Speech Recognition by Fuzzy-Neuro ANFIS Network and Genetic Algorithms

Amane TALEB

Abstract—This paper presents a neural fuzzy system ANFIS for speech recognition. The appropriate learning algorithm is performed on the TIMIT speech database supervised type, a pre-processing of the acoustic signal and then extracting the coefficients MFCCs parameters relevant to the recognition system. Optimization by genetic algorithms in two objectives: minimizing the number of input parameters of ANFIS, with Referring to a selection technique based on an expertise related to the implementation of the speech recognition and minimizing the error rate in classification in determining the optimal parameters of the GA by experiment, first choose the method of crossing then parameters of probabilities of crossing and mutation giving the most optimal values of fitness function. These parameters allow obtaining experimentally the highest recognition rate. An attempt to outline the features of GA and neural fuzzy system in terms of the genetic functionality of operators, the inherent capability of GA for solving complex problems, including continues speech recognition in noisy environment using the database NTIMIT.

Keywords— ANFIS, Genetic algorithms, Speech Recognition, TIMIT.

I. INTRODUCTION

The speech recognition is to extract from speech signal a phonetic transcription which represents the phonetic content of that signal. The signal is first filtered and then sampled. Pre-emphasis is performed to meet the high frequency signal is then segmented into frames, each consisting of N speech samples over a period of 23ms of speech (duration of stationarity of the signal). The signal is convolved by the Hamming window in which we're extracting 13 MFCC coefficients. (Fig.1) [5], [6]

If $x_1$ is $A_{11}$ and... $x_j$ is $A_{1j}$ and... $x_N$ is $A_{1N}$, then $f_1=p_{11}x_1+...+p_{1N}x_N+q_1$.
If $x_1$ is $A_{21}$ and... $x_j$ is $A_{2j}$ and... $x_N$ is $A_{2N}$, then $f_2=p_{21}x_1+...+p_{2N}x_N+q_2$.
If $x_1$ is $A_{31}$ and... $x_j$ is $A_{3j}$ and... $x_N$ is $A_{3N}$, then $f_3=p_{31}x_1+...+p_{3N}x_N+q_3$.

With $N=13$ (number of MFCC by interval). The premise part of the 3 rules represents simultaneously the overall membership degree of fuzzy subsets "Small", "Medium" and "Great".

The parameters are trained by the ANFIS, according diagram form on figure 2. [1], [3]

Layer 01 Layer 02 Layer 03 Layer 04 Layer 05

The first layer consists of membership Gaussian function $\mu_{ij}$ of MFCC’s standardized coefficients described by generalized bell function

$$\mu_{ij}(x) = e^{-\frac{1}{2} \left( \frac{x-c_i}{S_i} \right)^2} \quad (1)$$

Fuzzy subsets $A_{ij}$ representing the fuzzy linguistic labels {Small, Medium, Great}. The second layer generates on outputs the product of entries $w_i$, this product represents the the degree of activation of each rule.
The third layer calculates the normalized value $\overline{w_i}$ of membership degree of each rule.
The fourth layer makes connections filling the role of the consequence part of each rule by outputting the value of the function

$$\overline{w_i} \cdot f_i(x_1, x_2, ..., x_N).$$

The fifth layer outputs the sum of the outputs of the previous layer.

II. REPRESENTATION MODEL OF THE RECOGNITION SYSTEM

The representative model is a fuzzy inference system of Takagi and Sugeno three rules as follows:

TALEB Amane is associated with Signa-lMage PAre role Laboratory, Department of Computer Science, University of Sciences and Technology of Oran, Algeria
\[
f = \overline{w_1f_1} + \overline{w_2f_2} + \overline{w_3f_3} = \frac{w_1f_1 + w_2f_2 + w_3f_3}{w_1 + w_2 + w_3}
\]

(2)

The parameters of fuzzy inference system:
- Number of membership functions: nummf s = 39
- Number of function parameters = 42
- Number of Rules = 03
- Number of entries = 13 (number of MFCC parameters / interval)
- Type of membership functions: Gaussian

The network training is performed on the TIMIT database with 60 phonemes spoken by several speakers, with 31,514 of occurrences on training and 12,055 at the test phase. The ANFIS uses back propagation learn3.g to determine premise parameters (to learn the parameters related to membership functions) and least mean square estimation to determine the consequent parameters. It’s proposed on input of the ANFIS network the standardized MFCC parameters. Refer to “(3),”

\[
C_i = \frac{\text{Coef}_i}{\sum_{j=1}^{\text{coef}_j}}
\]

A step in the learning procedure has got two parts: In the first part the input patterns are propagated, and the optimal consequent parameters are estimated by an iterative least mean square procedure, while the premise parameters are assumed to be fixed for the current cycle through the training set. In the second part the patterns are propagated again, and in this epoch, back propagation is used to modify the premise parameters, while the consequent parameters remain fixed or recognition rate evaluated on the basis of test data TIMIT decreases. This procedure is then iterating.

III. OPTIMIZATION OF A RECOGNITION SYSTEM BY GA

Genetic Algorithms, because of their robustness and ability to provide global solutions, have been used as a tool by a number of researchers to identify parameters of fuzzy system, since GAs work on coding of the parameter set, and not on the derivative of a function, they are capable of solving a vast range of optimization problems. This paper uses GA to optimize parameters of FIS which consists of rules premises parameters (such as mean and variance for Gaussian membership function).

The fitness function is a multi-criteria function by two objective functions. It represents the fitness function by a weighted sum of objective functions.

\[
\text{Fitness} = \text{Minimize} \left( \alpha O_1 + \beta O_2 \right)
\]

where \( O_1 \): is the ratio between the number of selected primitives and the total number of primitives.

Its range of membership is [0, 1].

\( O_2 \): is the error rate achieved in classification using the ANFIS network. Its membership is interval [0,1]. Minimized this rate amounts to applying the operators of genetic algorithms to parameters \( \mu_{ij} \), (degree of membership in fuzzy sets).

\( \alpha \) and \( \beta \) are two parameters depending on the normalization and weighting for each objective. [4]

A. Selection of Primitives

To optimize the \( O_1 \) objective is to reduce the primitive (Hamming window) corresponding to one phoneme is represented as a matrix (N, M), N=13 (MFCC’s number in a window) M=160 (segment’s number of a phoneme).

We must therefore find L primitives, such that L<M, with the aim of adjusting the parameters of fuzzy inference system. The number of optimized primitives is at L = 64, approximate the average number of intervals required for a plosive consonant and a vowel.

The presence of bit 1 means that the primitive is selected and the presence of bit 0 means that the primitive is not.

The selection of primitives is done by the method formulated by the following equation:

\[
\min_k \left( \sum_{i=k}^{L-1} \sum_{j=1}^{N-1} \left| \text{Coef}_{ij} - \text{Coef}_{ij} + 1 \right| K \in \left[1, M - L + 1 \right] \right)
\]

With L=64, M=160 and N=13. This method says that the region where there is less variation between the values of the MFCC is the tranche in which the system is stable and relevant information contained. Selected primitives are coded to 1 else to 0 (fig. 3).

A second technique applied in the interest of improving the \( O_1 \) objective is to reduce the number of occurrences per phoneme of TIMIT database for training.

\[
M \text{ primitives}
\]

<table>
<thead>
<tr>
<th>P</th>
<th>P_1</th>
<th>P_2</th>
<th>P_1</th>
<th>P_{i+1}</th>
<th>P_{i+63}</th>
<th>P_{16}</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 3 Codification of selected primitives

The result of multiplying the binary vectors of two occurrences of the same phoneme, tells us the rate of overlap of the N primitive selected. If the overlap rate exceeds the ratio of 2/3 then the two occurrences are similar on the acoustic level and the second is not taken for learning, else it is included (fig.4). This process is repeated until exhaustion of occurrences of the phoneme stored in the data base TIMIT. This technique is applied to all phonemes and we get a compressed data base with a lesser number of occurrences, hence saving time result in a learning phase. Experimentally this technique reduced the number of occurrences of 55.96% for vowels of 25.32% for fricatives and 12.26% for plosives consonants. With an average reduction of 27.69%, this enabled a shift from a training set of 31,514 hits to a database of 22,788 occurrences.
B. Optimization of the Recognition Rate by GAs

In this phase we reduce the error rate in generalization phase of the ANFIS network or the realization of the O2 objective. To achieve this goal we must improve the network parameters. The parameters of the consequences parts of the rules are not affected by the application of GAs in that they depend on the parameters of premise parts, optimizing the membership functions $\mu_{ij}$ affect the parameters of linear functions $P_{ij}$, because minimizing the error amounts to solving the following equations. [7]

$$\frac{\partial e}{\partial c} = (f - out_d) \left( \frac{x - c}{s^2} \right) w_i (1 - w_i) p_i x + q_i y + r_i$$

$$\frac{\partial e}{\partial s} = (f - out_d) \left( \frac{(x - c)^2}{s^3} \right) w_i (1 - w_i) (p_i x + q_i y + r_i)$$

$$\frac{\partial e}{\partial p_i} = (f - out_d) w_i x$$

$$\frac{\partial e}{\partial q_i} = (f - out_d) w_i y$$

$$\frac{\partial e}{\partial r_i} = (f - out_d) \overline{w_i}$$

C. Coding of Chromosomes

The binary coding of chromosomes is of 6 bits, each $\mu_{ij}$ ($i \in [1, 13]$ and $j \in [1, 3]$) is coded by a binary value, with $\sum_{ij} \mu_{ij} = 1$

The chromosome consists of 39 genes (fig. 5).

<table>
<thead>
<tr>
<th>$\mu_{1.1}$</th>
<th>$\mu_{1.2}$</th>
<th>...</th>
<th>$\mu_{1.13}$</th>
<th>$\mu_{2.1}$</th>
<th>...</th>
<th>$\mu_{ij}$</th>
<th>...</th>
<th>$\mu_{3.13}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>110100</td>
<td>010011</td>
<td>...</td>
<td>010010</td>
<td>110011</td>
<td>...</td>
<td>111000</td>
<td>...</td>
<td>001111</td>
</tr>
</tbody>
</table>

Fig. 5 Binary coding of membership degrees

D. Choice of Parameters of the Genetic Algorithm

The determination of optimal parameters of the GA is performed experimentally, where it is necessary to first choose the method of crossing the most suitable. Experiments are conducted over three crossover mechanisms: The technique of one crossover point, two crossover points and uniform crossover with probability of crossover and mutation for each technique respectively $\{1, 0.8, 0.6, 0.4, 0.2, 0\}$ and $\{0, 0.1/L, 1/L, 10/L, 1\}$. [4]

So the best method of crossing is a one-point crossover with parameters of probability $p_{cross} = 0.8$ and $p_{mut} = 1/L$.

We summarize the average values of the fitness function Results on Table 1.

<table>
<thead>
<tr>
<th>$P_{mut}$</th>
<th>$P_{cross}$</th>
<th>$P_{mut}$</th>
<th>$P_{cross}$</th>
<th>$P_{mut}$</th>
<th>$P_{cross}$</th>
<th>$P_{mut}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>76.80</td>
<td>0.02</td>
<td>0.01</td>
<td>20.14</td>
<td>31.57</td>
</tr>
<tr>
<td>0.1/L</td>
<td>0.8</td>
<td>76.09</td>
<td>0.02</td>
<td>0.00</td>
<td>20.11</td>
<td>28.12</td>
</tr>
<tr>
<td>1/L</td>
<td>0.6</td>
<td>76.66</td>
<td>0.01</td>
<td>0.01</td>
<td>22.55</td>
<td>31.48</td>
</tr>
<tr>
<td>10/L</td>
<td>0.4</td>
<td>79.88</td>
<td>0.09</td>
<td>0.06</td>
<td>22.85</td>
<td>31.33</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>80.91</td>
<td>2.12</td>
<td>0.07</td>
<td>23.14</td>
<td>32.02</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>91.34</td>
<td>4.33</td>
<td>1.11</td>
<td>21.53</td>
<td>31.39</td>
</tr>
</tbody>
</table>

The curve in Figure 6 allows us to compare the values of adaptation function in the initial population $M = 160$ and compressed population $L = 64$ we find that the values converge after 700 generations in the cycle of AG.

![Comparison of the convergence the average value of the fitness](Image)

Our experiences have lead to the optimal settings of the GA as follows:

- Selection Technique : Roulette Wheel.
- Coding : Binary.
- Elitism : True.
- Population size : 64.
- Number of generations : 1000.
- Fitness Function : fitness = Minimize (α O₁ + β O₂).
- Size of chromosome : 39.
- Method of mutation : Random.
- Mutation probability : pmut = 1/L.
- Method of crossing : One point crossover.
- Probability of crossover : pcross = 0.8.
- Stopping criterion : Maximum number of generations or Maximum performance based on validation.

IV. RESULT

We summarize in table 2 the best results obtained experimentally, by presenting the recognition rates in generalization phase by phonetic class. Only the rates of one-crossover point technique with probability pcross = 0.8 are presented, we conclude that the highest rate obtained with mutation probability pmut = 1/L is equal to 72.43%.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>RECOGNITION RATE PER PHONETIC CLASS, USING THE ONE-POINT TECHNIQUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>phonetic classes</td>
<td>pmut=0.0</td>
</tr>
<tr>
<td>Vowels</td>
<td>32.89</td>
</tr>
<tr>
<td>Fricatives consonants</td>
<td>26.05</td>
</tr>
<tr>
<td>Plosive consonants</td>
<td>26.17</td>
</tr>
</tbody>
</table>

V. CONCLUSION

This work is oriented on the continuous speech recognition in no-noisy, with a base TIMIT of training and testing, despite the absence in the speech signal indicator on the boundaries of phonemes, by our tendency cons is directed towards studies on speech recognition in noisy environments, using the base NTIMIT, with robust representation parameters of speech signal and better suited to this kind of problem: the NPC.