Using a New Hybrid Models for Speech and Medical Pattern Classification

Lilia Lazli[1], Mounir Boukadoum[2], Abdennasser Chebira[3], Kurosh Madani[3]
[1]Laboratory of research in Computer Science (LRI/GRIA)
Badji Mokhar University
B.P.12 Sidi Amar 23000 Annaba – Algeria
[2]Univesité du Québec A Montréal (UQAM)
Canada
[3]Images, Signals and Intelligent Systems Laboratory (LISSI / EA 3956)
PARIS XII University, Senart-Fontainebleau Institute of Technology
Bat.A, Av. Pierre Point, F-77127 Lieusaint, France
l_lazli@yahoo.fr, Boukadoum-mounir@uqam.ca, {achebira, kmadani}@univ-paris12.fr

ABSTRACT: *The main goal of this paper is to compare the performance which can be achieved by two different hybrid approaches analyzing their applications' potentiality on real world paradigms (speech recognition and medical diagnosis). We compare the performance obtained with (1) Multinetwork RBF/LVQ structure, we use involves Learning Vector Quantization (LVQ) as a competitive decision processor and Radial Basis Function (RBF) neural models is used as classifier. (2) Hybrid HMM/MLP system using a Multi Layer Perceptron (MLP) to estimate the Hidden Markov Models (HMM) emission probabilities.*

Keywords: Speech recognition, Medical diagnosis, Hybrid RBF/LVQ model, Hybrid HMM/MLP model

Received: 24 December 2011, Revised 9 February 2012, Accepted 13 February 2012

1. Introduction

In many target (or pattern) classification problems the availability of multiple looks at an object can substantially improve robustness and reliability in decision making. The use of several aspects is motivated by the difficulty in distinguishing between different classes from a single view at an object [1]. It occurs frequently that returns from two different objects at certain orientations are so similar that they may easily be confused. Consequently, a more reliable decision about the presence and type of an object can be made based upon observations of the received signals or patterns at multiple aspect angles. This allows for more information to accumulate about the size, shape, composition and orientation of the objects,which in turn yields more accurate discrimination.

Moreover, when the feature space undergoes changes, owing to different operating and environmental conditions, multi aspect classification is almost a necessity in order to maintain the performance of the pattern recognition system.

In this paper, we propose two original hybrid approaches on classification of electrical signals (speech and biomedical signals). The classification of these signals presents some problems, because of the difficulty to distinguish one class of signal from the others. The results can be different for different test session for the same pattern (speaker or patient).

First, we have developed a serial multi-neural network approach that involves both Learning Vector Quantization (LVQ) and Radial Basis Function (RBF) Artificial Neural Networks (ANN). These two models of ANNs are particularly adapted for classification tasks. If it is admitted that techniques based on single neural network show a number of attractive futurs to solves problems for which classical solutions have been limited, it is also admitted that a flat neural structure doesn't represent the more appropriated way to approach "*intelligent behavior*". The approach we propose uses a Multi-Neural Network (MNN) architecture.

Second, in this paper we present the Hidden Markov Model (HMM) and apply them to complex pattern recognition problem. There are two reasons why the HMM exists. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. However, standard HMM require the assumption that adjacent feature vectors are statistically independent and identically distributed. These assumptions can be relaxed by introducing Neural Network (NN) in the HMM framework.

These NN estimate the posterior probabilities used by the HMM. Among these, the hybrid approach using the Multi-Layer Perceptron (MLP) to estimate HMM emission probabilities has recently been shown to be particularly efficient by example for French speech [2] and American English speech [3]. We then propose in second part a hybrid HMM/MLP model for speech recognition and biomedical diagnosis which makes it possible to join the discriminating capacities, resistance to the noise of MLP and the flexibilities of HMMs in order to obtain better performances than traditional HMM.

This paper is structured as follows. In the next section, we present the speech and biomedical DataBases (DB). Then we expose the two hybrid approaches structure (hybrid BRF/LVQ and hybrid HMM/MLP approaches). In section 5, we present the classification results we obtained by using a DBs. A comparison study with the classical approaches (RBF, LVQ ANN and HMM) has been made. Finally, we conclude and give the prospects that follow from our work.

2. Databases Construction

2.1. Speech Databases

Three speech DBs have been used in this work:

1) The first one referred to as DB1, the isolated digits task has 13 words in the vocabulary: 1, 2, 3, 4, 5, 6, 7, 8, 9, zero, oh, yes, no. They are spoken by 30 speakers, producing a total of 3900 utterances (each word should be marked 10 times). The digits DB has about 30,000 frames of training data. This first corpus consists of isolated digits collected over the microphone.

2) The second DB, referred to as DB2 contained about 30 speakers saying their last name, first name, the city of birth and the city of residence. Each word should be marked 10 times. The used training set in the following experiments consists of 2000 sounds.

3) The third DB, referred to as DB3, contained the 13 control words (i.e. View/new, save/save as/save all) so that each speaker pronounces each control word 10 times. The used training set in the following experiments consists of 3900 sounds saying by 30 speakers.

For each DB, 22 speakers were used for training (a non-overlapping subset of these were used for cross-validation used to adapt the learning rate of the MLP), while the remaining 8 speakers were used for testing.

The acoustic feature were quantized into independent codebooks according to the FCM algorithm respectively: [1]

- 128 clusters for the J-RASTA PLP vectors.
- 128 clusters for the first time derivative of cepstral vectors.
- 32 clusters for the first time derivative of energy.
- 32 clusters for the second time derivative of energy.

2.2 Biomedical Database

The used signals are called Potentials Evoked Auditory (PEA), examples of PEA signals are illustrated in figure 1. Indeed, the

exploration functional otoneurology possesses a technique permitting the objective survey of the nervous conduction along the auditory ways. The classification of the PEA is a first step in the development of a help tool to the diagnosis [4]. The main difficulty of this classification resides in the resemblance of signals corresponding to different pathologies, but also in the disparity of the signals within a same class. The results of the medical test can be indeed different for two different measures for the same patient.

The PEA signals descended of the examination and their associated pathology are defined in a DB containing the files of 11185 patients. We chose 3 categories of patients (3 classes) according to the type of their trouble. The categories of patients are:

1) Normal (N): the patients of this category have a normal audition (normal class).

2) Endocochlear (E): these patients suffer from disorders that touches the part of the ear situated before the cochlea (class endocochlear).

3) Retrocochlear (R): these patients suffer from disorders that touches the part of the ear situated to the level of the cochlea or after the cochlea. (class retrocochlear).

We selected 213 signals (correspondents to patients). So that every process (signal) contains 128 parameters. 92 among the 213 signals belong to the N class, 83 to the class E and 38 to the class R. The basis of training contains 24 signals, of which 11 correspondent to the class R, 6 to the class E and 7 to the N class.

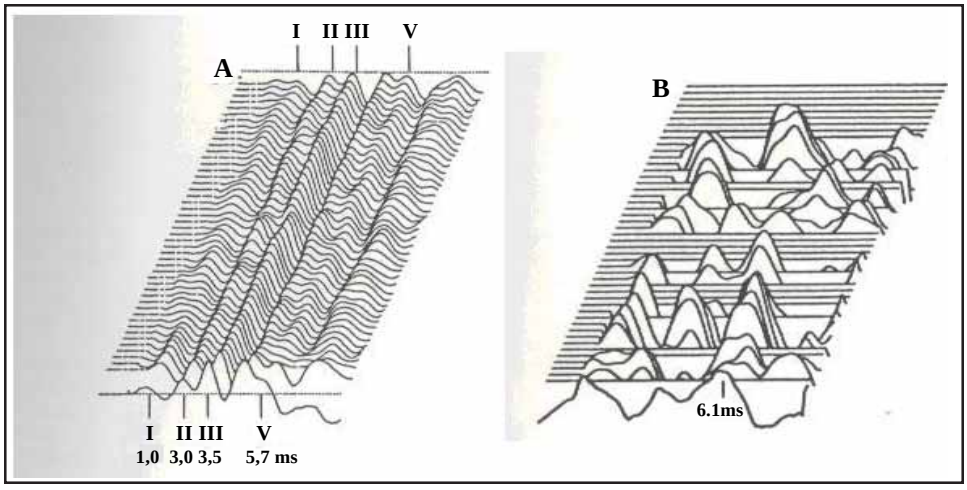

Figure 1. (A) PEA signal for normal patient, (B) Patient with auditory disorder

3. Multi-Neural Network Based Approach

The approach we proposed to solve th posed problem is based on MNN concept. A MNN could be seen as a neural structure including a set of similar neural networks (homogneous MNN architecture) or a set of diffrent neural nets (heterogeneous MNN architecture). On the other hand, both two above mentioned (homogeneous and heterogeneous MNN) could be organized in different manners. In a general point of view, three topologis [4] could characterize the MNN‘s organization:

- Parallel organization: in the case, ANN‘s are not inter-connected. The MNN input is dispatched to all neural networks composing such structure.

- Serial organization: in this case, the output of a given ANN composing the structure is the input of the following ANN.

- Serial/parallel organization: which combines the two structures above mentioned connections.

The main difficulty in classification of used signals is related, on the one hand, to a large varity of such signals for a same diagnosis result (the variation panel of corresponding our electrical signals could be very large), and on the other hand, to the close resemblance between such signals for two differents classification results. The serial homogeneous MNN is equivalent to a single neural network structure with a greater number of layers with different neuron activation functions. So the use of

homogeneous MNN with a serial organization is here out of real interest. In the parallel homogeneous MNN configuration, each neural net operates as some "*expert* " (learning a specefic characteristic of the feature space). So the interest of parallel homogeneous MNN appears when a decision stage, to process the results pointed out by the set of such "*expert* ", is associated to such MNN structure becomes then a serial/parallel MNN, needing an optimization procedure to determine the number of neural nets to be used.

We propose an intermediary solution: a two stage serial heterogeneous MNN structure combining a RBF based classifier (operating as the first processing stage) with a LVQ based decisionclassification stage.

The RBF model we use is a weighted-RBF model but a standard one and so, it performs the feature space mapping associating a set of "*categories*" (in our case a category corresponds to a possible pathological class for example for medical DB) to a set of "*areas*" of the feature space. The LVQ neural model belongs to the class of competitive neural network structure. It includes one hidden layer, called competitive layer. Even if the LVQ model has essentially been used for the classification tasks, the competitive nature of it's learning strategy (based on winner takes all strategy), makes it usable as a decisionclassification operator.

4. Hybrid HMM-ANN Models

Recent work at ICIS [6] and ARPA speaker independent Resource Management (RM) task [7] have provided us with further insight into the discriminant HMM, particularly in the light of recent work on transition based models [6]. This new perspective has motivated us to further develop the original discriminant HMM theory, in which an MLP is trained to optimize the full a posteriori probabilities of HMM given the acoustic data via conditional transition probabilities, i.e., probabilities of the next state given the current state and the current acoustic vector. This approach uses posterior probabilities at both local and global levels and is more discriminant in nature.

In this paper, we present experimental and theoretical results using a framework for training and modeling pattern recognition systems based on the theoretically optimal Maximum A Posteriori (MAP) criterion. This is in constrast tomost state of the art systems which are trained according to a Maximum Likelihood (ML) criterion. Although the algorithm is quite general, we applied it to a particular form of hybrid system combining HMM and ANN in particular, MLP in which MLP targets and weights are iteratively reestimated to guarantee the increase of the posterior probabilities of the correct model, hence actually minimizing the error rate.

4.1 Estimating HMM Likelihoods with ANN

ANN can be used to classify speech classes such as words. For statistical recognition systems, the role of the local estimator is to approximate probabilities or Probability Density Functions (PDF). Practically, given the basic HMM equations, we would like to estimate something like $p(x_n \setminus q_k)$, is the value of the PDF of the observed data vector given the hypothesized HMM state. The ANN can be trained to produce the posterior probability $p(q_k \setminus x_n)$ of the HMM state give the acoustic data (figure 3). This can be converted to emission PDF values using Bayes'rule.

Several authors [1; 4; 5; 6] have shown that ANN can be trained to estimate a posteriori probabilities of output classes conditioned on the input pattern.

Since the network outputs approximate Bayesian probabilities, $g_k(x_n, \Theta)$ is an estimate of:

$$p(q_k \setminus x_n) = \frac{p(x_n \setminus q_k)\, p(q_k)}{p(x_n)} \qquad (1)$$

Which implicitly contains the a priori class probability $p(q_k)$. It is thus possible to vary the class priors during classification without retraining, since these probabilities occur only as multiplicative terms in producing the network outputs. As a result, class probabilities can be adjusted during use of a classifier to compensate for training data with class probabilities that are not representative of actual use or test conditions.

Thus, scaled likelihoods $p(x_n \setminus q_k)$ for use as emission probabilities in standard HMM can be obtained by dividing the network outputs $g_k(x_n)$ by the training set, which gives us an estimate of :

$$\frac{p(x_n \setminus q_k)}{p(x_n)} \qquad (2)$$

During recognition, the seating factor $p(x_n)$ is a constant for all classes and will not change the classification. It could be argued that, when dividing by the priors, were using a scaled likelihood, which is no longer a discriminant criterion.

4.2 Motivations of ANN

ANN has several advantages that make them particularly attractive for pattern recognition [1; 5]:

- They can provide discriminant learning between pattern units or HMM states that are represented by ANN output classes.

- Because ANN can incorporate multiple constraints for classification, features do not need to be assumed independent. More generally, there is no need for strong assumptions about the statistical distributions of the input features (as is usually required in standard HMM).

- They have a very flexible architecture which easily accommodates contextual inputs and feedback, and both binary and continuous inputs.

- ANN is typically highly parallel and regular structures, which makes them especially amenable to high-performance architectures and hardware implementations.

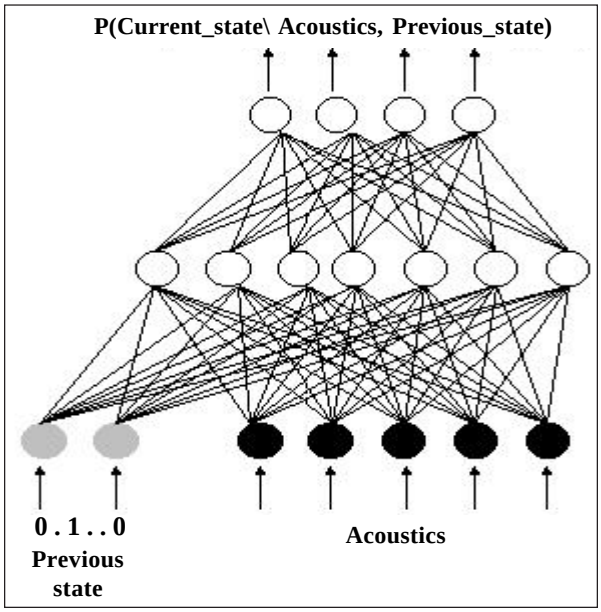

Figure 2. An MLP that estimates local conditional transition probabilities

5. Case Study and Experimental Results

5.1 Comparison study with single RBF and LVQ approaches

The structure of the RBF and LVQ for the respective single approaches are composed as follows by example for biomdical DB:

- The number of input neurons for RBF and LVQ corresponds to the number of components of the input vectors.

- The output layer of RBF and LVQ contains 3 neurons, corresponding to the 3 classes.

- For RBF, the number of neurons of the hidden layer (in this case, 22 neurons) has been determined so as to satisfy the following heuristic rule:

$$\textit{Number of hidden neurons} = (\textit{a number of entry neurons} * \textit{number of output neurons})^{1/2}$$

For LVQ, the number of hidden neurons (in this case, 10 neurons) has been determined by considering the number of subclasses we can count into the 3 classes.

For RBF, the learning DB contains 24 signals, 11 of them are R, 6 E and 7 N. For LVQ, the learning DB contains 20 signals, 6 of them are R, 7 E and 7 N.

In the two cases, the learning DB has successfully been learnt. All of the learnt vectors are well classified in the generalization phase.

The RBF network well classifies **62,3**% of the full DB. For the tree speech DBs, a rate of correct classification as follows: **51**% for the DB1, **52**% for the DB2, **63**% for the DB3.

The LVQ network well classifies **62**% of the full medical DB. For the tree speech DBs, a rate of correct classification as follows: **65**% for the DB1, **61**% for the DB2, **68**% for the DB3.

5.2 Results relative to RBF/LVQ based multineural network approach

We used the RBF/LVQ based serial heterogeneous MNN. Concerning the RBF ANN, by example for the biomedical DB, the number of input neurons (88) corresponding to the number of components of the input vectors, the output layer contain 3 neurons. The number of neurons of the hidden layer is 20 neurons has been determined by heuristic rule.

For the LVQ ANN, the number of input cells is equal to the number of output cells of the RBF ANN. The output layer of the LVQ ANN contains as many neurons as classes (3). The number of neurons in hiddn layer is 8 neurons has ben determined by considering the number of subclasses we can count into the 3 classes.

The learning DB contains 4 signals, 11 of them correspond to R disorders, 6 to E disoders and 7 to N. for the generalization phase, we use the full DB.

The learning DB has successfully been learnt. All of the learnt vectors are well classified in the generalization phase. We can see that this network well classifies **65**% of the full DB.

For the tree speech DBs, a rate of correct classification as follows: **79**% for the DB1, **86**% for the DB2, **72**% for the DB3.

Comparing to the two single approaches, with our proposed MNN technique, the MNN structure combines the advantages of both LVQ and RBF ANNs, these globally better results of our MNN technique are achived with low number of neurons in the ANNs architecture, taken into account the difficulty of our problem.

5.3 Results relative to discrete HMM and hybrid HMM/MLP approach

Further assume that for each class in the vocabulary we have a training set of *k* occurrences (instances) of each class where each instance of the categories constitutes an observation sequence. In order to build our tool, we perform the following operations.

1. For each class v in the vocabulary, we estimate the model parameters $\lambda^v(A, B, \pi)$ that optimize the likelihood of the training set for the v^{th} category.

2. For each unknown category to be recognized, the processing is carried out: measurement of the observation sequence $O = \{o_1, o_2, \ldots, o_T\}$, via a feature analysis of the signal corresponding to the class; the computation of model likelihoods for all possible models, $P(O / \lambda^v)$, $1 \leq v \leq V$; at the end the selection of the category with the highest likelihood.

All systems are trained with clean (original data collected) data and tested with data that are artificially corrupted with convolutional and / or additive noise. All hybrid trainings are bootstrapped from the DB that are trained on the same front-end feature.

In this work, the same HMM topology were used for all the experiments and the following HMM have been used:

5.3.1 Discrete HMM

For speech DBs, 10-state, strictly left-to-right, discrete HMM were used to model each basic unit. Noting only that the choice of 10 states per model was selected in an empirical manner. In this case, the acoustic feature were quantizied into 4 independent codebooks according to the FCM algorithm:

- 128 clusters for the J RASTA-PLP coefficients,
- 128 clusters for the first time derivative of cepstral vectors,
- 32 clusters for the first time derivative of energy,
- 32 clusters for the second time derivative of energy.

Table 1 gives the results of this experiment for Biomedical DB (BDB) and Speech DBs (SDB1..SDB3).

For the PEA signals, 5-state, strictly left-toright, discrete HMM were used to model each basic unit.

5.3.2 Discrete MLP with entries provided by the FCM algorithm

For this case, we compare the performance of the basis hybrid model with that of an hybridHMM/MLP model using in entry of the network an acoustic vector composed of real values which were obtained by applying the FCM algorithm with 2880 real components corresponding to the various membership degrees of the acoustic vectors to the classes of the "*codebook* ". We presented each cepstral parameter (JRASTA PLP, Δ J-RASTA PLP, ΔE, ΔΔE) by a real vector which the components definite the membership degrees of the parameter to the various classes of the "*code-book* ". For speech databases, for example, we reported the values used for DB2. 10-state, strictly left-to-right, word HMM with emission probabilities computed from an MLP with 9 frames of quantizied acoustic vectors at the input, i.e., the current acoustic vector preceded by the 4 acoustic vectors on the left context and followed by the 4 acoustic vectors on the right context. the entry layer is made up of a real vector with 2880 real components corresponding to the various membership degrees of the acoustic vectors to the classes of the "*code-book* ".

Thus a MLP with only one hidden layer including 2880 neurons at the entry, 30 neurons for the hidden layer and 10 output neurons was trained. The three-layer MLP is trained by using stochastic gradient descent, and relative entropy as the error criterion. A sigmoid function is applied to the hidden layer units, and softmax (exponential of the unit's weighted sum normalized by the sum of exponentials for the entire layer) is used as the output nonlinearity.

For the PEA signals, a MLP with 64 neurons at the entry, 18 neurons for the hidden layer and 5 output neurons was trained.

Table 2 gives the results of this experiment for biomedical DB and speech DBs.

DBs	BDB	SDB1	SDB2	SDB3
Rate%	84	87	90	76

Table 1. Discrete HMM results

	BDB	SDB1	SDB2	SDB3
Rate%	94	97	97	83

Table 2. Hybrid HMM/MLP results

For the majority of cases, we can draw a preliminary conclusion from the results reported in tables 1 and 2 for the speech recognition and the PEA signals diagnosis. The hybrid discrete HMM/MLP approach using FCM clustering always outperforms standard discrete HMM or RBF/LVQ ANN approach. Also, our interest to fuzzification clustering because the FCM algorithm produced very consistent segmentation results across all repetitions. The resulting label space for the FCM is a K-dimensional unit hypercube formed from the convex hull about the standard basis of for example, the discrete K-Means label space.

6. Conclusion and Future Work

In this paper, we presented the test of two types of hybrid models in the framework of speech recognition and medical diagnosis. The first hybrid model, the MNN structure we use involves LVQ and RBF neural models. The first neural net (RBF ANN) is used as a classifier, and the second one (LVQ ANN) as a competitive decision processor. So, the association of two RBF and LVQ neural models improves the global order of the non linear approximation capability of such global neural operator, comparing to each single neural structure constituting the MNN system.

For the second hybrid model, a discriminate training algorithm for hybrid HMM/MLP system based on the fuzzy clustering are described. Our results on isolated speech and biomedical signals recognition tasks show an increase in the estimates of the posterior probilities of the correct class after training, and significant decreases in error rates. These preliminary experiments have set a baseline performance for our hybrid HMM/MLP system. Better recognition rates were observed. From the effectiveness view point of the models, it seems obvious that the hybrid HMM/MLP model are more powerful than discrete HMM or multi-network RBF/LVQ structure for the PEA signals diagnosis and speech DB recognition.

We thus envisage improving the performance of the suggested system with the following points:

- In addition, for an extended speech vocabulary, it is interesting to use the phonemes models instead of words, which facilitates the training with relatively small bases.

- For the two types of signals recognition, the main idea is to define a fusion scheme: cooperation of HMM with the multi-network RBF/LVQ structure in order to succeed to a hybrid model and compared the perfomance with the HMM/MLP model proposed in this paper.

References

[1] Lazli, L., Sellami, M. (2003). Connectionist Probability Estimators in HMM Speech Recognition using Fuzzy Logic. MLDM 2003: the 3rd international conference on Machine Learning & Data Mining in pattern recognition, LNAI 2734, Springer-verlag, p.379-388, July 5-7, Leipzig, Germany.

[2] Deroo, O., Riis, C., Malfrere, F., Leich, H., Dupont, S., Fontaine,V., Boite, J. M. (1997). Hybrid HMM/ANN system for speaker independent continuous speech recognition in French. Thesis, Faculté polytechnique de Mons – TCTS, Belgium.

[3] Riis, S- K., Krogh, A. (1997). Hidden Neural Networks: A framework for HMM-NN hybrids. IEEE 1997, to appear in Proc. ICASSP-97, Apr 21-24, Munich, Germany.

[4]Motsh, J- F. (1987). La dynamique temporelle dutrons cérébral: Recueil, extraction et analyse optimale des potentiels évoqués auditifs dutronc cérébral. Thesis, University of Créteil, Paris XII.

[5] Lazli, L., Chebira, A- N., Madani, K. (2007). Hidden Markov Models for Complex Pattern Classification. Ninth International Conference on Pattern Recognition and Information Processing, PRIP'07. http://uiip.basnet. by/conf/prip2007/prip2007.php-id=200.htm, may 22-24, Minsk, Belarus.

[6] Morgan, N., Bourlard, H., Greenberg, S., Hermansky, H. (1994). Stochastic perceptual auditory-event-based models for speech recognition. *In: Proceedings Int. Conference on Spoken Language Processing*, 1943-1946, Yokohama, Japan.

[7] Renals, S., Morgan, N., Choen, M., Franco, H., Bourlard, H. (1992). Connectionist probability estimation in the DECIPHER speech recognition system. *In: Proceedings IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 601-604, San Francisco, California.