Speaker Identification in Emotional Environments

Ismail Shahin

Abstract—**The performance of speaker identification is almost perfect in the neutral environment. However, the performance is significantly deteriorated in emotional environments. In this work, three different and separate models have been used, tested and compared to identify speakers in each of the neutral and emotional environments (completely two separate environments). Our emotional environments in this work consist of five emotions. These emotions are: angry, sad, happy, disgust and fear. The three models are: Hidden Markov Models (HMMs), Second-Order Circular Hidden Markov Models (CHMM2s) and Suprasegmental Hidden Markov Models (SPHMMs). Our results show that the three models perform extremely well for speaker identification in the neutral environment. In emotional environments, the average speaker identification performance based on HMMs, CHMM2s and SPHMMs is 61.4%, 66.4% and 69.1%, respectively. Our results in this work are better than those obtained in subjective evaluation by human judges.**

Index Terms—**Emotional environments, hidden Markov models, neutral environment, second-order circular hidden Markov models, speaker identification, suprasegmental hidden Markov models.**

I. Introduction

Speaker recognition by machine (computer) is the process of recognizing he or she on the basis of the information obtained from his or her speech signal. Speaker recognition is divided into two categories: speaker identification and speaker verification (authentication). Speaker identification is the process of determining to which of the registered speakers a given utterance belongs. Speaker identification can be used in civil cases or for the media. These cases include calls to radio stations, local or other government authorities, insurance companies, or recorded conversations and many other applications [1]. Speaker verification is the process of accepting or rejecting the identity of the claimed speaker. The applications of speaker verification involve the use of voice as a key to confirm the identity claim of a speaker. Such services include banking transactions using a telephone network, database access services, security control for confidential information areas, remote access to computers and many other areas.

Based on the text to be spoken, speaker recognition methods typically can be grouped into text-dependent (fixed-text) or text-independent (free-text). In text-dependent, speaker recognition systems require the speaker to generate speech signals of the same text in both training and testing. On the other hand, in text-independent, speaker recognition systems do not require the speaker to generate speech signals of the same text in both training and testing. The process of speaker recognition can be divided into two sets: "open set" and "closed set". In the "open set", a reference model of the unknown speaker may not exist; whereas, in the "closed set", a reference model of the unknown speaker should be available.

Speaker recognition in emotional environments is one of research fields in human-computer interaction or affective computing [2]. Emotional environments can be defined as the environments where speakers produce their speech under the influence of emotional states such as sadness, anger and happiness. A major motivation comes from the desire to develop a human machine interface that is more adaptive and responsive to a user's identity in emotional environments. The main task of intelligent human-machine interaction is to empower a computer with the affective computing ability so that a computer can recognize the identity of the user in such environments for many different applications.

Speaker identification systems in emotional environments can be used in many applications. In telecommunications, emotional speaker identification systems can be used to enhance the telephone-based speech recognition performance, route emergency call services for high priority emergency calls and assess a caller's emotional state for telephone response services. Emotional speaker identification systems can also be used in the applications of emotionally intelligent automated systems in call-centers. In many cases, call-centers have a difficult task in managing customer disputes. It is very important for call-centers to take note of disputes using emotionally intelligent automated systems and successfully respond to these disputes to achieve the customers' satisfaction.

In literature, there are many studies that focus on speaker recognition in the neutral environment. Neutral environment is defined as the environment in which speech is produced assuming that speakers are not under the influence of any emotion. Furui focused on speaker feature extraction, recognition and processing in the neutral environment [1]. Zheng and Yuan implemented circular hidden Markov models for speaker identification in the neutral environment [3]. Shahin used second-order hidden Markov models to enhance speaker identification performance in the neutral environment [4]. Farrell *et al.* applied neural networks and conventional classifiers for speaker recognition systems in the neutral environment [5]. On the other hand, there are few studies that focus on speaker recognition in emotional environments. Zetterholm focused his attention on the prosody and voice quality in

Manuscript received December 6, 2007; revised August 12, 2008.

I. Shahin is with the Department of Electrical and Computer Engineering, University of Sharjah, Sharjah, United Arab Emirates. (e-mail: ismail@sharjah.ac.ae).

Publisher Item Identifier S 1682-0053(09)1675

the expression of emotions [6]. Koike *et al.* studied prosodic parameters in emotional speech [7]. Pereira and Watson studied some acoustic characteristics of emotion [8]. In one of our previous studies, we identified speakers using their emotions (emotion-dependent speaker identification) [9]. Wu *et al.* focused their study on the influence of emotion on the performance of a GMM-UBM based speaker verification system [10]. Tao *et al.* concentrated on prosody conversion from neutral speech to emotional speech [11].

Our contribution in this work includes studying and enhancing text-dependent speaker identification in each of the neutral and emotional environments based on each of hidden Markov models (HMMs), second-order circular hidden Markov models (CHMM2s) and suprasegmental hidden Markov models (SPHMMs). Our emotional environments in this work consist of five emotions. These emotions are angry, sad, happy, disgust and fear.

This paper is organized as follows. Section II focuses on second-order circular hidden Markov models. The details of suprasegmental hidden Markov models are given in Section III. Section IV describes the speech database used in this work. Section V discusses the algorithm that has been used for speaker identification in each of the neutral and emotional environments based on each of HMMs, CHMM2s and SPHMMs. Results and discussion are given in Section VI. Concluding remarks are drawn in Section VII.

II. Second-Order Circular Hidden Markov Models

HMMs have become one of the most successful and broadly used modeling techniques in the fields of speech and speaker recognition in the last three decades [12], [13]. HMMs provide a very useful paradigm to model the dynamics of speech signals. They provide a solid mathematical formulation for the problem of learning HMM parameters from speech observations. More details about HMMs are available in many references [12], [13].

CHMM2s have been proposed, implemented and tested in previous study by Shahin to enhance the performance of text-dependent speaker identification under the shouted talking condition [14]. CHMM2s have proven to be superior models over each of first-order left-to-right hidden Markov models (LTRHMM1s), second-order left-to-right hidden Markov models (LTRHMM2s) and first-order circular hidden Markov models (CHMM1s) [14]. This is because CHMM2s possess the characteristics of both CHMMs and HMM2s:

1. The underlying state sequence in HMM2s is a second-order Markov chain where the stochastic process is specified by a 3-D matrix because in these models the state-transition probability at time $t+1$ depends on the states of the Markov chain at the two times t and $t-1$. On the other hand, the underlying state sequence in HMM1s is a first-order Markov chain where the stochastic process is specified by a 2-D matrix because in these models it is assumed that the state-transition probability at time $t+1$ depends only on the state of the Markov chain at time t.
2. The Markov chain in CHMMs is more powerful than that in LTRHMMs in modeling the changing statistical characteristics that exist in the actual observations of speech signals.
3. The absorbing state in LTRHMMs governs the fact that the rest of a single observation sequence provides no further information about earlier states once the underlying Markov chain reaches the absorbing state. In speaker identification systems, it is true that a Markov chain should be able to revisit the earlier states because the states of HMMs reflect the vocal organic configuration of the speaker. Therefore, the vocal organic configuration of the speaker is reflected to states more conveniently using CHMMs than those using LTRHMMs.

The initial elements of the parameters in the training phase of CHMM2s are [14]:

The initial element of the probability of an initial state distribution is given by

$$v_k(i) = \frac{1}{N} \qquad N \geq i, k \geq 1 \tag{1}$$

where, N is the number of states.

The initial element of the forward probability of producing the observation O_1 is given by

$$\alpha_1(i,k) = v_k(i) b_{ki}(o_1) \qquad N \geq i, k \geq 1 \tag{2}$$

The initial element of a_{ijk} is given as

$$a_{ijk}^1 = \begin{cases} \frac{1}{3} & i = 1, j, k = 1,2,...,N \\ \frac{1}{3} & N-1 \geq i \geq 2, i+1 \geq j \geq i-1, N \geq k \geq 1 \\ \frac{1}{3} & i = N, j, k = 1,2,...,N \\ 0 & otherwise \end{cases} \tag{3}$$

The initial element of the observation symbol probability is given by

$$b_{ijk}^1 = \frac{1}{M} \qquad N \geq j, k \geq 1, M \geq i \geq 1 \tag{4}$$

where, M is the number of observation symbols.

The initial element of the backward probability of producing the observation O_T is given by

$$\beta_T(j,k) = \frac{1}{N} \qquad N \geq j, k \geq 1 \tag{5}$$

The probability of the observation vector O given the CHMM2s model Φ, can be calculated as

$$P(O \mid \Phi) = \sum_{k=1}^{N} \sum_{i=1}^{N} \alpha_T(i,k) \tag{6}$$

More details about the second-order circular hidden Markov models can be found in [14].

III. Suprasegmental Hidden Markov Models

SPHMMs allow us to summarize several states with HMMs into what is called suprasegmental state. Suprasegmental states are able to look at the observation sequence through a larger window to capture the prosodic properties. Such states allow observations at rates suitable for the situation of emotional modeling. For instance, prosodic information can not be observed at a rate that is

used for acoustic modeling. Therefore, prosodic events are modeled using suprasegmental states, while acoustic events are modeled using conventional hidden Markov states.

Suprasegmental information plays a major role in human decoding of speech. The term suprasegmental was introduced by Lehiste as a cover term to speech phenomena which are attributed to speech segments larger than phonemes [15]. Syllables, words and phrases are examples for such segments. To these segments we attribute perceived properties such as pitch, speaking rate, loudness, voice quality, rhythm, duration and pause. The prosodic functions which are generally considered to be the most important ones in human-human communication are phrase boundaries, accents and sentence mood. Lea proposed the use of prosodic information in the Automatic Speech Understanding (ASU) systems [16]. Seventeen years later, a German speech-to-speech translation system called VERBMOBIL was completed as the world wide first complete speech understanding system, where prosody was really used. It was shown by the VERBMOBIL system that the used and implemented prosody yielded remarkable performance improvement [16].

Polzin and Waibel showed how prosodic information can be combined and integrated with acoustic information within HMMs in detecting emotions in speech [17]. In our work, prosodic and acoustic information of speaker identification can be combined and integrated as given by the formula.

$$\log P(\lambda^v, \Psi^v \mid o) = (1-\alpha).\ \log P(\lambda^v \mid O) + \alpha.\ \log P(\Psi^v \mid O) \quad (7)$$

where α is a weighting factor that is chosen to be equal to 0.5 (so no biasing towards any model), λ^v is the acoustic model for the vth speaker, Ψ^v is the suprasegmental model for the vth speaker, and O is the observation vector or sequence of an utterance.

The previous formula shows that each time we leave a suprasegmental state we need to add the log probability of this suprasegmental state given the respective suprasegmental observations within the speech signal to the log probability of the current acoustic model given the respective acoustic observations within the speech signal.

IV. Speech Database

Our speech corpus was collected from 40 (20 males and 20 females) nonprofessional (therefore our speech database is closer to the real-life data than to the acted data) healthy adult Native speakers of American English. Each speaker uttered 8 sentences where each sentence was uttered 9 times under each emotion in emotional environments and under the neutral state in the neutral environment. Before uttering the sentences, the speakers listened to some recorded sentences that were uttered under each emotion. The 8 sentences were unbiased towards any emotion when uttered under the neutral state. These sentences were:

1. *He works five days a week.*
2. *The sun is shining.*
3. *The weather is fair.*
4. *The students study hard.*
5. *Assistant professors are looking for promotion.*
6. *University of Sharjah.*
7. *Electrical and Computer Engineering Department.*
8. *He has two sons and two daughters.*

Our speech database was captured by a speech acquisition board using a 16-bit linear coding A/D converter and sampled at a sampling rate of 16 kHz. Our database was a 16-bit per sample linear data.

In this work, our features representing the phonetic content of speech signals were called the short time log frequency power coefficients (LFPCs). LFPCs have proven to be superior features over each of the linear prediction cepstral coefficients (LPCCs) and the Mel-frequency cepstral coefficients (MFCCs) in the emotional speech and speaker recognition systems [18]. Our speech database in this work was a "closed set".

V. Speaker Identification Algorithm Based on Each IFF HMMS, CHMM2S, and SPHMMS

In this work, each of HMMs, CHMM2s and SPHMMs has been separately used for speaker identification in each of the neutral and emotional environments. There are many studies that focus on speaker recognition based on HMMs [1], [3], [19].

A. The Algorithm Based on HMMs

Our recognizer in this work adopted LFPC as feature parameters of each of neutral and emotional speech to represent energy distribution across the frequency spectrum and an LTRHMM was used as the classifier. The number of states, N, was 9 (this number is adequate for the used utterances). The number of mixture components, M, was 5 per state, with a continuous mixture observation density was selected for an LTRHMM as the recognizer. In the last three decades, the majority of the studies performed in the fields of speech and speaker recognition on HMMs have been done using LTRHMMs because phonemes follow strictly the left to right (LTR) sequence [12], [13].

In our training session based on HMMs for each of the neutral and emotional environments (completely two separate environments), one reference model for each speaker was derived using 5 of the 9 utterances per the same speaker per the same sentence under the neutral emotion. The reference models were common for both the neutral and emotional environments. The number of utterances in the training session was 1600.

In our testing (identification) session based on HMMs in each of the neutral and emotional environments, each one of the 40 speakers used 4 of the 9 utterances per the same sentence (text-dependent) in the neutral environment and 9 utterances per the same sentence under each of the 5 emotions in emotional environments. The number of utterances in the neutral and the emotional testing sessions was composed of 1280 and 14400, respectively. In each separate environment, the probability of generating every utterance was computed, the model with the highest probability was chosen as the output of speaker identification as given in the following formula

$$V^* = \arg\max_{40 \geq v \geq 1} \{P(O \mid \lambda^v)\} \quad (8)$$

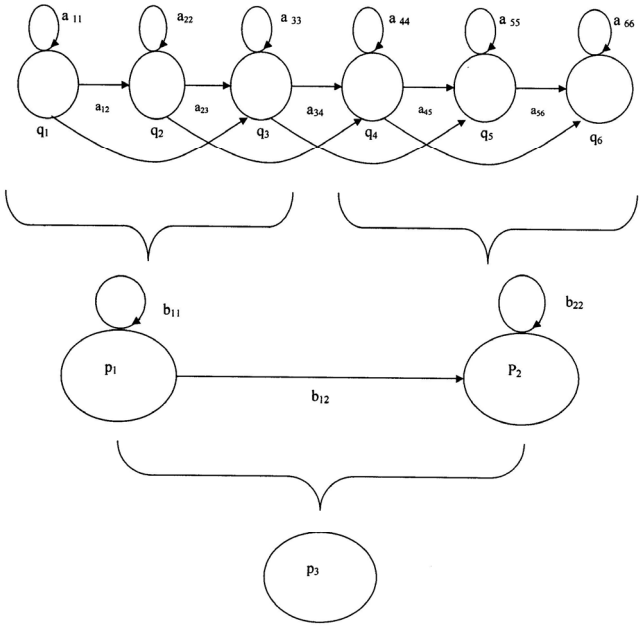

Fig. 1. Basic structure of LTRSPHMM derived from LTRHMM.

where, V^* is the index of the identified speaker, O is the observation vector or sequence that belongs to the unknown speaker and $P(O|\lambda^v)$ is the probability of the observation sequence O given the vth acoustic model λ^v.

B. The Algorithm Based on CHMM2s

In this work, LFPCs were used as feature parameters of each of neutral and emotional speech to represent energy distribution across the frequency spectrum and an CHMM2 was used as the classifier. Based on these models, the number of states was 9. The number of mixture components was 5 per state, with a continuous mixture observation density was selected for an CHMM2 as the recognizer.

In these models, the number of utterances in the training session was 1600. The number of utterances in the neutral and emotional testing sessions was made up of 1280 and 14400, respectively. In each environment, the probability of generating every utterance was computed as given in the following formula

$$V^* = \arg\max_{40\geq v\geq 1} \{P(O|\Phi^v)\} \quad (9)$$

where, $P(O|\Phi^v)$ is the probability of the observation sequence O given the vth CHMM2s model Φ^v.

C. The Algorithm Based on SPHMMs

Our recognizer in this work used LFPC as feature parameters of speech signals in each of the neutral and emotional environments and a left-to-right suprasegmental hidden Markov model (LTRSPHMM) was used as the classifier. The number of states was 9. The number of mixture components was 5 per state, with a continuous mixture observation density was selected for an LTRSPHMM as the recognizer. In this work, LTRSPHHM was derived from LTRHMM. Fig. 1 shows our adopted structure of LTRSPHMM that was derived from LTRHMM. In this figure, $q_1, q_2, ..., q_6$ are hidden Markov states. p_1 is a suprasegmental state (e.g. phone) that consists of q_1, q_2 and q_3. p_2 is a suprasegmental state (e.g. phone) that is made up of q_4, q_5 and q_6. p_3 is a suprasegmental state (e.g., syllable) that is composed of p_1 and p_2. a_{ij} is the transition probability between the ith hidden Markov state and the jth hidden Markov state, while b_{ij} is the transition probability between the ith suprasegmental state and the jth suprasegmental state.

Our training session of SPHMMs was similar to the training session of the conventional HMMs. In our training session of SPHMMs, suprasegmental models were trained on top of acoustic models. In each of the neutral and emotional environments, one reference model was derived using 5 of the 9 utterances per the same speaker per the same sentence under the neutral emotion. The number of utterances in the training session of each of the neutral and emotional environments was 1600.

In the testing session, each one of the 40 speakers used 4 of the 9 utterances per the same sentence in the neutral environment and 9 utterances per the same sentence under each of the 5 emotions in emotional environments. The number of utterances in the neutral and emotional testing sessions was composed of 1280 and 14400, respectively. The probability of generating every utterance was computed as given in the following formula

$$V^* = \arg\max_{40\geq v\geq 1} \{P(O \mid \lambda^v, \Psi^v)\} \quad (10)$$

TABLE I
SPEAKER IDENTIFICATION PERFORMANCE IN THE NEUTRAL ENVIRONMENT BASED ON HMMS

Gender	Speaker identification performance (%)
Male	99
Female	99
Average	99

TABLE II
SPEAKER IDENTIFICATION PERFORMANCE IN THE NEUTRAL ENVIRONMENT BASED ON CHMM2S

Gender	Speaker identification performance (%)
Male	99
Female	99
Average	99

TABLE III
SPEAKER IDENTIFICATION PERFORMANCE IN THE NEUTRAL ENVIRONMENT BASED ON SPHMMS

Gender	Speaker identification performance (%)
Male	99
Female	99
Average	99

VI. RESULTS AND DISCUSSION

Tables I-III summarize the results of speaker identification performance in the neutral environment based on HMMs, CHMM2s and SPHMMs, respectively. It is evident from the three tables that each of HMMs, CHMM2 and SPHMMs perform almost perfect in this environment. Therefore, the three models are extremely powerful in such an environment. Our results in this environment are consistent with the results obtained in previous studies. Based on Gaussian mixture models (GMMs), Reynolds achieved speaker identification performance of 99.5% using TIMIT database [20]. Shahin and Botros obtained speaker identification performance of 100% based on dynamic time warping (DTW) [21].

TABLE IV
SPEAKER IDENTIFICATION PERFORMANCE IN EMOTIONAL ENVIRONMENTS BASED ON HMMS

Emotion	Speaker Identification Performance (%)		Average (%)
	Male	Female	
Angry	54	54	54
Sad	61	62	61.5
Happy	65	64	64.5
Disgust	62	64	63
Fear	63	65	64

TABLE V
SPEAKER IDENTIFICATION PERFORMANCE IN EMOTIONAL ENVIRONMENTS BASED ON CHMM2S

Emotion	Speaker identification performance (%)		Average (%)
	Male	Female	
Angry	58	59	58.5
Sad	66	66	66
Happy	70	70	70
Disgust	68	69	68.5
Fear	69	69	69

In emotional environments, HMMs perform poorly as shown clearly in Table IV. The performance of speaker identification based on HMMs in such environments has been significantly deteriorated compared to that in the neutral environment. Table IV shows that the average speaker identification performance in emotional environments based on HMMs is 61.4%.

Table V shows apparently that speaker identification performance in emotional environments based on CHMM2s has been greatly improved compared to that based on HMMs. This is because CHMM2s possess the characteristics of both CHMMs and HMM2s. This table shows that the average speaker identification performance in such environments is 66.4%.

Speaker identification performance in emotional environments based on SPHMMs is given in Table VI. This table shows that the average speaker identification performance in these environments based on SPHMMs is 69.1%. Therefore, SPHMMs are superior models over each of HMMs and CHMM2s in such environments. This may be attributed to the following reasons:

1. SPHMMs are convenient models to integrate observations from emotional modality because such models allow for observations at a rate appropriate for emotional modality.
2. SPHMMs possess more ability than each of HMMs and CHMM2s in capturing prosodic properties, which can reflect more emotional properties of speech signals. Speech signals in emotional environments differ from those in the neutral environment in many aspects including intonation, speaking rate and intensity.
3. Emotional environments are communicated by a subtle combination of features at all three levels of speech abstraction. These three levels are suprasegmental, segmental and intrasegmental [22].

It is noticeable from Tables IV, V, and VI that the least speaker identification performance in emotional environments occurs when speakers speak in the angry emotion. This is because the angry emotion can not be entirely separated from the shouted talking condition in our daily life [23]. It is well known that speaker identification

TABLE VI
SPEAKER IDENTIFICATION PERFORMANCE IN EMOTIONAL ENVIRONMENTS BASED ON SPHMMS

Emotion	Speaker identification performance (%)		Average (%)
	Male	Female	
Angry	61	62	61.5
Sad	68	69	68.5
Happy	72	73	72.5
Disgust	71	71	71
Fear	72	72	72

TABLE VII
RELATIVE IMPROVEMENT OF USING SPHMMS OVER USING EACH OF HMMS AND CHMM2S PER EMOTION

Emotion	Relative improvement of using SPHMMs over HMMs (%)	Relative improvement of using SPHMMs over CHMM2s (%)
Neutral	0	0
Angry	13.9	5.1
Sad	11.4	3.8
Happy	12.4	3.6
Disgust	12.7	3.7
Fear	12.5	4.3

performance under the shouted talking condition has been sharply degraded [14], [23]. These tables also show that the highest speaker identification performance in such environments happens when speakers speak in the happy emotion.

The relative improvement of using SPHMMs over using each of HMMs and CHMM2s per emotion is summarized in Table VII. This table shows that the highest relative improvement of using SPHMMs over using each of HMMs and CHMM2s happens under the angry emotion.

Comparing each of Table I with Table IV, Table II with Table V and Table III with Table VI, it is evident that speaker identification performs extremely well in the neutral environment; however, the performance decreases sharply in emotional environments. Therefore, the emotional states of speakers have significantly negative impact on the performance of speaker identification in emotional environments.

An informal subjective evaluation was carried out with 10 nonprofessional listeners (human judges). A total of 160 utterances (40 speakers with only 4 sentences each) were used in the neutral environment evaluation. In the emotional environment evaluation, a total of 800 utterances (40 speakers, 5 emotions and 4 sentences only) were used. The results of the evaluation were satisfactory and encouraging. The average speaker identification performance in the neutral and emotional environments is 98.5% and 60.5%, respectively. These averages are less than our achieved averages in this work.

VII. CONCLUDING REMARKS

Some conclusions can be drawn from this work. First, speaker identification in the neutral environment is very close to 100% based on each of HMMs, CHMM2s and SPHMMs. Based on each of these three models separately, speaker identification in emotional environments has been significantly improved. Our results show evidently that SPHMMs are superior models over each of HMMs and CHMM2s for speaker identification in emotional environments. Second, prosodic features are very important

for speech signals in emotional environments; whereas, acoustic features are convenient for speech signals in the neutral environment. Third, the field of human emotions is an enormously complicated field of study. This may be attributed to a number of considerations which include: systems in emotional environments depend on many areas such as signal processing and analysis techniques, psychology, physiology and linguistics. Finally, speaker identification performance in emotional environments based on SPHMMs is limited. This limitation needs to be studied thoroughly in future work.

REFERENCES

[1] S. Furui, "Speaker-dependent-feature-extraction, recognition, and processing techniques," *Speech Communication*, vol. 10, pp. 505-520, Mar. 1991.

[2] R. W. Picard, *Affective Computing*, MIT Media Lab Perceptual Computing Section Tech. Rep., no. 321, 1995.

[3] C. Zheng, and B. Z. Yuan, "Text-dependent speaker identification using circular hidden Markov models," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 580-582, Mar. 1988.

[4] I. Shahin, "Using second-order hidden Markov model to improve speaker identification recognition performance under neutral condition," *in Proc. 10th IEEE Int. Conf. on Electronics, Circuits and Systems, ICECS'03*, pp. 124-127, Sharjah, United Arab Emirates, Dec. 2003,.

[5] K. R. Farrell , R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, Pt. II, pp. 194-205, Jan. 1994.

[6] E. Zetterholm, "Prosody and voice quality in the expression of emotions," in *Proc. of Int. Conf. on Spoken Language Processing, ICSLP'98*, 30 Nov.-4 Dec. 1998.

[7] K. Koike, H. Suzuki, and H. Saito, "Prosodic parameters in emotional speech," in *Proc. of Int. Conf. on Spoken Language Processing ICSLP'98*, pp. 679-682, 30 Nov.-4 Dec. 1998.

[8] C. Pereira and C. Watson, "Some acoustic characteristics of emotion," in *Proc. of Int. Conf. on Spoken Language Processing, ICSLP'98*, pp. 927-930, 30 Nov.-4 Dec. 1998.

[9] I. Shahin, "Using emotions to identify speakers," *The 5thP Int. Workshop on Signal Processing and its Applications, WoSPA'08*, Sharjah, United Arab Emirates, Mar. 2008.

[10] W. Wu, T. F. Zheng, M. X. Xu, and H. J. Bao, "Study on speaker verification on emotional speech," in *Proc. of Int. Conf. on Spoken Language Processing, INTERSPEECH 2006*, pp. 2102-2105, Sep. 2006.

[11] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145-1154, Jul. 2006.

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

[13] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Eaglewood Cliffs, New Jersey, 1993.

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

[15] I. Lehiste, *Surasegmentals*, MIT Press, Cambridge, MA, 1970.

[16] W. Lea, *Prosodic aids to speech recognition. In W. Lea, editor, Trends in Speech Recognition*, pp. 166-205, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1980.

[17] T. S. Polzin and A. H. Waibel, "Detecting emotions in Speech," in *Proc. Second Int. Conf. Cooperative Multimodal Communication,* 1998.

[18] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, issue 4, pp. 603-623, Nov. 2003.

[19] I. Shahin, "Enhancing speaker authentication systems using circular hidden Markov models," in *Proc. Eighth Int. Symp. on Signal Processing and Its Applications, ISSPA'05*, vol. 2, pp. 703-706, Sydney, Australia, Aug. 2005.

[20] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, issues 1-2, pp. 91-108, Aug. 1995.

[21] I. Shahin and N. Botros, "Speaker identification using dynamic time warping with stress compensation technique," in *Proc. IEEE SOUTHEASTCON'98*, pp. 65-68, Orlando, FL, US, Apr. 1998.

[22] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion," *J. of the Acoustic Society of America*, vol. 93, no. 2, pp. 1097-1108, 1993.

[23] I. Shahin, "Improving speaker identification performance under the shouted talking condition using the second-order hidden Markov models," *EURASIP J. on Applied Signal Processing*, vol. 5, no. 4, pp. 482-486, Mar. 2005.

Ismail Shahin was born in Hebron, Palestine, on June 30, 1966. He received his B.Sc., M.Sc., and Ph.D. degrees in electrical engineering in 1992, 1994, and 1998, respectively, from Southern Illinois University at Carbondale, US.

From 1998 to 1999, he was a Visiting Instructor in the Department of Electrical Engineering and the Computer Science Department at Southern Illinois University at Carbondale. Since 1999, he is an assistant professor in the Electrical and Computer Engineering Department at the University of Sharjah in the United Arab Emirates. His research interests include speech processing, speech recognition, speaker recognition in each of the neutral, stressful and emotional environments, emotion and talking condition recognition, gender recognition using voice and accent recognition. He has 26 journal and conference publications. He has remarkable contribution in organizing many conferences, symposiums and workshops.

Dr. Shahin is a member of the Institute of Electrical and Electronics Engineers (IEEE) and a member of the editorial board of the International Journal of Applied Engineering Research (IJAER).