Hybrid Fuzzy HMM System for Arabic Connectionist Speech Recognition

SINOUT D. SHENOUDA
Computer Science
Department
American University in Cairo
P. O. Box 2511 Cairo
EGYPT

DR. FAYEZ W. ZAKI
Electronics and Communications
Engineering
Mansoura University
Faculty of Engineering
EGYPT

DR. AMR GONEID
Computer Science
Department
American University in Cairo
P. O. Box 2511 Cairo
EGYPT

Abstract:- In this paper, a new Arabic connectionist speech recognition system is presented. This recognition system is based on the combination of the fuzzy integral and measure theory [1] and Hidden Markov Model (HMM) [2] using the CSLU toolkit. The CSLU toolkit [3] is a research and development software environment that provides a powerful and flexible tool for research in the field of spoken language understanding.

The objective of this paper is to design a hybrid Fuzzy HMM (FHMM) system for Arabic speech recognition. This system is based on a novel Hidden Markov Model with fuzzy logic and fuzzy integral theory. In this context, the fuzzy integral is used to relax the independence assumptions that are necessary with probability functions. Interestingly, it should be noted that one particular case in the choice of fuzzy integral (the Choquet integral), fuzzy measure (probability measure), and fuzzy intersection operator (multiplication), reduces the generalized fuzzy HMM to the classical HMM. The traditional HMM and the proposed Fuzzy HMM systems were implemented by computer simulation and a performance comparison was carried out.

It is noticed that, there are some improvements in recognition accuracy in case of the Fuzzy HMM (FHMM) system over the classical HMM recognition system. The FHMM recognition system accuracy varies from 93.36% to 98.36% depending on the data set used whereas the classical HMM' accuracy varies from 91.27% to 94.60% for the same data sets.

Key-Words: - Speech recognition – Signal processing – Speech processing – Hidden Markov Model – Fuzzy logic – Fuzzy integral theory – Fuzzy measure - Man-machine communications.

1 Inroduction

The main goal of Automatic Speech Recognition (ASR) is to develop techniques and systems that enable computers to accept speech as input. Speech is the human's most efficient communication media. Using speech in man-machine communication is more comfortable and position independent than any other media that require more concentration, and restrict movement. In speech recognition techniques, pattern comparison can be performed in several ways depending on the specifics of the recognition system. There are many approaches to ASR. The main approaches are template-based, knowledge-based, and stochastic-based approaches. There are other new approaches, which are Artificial Neural Network (ANN), and fuzzy logic. The main problem in speech recognition, as with other complex tasks that require some form of intelligence, is the amount of information that must be examined before making a classification or decision. Speech recognition is an extremely complex pattern-matching problem. The complexity arises from the variability in speech rate, pitch, volume, and emotion. Together with the natural differences in individual human voice production systems, these factors produce variable and nonlinear waveforms. As if these challenges were not enough, a speech recognition system must also deal with non-speech sounds and environmental noise.

There are still many research problems to resolve in speech recognition, as it is still often not completely robust or efficient for certain applications. Nevertheless, speech recognition systems today can obtain high accuracy with the utilization of neural networks, fuzzy logic and hidden Markov models (HMM). Today, the HMM is the most widely used, and its strong mathematical base allows many new studies to improve its efficiency. Recently, a novel generalization of HMM has been introduced [4] and successfully applied to hand-written recognition [5].

This new model uses a fuzzy approach instead of a stochastic one for the classical HMM. This requires a new computation of integrals and summations that use fuzzy integrals and fuzzy measures.

In this work, the main contribution is that this new model is applied for the first time to Arabic speech recognition, in order to show the feasibility of the new Fuzzy HMM (FHMM) for this application, and to compare its performance with the classical HMM. This method is expected to reduce the word error rate of the system and hence increase the recognition rate and system accuracy.

In this paper, firstly a brief introduction to HMM is given in section 2, then in section 3 the block diagram of the proposed system is explained. In section 4, the fuzzy hidden Markov model is introduced, and results in speech recognition are given in section 5. Then the conclusions is given in section 6.

2 Hidden Markov Model

The Hidden Markov Model (HMM) has proved to be a most successful modeling approach for speech recognition [6]. The HMM is typically defined as a stochastic finite state machine which is assumed to be built up from a finite set of possible states Q = $\{q_1, q_2, \ldots.., q_k, \ldots.q_K\}$. Each of these states being associated with a specific probability distribution.

A specific HMM M_i will then be represented by a stochastic finite state machine with L_i states $S_i = \{S_1, S_2, \ldots.., S_l, \ldots.S_{Li}\}$ where $S_l \in Q$. S may only contain a subset of Q, while also having the same state appearing at different nodes of the finite state. According to this definition, the HMM models the feature vector X = $\{x_1, x_2, \ldots.., x_n, \ldots.x_N\}$ as a piecewise stationary process for which each stationary segment will be associated with a specific HMM state. For a given model M, an utterance X = $\{x_1, x_2, \ldots.., x_n, \ldots.x_N\}$ is modeled as a succession of discrete stationary states S = $\{S_1, S_2, \ldots.., S_l, \ldots.S_L\}$, $L \leq N$, with instantaneous transitions between these states. An example of a simple three–state HMM is shown in Fig. 1. This could be the model of a short word assumed to be composed of three stationary parts.

HMMs are commonly specified [7] by a set of states q_i, an emission probability density $P(x_n|q_i)$ associated with each state, and transition probabilities $P(q_j|q_i)$ for each permissible transition from state q_i to state q_j .

Ideally, there should be a HMM for every possible utterance, but this is clearly infeasible. To reduce the number of possible models, there are many ways. First, a sentence is modeled as a sequence of words. Second, word models are comprised of concatenated sub–word units. These units may be syllables, semi-syllables, phones or phonemes. Phonemes are speech sound categories that are sufficient to differentiate between different words in a specific language. For example in Arabic language there are 36 phonemes [8].

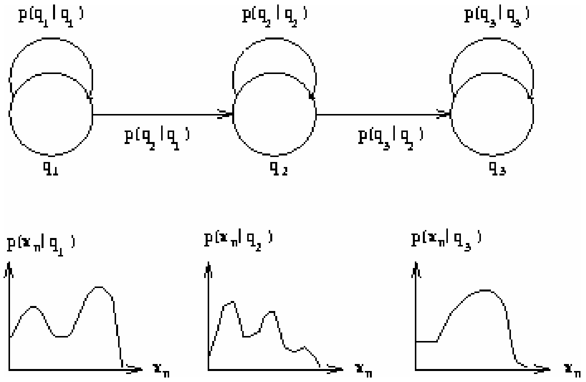


Fig. 1. A three-state Hidden Markov Model (HMM).

3 Fuzzy HMM System Block Diagram

In this section, the design and implementation of the proposed system will be considered in brief. Figure 2, shows a block diagram illustration for the proposed system.

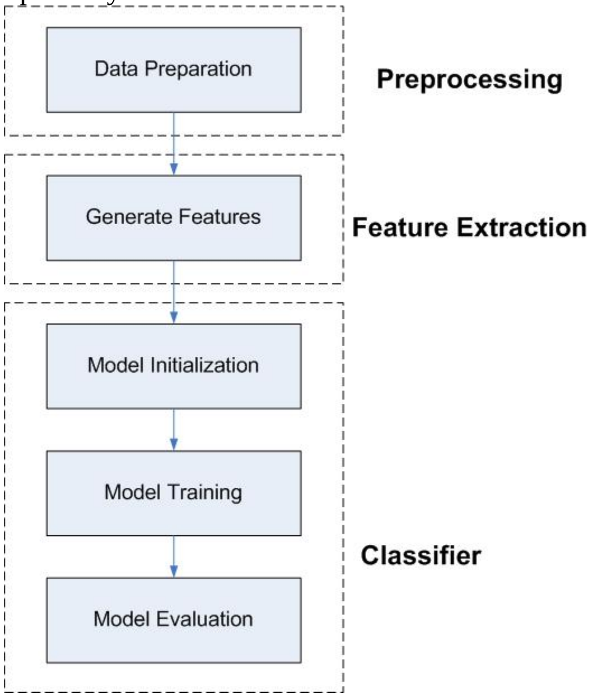


Fig. 2. Block diagram of the proposed system

The system consists of three main blocks: Pre-processing, feature extraction, and classification. The preprocessing stage consists of data recording, and pre-emphasis of the data to enhance the high frequency content of the speech signal. Data selection and recording is an important task in speech recognition. In this work two Arabic data sets are developed by author. The development of corpus or data set includes recording the speech file, transcripted it to its text contents and labeling them to their corresponding phoneme files. In the second stage, speech compression is performed through feature extraction. The last stage is the classification stage in which, initialization, training, transcription, and classification are performed. The classification model of this work is Fuzzy Hidden Markov Model (FHMM).

In this paper, the focus is only given to the third stage which is the classifier. The implementation of the FHMM classifier will be discussed in the following section.

4. Proposed Model Structure

The fuzzy HMM is defined with the same type of model parameters of the classical HMM but with a different mathematical basis. Fuzzy logic replaces probability theory and this leads to a new definition of these model variables [9] [10]. The structure in terms of states and observations, of course, remains the same. The features used in this model are the feature vector V which consists of thirty-nine values computed every frame (ten milliseconds). These values represent the 13 Mel scale cepstral coefficients, and its first and second-order time derivatives.

4.1. Fuzzy HMM Model Initialization

In this stage, the data collection script is used to initialize the Fuzzy HMM (FHMM) model. The original data segments of the HMM model are used to create the label output file that contains the segment boundary and label information for the selected segments. The number of examples selected for each Fuzzy HMM model is listed in the counter output file.

Using this counter output file, initial parameter estimates can be computed using Vector Quantization and Viterbi realignment [11]. The model initialization script reads the features for each speech waveform file. The segment boundary information is then used to extract the set of features associated with the Fuzzy HMM model to be trained. The number of iterations for Fuzzy HMM parameter initialization is sixteen iterations in two phases. These are used to control the number of Viterbi realignment.

4.2. Model Training

Fuzzy logic replaces probability theory and this leads to a new definition of these model variables or parameters [12]. These parameters are fuzzy transition matrix $\widehat{A}$, fuzzy observation matrix $\widehat{B}$ and, the initial state fuzzy density $\widehat{\pi}$. Similar to the model initialization of the HMM model, the initial parameter estimates of the state transition probability matrix is computed as:

$$\widehat{\lambda} = (\widehat{A}, \widehat{B}, \widehat{\pi}) \qquad (1)$$

The aim of this step is to find a set of model parameters $\widehat{\lambda}$ which maximizes the likelihood of the models. The training should maximize the posterior probability of the correct model given the sequence of feature vectors.

The training script reads the model transcription file and selects data for each model defined. The configuration file contains the word list used by this model and their labels. It uses these parameters to search the training data set. It updates the phoneme list output file which contains the phonemes and non-speech segments list used in data set.

4.3. Initial Model Evaluation

Again, the performance of the initial model parameter estimates is evaluated. The search script of FHMM model, creates a finite state grammar search, using the grammar definition and word pronunciation models as described in the configuration file. The model configuration script creates the lookup table. The left context of the word initials and the right context of the word terminals are assumed to expand to the silence model.

Since each of these models is essentially monophone models rather than the expected triphone models, each triphone model will be tied to its corresponding monophone model. Given the word pronunciation models and grammar definition as well as the triphone lookup table, the finite state search network can be created in the search output file. The search script file applies the Viterbi decoder to the development test set of data. The first best answers are written to the answer output file contains the hypothesized transcription. The performance of these models can be computed using these output files.

4.4. Automatic Transcription

After the initial FHMM model is evaluated, the word transcriptions of the training set are used to create phonetic transcriptions. Using the word list generation script, the unique set of words in the training corpus is created. In this work, the word list is the same as the words in the proposed recognition system. For larger more general-purpose recognition tasks, the words in the training vocabulary typically do not match the words of the recognition task [13].

A pronunciation dictionary can be created using the word database script from the word list, which is accessed by the forced alignment script. It is needed to select the pronunciation which best matches the acoustics. The alignment process also determines whether there are silences between words or not. By the same manner as HMM model, two output files are created using this script. These two files contain the phoneme and word force alignments.

4.5. Parameters Re-estimation

Similar to the model parameters, the forward and backward prediction variables can be easily extended to the fuzzy case: $\hat{\alpha}_i(t)$ measures the grade of certainty of $O_1 O_2 \ldots O_t$ and x_i at time t. The new algorithm uses fuzzy operators for the computation and fuzzy integrals for the summation over the states [14]. Initialization becomes:

$$\hat{\alpha}_1(i) = \hat{\pi}_i \wedge \hat{b}_i(O_1) \qquad (2)$$

and induction becomes:

$$\hat{\alpha}_{t+1}(i) = \int_x \hat{a}_{\mathrm{Y}}(\{y_i\} \mid x) \circ \hat{\alpha}_{\Omega_x}(\{O_1 \ldots O_t\}) \wedge \hat{b}_i(O_{t+1}) \quad (3)$$

where $\hat{a}_{\mathrm{Y}}$ is the transition fuzzy measure on Y, $\widehat{b_i}$ is the symbol fuzzy measure and $\hat{\alpha}_{\Omega_x}(\{O_1, O_2, \cdots O_l\})$ is the forward prediction fuzzy measure over the set of classes of X.

In this case, the embedded parameter estimation creates a composite model from the concatenation of the model transcriptions. First the automatic phoneme transcriptions are mapped to the FHMM model transcriptions. The mapped transcriptions can be used to create the composite model. In this phase the segmentation information is ignored, and only the transcriptions are used. Embedded parameter estimation uses the Fuzzy form of Forward-backward algorithm, to compute a probabilistic segmentation between models. This algorithm uses fuzzy operators for the computation and fuzzy integrals for the summation over the states as seen in equations (2) and (3). To implement the fuzzy HMM without using the probability measure, the possibility measure is used instead. Then the FHMM can be developed by the introduction of this new fuzzy component. Once these variables are obtained, the problem of computation of an observation probability is solved using the formula:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \hat{\alpha}_T(i) \qquad (4)$$

or by using the backward variable:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \hat{\beta}_1(i) * \hat{b}_i(O_1) \qquad (5)$$

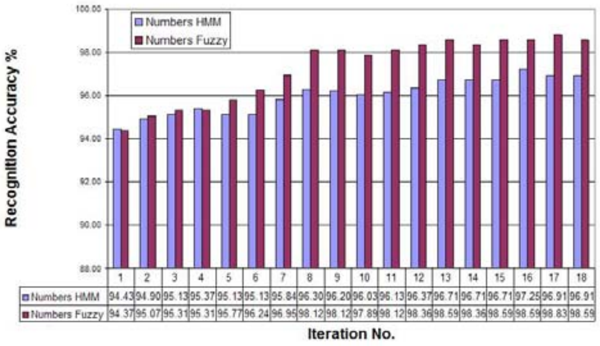
Hence, the recognition is achieved by choosing the model which gives the highest grade of certainty for the sequence. This fuzzy component will also be the major innovation for the other basic problems of HMM. Then the forward and backward prediction variables can be easily computed if the multiplication stands for the operator of fuzzy intersection and Choquet integral for the fuzzy integral. These two variables are shown in equations (4) and (5). The embedded parameter re-estimation script is used in this step to obtain the required FHMM models and stored it in the output files. These models can be evaluated to select the best model using the model scoring script for each of the models obtained.

5. Results and Analysis

In order to evaluate the performance of the proposed Fuzzy HMM approach as well HMM-based approach, a series of experiments with different data sets in the speaker-independent automatic speech recognition (ASR) environments was performed. These experiments were performed for three data sets, containing different sets of vocabularies under the recording and experimental conditions.

5.1. Task A: 30K CSLU Numbers Corpus

In this task, 30 K numbers corpus from CSLU data set was used to assess the performance of the proposed system. This data set contains 480 speech files. The total number of words is 2164 words (numbers). The number of training samples was 288 speech files and 1325 words. The number of development samples was 96 speech files and 413 words. The number of test samples was 96 speech files and 426 words. It was seen that the FHMM system provides recognition improvement over classical HMM by about 1.5%. That is, the recognition accuracy of FHMM is 94.4% to 97.3% whereas HMM accuracy is 94.4% to 98.8% as shown in Fig. 3.



Fig, 3 Recognition accuracy of 30 K CSLU data set for HMM and Fuzzy HMM systems.

5.2. Task B: Arabic Numbers AR_NUM Corpus

In this task, Arabic numbers (AR_NUM) corpus developed by the author were used to test the performance of the proposed system. This data set contains 406 speech files. The total number of words is 4763 words (numbers). The number of training samples was 244 speech files and 2893 words. The number of development samples was 80 speech files and 954 words. The number of test samples was 82 speech files and 916 words. From the recognition results of this experiment, it was noticed that the recognition accuracy of FHMM is 91.9 % to 95.2% whereas HMM accuracy is 91.9% to 96.7% as shown in Fig. 4. This means that the FHMM system provides recognition improvement over classical HMM by about 1.5%.

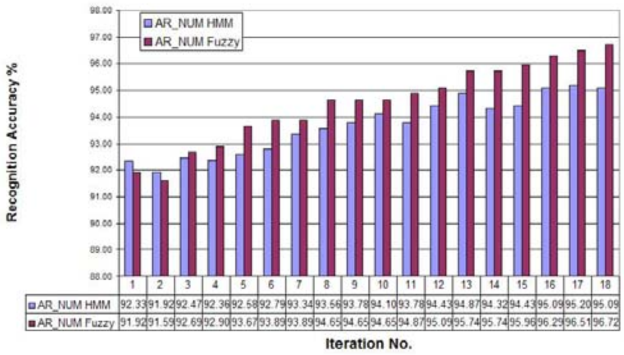


Fig. 4. Recognition accuracy of AR_NUM data set for HMM and Fuzzy HMM systems.

5.3. Task C: Arabic Sentences AR_ST Corpus

In this task, the performance of the proposed system was tested using the Arabic sentences (AR_ST) corpus. This data set contains 1104 speech files prepared by the author. The total number of words is 3847 words. The number of training samples was 660 speech files and 2283 words. The number of development samples was 220 speech files and 751 words. The number of test samples was 224 speech files and 813 words.

From this experiment, it was seen that the recognition accuracy of the fuzzy HMM system out perform the recognition accuracy of the classical HMM system by about 1.6%. In other words, the FHMM accuracy reaches 92.5 % and the HMM accuracy reaches 94.1% as shown in Fig. 5.

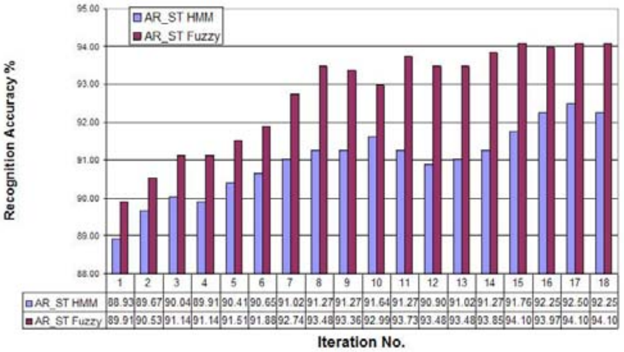


Fig. 5. Recognition accuracy of AR_ST data set for HMM and Fuzzy HMM systems.

5.4. Result Comparison:

The system recognition accuracy for all tasks was improved for the fuzzy HMM system than for the classical HMM system. The variabilities between the different tasks are due to the vocabulary size and the language used. Fig. 6. summarizes the results obtained above.

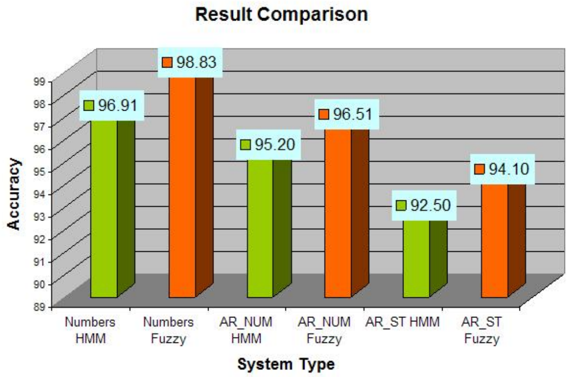


Fig. 6. Recognition accuracy of all data set for HMM and Fuzzy HMM systems.

6. Conclusions

The main contribution of this work is the new classification technique introduced. The technique is based on a new Hidden Markov Model based on fuzzy logic and fuzzy integral theory. In this method, the fuzzy integral is used to relax one of the two assumptions that one had with the classical HMM. The main advantage of the new model is the fuzziness of the model in terms of a lower computation time. A performance comparison study was carried out by computer simulation on both the Fuzzy HMM and the classical HMM systems. The Fuzzy HMM system provided about 4% increase in recognition rate over the classical HMM system.

For the future work, it is recommended to study the effect of noise on the proposed system. Although, some preliminary experiments were conducted in this direction, but it needs more investigations, that are dedicated for this important issue [15].

References:

[1] M. Sugeno, “Fuzzy measure and fuzzy integral,” Transactions of the Society of Instrument and Control Engineers, vol. 8, pp. 218–230, 1972.

[2] 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, February 1989.

[3] J. Schalkwyk, P. Hosom, Ed Kaiser, K. Shobaki, “CSLU-HMM: The CSLU Hidden Markov Modeling Environment”, Center for Spoken Language Understanding (CSLU), Oregon Graduate Institute of Science & Technology, March 2000.

[4] M. Mohamed and P. Gader, “Generalized hidden markov models. i. theoretical frameworks”, IEEE Transactions on Fuzzy Systems, vol. 8, no. 1, pp. 67–81, 2000.

[5] M. Mohamed and P. Gader, “Generalized hidden markov models. ii. application to handwritten word recognition,” IEEE Transactions on Fuzzy Systems, vol. 8, no. 1, pp. 82–94, 2000.

[6] Rabiner, L.R., “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-285, 1989.

[7] Cole, Ronald A. “Survey of the State of the Art in Human Language Technology”, Technical Report, http://cslu.cse.ogi.edu/tts/publications/index.html, 1995.

[8] Shenouda, Sinout D., Elawady, Rashid M., and Zaki, Fayez W. “A Study Of Speech Recognition and its Application to Arabic Speech”, M.Sc. Thesis. Mansoura University, Faculty of Engineering, 1991.

[9] Mills, Patrick M. “Fuzzy Speech Recognition” Thesis. University of South Carolina, 1996.

[10] A. Goneid, F. W. Zaki, Sinout D. Shenouda, “Fuzzy Arabic Speech Recognition System Based on Hybrid HMM/ANN System” Proceeding of The 10th International Conference on Artificial Intelligence Application (ICAIA), Feb 2002.

[11] H. Bahi, and M. Sellami, “Combination of Vector Quantization and Hidden Markov Models for Arabic Speech Recognition”, Proceedings of ACS/IEEE International Conference on Computer Systems and Applications (AICCSA’01), 2001.

[12] George Klir, and Bo Yuan, “Fuzzy Sets and Fuzzy Logic: Theory and Applications” Prentice Hall, New Jersey, 1995.

[13] Mustafa N. Kaynak, Qi Zhi, Adrian D. Cheok, Kuntal Sengupta, Zhang Jian and Ko Chi Chung, "Analysis of Lip Geometric Features for Audio-Visual Speech Recognition", IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems & Humans, vol. 34, no. 4, pp. 564–570, July 2004.

[14] Adrian David Cheok, Sylvain Chevalier, Mustafa Kaynak, Kuntal Sengupta, and Ko Chi Chung, “Use of a Novel Generalized Fuzzy Hidden Markov Model for Speech Recognition”, Proc. of the IEEE Fuzzy Systems Conf., vol. 3, pp. 1207-1210, Dec. 2001.

[15] Mustafa N. Kaynak, Tolga M. Duman and Erozan M. Kurtas, “Noise Predictive Belief Propagation,” Proc. of the IEEE International Conf. on Communications, vol. 1, pp. 704-708, May 2005.