 1. represent the system of automatic speech verification has been used in this analysis. Assume that $p_A(z|H_0)$ is the conditional density function of the other speakers' score $z(H_0 \text{hypothesis})$ and that $p_A(z|H_1)$ is the conditional density function of the claimed speaker's score $z(H_0 \text{hypothesis})$ $(H_1 \text{hypothesis})$. The Bayes test with identical misclassification costs for speaker A is based on the likelihood ratio if both con- ditional density functions are known $\lambda_A(z)$ [22].

$$\lambda_A(z) = \frac{p_A(z|H_0)}{p_A(z|H_1)} \tag{2}$$

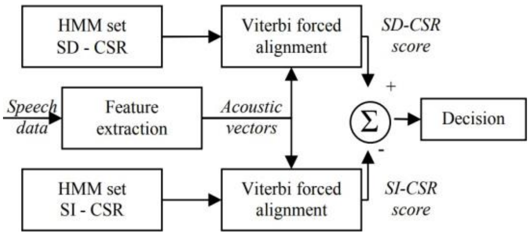

Figure 1. The architecture of the ASV system

By comparing the overlapping areas of the two cumulative distribution function (CDF), the error probability is reduced to its bare mini- mum (CDF). The ASV system's chance of inaccuracy decreases with decreasing surface area. Experimentation may be used to estimate the unknown density functions (for the client speaker and the other speakers):

a. Given the client speaker A, the conditional pdf $p_A(z|H_1)$ is calculated using the client speaker's acoustic scores, using its own model;

b. Other speakers' scores are calculated using the speakerA′smodel, and this is the condition pdf supplied to impostors $p_A(z|H_0)$.

So, we can now calculate the speaker's likelihood ratio $\lambda_A(z)$. When the threshold T is selected, classification returns to determining the classification, and the decision rule is changed to;

$$\lambda_A(z) = \begin{matrix} \geq T\ choose\ H_0 \\ \leq T\ choose\ H_1 \end{matrix}$$

the criterion used to define threshold T is as;

c. Aiming for a minimal error performance by setting T to the user's a priori probability of being an imposter vs the actual speaker.

d. selecting T in order to meet a certain FA or face recognition (FR) requirement;

e. adjusting T until the required FR/FA ratio is achieved.

Threshold T may be customized for each individual customer in a database, or it can be dynamically adjusted based on a variety of factors. FR and FA mistakes each have a range of possible values. In other words, as the threshold rises, so do the number of bogus acceptances and denials. FA and FR mistakes are typically balanced based on a threshold value [23].

3. WORD RECOGNITION FOR THE LONG TERM

Based on an earlier french language speech recognition system, the ASV system uses a continuous speech recognizer. Thesaurus for the English language. in this study have been made use of a database that was created in an office setting. More than ten hours of read speech by 11 different speakers are included in this set (8 males and 5 females). A total of 4100 phrases and over 3000 popular terms from different disciplines such as education, athletics, politics, and so on are included in the texts. There are two sets for each speaker: training and testing purposes.

Extrapolation of characteristics. Using a cardio id desktop microphone (32 Hz 8 kHz) and a standard 32-bit PC sound card with 8 kHz/8-bit sampling, SNR=31 dB was used to record the waveforms. The waveform was pre-emphasized with a coefficient of 0.91 before being parameterized using cepstral algorithm on a 25 ms (HM) Hamming window to get the final outcome (overlapping 41 percent). The first and second derivatives of 13 fundamental Mel-frequency cepstral coefficients (MFCCs) have been extracted and added (a total of 37 parameters) [24], [25].

Models based on the sound waves. Using a single mixed Gaussian continuous distribution, a three-state HMM was used to describe each french phoneme. The covariance matrices are diagonal to save on computing re- sources while calculating the output probability. All models are created identically to begin with, and then the Baum-Welch process is used, which is also known as embedded training. We calculate the global speech mean and covariance (across all training utterances), and we utilize these values to initialize the whole collection of HMMs, all of which are identical. Embedded training is also utilized to distinguish between models. A composite model is created for each training phrase by concatenating distinct models based on phrase labels and a phonetic dictionary.

The following are the major stages in the context-free models' training procedure:

a. Initialization of all HMMs is the same.

b. Re-estimating the Baum-Welch parameter takes 4 to 6 iterations for composite models (with a convergence criterion of 0.01 in log-likelihood).

c. When numerous pronunciations are present in the training lexicon, Viterbi forced alignment is used to choose the candidate with the highest alignment score.

d. Re-estimation of the Baum-Welch parameter takes 4 to 6 iterations to complete.

4. RESULT AND DISCUSSION

A random speaker from the pool of speakers was selected to be the subject of the studies. A speaker dependent continuous speech recognizer was used to create a client model for this speaker (SD-CSR3). Next, a world model was created for the database's other speakers, using a speaker-independent continuous speech recognition system (CSR-SI). A client model and a world model both have a similar set of structural and acous- tical properties; only the training data distinguishes between the two models.

The phrase being tested is aligned using a Viterbi forced alignment process in the verification step, and the normalized score is calculated for both systems in (1). Figure 1 shows the ASV architecture. Then, using the difference between the normalized score and threshold, a judgment is made. Take a look at some of the results in Table 1 for the test phrase ”S0001,” which was said by the customer (client) and an imposter(speaker #1) who were identically dressed (male). There is a difference between the normalized scores for the client and imposter phrases, with the client scoring more than zero and the impostor scoring lower.

To evaluate the ASV system's capabilities, a normalized score is generated for each of the test phrases. When determining the grade, the average word score for a certain test phrase is used. The normalized score from the test sentences may be shown in Figure 2. Each set of 300 speaker phrases consists of phrases from all of the speakers in the group. Speaking of speaker #3 (client), the phrase indices vary from 601 to 900 for example. The fact that client phrases score higher than those of other speakers means we can use it as a criterion for accepting or rejecting a proposal.

Table 1. For the same sentences, impostor and client bought acoustic degree resultantly

	Acoustic score Z (log-likelihood)					
	client			Speaker (impostor		
phrase	SD score	SI score	Norm score	SD score	ST score	N score
ARI	−59.61	−73.27	−12.65	−64.22	−63.11	1.09
ODA	−60.61	−63.87	−2.25	−66.22	−60.22	5.15
SAR	−62.26	−65.66	−2.39	−68.81	−64.01	3.79
VELM	−59.58	−66.79	−6.52	−61.31	−58.20	2.10
DA	−55.58	−56.46	−0.77	−64.35	−60.24	3.10
RDCI	−64.12	−66.96	−1.84	−69.20	−61.42	6.77

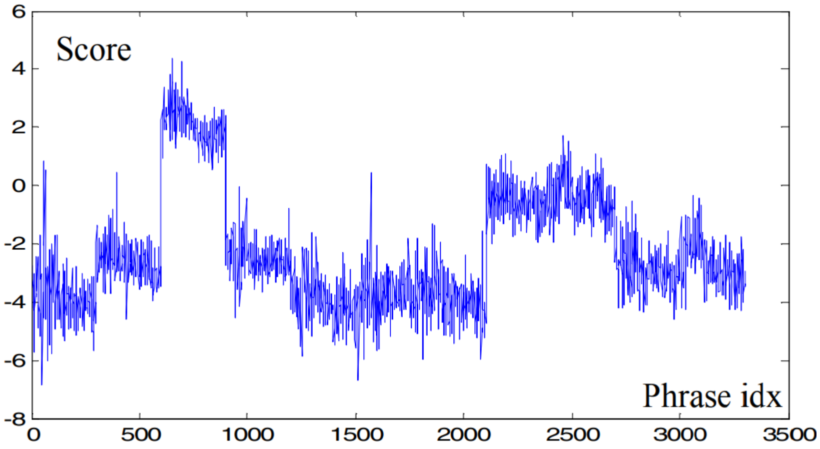

Figure 2. Calculated average score for all of the test-phrases

The following actions are used to assess system performance:

a. Analyze ASV system results for speaker #3 for similar speaker phrases to estimate probability function $P_3(z\lambda H_1)$.

b. Speaker #3's score on the other speaker phrases was used to estimate probability function $P_3(z\lambda H_0)$. With this information, we can now compute the probability function distributions (pdf) for speaker $P_3(z\lambda H_0)$, as well as speaker $P_3(z\lambda H_1)$.

c. Using T as a judgment threshold, determine the false rejection rate (FRR).

$$FRR = (T) = \frac{N_C(T)}{N_{TC}} \qquad (3)$$

ASV rejects client phrases if T is more than or equal to NC(T), where NT C is the total number of client phrases.

d. Using T as a judgment threshold, calculate the false acceptance rate (FAR):

$$FAR(T) = \frac{N_1(T)}{N_{Tr}} \quad (4)$$

where T is the threshold at which the ASV system accepts the imposter phrases $N_1(T)$, and N_{TI} is the total amount of impostor phrases N_{TI};

e. Calculate the values of both FAR (T) and FRR (T) (Figure 3). The lowest and highest score achieved across the whole phrase test set restrict the variation range of the threshold T.

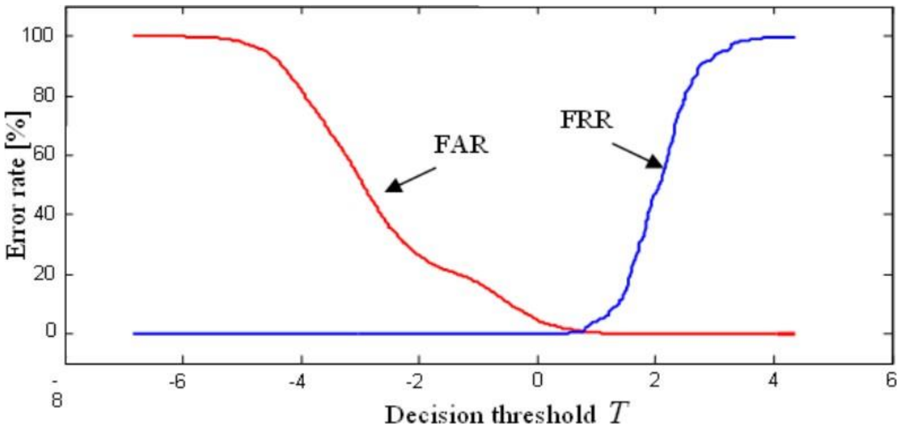

Figure 3. both the FAR and the FRR

f. Calculate T based on either FAR or FRR criteria, or a combination of the two. In Table 2, numerous criteria are presented together with their respective error rates and thresholds. It's possible to see that a higher threshold $(T = 1.75)$is required to achieve a 0 false acceptance rate, resulting in a large percentage of false rejections (FRR percent=1). It's possible that this is the case for access applications since the system shouldn't let phonies through. The number of tries might be raised in order to reduce the FRR. Each speaker model was put to the test against the phrases of the other speakers in order to get a sense of how well it performed. It was decided not to integrate the client's terms in the global model.

Table 2. Threshold, far and farthest-reaching criteria for various decision-making

Criterion	Threshold	FRR [%]	FAR [%]
Minimum FRR	0.51	1.45	0.0
Minimum FRR x FAR	0.76	0.5	0.80
Minimum FAR	1.75	0.0	42

5. CONCLUSION

This study proposes a continuous speech recognizer-based automated speech verification system. Two recognizes are learned in the training stage: one for the client model is reliant on the speaker, while the other is not dependent on the speaker for the world model. Each of the two recognizes uses the same set of parameters and has the same structure. A forced alignment Viterbi algorithm is used at the input phrase verification step. You'll need to compare the normalized acoustic score to an acceptance threshold to see whether it meets the requirements.

This method's experimental findings show that it can achieve error rates of less than 1%. Using a forced alignment approach reduces recognition time and hence computing cost since the search space limitation is taken into consideration. The experiments used a 1,5 GHz CPU, with a mean pro- cessing time of less than 1 second. Even though the verification device is inexpensive to deploy, significant resources are required during the enrollment phase to build up the client's speaker-dependent recognizer from scratch. This problem may be overcome by using standard adaptation methods such as MAP or MLLR to modify world model parameters to the new speaker.

REFERENCES

[1] F. Bimbot *et al*., “A tutorial on text-independent speaker verification,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, Dec. 2004, doi: 10.1155/S1110865704310024.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, Jul. 2008, doi: 10.1109/TASL.2008.925147.

[3] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, p. 101027, Mar. 2020, doi: 10.1016/j.csl.2019.101027.

[4] S. B. B. Rodzman, N. K. Ismail, N. A. Rahman, S. A. Aljunid, Z. M. Nor, and A. Y. M. Noor, “Domain specific concept ontologies and text summarization as hierarchical fuzzy logic ranking indicator on malay text corpus,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 3, p. 1527, Sep. 2019, doi: 10.11591/ijeecs.v15.i3.pp1527-1534.

[5] J. S. Hussein, A. A. Salman, and T. R. Saeed, “Arabic speaker recognition using HMM,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 1212–1218, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp1212-1218.

[6] M. Sahidullah and G. Saha, “A novel windowing technique for efficient computation of MFCC for speaker recognition,” *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 149–152, Feb. 2013, doi: 10.1109/LSP.2012.2235067.

[7] Z. S. Mahmood, A. najdet nasret Coran, A. E. Kamal, and A. B. Noori, “Dynamic Spectrum Sharing is the Best Way to Modify Spectrum Resources,” in *2021 Asian Conference on Innovation in Technology, ASIANCON 2021*, Aug. 2021, pp. 1–5, doi: 10.1109/ASIANCON51346.2021.9544912.

[8] A. M. Aaref and Z. S. Mahmood, “Optimization the accuracy of FFNN based speaker recognition system using PSO Algorithm,” *International Journal on Communications Antenna and Propagation (IRECAP)*, vol. 11, no. 4, p. 253, Aug. 2021, doi: 10.15866/irecap.v11i4.19883.

[9] R. T. Al-Hassani, D. C. Atilla, and Ç. Aydin, “Development of high accuracy classifier for the speaker recognition system,” *Applied Bionics and Biomechanics*, vol. 2021, pp. 1–10, May 2021, doi: 10.1155/2021/5559616.

[10] K. S. M. H. Ibrahim, Y. F. Huang, A. N. Ahmed, C. H. Koo, and A. El-Shafie, “A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting,” *Alexandria Engineering Journal*, vol. 61, no. 1, pp. 279–303, Jan. 2022, doi: 10.1016/j.aej.2021.04.100.

[11] A. N. N. Coran, Z. S. Mahmood, and A. E. Kamal, “Classification of acoustic data using the FF neural network and random forest method,” in *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Oct. 2021, pp. 1–4, doi: 10.1109/SMARTGENCON51891.2021.9645847.

[12] V. Sankaravadivel and S. Thalavaipillai, “Symptoms based endometriosis prediction using machine learning,” *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 6, pp. 3102–3109, Dec. 2021, doi: 10.11591/eei.v10i6.3254.

[13] J. A. M. Guiu *et al*., “Identification of the main components of spontaneous speech in primary progressive aphasia and their neural underpinnings using multimodal MRI and FDG-PET imaging,” *Cortex*, vol. 146, pp. 141–160, Jan. 2022, doi: 10.1016/j.cortex.2021.10.010.

[14] Y. Bestgen, “A simple language-agnostic yet very strong baseline system for hate speech and offensive content identification,” *arxiv pre-prints*, Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.02511.

[15] A. N. Nasret, A. B. Noori, A. A. Mohammed, and Z. S. Mahmood, “Design of automatic speech recognition in noisy environments enhancement and modification,” *Periodicals of Engineering and Natural Sciences*, vol. 10, no. 1, pp. 71–77, Dec. 2022, doi: 10.21533/pen.v10i1.2575.

[16] H. Cabrera, S. M. Jiménez, and E. S. Tellez, “INFOTEC-LaBD at PAN@CLEF21: Profiling hate speech spreaders on twitter through emotion-based representations,” in *CEUR Workshop Proceedings*, 2021, vol. 2936, pp. 1858–1870.

[17] J. A. Hassan and B. H. Jasim, “Design and implementation of internet of things-based electrical monitoring system,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3052–3063, Dec. 2021, doi: 10.11591/eei.v10i6.3155.

[18] B. Mouaz, C. Walid, B.-H. Abderrahim, and E. Abdelmajid, “A new framework based on KNN and DT for speech identification through emphatic letters in Moroccan dialect,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, p. 1417, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1417-1423.

[19] P. Cerva, L. Mateju, F. Kynych, J. Zdansky, and J. Nouza, “Identification of scandinavian languages from speech using bottleneck features and X-Vectors,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12848 LNAI, 2021, pp. 371–381, doi: 10.1007/978-3-030-83527-9_31.

[20] R. Kolinsky *et al*., “The impact of alphabetic literacy on the perception of speech sounds,” *Cognition*, vol. 213, p. 104687, Aug. 2021, doi: 10.1016/j.cognition.2021.104687.

[21] M. D. Hassan, A. N. Nasret, M. R. Baker, and Z. S. Mahmood, “Enhancement automatic speech recognition by deep neural networks,” *Periodicals of Engineering and Natural Sciences*, vol. 9, no. 4, pp. 921–927, Nov. 2021, doi: 10.21533/pen.v9i4.2450.

[22] P. M. Diaz and M. J. E. Jiju, “A comparative analysis of meta-heuristic optimization algorithms for feature selection and feature weighting in neural networks,” *Evolutionary Intelligence*, pp. 1–20, Jul. 2021, doi: 10.1007/s12065-021-00634-6.

[23] B. Jolad and R. Khanai, “ANNs for automatic speech recognition-a survey,” in *Lecture Notes in Networks and Systems*, vol. 209, 2022, pp. 35–48, doi: 10.1007/978-3-030-83527-9_31.

[24] T. A. Farrag and E. E. Elattar, “Optimized deep stacked long short-term memory network for long-term load forecasting,” *IEEE Access*, vol. 9, pp. 68511–68522, 2021, doi: 10.1109/ACCESS.2021.3077275.

[25] M. T. S. Al-Kaltakchi, H. A. A.Taha, M. A. Shehab, and M. A. M. Abdullah, “Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database*,” Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 2, pp. 782-789, May 2020, doi: 10.11591/ijeecs.v18.i2.pp782-789.

BIOGRAPHIES OF AUTHORS

Imad Burhan Kadhim received a B.S. in Electronics and Control Engineering from Kirkuk tech- nical college, Iraq, in 2013, and M.S. degree in Electrical and Electronic Engineering from Altinbas University, Turkey; in 2020, He is a lecturer in northern technical university Kirkuk institute from 2017 to the present. He can be contacted at email: emadburhan86@ntu.edu.iq.

Ali Najdet Nasret received a B.S. in Electrical Engineering from Al-mustansriya Uni- veristy, Iraq, in 2010, and M.S. degree in Electronic and Communication Engineering from Cankaya University, Turkey, in 2014. He is an lecturer in northern technical university Kirkuk technical institute from 2017 to the present. He can be contacted at email: alinajdet@ntu.edu.iq.

Zuhair Shakor Mahmood received a B.S. in Electronics and Control Engineering from Kirkuk technical college, Iraq, in 2005, and M.S. degree in Electrical and Electronic Engineering from Eskishir Osmangazi University, Turkey; in 2012, He is a lecturer in northern technical university Kirkuk institute from 2017 to the present. He can be contacted at email: email: zuherkazanci@ntu.edu.iq.