Arabic speaker recognition using HMM

Jabbar S. Hussein[1], Abdulkadhim A. Salman[2], Thmer R. Saeed[3]
[1]Karbala University, college of Engineering, Karbala, Iraq
[2]Technical Institute of Karbala, Al-Furat Al-Awsat Technical University, Karbala, Iraq
[3]Department of Electrical Engineering, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received Jan 12, 2021
Revised Jun 30, 2021
Accepted Jul 13, 2021

Keywords:

Arabic language
Hidden markov model
Speaker recognition

ABSTRACT

In this paper, a new suggested system for speaker recognition by using hidden markov model (HHM) algorithm. Many researches have been written in this subject, especially by HMM. Arabic language is one of the difficult languages and the work with it is very little, also, the work has been done for text dependent system where HMM is very effective and the algorithm trained at the word level. One the problems in such systems is the noise, so we take it in consideration by adding additive white gaussian noise (AWGN) to the speech signals to see its effect. Here, we used HMM with new algorithm with one state, where two of these components, i.e. (π and A) are removed. This give extremely accelerates the training and testing stages of recognition speeds with lowest memory usage, as seen in the work. The results show an excellent outcome. 100% recognition rate for the tested data, about 91.6% recognition rate with AWGN noise.

Corresponding Author:

Jabbar S. Hussein
Department of Prosthetic & Orthetic Engineering
Karbala University, college of Engineering
Karbala, Iraq
jabbar.salman@uokerbala.edu.iq

1. INTRODUCTION

These days, the upheaval in the hardware innovation gives a wide region to the specialists for taking care of complex issues, for example, speaker recognition (SR) with noisy environment. The significance of SR can be seen through its applications in security and reconnaissance frameworks. In the writing, different procedures for SR have been illustrated, hidden markov model (HMM) [1], support vector machine [2], linear discriminant analysis [3], independent component analysis [4], principal component analysis [5], quantization of mel-frequency cepstral coefficients [6] and artificial neural network [7]. In spite of different procedures which need to retrain the framework if there should be an occurrence of refreshing the database, the HMM can be utilized so that each model is independently prepared. In different words, adding or expelling any individual to/from the framework can be effectively performed without the need to retrain the framework.

The classification and recognition of the Arabic langue words is the motivating topic in the applications of the Arabic computer interface. The computer interface is a significant means in the intelligent structures and the technologies. The Linguistic recognition is talking recognition, and it is characterized such as the method to varying over acoustic discourse signals to its connecting set of words or other language units [8]-[12].

Speaker recognition is a multi-disciplinary innovation which utilizes the vocal features of speakers to infer data about their characters. It is a part of biometrics that might be utilized for distinguishing proof, check, and recognition of individual speakers, with the capacity of detection, tracking, and segmentation by extension. Speaker recognition and speaker check structure a bigger control of speaker classification [13]. Speaker recognition attempts to figure out which speaker produced a discourse signal though speaker check affirms if the part of the discourse has a place with the person who allegation it. It ought to be noticed that there are two sorts of speaker recognition, which are; text independent and text dependent [14]. This paper will anyway concentrate on text dependent speaker recognition. Present content text dependent produces sensible outcomes, yet at the same time do not have the fundamental execution on the off chance that they are to be utilized by the overall population (for example live testing).

So as to lessen the complicated nature of SR framework that utilizes HMM, a few procedures have been attempted, where the most transcendent strategy is the decrease of speaker file size utilizing one of the transformation strategies, for example, discrete wavelet transformation (DWT) [1] and discrete cosine transformation [15]. Then again, the downside of accomplishing further decrease in the framework's unpredictability is the improper number of HMM states utilized [16], [17], where this disadvantage is understood by utilizing one-state HMM. In discovering this topic, primary, a theory part covering the concept of MHH with one state [17] and the method of decreasing the size of the spoken word, discrete wavelet transformation (DWT). Then the Methodology of the work with its steps, finally, the outcomes and the conclusion of the speaker recognition utilizing the one-state Hidden.

2. METHODOLOGY

The work done through the following steps; i) recording Arabic words; ii) pre-processing; iii) features extraction and iv) recognition, with two phases; training, testing, and experiments:

2.1. Data sets

Arabic words are recorded using a microphone, with persons live around us, and from learning program for Arabic language, all that have been done with real environments, not in especial environments like in [6], then to the computer through the audio port, that is accomplished with 8000 Hz as sampling frequency and 16bit resolution for clear recording and single channel. The recoding process revealed that using the microphone results in good quality output signals. However, it might be a difficult process, due to the noise effect as well as unstable distance between the speakers and the microphone.

2.2. Pre-processing

After converting the audio signal to digitized form, the pre-processing stage starts, the original signal consists of two parts, i.e., information part and silent part with 8000 double samples. It should be mentioned that silent part must be removed that give a signal about 4000 doable, such as in [18]. Normalization parts necessary for making the signal smoother for next operations. Pre-emphasis part amends the loss of higher frequencies that have been lost through the propagation and radiation form voice source to the microphone, improving efficiency for the next stages, the framing and windowing are accomplished. As the human speech signal is varying slowly in time, it is normally divided into frames, which are overlapping with each other. While windowing process includes dividing the frames with a window, such a Hamming, such process decreases the effects of discontinuity that is produced by framing process. Finally resizing of the spoken file by using discrete wavelet transform (DWT), hence, with 1st level of DWT we get about 1000 double samples, while with more levels, the data will lose the mean part of it.

2.3. Features extraction

In the whole work, feature extraction and recognition were implemented in MATLAB2017b software. Each speech signal corresponding to any word is put in a specific file. Many speech features have been studied, for considering the spoken Arabic words as audio signal, and from that an audio features can be extracted and broadly classified based on their semantic interpretation as perceptual and physical features. Moreover, statistical features including, mean value, root mean square (RMS), standard deviation, median value, covariance, variance value, maximum value and minimum value. In this work, the statistical features (mean value and covariance) are the depended features because the statistical features represent the core of the signal and reduce the required size and the processing time.

2.4. Hidden markov model

HMM [19] is a stochastic system used to foresee a future occasions dependent on a previous data. The system includes an assortment of states, where just the yields of the states can be viewed and all the changes among the states are unknown. HMM can be grouped into two classes as indicated by the knowledge of the yields: discrete HMM and continues HMM [20], Figure 1 shows the state diagram 3-state left-to-right HMM.

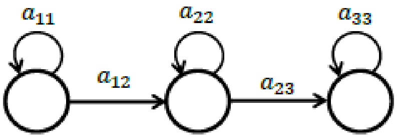

Figure 1. State diagram of 3-state left-to-right HMM

Discrete HMM, this type manages discrete codes that are transmitted from the states and the system λ is demonstrated by the three boundaries (π, A, B). Continuous HMM, The expression "continuous" indicates the idea of the yield densities of the masked states. Like a Gaussian capacity, these yields track the probability density function (PDF), where it is a symmetric bend framing a form resembles a chime. PDF of the perception vector O is determined by the accompanying condition [20]:

$$P(O) = \sum_{n=1}^{k} \frac{w_n}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(O-\mu_n)^2}{\sigma_n^2}\right] \tag{1}$$

Where; wn, σn and µn are, individually, the weight, standard deviation and mean of the nth Gaussian blend. It is important that the covariance (∑) of a vector is equivalent to the square of the standard deviation and thus, the continuous HMM is represented as in the associated tuple:

$$\lambda = (\pi\,, A, \mu\,, \Sigma) \tag{2}$$

The following points give an overview of its construction: symbols
N: States number in each system.
M: Code number in the yields.
π: The fundamental state probability parameter of size N × 1.
A: The change probability framework of size N × N.
B: The release probability framework of size N × M.

The contrast between the continuous and discrete HMMs, concerning the HMM boundaries, is in the discharge boundary, where in continuous HMM; it is indicated by the covariance and mean rather than discrete codes.

2.5. Recognition

Recognition has two parts: training and testing,

2.5.1. Training

For every spoken word, an array is created by linking all the sequences got from the training word as clarified earlier. When the array is framed, it is delivered to the HMM for training. HMM utilized in the proposed work is a unique system that contains just one state with continuous yield densities. Neither starting vector π nor transformation matrix A, occurs in one-state system and, for this situation, they are equivalent to one. In this manner, the system λ is commonly founded on the µ and ∑ of the perception vectors, as shown in Figure 2. The Baum-Welch calculation [10] with a one iteration is utilized to train the system of every word. Just one Gaussian mixture is utilized and the PDFs are determined as in (1), where $P(O) = [P1, P2, P3, \ldots, PM]$. Figure 2 shows the state chart of the COSM.

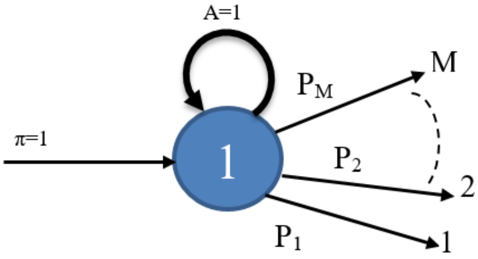

Figure 2. State diagram of the COSM

2.5.2. Testing

All spoken words that are not utilized in training of the HMM track the corresponding above points for testing, where each word is independently treated. The μ and ∑ of the perception vectors are determined and the Viterbi calculation [10] is utilized to find their probabilities by all PDFs that are gotten in the training procedure. Subsequently, the index of the most extreme probability might be utilized to distinguish the unknown word.

2.5.3. Case study

The experiments are performed on the work databases: where for five persons, one hundred patterns for each one, 70 words for training, and 30 words for testing.

With HMM, the experiments show that the technique of using the mean value and conference are the fine one. So, this method is examined using continues HMM, and the following specifications are employed:

1. Preprocessing: For the words database: 1st stage of DWT produces vector of size (2002 ×1)
2. 75% overlapped Hamming window of length n=100
3. Feature extraction $C= [MN\ MV]$
4. Training

After the information gathering, we tried our learning calculation as takes after:

- Randomly pick 70
- Test on the rest of the 30
- Repeat stages 1 and 2 ordinarily

Where stage (c) is added to diminish the variety from the decision of the preparation set.

3. RESULTS

The results shown in Table 1, is founded for five persons each one has 100 patters (words), 70 one for training and 30 patterns for test. With Figure 3, we take the patterns for one person, and also began with 70 one for training and 30 patterns for test, then in step of five patters we reduced the training patters and increased the test one, our goal to see the effect of the number of patters on the recognition rate and HMM algorithm, as shown in Figure 4, the recognition rate decreased with decreasing the training patters and that is a natural result with such algorithm.

Table1. HMM recognition rate

Speakers	Training words	Test words	Recognition rate %
1	70	30	100
2	70	30	100
3	70	30	100
4	70	30	100
5	70	30	100

To simulate the effects of error or noise on the performance of the recognition system, an additive white gaussian noise (AWGN) was added to the words patterns, training and test ones, because such a noise caver all the spectrum, the results show good outcomes, as shown in Table 2. While Figure 2 show the effect of additive noise for one person recognition. With less noise levels, one can get better results, such as in [21]-[23].

Table 2. HMM recognition rate with additive noise

Speakers	Training words	Test words	Recognition rate %
1	70	30	91.6
2	70	30	83.3
3	70	30	91.6
4	70	30	83.3
5	70	30	83.3

For comprising with other technologies, like neural network and ordinary HMM with more than one state), and as shown in Table 3, HMM with one, two and three state are shown, one can note from the results, that the one state HMM has better output than the others, and that also have been published as in [24]. While the comparison with NN, like multi-layer feed forward neural network (MLFFNN), and as shown in Table 4, still the HMM has better results.

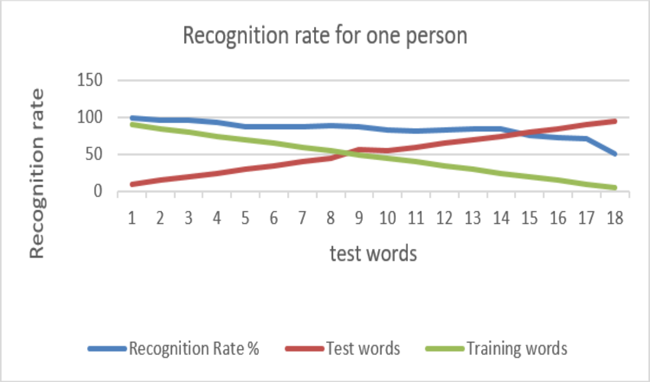

Figure 3. Recognition rate for one person with variables training and test words

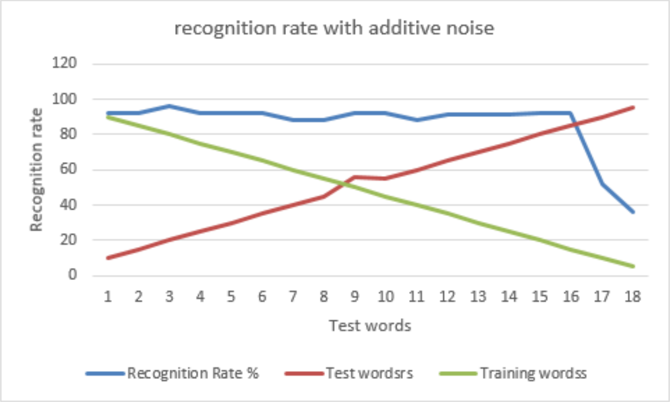

Figure 4. Recognition rate for one person with variables training and test words and additive noise

Table 3. Comparison one state HMM without state with, two and three state HMM

Speakers	Training words	Test words	Recognition rate % one state	Recognition rate % two states	Recognition rate % three states
1	70	30	100	86.66	76.66
2	70	30	100	86.66	76.66
3	70	30	100	86.66	76.66
4	70	30	100	86.66	76.66
5	70	30	100	86.66	76.66

Table 4. Comparison HMM without state with MLFFNN

Speakers	Training words	Test words	Recognition rate % one state HMM	Recognition rate % MLFFNN [25]
1	70	30	100	90
2	70	30	100	90
3	70	30	100	90
4	70	30	100	90
5	70	30	100	90

4. CONCLUSIONS

Speaker recognition is the use of a machine to recognize a person from a spoken phrase. Speaker-recognition systems can be used to identify a particular person or to verify a person’s claimed identity. Speech processing, speech production, and features and pattern matching for speaker recognition were introduced. A unique technique is established for recognizing spoken word by means of continuous one-state system in combination with a DWT. Dissimilar to other methods that depend on all parameters of HMM, the suggested work removes two of these components, i.e. π and A, and the recognition be determined by simply the PDF of one Gaussian mixture element. Environment noise is accordingly detached from the words via the pre-processing and the DWT. The 1st level of DWT was applied to the spoken word of the words databases, which in turn decrease the spoken word size, where constructing feature vectors by DWT is a very hopeful method for the task of spoken word. Also, the utilizing of one state extremely accelerates the training and testing stages of recognition speeds with lowest memory usage. The experimental outcomes display that the precision of the suggested work is around 100% in spite of additive noise, that affect the spoken word recognition but with minimum effect. The influence of the quantity of state depends on the data scope. For minor data, small quantity of states has greater recognition percentage. For bigger data, the quantity of states has very minor result on recognition rate

REFERENCES

[1] E. Abbas and H. Farhan, “Face recognition using DWT with HMM,” *Engineering & Technology Journal*, vol. 30, no. 1, pp. 142-154, 2012.

[2] Z. Li and X. Tang, “Using support vector machines to enhance the performance of Bayesian face recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 2, pp. 174-180, 2007, doi: 10.1109/TIFS.2007.897247.

[3] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Face recognition using LDA-based algorithms,” *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195-200, 2003, doi: 10.1109/TNN.2002.806647.

[4] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face recognition by independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450-1464, 2002, doi: 10.1109/TNN.2002.804287.

[5] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, “Two-dimensional PCA: a new approach to appearancebased face representation and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131-137, 2004, doi: 10.1109/TPAMI.2004.1261097.

[6] V. N. K. R. Devana and T. Rajesh, “A High Bit-Rate Speech Recognition System through Quantization of Mel-Frequency Cepstral Coefficients,” *International Journal of Electrical, Electronics and Computer Systems (IJEECS)*, vol. 2, no. 8-9, 2014.

[7] M. Owayjan, R. Achkar, and M. Iskandar, “Face detection with expression recognition using artificial neural networks,” *2016 3rd Middle East Conference on Biomedical Engineering (MECBME)*, Beirut, 6-7 Oct. 2016, pp. 115-119, 2016, doi: 10.1109/MECBME.2016.7745421.

[8] Kh. M. O. Nahar, Nahar, M. Elshafei, W. G, Al-Khatib, and H. Al-Muhtaseb, “Statistical Analysis of Arabic Phonemes for Continuous Arabic Speech Recognition,” *International Journal of Computer and Information Technology*, vol. 01, no. 02, Nov. 2012.

[9] J. S. Hussein, A. H. Ali, and Th. R. Saeed, “Improve the Recognition of Arabic Sign Languages Based on Statistical Features,” *Iraqi Journal of Computers, Communications, Control & Systems Engineering (IJCCCE)*, vol. 18, no. 3, pp. 26-32, 2018.

[10] Th. R. Saeed, J. S. Hussein, and A. H. Ali, “Classification improvement of spoken arabic language based on radial basis function,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 402-408, Feb. 2019, doi: 10.11591/ijece.v9i1.pp402-408.

[11] A. Hussein, S. Watanabe, and A. Alia, “Arabic Speech Recognition by End-to-End, Modular Systems and Human,” *Journal of Computer Speech and Language*, arXiv preprint arXiv:2101.08454, 2021.

[12] M. A. Ahmad and R. M. El Awady, “Phonetic Recognition of Arabic Alphabet letters using Neural Networks,” *International Journal of Electric & Computer Sciences*, vol. 11, no. 1, pp. 44-49, 2011.

[13] R. M. Hanifa1, Kh. Isa and Sh. Mohamad, “Speaker ethnic identification for continuous speech inMalay language using pitch and MFCC,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 19, no. 1, pp. 207-214, 2020, doi: 10.11591/ijeecs.v19.i1.pp207-214.

[14] 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, 2020, doi: 10.11591/ijeecs.v18.i2.pp782-789.

[15] K. Singh, M. Zaveri, and M. Raghuwanshi, “Recognizing faces under varying poses with three states hidden Markov model,” in *Proc. of IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, Zhangjiajie, China, vol. 2, pp. 359-363, 2012, doi: 10.1109/CSAE.2012.6272792.

[16] H. Farhan, M. Al-Muifraje, and T. Saeed, “Using only two states of discrete HMM for high-speed face recognition,” in *2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA)*, Baghdad, Iraq, 2016, pp. 1-5, doi: 10.1109/AIC-MITCSA.2016.7759939.

[17] H. Farhan, M. Al-Muifraje, and T. Saeed, “A Novel Face Recognition Method based on One State of Discrete Hidden Markov Model,” *2017 Annual Conference on New Trends in formation & Communications Technology Applications (NTICT)*, Baghdad, 2017, pp. 252-257, doi: 10.1109/NTICT.2017.7976152.

[18] R. M. Hanifa, *et al.*, “Voiced and unvoiced separation in malay speech using zero crossing rate and energy,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 2, pp. 775-780, 2019, doi: 10.11591/ijeecs.v16.i2.pp775-780.

[19] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition**,**” in *Proc.of the IEEE*, vol. 77, no. 2, 1989, pp. 257-286, doi: 10.1109/5.18626.

[20] L. Shi, I. Ahmad, Y. He, and K. Chang," Hidden Markov Model based Drone Sound Recognition using MFCC Technique in Practical Noisy Environments," *Journal of Communications and Networks*, vol. 20, no. 5, pp. 509-518, 2018, doi: 10.1109/JCN.2018.000075.

[21] Mukherjee and A. Sengupta, "Estimating the probability density function of a nonstationary non-Gaussian noise," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 4, pp. 1429-1435, 2010, doi: 10.1109/TIE.2009.2039451.

[22] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, M. A. M. Abdullah, and H. N. Abdullah, "Thorough evaluation of TIMIT database speaker identification performance under noise with and without the G.712 type handset," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 851-863, 2019, doi: 10.1007/s10772-019-09630-9.

[23] O. A. Noor, "Robust speaker verification in band-localized noise conditions," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 2, pp. 499-506, 2019, doi: 10.11591/ijeecs.v13.i2.pp499-506.

[24] M. Alsulaiman, Y. Alotaibi, M. Ghulam, M. A. Bencherif, and A. Mahmoud, "Arabic Speaker Recognition: Babylon Levantine Subset Case Study," *Journal of Computer Science*, vol. 6, no. 4, pp. 381-385, 2010, doi: 10.3844/jcssp.2010.381.385.

[25] D. Mengistu and D. M. Alemayehu, "Speech Processing for Text Independent Amharic Language Dialect Recognition," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 5, no. 1, pp. 115 -122, 2017, doi: 10.11591/ijeecs.v5.i1.pp115-122.

BIOGRAPHIES OF AUTHORS

Jabbar Salman Hussein was born in Bagdad Iraq in 1974. He received the B.S., M.S., and Ph. D. degrees in Electronic Engineering from University of Technology in 1998, 2001 and 2019 respectively. From 2000 to 2007 he worked at University of Technology / Electrical Engineering Department as lecturer. Since 2007, he has been working as a lecturer in University of Kerbala / Collage of Engineering. He worked also in many companies as a Broadcasting engineer. During this period many research were published in international and local conferences and journals in the field of Speech Recognitions, Prosthesis Control, Antenna Design and Mobile Radiation Protection Systems.

Dr. Abdulkadhim A. Salman was born in Najaf, Iraq in 1964. He received the B.S., M.S. and Ph. D. degrees in Electronic Engineering from University of Technology in 1987, 2002 and 2018 respectively. From 1987 to 1993, he worked at the Atomic Energy Organization in the field of designing electronic circuits for generating and measuring high-speed current pulses as well as in the use of digital and logical circuits in the designs of control systems. Since 1994, he has been working as a lecturer at Technical Institute of Karbala at Al-Furat Al-Awsat Technical University. During this period many researches were published in international and local conferences and journals in the field of filter design, switching capacitor technique, channel model, power line communications (PLC), OFDM modulation, automation using FPGA and Microcontroller.

Thamir Rashed Saeed was Born in Baghdad, Iraq, on February 10, 1965. He received the B.Sc. Degree from military engineering college in Baghdad in 1987, the M.Sc. Degree from military engineering college in Baghdad in 1994 and Ph.D. degree from AL-Rashed college of engineering and Secinec in Baghdad 2003. From 1994 to 2003, he worked with military engineering college in Baghdad as a member of teaching staff. From 2003 till now, he worked with the University of Technology in Baghdad as a member of teaching staff. Currently, he is the Asst. Professor of electrical engineering at university of Technology and a head of radar research group in the Electrical Eng. Dept. His major interests are in digital signal processing, digital circuit design for DSP based on FPGA, sensor network and Pattern Recognition.