EURASIP Journal on Applied Signal Processing 2005:4, 482–486

# Improving Speaker Identification Performance Under the Shouted Talking Condition Using the Second-Order Hidden Markov Models

**Ismail Shahin**

*Electrical/Electronics and Computer Engineering Department, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates*
*Email: ismail@sharjah.ac.ae*

*Received 13 June 2004; Revised 22 September 2004; Recommended for Publication by Chin-Hui Lee*

Speaker identification systems perform well under the neutral talking condition; however, they suffer sharp degradation under the shouted talking condition. In this paper, the second-order hidden Markov models (HMM2s) have been used to improve the recognition performance of isolated-word text-dependent speaker identification systems under the shouted talking condition. Our results show that HMM2s significantly improve the speaker identification performance compared to the first-order hidden Markov models (HMM1s). The average speaker identification performance under the shouted talking condition based on HMM1s is 23%. On the other hand, the average speaker identification performance based on HMM2s is 59%.

**Keywords and phrases:** first-order hidden Markov models, second-order hidden Markov models, shouted talking condition, speaker identification performance.

## 1. MOTIVATION

Stressful talking conditions are defined as talking conditions that cause a speaker to vary his/her production of speech from the neutral talking condition. The neutral talking condition is defined as the talking condition in which speech is produced assuming that the speaker is in a "quiet room" with no task obligations.

Some talking conditions are designed to simulate speech produced by different speakers under real stressful talking conditions. Hansen, Cummings, and Clements used speech under simulated and actual stress (SUSAS) database in which eight talking styles are used to simulate the speech produced under real stressful talking conditions and three real talking conditions [1, 2, 3]. The eight conditions are as follows: neutral, loud, soft, angry, fast, slow, clear, and question. The three conditions are 50% task, 70% task and Lombard. Chen used six talking conditions to simulate speech under real stressful talking conditions [4]. These conditions are as follows: neutral, fast, loud, Lombard, soft, and shouted.

Most published works in the areas of speech recognition and speaker recognition focus on speech under the neutral talking condition and few published works focus on speech under stressful talking conditions. The vast majority of the studies that focus on speech under stressful talking conditions ignore the shouted talking condition [4, 5, 6]. The shouted talking condition can be defined as follows: when a speaker shouts, his/her object is to produce a very loud acoustic signal to increase either its range (distance) of transmission or its ratio to background noise.

## 2. INTRODUCTION

Hidden Markov model (HMM) is one of the most widely used modeling techniques in the fields of speech recognition and speaker recognition [7]. HMMs use Markov chain to model the changing statistical characteristics that exist in the actual observations of speech signals. The Markov process is a double stochastic process where there is an unobservable Markov chain defined by a state transition matrix. Each state of the Markov chain is associated with either a discrete output probability distribution (discrete HMMs) or a continuous output probability density function (continuous HMMs) [8].

HMMs are powerful models in optimizing the parameters that are used in modeling speech signals. This optimization decreases the computational complexity in the decoding procedure and improves the recognition accuracy [8]. Most of the work performed in the fields of speech recognition and speaker recognition using HMMs has been done using HMM1s [4, 7, 9, 10]. Despite the success of using HMM1s, experimental evidence suggests that using HMM2s in the training and testing phases of isolated-word text-dependent speaker identification systems gives better speaker identification performance than HMM1s under the shouted talking condition.

Despite the success of using HMM1s, it is still worth investigating if some of the drawbacks of HMM1s can be overcome by using higher-order Markov processes (like the proposed HMM2s in this work). HMM1s suffer from the following drawbacks [11].

(i) The frames for a particular state are assumed to be independent.

(ii) The dependencies of adjacent frames for a particular state are not incorporated by the model.

In this paper, HMM2s are used in the training and testing phases of isolated-word text-dependent speaker identification systems under each of the neutral and shouted talking conditions.

Our work differs from the work in [11, 12] is that our work focuses on isolated-word text-dependent speaker identification systems under the shouted talking condition based on HMM2s, while the work in [11, 12] focuses on describing a connected word recognition system under the neutral talking condition based on HMM2s. The work in [11, 12] shows that the recognition performance using HMM2s yields better results than using HMM1s.

## 3. BRIEF OVERVIEW OF HIDDEN MARKOV MODELS

HMMs can be described as being in one of the $N$ distinct states, $1, 2, 3, \ldots, N$, at any discrete time instant $t$. The individual states are denoted as

$$s = \{s_1, s_2, s_3, \ldots, s_N\}. \tag{1}$$

HMMs are generators of a state sequence $q_t$, where at any time $t : q = \{q_1, q_2, q_3, \ldots, q_T\}$, $T$ is the length or duration of an observation sequence $O$ and is equal to the total number of frames.

At any discrete time $t$, the model is in a state $q_t$. At the discrete time $t$, the model makes a random transition to a state $q_t$. The state transition probability matrix **A** determines the probability of the next transition between states:

$$\mathbf{A} = [a_{ij}], \quad i, j = 1, 2, \ldots, N, \tag{2}$$

where $a_{ij}$ denotes the transition probability from a state $i$ to a state $j$.

The first state $s_1$ is selected randomly according to the initial state probability:

$$\pi = [\pi_i] = \text{Prob}\,(q_1 = s_i). \tag{3}$$

The states that are unobservable directly are observable via a sequence of outputs or an observation sequence given as

$$O = \{O_1, O_2, O_3, \ldots, O_T\} \tag{4}$$

which are taken from a finite discrete set of observation symbols

$$V = \{V_1, V_2, V_3, \ldots, V_k\}, \quad O_t \in V. \tag{5}$$

When the model is in any state, say a state $s_j$, the selection of an output discrete symbol $V_k$ is governed according to the observation symbol probability given as

$$B = \{b_j(V_k)\} = \text{Prob}\,(V_k \text{ emitted at } t | q_{t-1} = s_i), \quad N \geq j \geq 1,\ K \geq k \geq 1. \tag{6}$$

## 4. SECOND-ORDER HIDDEN MARKOV MODELS

In HMM1s, the underlying state sequence is a first-order Markov chain where the stochastic process is specified by a 2D matrix of a priori transition probabilities $(a_{ij})$ between states $s_i$ and $s_j$ where $a_{ij}$ are given as

$$a_{ij} = \text{Prob}\,(q_t = s_j | q_{t-1} = s_i). \tag{7}$$

Many researchers have noticed that the transition probabilities of HMM1s have a negligible impact on the recognition performance of systems and can be ignored [12].

In HMM2s, the underlying state sequence is a second-order Markov chain where the stochastic process is specified by a 3D matrix $(a_{ijk})$. Therefore, the transition probabilities in HMM2s are given as [11]

$$a_{ijk} = \text{Prob}\,(q_t = s_k | q_{t-1} = s_j,\ q_{t-2} = s_i) \tag{8}$$

with the constraints

$$\sum_{k=1}^{N} a_{ijk} = 1, \quad N \geq i,\ j \geq 1. \tag{9}$$

The probability of the state sequence, $Q \triangleq q_1, q_2, \ldots, q_T$, is defined as

$$\text{Prob}(Q) = \Psi_{q_1} a_{q_1 q_2} \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t}, \tag{10}$$

where $\Psi_i$ is the probability of a state $s_i$ at time $t = 1$, $a_{ij}$ is the probability of the transition from a state $s_i$ to a state $s_j$ at time $t = 2$.

Each state $s_i$ is associated with a mixture of Gaussian distributions:

$$b_i(O_t) \triangleq \sum_{m=1}^{M} c_{\text{im}} N\Big(O_t, \mu_{\text{im}}, \sum_{\text{im}}\Big), \quad \sum_{m=1}^{M} c_{\text{im}} = 1, \tag{11}$$

where the vector $O_t$ is the input vector at time $t$.

Given a sequence of observed vectors, $O \triangleq O_1, O_2, \ldots, O_T$, the joint state-output probability is defined as

$$\text{Prob}\,(Q, O | \lambda) = \Psi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} b_{q_t}(O_t). \tag{12}$$

Table 1: Speech database under the neutral and shouted talking conditions.

| Models | Session | Total number of utterances under the neutral talking condition | Total number of utterances under the shouted talking condition |
|---|---|---|---|
| HMM1s | First training session | 1000 male utterances<br>1000 female utterances | 0 male utterance<br>0 female utterance |
| | Second testing session | 800 male utterances<br>800 female utterances | 1800 male utterances<br>1800 female utterances |
| HMM2s | First training session | 1000 male utterances<br>1000 female utterances | 0 male utterance<br>0 female utterance |
| | Second testing session | 800 male utterances<br>800 female utterances | 1800 male utterances<br>1800 female utterances |

## 5. EXTENDED VITERBI AND BAUM-WELCH ALGORITHMS

The most likely state sequence can be found by using the probability of the partial alignment ending at a transition $(s_j, s_k)$ at times $(t-1, t)$:

$$\delta_t(j,k) \triangleq \text{Prob}\,(q_1,\dots,q_{t-1} = s_j,\ q_t = s_k,\ O_1, O_2,\dots,O_t|\lambda),\\ T \geq t \geq 2,\ N \geq j,\ k \geq 1. \tag{13}$$

Recursive computation is given by

$$\delta_t(j,k) = \max_{N\geq i\geq 1}\{\delta_{t-1}(i,j)\cdot a_{ijk}\}\cdot b_k(O_t),\\ T \geq t \geq 3,\ N \geq j,\ k \geq 1. \tag{14}$$

The forward function $\alpha_t(j,k)$ defines the probability of the partial observation sequence, $O_1, O_2,\dots,O_t$, and the transition $(s_j, s_k)$ between times $t-1$ and $t$ is given by

$$\alpha_t(j,k) \triangleq \text{Prob}\,(O_1,\dots,O_t,\ q_{t-1} = s_j,\ q_t = s_k|\lambda),\\ T \geq t \geq 2,\ N \geq j,\ k \geq 1. \tag{15}$$

$\alpha_t(j,k)$ can be computed from the two transitions $(s_i, s_j)$ and $(s_j, s_k)$ between states $s_i$ and $s_k$ as

$$\alpha_{t+1}(j,k) = \sum_{i=1}^{N}\alpha_t(i,j)\cdot a_{ijk}\cdot b_k(O_{t+1}),\\ T-1 \geq t \geq 2,\ N \geq j,\ k \geq 1. \tag{16}$$

The backward function $\beta_t(i,j)$ can be expressed as

$$\beta_t(i,j) \triangleq \text{Prob}\,(O_{t+1},\dots,O_T|q_{t-1} = s_i,\ q_t = s_j,\lambda),\\ T-1 \geq t \geq 2,\ N \geq i,\ j \geq 1, \tag{17}$$

where $\beta_t(i,j)$ is defined as the probability of the partial observation sequence from $t$+1 to $T$, given the model $\lambda$ and the transition $(s_i, s_j)$ between times $t-1$ and $t$.

## 6. SPEECH DATABASE

In this work, our speech database consists of 40 different speakers (20 adult males and 20 adult females). Each speaker utters the same 10 different isolated words under each of the neutral and shouted talking conditions. These words are alphabet, eat, fix, meat, nine, order, processing, school, six, yahoo. The length of these words ranges from 1 to 3 seconds.

In the first session (training session), each speaker utters each word 5 times (5 utterances per word) under the neutral talking condition. In this session, one reference model per speaker per word is derived using the 5 utterances per the same speaker per the same word. Training of models in this session has been done based on HMM1s.

In another different session (testing or recognition session), each one of the 40 speakers utters the same word (text-dependent) 4 times under the neutral talking condition and 9 times under the shouted talking condition. The recognition phase in this session has been done based on HMM1s.

The second training session has been done like the first training session but based on HMM2s. The second testing session has been done like the first testing session but based on HMM2s.

Training of models in the two sessions uses the forward-backward algorithm, whereas recognition in the two sessions uses the Viterbi decoding algorithm. Our speech database is summarized in Table 1.

Our speech database was captured by a speech acquisition board using a 10-bit linear coding A/D converter (we believe that a 10-bit linear coding A/D converter is sufficient to convert an analog speech signal to a digital speech signal) and sampled at a sampling rate of 8 kHz. Our database consists of a 10-bit per sample linear data. A high emphasis filter, $H(z) = 1 - 0.95\,z^{-1}$, was applied to the speech signals, and a 30 milliseconds Hamming window was applied to the emphasized speech signals every 10 milliseconds. Twelfth-order linear prediction (LP) coefficients were extracted from each frame by the autocorrelation method. The 12 LP coefficients were transformed into 12 LP cepstral coefficients (LPCCs).

In each of HMM1s and HMM2s, LPCC feature analysis was used to form the observation vectors. The number of states, $N$, was 5. The number of mixture components, $M$, was 5 per state, with a continuous mixture observation density selected for each of HMM1s and HMM2s.

Table 2: Speaker identification performance for 20 male speakers, 20 female speakers, and their averages under each of the neutral and shouted talking conditions based on each of HMM2s and HMM1s.

| Models | Gender | Neutral | Shouted |
|---|---|---|---|
| HMM2s | Males | 92% | 57% |
| | Females | 96% | 61% |
| | Average | 94% | 59% |
| HMM1s | Males | 89% | 21% |
| | Females | 91% | 25% |
| | Average | 90% | 23% |

## 7. RESULTS

Based on the probability of generating an utterance, the model with the highest probability was chosen as the output of the speaker identification system.

Table 2 summarizes the results of the speaker identification performance for 20 male speakers, 20 female speakers, and their averages of 10 different isolated words under each of the neutral and shouted talking conditions based on each of HMM2s and HMM1s. Our results show that using HMM2s in the training and testing phases of isolated-word text-dependent speaker identification systems under the shouted talking condition significantly improves the identification performance compared to that using HMM1s.

## 8. DISCUSSION AND CONCLUSIONS

This work is based on an isolated-word text-dependent HMM2 speaker identifier trained by speech uttered under the neutral talking condition and tested by speech uttered under each of the neutral and shouted talking conditions. This is the first known investigation into HMM2s evaluated under the shouted talking condition for speaker identification systems.

This work shows that HMM2s significantly improve the recognition performance of isolated-word text-dependent speaker identification systems under the shouted talking condition. The average speaker identification performance under the shouted talking condition has been improved from 23% based on HMM1s to 59% based on HMM2s. The experimental evidence suggests that HMM2s outperform HMM1s under such a condition. This may be attributed to a number of considerations.

(1) In HMM2s, the state-transition probability at time $t+1$ depends on the states of the Markov chain at times $t$ and $t-1$. Therefore, the underlying state sequence in HMM2s is a second-order Markov chain where the stochastic process is specified by a 3D matrix. On the other hand, in HMM1s, it is assumed that the state-transition probability at time $t+1$ depends only on the state of the Markov chain at time $t$. Therefore, in HMM1s, the underlying state sequence is a first-order Markov chain where the stochastic process is specified by a 2D matrix. The stochastic process that is specified by a 3D matrix gives more accurate recognition performance than that specified by a 2D matrix.

(2) HMM2s eliminate singular alignments given by the Viterbi algorithm in the recognition process when a state captures just one frame and all other frames fall into the neighboring states. Thus, the trajectory of speech under the shouted talking condition, in terms of a state sequence, is better modeled by HMM2s than that by HMM1s.

Table 3: Speaker identification performance based on each of HMM2s and HMM1s for 9 male speakers under each of the neutral and angry talking conditions using SUSAS database.

| Models | Neutral | Angry |
|---|---|---|
| HMM2s | 93% | 58% |
| HMM1s | 91% | 25% |

In this work, the average speaker identification performance under the neutral talking condition has been improved slightly based on HMM2s compared to that based on HMM1s. Our results show that the average speaker identification performance under the neutral talking condition has been improved from 90% based on HMM1s to 94% based on HMM2s. In another work, the average speaker identification performance under the same talking condition was 90% based on HMM1s and 98% based on HMM2s [13].

Table 2 shows that the average speaker identification performance under the neutral talking condition based on HMM1s is 90%. On the other hand, the average speaker identification performance under the shouted talking condition based on HMM1s is 23%. Therefore, HMM1s are not powerful models under the shouted talking condition.

More extensive experiments have been conducted to show that HMM2s work better than HMM1s under the shouted talking condition. The following two experiments have been conducted in this work.

(1) Since the shouted talking condition can not be entirely separated from the angry talking condition in real life, HMM2s have been used to train and test speaker identification systems under the angry talking condition. SUSAS database has been used in the training and testing phases of isolated-word text-dependent speaker identification systems under the neutral and angry talking conditions (part of this database consists of 9 male speakers uttering words under these two talking conditions). Table 3 summarizes the results of the speaker identification performance based on each of HMM2s and HMM1s under each of the neutral and angry talking conditions using SUSAS database. Our results show that using HMM2s under the angry talking condition significantly improves the speaker identification performance compared to that using HMM1s.

Table 4: Speaker identification performance for 20 male speakers, 20 female speakers, and their averages under each of the neutral and shouted talking conditions based on HMM1s using the cepstral mean subtraction technique.

| Gender | Neutral | Shouted |
|---|---|---|
| Males | 89% | 39% |
| Females | 91% | 41% |
| Average | 90% | 40% |

A comparison between Table 2 and Table 3 shows that HMM2s dramatically improve the speaker identification performance under the shouted and angry talking conditions.

(2) An experiment has been conducted to compare the speaker identification performance based on HMM2s with that based on HMM1s using the stress compensation technique. It is well known that spectral tilt exhibits a large variation when a speaker utters a word under the shouted talking condition [4]. Such a variation usually contaminates the distance measure and is one of the most significant causes of degradation in the speaker identification performance. One of the stress compensation techniques that removes the spectral tilt and improves the speaker identification performance is the cepstral mean subtraction technique [14]. Table 4 summarizes the results of the speaker identification performance for the 20 male speakers, 20 female speakers, and their averages under each of the neutral and shouted talking conditions based on HMM1s using the cepstral mean subtraction technique.

Comparing Table 2 with Table 4, it is clear that HMM2s yield better speaker identification performance than HMM1s using the cepstral mean subtraction technique.

## REFERENCES

[1] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *J. Acoust. Soc. Amer.*, vol. 98, no. 1, pp. 88–98, 1995.

[2] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech, and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[3] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech, and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[4] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, no. 4, pp. 433–439, 1988.

[5] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 2, pp. 151–170, 1996, Special issue on speech under stress.

[6] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," *J. Acoust. Soc. Amer.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[7] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

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

[9] J. Dai, "Isolated word recognition using Markov chain models," *IEEE Trans. Speech, and Audio Processing*, vol. 3, no. 6, pp. 458–463, 1995.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] J. F. Mari, J. P. Haton, and A. Kriouile, "Automatic word recognition based on second-order hidden Markov models," *IEEE Trans. Speech, and Audio Processing*, vol. 5, no. 1, pp. 22–25, 1997.

[12] J. F. Mari, F. D. Fohr, and J. C. Junqua, "A second-order HMM for high performance word and phoneme-based continuous speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 1, pp. 435–438, Atlanta, Ga, USA, May 1996.

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

[14] I. Shahin and N. Botros, "Text-dependent speaker identification using hidden Markov model with stress compensation technique," in *Proc. IEEE Southeastcon '98*, pp. 61–64, Orlando, Fla, USA, April 1998.

**Ismail Shahin** was born in Hebron, Palestine, on June 30, 1966. He received his B.S., M.S., and Ph.D. degrees in electrical engineering in 1992, 1994, and 1998, respectively, from Southern Illinois University at Carbondale, USA. From 1998 to 1999, he was a Visiting Instructor in the Department of Electrical Engineering and the Computer Science, Southern Illinois University at Carbondale. Since 1999, he has been an Assistant Professor in the Electrical/Electronics and Computer Engineering Department, University of Sharjah, the United Arab Emirates. His research interests include speech processing, speech recognition, and speaker recognition (speaker identification and speaker authentication) under the neutral and stressful talking conditions.