2008 International Symposium on Intelligent Signal
Processing and Communication Systems (ISPACS2008)
Swissôtel Le Concorde,Bangkok,Thailand

ISOLATED PERSIAN DIGIT RECOGNITION USING A HYBRID HMM-SVM

S. A. Hejazi, R. Kazemi, and S. Ghaemmaghami
Sharif University of Technology, Tehran, Iran
s_amir_hejazi@sharif.edu, rezakazemi@ieee.org, ghaemmag@sharif.edu

***Abstract*-This paper introduces a new method for solving a traditional problem in isolated digits recognition in Persian language. The problem arises from pronunciation similarity of some Persian digits that are composed of very similar phonetic and spectral components. The process of recognition introduced here consists of three stages. First, the word is decomposed into small parts using efficient algorithms in order to make its Hidden Markov Model (HMM). Subsequently, based on this model, the most relevant candidates are chosen and introduced to a Support Vector Machine (SVM) based recognizer. At the final stage, the recognition is finalized by the SVM, with the aid of a novel idea that is to segment the input word and find an entry with the maximum number of similar segments. Experimental results show that the proposed method significantly improves both the recognition accuracy and the computational complexity in isolated Persian word recognition systems.**

I. INTRODUCTION

Introducing the Hidden Markov model (HMM) [1] was a breakthrough in recognition systems. Thanks to this method, many practical approaches utilizing different statistical models were then developed, which introduced noticeable improvements to speed and accuracy [2]. The strength of this probabilistic model is in classification of a wide range of data based on their statistical characteristics. However, the HMM may fail to classify sounds of highly similar phonetic structure [3], as is the case in Persian digits recognition.

The problem of HMM-based speech recognizers has been remarkably overcome in [4] and [5] using hybrid HMM/SVM systems. The SVM [6], relying on the idea of hyper-planes, has introduced a powerful tool for binary classification. The ability of this learning machine in classifying similar data has been employed in hybrid systems that have significantly improved the recognition accuracy.

The algorithms mentioned above are all designed and developed for English language and we found them inadequate to resolve the difficulty with the Persian digits. The main problem is due to very similar structure and pronunciations of some certain pairs of Persian digits that are identified erroneously in at least 15% of cases using the conventional recognition algorithms.

In this paper, we introduce a novel hybrid HMM/SVM system, effective for isolated Persian digit recognition, which is specifically capable of discerning words of similar phonetic structure with similar vowels and the same number and location of consonants. Speech recognition is a three-step process in this system. Using the statistical characteristics of the speech samples, the efficiencies of the HMM and the SVM models are improved. The proposed method is compared to the conventional HMM based systems in both clean and noisy environments. The results demonstrate significantly higher accuracy and robustness of the method in speaker independent Persian digits recognition.

The rest of the paper is organized as follows. Section 2 gives a brief discussion about the HMM model. The proposed hybrid HMM-SVM algorithm is presented in section 3. Section 4 introduces the Experimental results and a conclusion is made in section 5.

II. HMM IN A NUTSHELL

A Markov model is a finite state machine which changes the state once every time unit. Each time t that a state j is entered, a speech vector O_t is generated with the probability density $b_j(O_t)$. The transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. Figure 1 illustrates this process, where a six-state model moves through the state sequence X=1,2,2,3,4,4,5,6 in order to generate the sequence O_1 to O_6.

The joint probability that O is generated by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities and the output probabilities. Therefore, for the state sequence X in figure 1, we have:

$$P(O,X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)\ldots \qquad (1)$$

In practice, only the observation sequence O is known and the underlying state sequence X is hidden. This is why it is called a Hidden Markov Model.

Given that X is unknown; the required likelihood is computed by summing over all possible state sequences $X = x(1), x(2), x(3), \ldots, x(T)$, that is:

$$P(O|M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(O_t) a_{x(t)x(t+1)} \qquad (2)$$

where $x(0)$ is constrained to be the model entry state and $x(T+1)$ is constrained to be the model exit state.

This representation assumes that the parameters $\{a_{ij}\}$ and $\{b_j(o_t)\}$ are known for each model M_i. The elegance and power of the HMM framework lies in that, given a set of training examples corresponding to a particular model, the parameters of the model can be determined automatically by a robust and efficient re-estimation procedure. Thus, provided that a sufficient number of representative examples of each word can be collected, then an HMM can be constructed such that implicitly models all of the many sources of variability inherent in the real speech.

978-1-4244-2565-5/08/$25.00 ©2008 IEEE

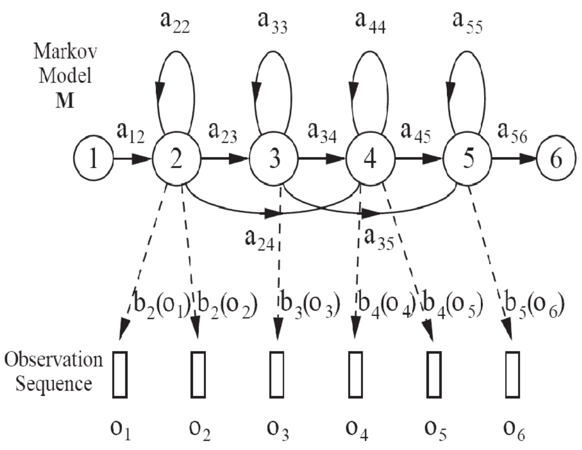

Figure 1. Markov generation model [7].

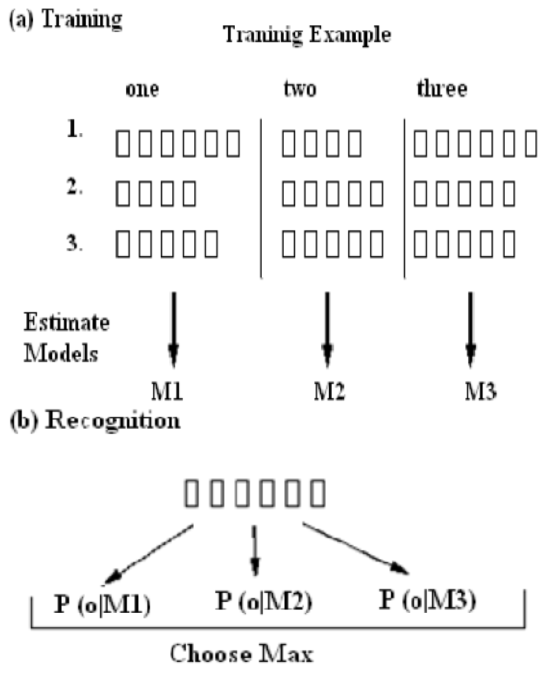

Figure 2. Using HMMs for isolated word recognition.

III. THE PROPOSED METHOD

As mentioned earlier, the main problem with Persian digit recognition is the pronunciation and structural similarities between certain digits that are of similar time-frequency patterns as well. An example is the similarity between digits two and nine in Persian, which are pronounced as /dʊ/ and /ηʊ/ respectively. Figure 3 illustrates the spectrogram of samples of these digits. It can be deduced from these spectrograms how close the spectral-temporal characteristics of these two numbers are. The energy distribution of these digits over the time-frequency plane is approximately the same. Moreover, the maximums and minimums of energies are very close and they appear to follow the same patterns. In addition to these clues, our experiments also confirm these similarities. The conventional HMM recognizers mistake these two numbers almost 15% of the time. The same clues are true for other digit groups, such as digits seven and eight, as well as one and three, due to their very similar phonetic structure.

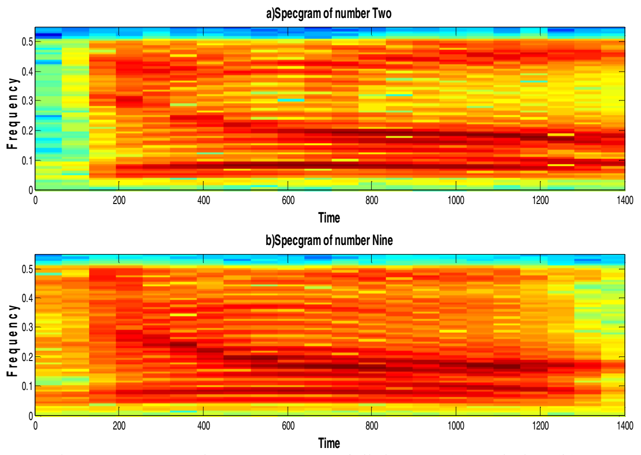

Figure 3. A sample spectrogram of digits two (a) and nine (b).

In order to overcome this recognition problem, we developed a new method based on a hybrid HMM/SVM system. In this new recognition machine, the HMM is used as a global classifier, due to its high precision in classification of huge amount of data to reduce the search space. Then, a SVM classifier, equipped with a novel segmentation technique, completes the recognition process. The main idea behind this is that the SVM classifiers have a nice record in classification of a limited number of data. The Following sections give details about the HMM and the SVM constructed in this research.

A. *THE PROPOSED HMM STRUCTURE*

Figure 2 summarizes the use of HMMs for isolated word recognition. Firstly, an HMM is trained for each vocabulary word using a number of representative examples of that word. In this case, the vocabulary consists of just three words: "one", "two" and "three". Secondly, to recognize some unknown words, the likelihood of each model generating that word is calculated and the word is then identified by the most likely model.

In figure 2, the rectangles denote feature vectors which can be Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC) coefficients, or any appropriate parameter set. However, the problem is that these feature vectors are such distributed that they could not be grouped together and inherently could not be assigned appropriate probability densities. To overcome this problem, one can use the K-means or the LBG algorithm [8,9] to split the vector space into smaller divisions, in which there exists a representative that could be assigned to the subspace probability density and could be used to measure the distance of any desired vector from the desired subspace.

The number of subspaces and the metric that is used to estimate the spectral distances are arbitrary parameters which are chosen according to the volume of the desired database.

B. *SUPPORT VECTOR MACHINE (SVM)*

In [6], vapnik introduces a new type of learning machines so-called support vector machines. In this method, the input data maps to a high dimension space called feature space by means of $\phi : \Re^n \rightarrow \Re^N$, where n is the dimension of the input data space and N is the feature space dimension. Now, the problem is to find a hyper-plane $w \cdot \phi(x) + b = 0$ corresponding to decision functions $f(x) = \text{sign}((w.x) + b)$ which can maximize the margin of separation between the two classes. If the input data is fully separable, we can rescale w and b such that the distance of the closest point to the hyper-plane satisfies $|((w.x) + b| = 1$. Then, the following inequality holds for all input data:

$$y_i(w\cdot\phi(x_i) + b) \geq 1 \qquad (3)$$

where y_i take values in the range of $\{-1,1\}$, according to the class to which the input data belongs. It is clear that maximizing the margin in this essence leads to minimizing the $\|w\|$, subject to some constraints.

In real world problems, the training data may not be separated without error. In this case, the decision hyper-plane is chosen to minimize the function:

$$\frac{1}{2}w^2 + C\cdot F\left(\sum_{i=1}^{l}\xi_i^{\sigma}\right) \qquad (4)$$

subject to the following constraints, where $F(.)$ is a monotonic convex function, C is a constant, and $\sum_{i=1}^{l}\xi_i^{\sigma}$ is a measure of the error with ξ_i and σ chosen to be positive variables.

$\xi_i \geq 0, \quad i = 1, \dots, l$

$$y_i(w\cdot\phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \dots, l \qquad (5)$$

where l is the number of the support vectors.

In [6], it is shown that the optimal w can be written as:

$$w_o = \sum_{i=1}^{l}\alpha_i y_i \phi(x_i) \qquad (6)$$

where α_i are the coefficients that can be found by solving a quadratic problem. Concluding from all of the above, the classification function f can be written as:

$$f(x) = y(w_o\cdot\phi(x) + b_o) = \sum_{i=1}^{l}\alpha_i y_i \phi(x).\phi(x_i) + b \qquad (7)$$

which clearly depends on some dot products. Under Hilbert-Schmidt Theory [10], we can use kernel functions that satisfy the Merser's condition as dot-products, which clearly reduces the computational complexity and lead to the decision function, as:

$$f(x) = \sum_{i=1}^{l}\alpha_i y_i K(x, x_i) + b \qquad (8)$$

where $K(.,.)$ is the kernel function.

C. HYBRID HMM/SVM MACHINE

In this section we introduce a novel method for isolated speech recognition. As depicted in figure 4, the recognition process in this new method is a three-step approach. First, an adaptive method similar to Voice Activity Detection (VAD), is employed to omit the silence segment of the speech. The resulting speech is then fed as an input to the HMM part of the speech recognizer. The HMM model is then constructed by splitting this speech sample into some equal parts, say five, so as to play the role of states (A). Then the MFCC parameters are computed for each of these segments as inner states (B). Now, the HMM model of the input data is compared to the training data of the HMM model which has already been constructed according to the method stated in section 3.1. Three candidates in descending order, according to their probabilistic chance of occurrence in the HMM model, are then introduced as the outputs of this process. These candidates are data with the HMM models which best match the input entry into the HMM model. In fact, we have employed the HMM part to reduce our search space to a limited number of classes.

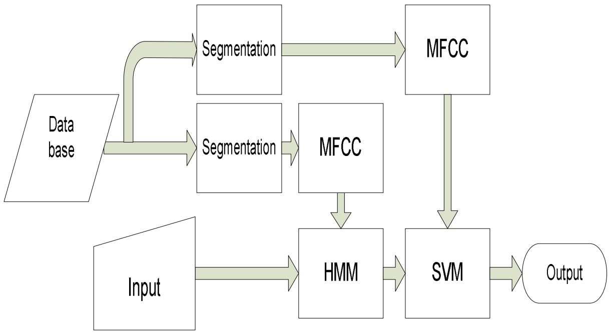

Figure 4. The Proposed hybrid system

As stated earlier, the SVMs in structures having a confined search area, are more accurate than other classifiers. As a glance at this property, we make use of the HMM as a global recognizer which provides the SVM with fewer numbers of candidates to choose from. According to our experiments, this reduction introduces no errors in the recognition process.

Now that we have narrowed our vision into three elements, the process is finalized using the SVM classifier. In order to prepare the input data for the SVM, we first split the input data into a number of segments, as specified earlier, and compute the MFCC parameters for each segment. Any SVM unit needs input data with the same number of space dimension. Consequently, we use the MFCC parameters which can be set to have the same space dimension; in addition to that they can better represent the features of the speech segment [11]. The estimated parameters for each segment are then used with the SVM unit to determine the closest corresponding segment at the outputs of the HMM model. At the end, we calculate for each of the outputs of the HMM part the numbers of segments, due to the output that are thought to best match the according segment of the input data. The element with the statistically biggest number of similar segments is then introduced as the final output of the recognizer (note that in the rare condition of a tie, the first output of the HMM part is introduced as the final recognition result).

IV. SIMULATION RESULTS

In our experiments, we have used approximately 400 Persian male speech samples and 50 Persian female speech samples of different age groups for each of the digits one through nine. We used 330 speech samples for training and the remaining 120 samples for testing our recognizer. These speech samples are recorded in the speech laboratory of the Electronics Research Center (ERC) of Sharif University of Technology with the sampling frequency of 11025 Hz. The MFCC parameters are chosen to be vectors of length 13. These parameters are calculated using the Bark frequency model with 13 linear and 27 logarithmic triangular (conventional) filters.

The employed SVM unit is LS_SVMLab [12]. The classification is made under the binary classification model using the Error Radial Basis Functions (ERBF) as the kernel functions. The SVM based recognizers have worse records, as compared to HMM based recognizers, and it is only in the hybrid systems that one can have improvement using the SVM. Therefore, we compare our results to the conventional HMM based models.

As illustrated in table 1, our experimental results show a significant improvement, as compared to the conventional HMM recognizers. In fact, the output of the HMM based recognition system can rarely exceed an accuracy of 92%. This is mainly due to the problem we already mentioned in the introduction. In Persian language, there are some groups of digits that are such similar in pronunciation that can even be mixed up by a human listener.

As shown in tables 1 and 2, the proposed method yields a better performance in both clean and noisy conditions. The higher accuracy of this method is well demonstrated when we face the outstanding errorless detections of some digits. In this method, the ability of the HMM in separating a huge amount of data is combined with the concise classification of the SVM in confined areas. As confirmed by the results shown in table 2, the novel idea of the segmentation in the SVM part has improved the results. Figure 5 also gives a comparison between the outcomes of the recognition machine with different numbers of segments in different noisy environments. As we increase the number of segments, the results improves until we reach the total of 13 segments. In this condition, the segmentation shows a noticeable improvement over conventional single-segment hybrid systems.

A major advantage of this segmentation is to get rid of the noisy segments in the initial and final part of the speech samples that can affect our decision in the conventional methods. In addition, the resulting segments became more stationary, where it is more convenient to classify the data. Because the similar samples have now less diversity as compared to those in the non-stationary case, they better concentrate in the feature space. This concentration of the data, makes it more easily divisible into some smaller groups.

V. CONCLUSION

In this paper, we have addressed a traditional problem in isolated Persian digits recognition. Amongst the digits one through nine, there exist groups of digits with the same phonetic structures and time-frequency patterns that make the recognition process a rather erroneous task. To overcome this problem, a novel isolated word recognition system has been proposed. The system is based on a three-step hybrid HMM/SVM classification, which is appropriately chosen so as to increase the statistical efficiency of both the HMM and the SVM subsystems. The HMM serves as a global classifier and the SVM completes the final recognition step in the confined search space.

Table 1. Proposed method results, as compared to results using a conventional HMM based speech recognizer.

	Conventional HMM based method		**Proposed method**	
	Mean	**STD**	**Mean**	**STD**
Clean	77.25	15.89	98.59	3.09
30dB	16.96	25.91	68.45	30.00
20dB	14.66	28.80	49.02	22.50
10dB	9.39	19.20	33.31	25.14

Table 2. Proposed method results for different number of segmentation in the SVM part.

	One segment		**Five segments**		**Thirteen segments**	
	Mean	**STD**	**Mean**	**STD**	**Mean**	**STD**
Clean	97.81	3.15	98.17	2.99	98.59	3.09
30dB	46.58	22.35	56.46	24.91	68.45	30.00
20dB	24.84	24.18	31.04	17.86	49.02	22.50
10dB	17.99	25.73	18.08	19.33	33.31	25.14

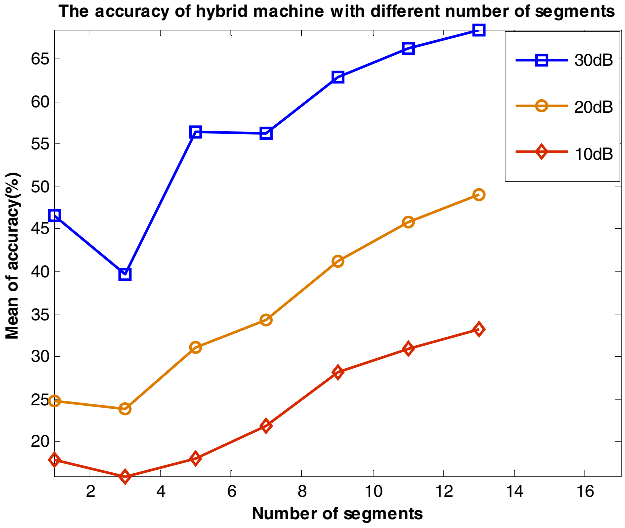

Figure 5. The accuracy of hybrid machine with different segmentations.

It is shown that by means of a novel segmentation process, the efficiency of the recognition is increased dramatically. The segmentation idea provides the SVM with more stationarity and therefore increases the classification precision of this machine. The simulation results also confirm that the proposed method achieves a significant improvement over conventional methods in both clean and noisy environments.

REFERENCES

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the* IEEE, vol.77, no.2, p. 257–286, February 1989.

[2] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, vol.13, No.5, pp.45-57, 1996.

[3] R. Chen, M. Tanaka, D. Wu, L. Olorenshaw, and Ma. Amador, "*A Four Layer Sharing HMM System for very large Vocabulary Isolated Word Recognition*," 5th *International Conference on Spoken Language Processing*., Sydney ,Australia, Dec. 1998.

[4] A. Ganapathiraju, J. Hamaker, and J. Picone, "*Applications of support vector machines to speech recognition," IEEE Trans. Signal Processing*, vol. 52,pp. 2348–2355, Aug 2004.

[5] M.Gurban, J.Thiran " *Audio-Visual Speech Recognition with a Hybrid SVM/HMM System", in 13th European Signal Processing Conference (EUSIPCO), 2005.*

[6] V. Vapnik, The *Nature of Statistical Learning Theory*: Springer-Verlag, New York, USA,1995.

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*: Cambridge University Engineering Department, 2001.

[8] J. A. Hartigan and M. A. Wong "*A K-Means Clustering Algorithm*", *Applied Statistics*, Vol. 28, No. 1, pp.100-108, 1979.

[9] J. B. MacQueen "*Some Methods for classification and Analysis of Multivariate Observations*", *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, pp.281-297, 1967.

[10] R. Cournat, D. Hilbert *Methods of Mathematical Physics*: Interscience, New York, 1953.

[11] J. R. Deller, J. H. L. Hansen, J. G. Proakis *Discrete-time processing of speech signal*: Springer-Verlag, New York, 2000.

[12] http://www.esar.kuleuven.ac.be