Emirati-Accented Emotion Verification based on HMM3s, HMM2s, and HMM1s

Ismail Shahin
Department of Electrical Engineering
University of Sharjah
Sharjah, United Arab Emirates
ismail@sharjah.ac.ae

Noor Ahmad Al Hindawi
Department of Electrical Engineering
University of Sharjah
Sharjah, United Arab Emirates
u18105940@sharjah.ac.ae

Abstract— **The proposed research is dedicated to verifying the claimed emotion of speaker-independent and text-independent formed on three dissimilar classifiers. The HMM3 short for Third-Order Hidden Markov Model, HMM2 short for Second-Order Hidden Markov Model, and HMM1 short for First-Order Hidden Markov Model are the three classifiers utilized in this study. Our work has been evaluated on our collected Emirati-accented speech corpus which entails 50 speakers of Emirati origin (25 female and 25 male) uttering sentences in six emotions by means of the extracted features by Mel-Frequency Cepstral Coefficients (MFCCs). Our outcomes prove that HMM3 is superior to each of HMM1 and HMM2 to authenticate the claimed emotion. The achieved results formed on HMM3 are very similar to the outcomes attained in the subjective valuation by Arab listeners.**

Keywords—"Emirati-accented speech corpus, emotion verification, first-order hidden Markov model, second-order hidden Markov model, third-order hidden Markov model"

I. Introduction

Emotion recognition appears in the following distinct types: emotion verification (authentication) and emotion identification. When an audio model out of the unidentified emotion samples is examined and compared against audio samples of identified emotions, this is called emotion identification. The emotion whose model best matches the input audio model is known as the anonymous emotion. Specifying if an emotion fits with specific identified emotion or with the unidentified emotions is known as emotion verification. A true emotion becomes when the emotion recognition model successfully recognizes an emotion that is properly claiming its identity. While, false emotion appears when the emotion is unrecognized to the system that is posturing as a recognized emotion. Target emotion is another denotation of a recognized emotion, whereas background emotion is the denotation of a false emotion [1].

There are two forms of errors in emotion verificatio0n. systems: false rejection, where a true emotion has been denied, and false acceptance, where a false emotion has been admitted.

Emotion verification has two forms of texts: text-dependent and text-independent. When emotions should provide the same text (utterances/sentences) for both testing and training audios, then the text form is text-dependent. When emotions should not be provided by the same text for both the training and the testing stages, then the text form is text-independent. The two classes of sets of emotion verification are: closed set and open set. Closed set class is when a reference sample for a test emotion must be available; however, in the open set class, a reference sample for a test emotion might not be accessible.

Emotion recognition has extensive applications that range from "emerging in smart call centers, to human robotic interfaces, telecommunications, and smart verbal tutoring schemes. Emotion recognition is involved in various fields, where evaluating a callers' emotion for phone answer services is in telecommunications field. Another field is human robotics interfaces, where emotions are identified by trained robots in order to intermingle with people and identify people emotions. Additional applications in different fields can be exposed in intelligent call centers, which define the issues occurring in unfortunate communications that are detected through emotion recognition. Emotion recognition is as well utilized in intelligent voiced teaching in order to perceive and modify students' emotions when students went through a dull condition throughout tutoring meetings" [2], [3], [4].

II. Literature Review

There is a growing research work on emotion recognition in the last few decades. Many studies focused on this area using English databases [1, 5-11], while there are few studies that made effort on such field using Arabic datasets [12-14].

Emotion recognition using English datasets has been studied in many studies [1, 5-11]. Yogesh et.al [5] came up with "a recent Particle Swarm Optimization (PSO) aided population-based method to select features. BES short for Berlin Emotional Speech corpus, and SUSAS short for Speech Under Simulated Actual Stress database, and SAVEE short for Surrey Audio-Visual Expressed Emotion corpus, were used and executed in their experiments. Shahin payed attention [6] on the study and improvement of speaker-independent and text-independent under both emotional and stressful environments for talking condition identification formed on three various classifiers: Supra-segmental Hidden Markov Models (SPHMMs), Hidden Markov Models (HMMs), and Second-Order Circular Hidden Markov Models (CHMM2s). His work proved that SPHMMs surpasses both CHMM2s and HMMs for emotion recognition in both emotional and stressful talking environments [6]. The study by Shahin and Ba-Hutair [7] shed the light on the improvement of speaker-independent and text-independent talking states in both emotional and stressful environments formed on CSPHMM2s which is short for Second-Order Circular Suprasegmental Hidden Markov Models. Moreover, one of the great objectives was to tell apart between emotional speaking condition and stressful speaking condition formed on CSPHMM2s. The approached decision is to recognize the speaking emotion along with the stressful surrounding environments formed on CSPHMM2s surpasses the ones formed on all of CHMM2s, SPHMMs, and HMMs. In one of his foregoing research [8], Shahin used emotions to determine the unidentified speakers. Shahin put forward a new model for the recognition of HMM-based speakers from the corresponding emotions. In yet another work by Shahin [9], he presented, enforced, and evaluated "text-dependent and speaker-dependent speaking style authentication system that permits or rejects the affirmed identity of a speaking style derived from SPHMMs". Shahins' findings, formed on SPHMMs, revealed that ''average speaking style authentication efficiency exist as, 85%, 60%, 57%, 59%, 61%, 41%, 99%, 61%, and 37%, respectively, referring to the speaking states; slow, loud, fear, fast, happy, angry, neutral, soft, and shout". Shahin [10] recommended a two-stage method that utilizes the speaker's emotion cues (emotion-dependent and text-independent speaker verification issue) supported by HMMs and SPHMMs as classifiers, for improvement. His model consists of ''cascaded phases that incorporate and merge the emotion recognizer accompanied by a speaker recognizer into one recognizer. His analysis indicated that his method produced better outcomes with a dramatic change over prior research as well as other methods including "emotion-independent speaker verification approach" as well as "emotion-dependent speaker verification method largely formed on HMMs." In another work by Shahin and Nassif[11], the aim was to increase the precision of emotion recognition focused on "a classifier named Third-Order Hidden Markov Models (HMM3s)." The thesis was measured on the EPST corpus. "Extract features of the EPST database are Mel-Frequency Cepstral Coefficients (MFCCs)." The outcomes provided 71.8 per cent as an average of emotional recognition precision. Shahin [1] concentrated on determining the unidentified emotion formed on the 'Third-Order Circular Suprasegmental Hidden Markov Model (CSPHMM3) as a classifier.' His dissertation was checked in the EPST database. The MFCCs were the derived features of the EPST database. The findings provided an average of 77.8 per cent in emotion recognition precision that is dependent on the CSPHMM3. The tests of Shahin's research have shown that "CSPHMM3 is superior to Gaussian Mixture Model (GMM), HMM3, Vector Quantization (VQ), and Support Vector Machine (SVM) by 4.9 per cent, 6.0 per cent, 5.4 per cent and 3.5 per cent, respectively, for emotional recognition" [1].

Contrastingly, only few number of studies concentrated on the recognition of emotions utilizing Arabic corpora [12-14]. Klaylat et. al [12] proposed two-stages for the improvement of an emotion recognition system. Their system identifies three distinct emotions: surprised, angry and happy using an Arabic speech dataset. Moreover, they applied 35 categorization representations as well as a Sequential Minimal Optimization (SMO) classifier in the proposed research. The outcomes displayed a percentage of 95.52% as an average for emotion recognition precision [12]. Concerning El Gohary et .al[13] is particularly with the detection of emotion in the Arabic text. They are mainly focused on moderate Arabic vocabulary of emotions used only to encode Arabic youngsters stories of six different emotions: fear, anger, joy, sadness, disgust, and surprise; achieving a precision of 65% emotion detection. Shahin et. al. [14] spotlighted on ''recognizing speaker-independent and text-independent emotions utilizing Arabic speech corpus in Emirati accent centered on an existing hybrid classifier named cascaded Gaussian Mixture Model and Deep Neural Network, GMM-DNN, (GMM followed by DNN). In their work, six distinct emotions were utilized. These emotions are: sadness, happiness, neutrality, anger, fear and disgust. They reported an average of 83.97 per cent as an emotion recognition precision utilizing novel GMM-DNN classifier" [14].

In this proposed research, we focus on verifying the claimed emotion of text-independent and speaker-independent system formed on three different classifiers. First-Order Hidden Markov Model (HMM1), Second-Order Hidden Markov Model (HMM2), and Third-Order Hidden Markov Model (HMM3) are the classifiers used. Our work was assessed on the Emirati-accented speech dataset collection which is conceded of twenty five women and twenty five men Emirati speakers speaking in six diverse emotions that utilize MFCCs as extracted features.

The rest of the document is structured as: The fundamentals of HMM1, HMM2, and HMM3 are shown in Section III. The dataset used as well as the extraction of features are shown in Section IV. The emotion verification algorithm characterized by three classifiers as well as the tests are given in Section V. The decision threshold is set out in Section VI. The obtained results as well as the experimentation are addressed in Section VII. The proposed work outcome is given in Section VIII.

III. FUNDAMENTALS OF HMM1, HMM2, AND HMM3

A. Basics of HMM1

In HMM1, the state sequence is a first-order Markov chain where the stochastic process is represented by a 2-D matrix of a priori transition probabilities (a_{ij}) between states s_i and s_j where a_{ij} are provided by:

$$a_{ij} = Prob\left(q_t = s_j | q_{t-1} = s_i\right) \qquad (1)$$

In this model, $(t+1)$ is the time at which the state-transition probability is dependent upon the state of the Markov chain at time (t). More information could be reached about HMM1 from the references [15, 16].

B. Basics of HMM2

In HMM2, "the state sequence is a second-order Markov chain where the stochastic procedure is defined by a 3-D matrix (a_{ijk}). Therefore, the transition probabilities in HMM2 are stated as" [17]:

$$a_{ijk} = Prob\left(q_t = s_k \middle| q_{t-1} = s_j, q_{t-2} = s_i\right) \qquad (2)$$

by means of the restraints,

$$\sum_{k=1}^{N} a_{ijk} = 1 \qquad N \geq i, j \geq 1$$

The state-transition probability in HMM2 at time *t+1* depends on the states of the Markov chain at times *t* and *t-1*. More information about HMM2 could be acquired from the reference [17, 18, 19].

C. HMM3 Basics

In HMM3, "the underlying state sequence is a third-order Markov chain where the stochastic procedure is expressed by a 4-D matrix (a_{ijkw}). Thus, the transition probabilities in HMM3 were also provided as" [20],

$$a_{ijkw} = Prob\left(q_t = s_w \middle| q_{t-1} = s_k, q_{t-2} = s_j, q_{t-3} = s_i\right) (3)$$

with the restraints,

$$\sum_{w=1}^{N} a_{ijkw} = 1 \qquad N \geq i, j, k \geq 1$$

The state sequence probability, $Q \underline{\Delta} q_1, q_2, \ldots, q_T,$ is defined as:

$$Prob(Q) = \Psi_{q_1} a_{q_1 q_2 q_3} \prod_{t=4}^{T} a_{q_{t-3} q_{t-2} q_{t-1} q_t} \qquad (4)$$

where Ψ_i at time $t = 1$ is the state s_i probability, and a_{ijk} at time t = 3 is the transition to a state s_k from a state s_i probability.

Given a sequence of observed vectors, $O \underline{\Delta} O_1, O_2, \ldots, O_T,$ the probability of joint state-output is stated as:

$$Prob(Q, O|\lambda) = \Psi_{q_1} b_{q_1}(O_1) a_{q_1 q_2 q_3} b_{q_3}(O_3). \prod_{t=4}^{T} a_{q_{t-3}\, q_{t-2} q_{t-1} q_t} b_{q_t}(O_t) \qquad (5)$$

Researchers can find additional clarifications and data about this model from [20], [21].

IV. Speech Corpus and Extraction of Features

A. Speech corpus

An Emirati-accented Arabic corpus has been used to test our work. This corpus has been collected from 30 local Emirati speakers (15 female and 15 male), where 10 speakers per gender are consumed for the training phase and the remainder are consumed for the testing phase. These utterers utter eight habitually Emirati utterances in the UAE culture. Each utterance is spoken and recorded by each speaker, where each utter is repeated nine times under each of the following emotions: disgust, happy, angry, sad, fear, and neutral. Table 1 demonstrates Arabic Emirati corpus that is used in this study. The right column demonstrates the sentences in Arabic Emirati dialect, whereas the left column demonstrates the English translation of the Arabic Emirati sentences. This corpus was captured in two disconnected phases: testing phase and training phase. The corpus was collected in a noise-free environment in the College of Communication, University of Sharjah, United Arab Emirates by a group of dedicated engineers. The corpus was recorded by a speech acquisition board using a 16-bit linear coding A/D converter and sampled at a sampling rate of 44.6 kHz.

B. Features Extraction

Delta Mel-Frequency Cepstral Coefficients (MFCCs - delta) and static Mel-Frequency Cepstral Coefficients (MFCCs-static) are the phonetic content representation of our captured speech signals. Such coefficients are commonly utilized in numerous emotion and speaker recognition research [6], [7], [8], [10], [14], [22], [23]. In this research, the purpose of utilizing MFCCs is to create the observation vectors in HMM1, HMM2, and HMM3. Sixteen static MFCCs and sixteen delta MFCCs are merged together to form 32-dimension MFCC feature vectors in every model of HMM1, HMM2, and HMM3".

V. Experiments and Emotion Verification Algorithm formed on HMM1, HMM2, and HMM3

The training stage of HMM3s, HMM2s, and HMM1s (three distinct and independent stages), the *v*th emotion was signified by a *v*th model vector. The *v*th model is produced using 10 speakers per gender uttering the first four sentences with a replicate of nine utterances per sentence of the corpus. The total number of sentences utilized to establish each emotion model is 720 (10 speakers per gender × 4 sentences × 9 repetitions per utterance).

In the verification stage in each of HMM3s, HMM2s, and HMM1s, each one of the remaining 5 speakers per gender used nine sentences/utterance of the last four sentences upon each emotion "speaker-independent and text-independent experiments". The entire number of sentences utilized in this stage is 2160 (five speakers per gender × four utterances × nine utterances per sentence × six emotions)".

In order to verify the claimed emotion, the log-likelihood ratio has been computed in the domain of log formed distinctly on each of HMM3s, HMM2s, and HMM1s, as specified in the following equation [24],

$$\Lambda_{model}(O) = log\left[P\left(O \middle| \lambda_{model,C}\right)\right] - log\left[P\left(O \middle| \lambda_{model,\bar{C}}\right)\right] \quad (6)$$

where, the log-likelihood ratio Λ_{model} (*O*), which is computed in the domain of log, $P\left(O\middle|\lambda_{model,C}\right)$ is the sequence of observation *O* probability assumed it was derived from the stated emotion. $P\left(O\middle|\lambda_{model,\bar{C}}\right)$ which is the observation sequence *O* probability assumed it was not derived from the stated emotion, and representation signifies either HMM3s, HMM2s, or HMM1s.

The sequence of observation *O* probability assumed it arises from the stated emotion is expressed as [24],

$$log\, P\left(O \middle| \lambda_{model,C}\right) = \frac{1}{T}\sum_{t=1}^{T} log\, P\left(o_t \middle| \lambda_{model,C}\right) \qquad (7)$$

where, $O = o_1 o_2 \ldots\ o_t \ldots o_T$ and *T* is the sentence period.

The probability of the sequence of observation *O* as it has not come from the stated emotion could be measured utilizing a set of *B* imposter emotion representations: $\{\lambda_{model,\bar{C}_1}, \lambda_{model,\bar{C}_2}, \ldots, \lambda_{model,\bar{C}_B}\}$ as,

$$log\ P\left(O \mid \lambda_{model,\bar{C}}\right) = \left\{\frac{1}{B}\sum_{b=1}^{B} log\left[P\left(O \mid \lambda_{model,\bar{C}_b}\right)\right]\right\} \quad (8)$$

where $P\left(O \mid \lambda_{model,\bar{C}_b}\right)$ could be calculated utilizing Eq. (7).

VI. DECISION THRESHOLD

Two distinct forms of fault can happen in emotion verification. These two types are false acceptance and false rejection. When a fraud 's identity claim is acknowledged, a false acceptance error is considered. On the other hand, when a genuine identity claim is denied, it is called a false rejection error.

Emotion verification issue involves the development of a ‘‘binary decision focused on two hypothesis: Hypothesis H$_0$ is if analysis series *O* is focused on the stated emotion or Hypothesis H$_1$ if the analysis series *O* is not focused on the stated emotion’’.

A comparison between the log-likelihood ratio and the threshold θ should be made in the final stage of the verification process in order to require or reverse the reported emotion [24].

$$Accept\ the\ claimed\ emotion\ if\ \Lambda(O) \geq \theta$$

$$Reject\ the\ claimed\ emotion\ if\ \Lambda(O) < \theta$$

“Open set emotion testing uses thresholding to build a choice when an emotion exceeds the range. Both types of error in emotion verification depend upon the threshold utilized in decision-making. A strong threshold value toughens the chance of false emotions being wrongly regarded, but at the cost of falsely disproving true emotions. Contrastingly, an eased threshold value relieves true emotions to be enrolled every time falsely admitting false emotions are spent. In order to achieve an adequate threshold value that consistently meets the level of true disapproval of emotion and false affirmation of emotion, it is necessary to acknowledge the allocation of true emotions and false emotions. Appropriate practice for setting a threshold value is to begin with a loose initial threshold value and then allow it to be modified by setting the average of the latest test scores. This eased threshold value gives inadequate security toward false emotion attempts’’.

VII. RESULTS, EXPERIMENTS AND THEIR DISCUSSIONS

In this work, three distinct classifiers have been utilized to verify the claimed emotion using “Emirati-accented speech corpus. These classifiers are HMM1, HMM2, and HMM3. We evaluated our work using six dissimilar emotions. The proposed emotions include: neutral, happy, sad, disgust, fear, and angry”.

Table 2 demonstrates “percentage Equal Error Rate (EER) of emotion verification formed on HMM1, HMM2, and HMM3”. Average percentage of EER based on HMM3, HMM1, and HMM2 is 24.4%. 30.4%, 28.0%, respectively. Therefore, it is apparent that HMM3 outperforms each of HMM2 and HMM1 for emotion verification. This table evidently reveals the minimum percentage EER happens when the stated emotion is neutral; on the other hand, the highest percentage EER occurs when the stated emotion is anger.

To authenticate whether the distinctions in EER (EER formed on HMM3 and premised on each of HMM1 and HMM2) are real or based solely on statistical variations, a suitable statistical test of significance was performed. This examination was used on the basis of the Student's t Distribution Test:

$$t_{model\ x, model\ y} = \frac{\bar{x}_{model\ x} - \bar{x}_{model\ y}}{SD_{pooled}} \quad (9)$$

where “$\bar{x}_{model\ x}$ defined as the mean of the first sample (model *x*) of size *n*, $\bar{x}_{model\ y}$ is defined as the mean of the second sample (model *y*) of the same size, and SD pooled is defined as the pooled standard deviation of the two samples (models *x* and *y*)” specified as,

$$SD_{pooled} = \sqrt{\frac{SD^2_{model\ x} + SD^2_{model\ y}}{2}} \quad (10)$$

where “SD $_{\text{model x}}$: is an estimate of the standard deviation of the average of the first sample (model *x*) of size *n* and SD$_{\text{modely}}$ is an estimate of the standard deviation of the average of the second sample (model *y*) of equal size”.

The value of *t* evaluated between HMM3 and each of HMM1 and HMM2 is calculated formed on Table 2. The computed values are t$_{\text{HMM3, HMM1}}$ = 1.899 and t$_{\text{HMM3, HMM2}}$ = 1.737. Each computed value is higher than the "Tabulated Critical Value t $_{0.05}$ = 1.645 at a relevant level of 0.05." It is therefore clear that HMM3 leads each of HMM1 and HMM2 for emotion verification.

Emotion verification accuracy premised on HMM3 is competed with that premised on "state-of-the-art classifiers and models including Vector Quantization (VQ) [26], Gaussian Mixture Model (GMM) [24] and Support Vector Machine (SVM) [25]". The average EER for the "GMM, SVM, and VQ" emotion verification is 27.3 percent, 26.1 percent and 25.7 percent, respectively. It is apparent from this experiment that HMM3 produces less EER than all of these three classifiers.

An ‘‘informal subjective evaluation of emotion verification using our captured dataset was carried out employing 10 human unprofessional adult listeners. A sum of 540 utterances (30 speakers × 6 feelings × 3 repetitions) has been included in the evaluation. These listeners are required to confirm the emotion that has been claimed. Formed on this analysis, the average EER achieved is 25.3 per cent. This average EER is near the average found formed on HMM3 (24.4 per cent)’’.

VIII. Conclusion

In this work, HMM1, HMM2, and HMM3 are used as classifiers to check the stated emotion spoken in the "Arabic Emirate" dialect. In this work, some concluding remarks can be drawn. Firstly, "HMM3 outperforms each of HMM1, HMM2, GMM, SVM, and VQ" in verifying the claimed emotion. Secondly, the greatest emotion verification accuracy takes place when the claimed emotion is uttered neutrally. Finally, the smallest emotion verification accuracy happens when the claimed emotion is expressed angrily.

This work has some limitations. First, our corpus is constrained to six emotions only. Second, the obtained emotion verification accuracy premised on HMM3 is not optimal. Our next tactic is to enforce the Deep Neural Network (DNN) in order to achieve better results [27]. Besides, our strategy is to study and research Emirati-accented emotional verification in bias-based conversational environments [28], [29].

Acknowledgment

We would like to express our thanks to the "University of Sharjah for supporting this work through the competitive research project entitled Emirati-Accented Speaker and Emotion Recognition Formed on Deep Neural Network.

References

[1] I. Shahin, "Emotion recognition formed on third-order circular suprasegmental hidden Markov model," The 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), April 2019, Amman, Jordan, pp. 599-604.

[2] V.A. Petrushin, "Emotion recognition in speech signal: experimental study, development, and application," Proceedings of International Conference on Spoken Language Processing, ICSLP 2000.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis S. Collias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, 18 (1), 2001, pp. 32-80.

[4] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," Neural Networks, special issue (18), 2005, pp. 389-405.

[5] C. K. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, C. Berkai, and K. Polat, "A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal," Expert Systems with Applications, Vol. 69, 2017, pp.149-158.

[6] I. Shahin, "Studying and enhancing talking condition recognition in stressful and emotional talking environments formed on HMMs, CHMM2s and SPHMMs," Journal on Multimodal User Interfaces, Vol. 6, issue 1, June 2012, pp. 59-71, DOI: 10.1007/s12193-011-0082-4.

[7] I. Shahin and Mohammed Nasser Ba-Hutair, "Talking condition recognition in stressful and emotional talking environments formed on CSPHMM2s," International Journal of Speech Technology, Vol. 18, issue 1, March 2015, pp. 77-90, DOI: 10.1007/s10772-014-9251-7.

[8] I. Shahin, "Using emotions to identify speakers," The 5th International Workshop on Signal Processing and its Applications (WoSPA 2008), Sharjah, United Arab Emirates, March 2008.

[9] I. Shahin, "Speaking style authentication using suprasegmental hidden Markov models," University of Sharjah Journal of Pure and Applied Sciences, Vol. 5, No. 2, June 2008, pp. 41-65.

[10] I. Shahin, "Employing emotion cues to verify speakers in emotional talking environments," Journal of Intelligent Systems, Special Issue on Intelligent Healthcare Systems, DOI: 10.1515/jisys-2014-0118, Vol. 25, issue 1, January 2016, pp. 3-17.

[11] I. Shahin and A. B. Nassif, "Utilizing third-order hidden Markov models for emotional talking condition recognition," 14th IEEE International Conference on Signal Processing (ICSP2018) 12-16 August 2018, Beijing, China, pp. 250-254.

[12] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Enhancement of an Arabic speech emotion recognition system," International Journal of Applied Engineering Research, Vol. 13, issue 5, 2018, pp. 2380–2389.

[13] A. F. El-Gohary, T. I. Sultan, M. A. Hana, and M. M. El Dosoky, "A Computational approach for analyzing and detecting emotions in Arabic text," International Journal of Engineering Research and Applications (IJERA), 2013, Vol. 3, pp. 100-107.

[14] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion Recognition using Hybrid Gaussian Mixture Model and Deep Neural Network," IEEE Access, Vol. 7, March 2019, pp. 26777 - 26787, DOI 10.1109/ACCESS.2019.2901352.

[15] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Eaglewood Cliffs, New Jersey, 1983.

[16] X. D. Huang, Y. Ariki and M. A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Great Britain, 1990.

[17] J. F. Mari, J. P. Haton and A. Kriouile, "Automatic word recognition formed on second-order hidden Markov models," IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 1, January 1997, pp. 22-25.

[18] I. Shahin, "Employing second-order circular suprasegmental hidden Markov models to enhance speaker identification performance in shouted talking environments," EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2010, Article ID 862138, 10 pages, June 2010. doi:10.1155/2010/862138.

[19] I. Shahin, "Enhancing speaker identification performance under the shouted talking condition using second-order circular hidden Markov models" Speech Communication, Vol. 48, issue 8, August 2006, pp. 1047-1055.

[20] I. Shahin, "Novel third-order hidden Markov models for speaker identification in shouted talking environments," Engineering Applications of Artificial Intelligence, Vol. 35, October 2014, pp. 316-323, DOI: 10.1016/j.engappai.2014.07.006.

[21] I. Shahin, “Speaker identification in a shouted talking environment formed on novel third-order circular suprasegmental hidden Markov models,” Circuits, Systems and Signal Processing, DOI: 10.1007/s00034-015-0220-4, Vol. 35, issue 10, October 2016, pp. 3770-3792.

[22] I. Shahin, “Employing both gender and emotion cues to enhance speaker identification performance in emotional talking environments,” International Journal of Speech Technology, Vol. 16, issue 3, September 2013, pp. 341-351, DOI: 10.1007/s10772-013-9188-2.

[23] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," Speech Communication, Vol. 49, issue 2, February 2007, pp. 98-112.

[24] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communication, Vol. 17, 1995, pp. 91-108.

[25] V. Wan and W. M. Campbell, “Support vector machines for speaker verification and identification,” Neural Networks for Signal Processing X, Proceedings of the 2000 IEEE Signal Processing Workshop, Vol. 2, 2000, pp. 775 – 784.

[26] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech Communication, Vol. 52, No. 1, January 2010, pp. 12 – 40.

[27] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: a Systematic Review,” IEEE Access, Vol. 7, February 2019, pp. 19143 - 19165, DOI: 10.1109/ACCESS.2019.2896880.

[28] I. Shahin, “Speaker identification investigation and analysis in unbiased and biased emotional talking environments,” International Journal of Speech Technology, Vol. 15, issue 3, September 2012, pp. 325-334, DOI: 10.1007/s10772-012-9156-2.

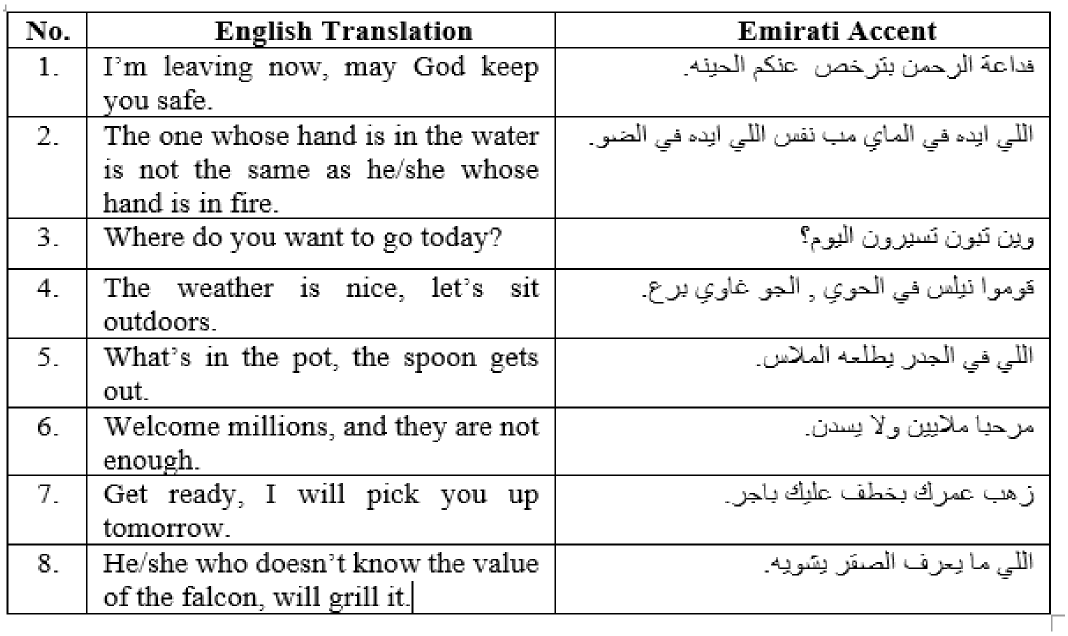
[29] I. Shahin, “‘Analysis and investigation of emotion identification in biased emotional talking environments,” IET Signal Processing, Vol. 5, No. 5, August 2011, pp. 461 – 470, DOI: 10.1049/iet-spr.2010.0059.

Author Contributions:

Ismail Shahin has implemented the simulation, and major parts of paper writing and correction.

Noor Ahmad wrote some sections in the paper along with the paper organization and some parts of simulation.

TABLE I. EMIRATI DATASET AND ITS ENGLISH TRANSLATION.



No.	English Translation	Emirati Accent
1.	I’m leaving now, may God keep you safe.	فداعة الرحمن بترخص عنكم الحينه.
2.	The one whose hand is in the water is not the same as he/she whose hand is in fire.	اللي ايده في الماي مب نفس اللي ايده في الضو.
3.	Where do you want to go today?	وين تبون تسيرون اليوم؟
4.	The weather is nice, let’s sit outdoors.	قوموا نيلس في الحوي , الجو غاوي برع.
5.	What’s in the pot, the spoon gets out.	اللي في الجدر يطلعه الملاس.
6.	Welcome millions, and they are not enough.	مرحبا ملايين ولا يسدن.
7.	Get ready, I will pick you up tomorrow.	زهب عمرك بخطف عليك باجر.
8.	He/she who doesn’t know the value of the falcon, will grill it.	اللي ما يعرف الصقر يشويه.

TABLE II. PERCENTAGE EER OF EMOTION VERIFICATION FORMED ON HMM1, HMM2, AND HMM3.

Emotion	Percentage EER of emotion verification formed on:		
	HMM1	HMM2	HMM3
Neutral	14.6	13.7	10.4
Happy	26.7	24.4	20.6
Sad	29.8	26.1	22.5
Disgust	36.9	34.0	30.6
Angry	44.2	41.6	37.4
Fear	30.3	28.2	24.7