SPEAKER RECOGNITION USING AN INTELLIGENT APPROACH

Prof.Y.Rajeshwari
HOD,ICE Department
G.Narayanamma Institute of
Tech. &Science ,Hyderabad

V.Rajeswari
Assistant Professor,ICE Department
G.Narayanamma Institute of
Tech. &Science ,Hyderabad

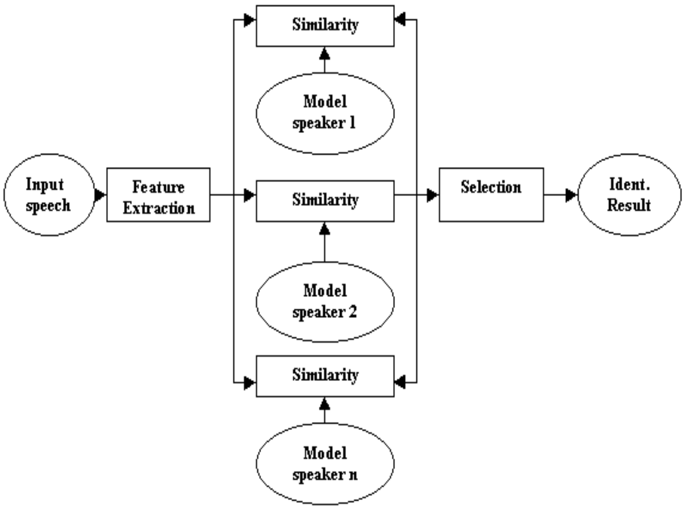
Abstract—This paper describes the analysis of sound signals with the help of intelligent techniques, such as the neural networks and fuzzy systems for specific speaker recognition. In the first step, we use the neural networks for analyzing the sound signal of an unknown speaker, and then, a set of type-2 fuzzy rules are used for decision making. Here we use fuzzy logic due to the uncertainty of the decision process. And to optimize the architecture of the neural networks we make use of genetic algorithms. In this study, we illustrate our approach with a sample of sound signals from real speakers in our institution.

Keywords-Recognition,Fuzzysystems,Cepstralco-efficients,Speaker,identification,Linearpredictive coding.

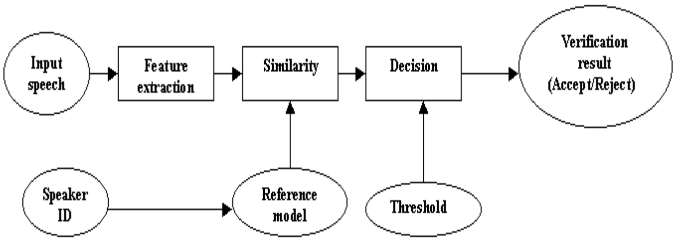
I INTRODUCTION

Speech recognition is a two step process, identification and verification, which involves in automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and real-time control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers [10]. Fig. 1 shows the basic components of speaker identification

Text-dependent and text-independent are two speaker recognition methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken [2].



(a) Speaker identification



(b) Speaker Verification

Fig. 1. Basic structure of speaker recognition systems.

Both text-dependent and independent methods can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used. Yet even this method is not completely reliable, since it can be deceived with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted speaker recognition method has recently been proposed by [7].

II. TRADITIONAL METHODS FOR SPEAKER RECOGNITION

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments).

The most common short-term spectral measurements currently used are Linear Predictive Coding (LPC)-derived cepstral coefficients and their regression coefficients. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed from LPC coefficients. Therefore it provides a stabler representation from one repetition to another of a particular speaker's utterances.

A. Normalization Techniques

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from trial to trial. Variations arise from the speaker themselves, from differences in recording and transmission conditions, and from background noise. It is important for speaker recognition systems to accommodate to these variations. Two types of normalization techniques have been tried; one in the parameter domain, and the other in the distance/similarity domain.

B. Parameter-Domain Normalization

Spectral equalization, the so-called blind equalization method, is a typical normalization technique effective in reducing linear channel effects and long-term spectral variation [2] for text-dependent speaker recognition applications. Cepstral coefficients are averaged over the duration of an entire utterance and the averaged values subtracted from the cepstral coefficients of each frame. However, it unavoidably removes some text-dependent and speaker specific features; therefore it is inappropriate for short utterances in speaker recognition applications.

C. Distance/Similarity-Domain Normalization

A normalization method for distance values using a likelihood ratio, which is defined as the ratio of two conditional probabilities of the observed measurements of the utterance. The first probability is the likelihood of the acoustic data given the claimed identity of the speaker, and the second is the likelihood given that the speaker is an imposter. The likelihood ratio normalization approximates optimal scoring in the Bayes sense. It improves speaker separability and reduces the need for speaker-dependent or text-dependent thresholding.

D. Text-Dependent Speaker Recognition Methods

Text-dependent methods are usually based on template-matching techniques. The input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated.The hidden Markov model (HMM) can efficiently model statistical variation in spectral features. Therefore, HMM-based methods were introduced as extensions of the DTW-based methods, and have achieved significantly better recognition accuracies [3].

E. Text-Independent Speaker Recognition Methods

One of the most successful text-independent recognition methods is based on vector quantization (VQ). In this method, VQ code-books consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific features. A speaker-specific code-book is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the code-book of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

F. Text-Prompted Speaker Recognition Method

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested. Not only can this method accurately recognize speakers, but it can also reject utterances whose text differs from the prompted text, even if it is spoken by the registered speaker.This method is facilitated by using speaker-specific phoneme models, as basic acoustic units. The phoneme models are represented by Gaussian-mixture continuous HMMs or tied-mixture HMMs, and they are made by adapting speaker-independent phoneme models to each speaker's voice. In the recognition stage, the system concatenates the phoneme models of each registered speaker to create a sentence HMM, according to the prompted text. Then the likelihood of the input speech matching the sentence model is calculated and used for the speaker recognition decision. If the likelihood is high enough, the speaker is accepted as the claimed speaker. Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises.

G. Speaker Verification

The speaker-specific characteristics of speech are due to differences in physiological and behavioral aspects of the speech production system in humans. The main physiological aspect of the human speech production system is the vocal tract shape. The vocal tract modifies the spectral content of an acoustic wave as it passes through it, thereby producing speech. Hence, it is common in speaker verification systems to make use of features derived only from the vocal tract.

Using cepstral analysis, an utterance may be represented as a sequence of feature vectors. The purpose of voice modeling is to build a model that captures these variations in the extracted set of features. There are two types of models that have been used extensively in speaker verification and speech recognition systems: stochastic models and template models. However, recent work in stochastic models has demonstrated that these models are more flexible and hence allow for better modeling of the speech production process. A very popular stochastic model for modeling the speech production process is the Hidden Markov Model (HMM).The pattern matching process involves the comparison of a given set of input feature vectors against the speaker model for the claimed identity and computing a matching score. We show in Figure 2 a schematic diagram of a typical speaker recognition system.

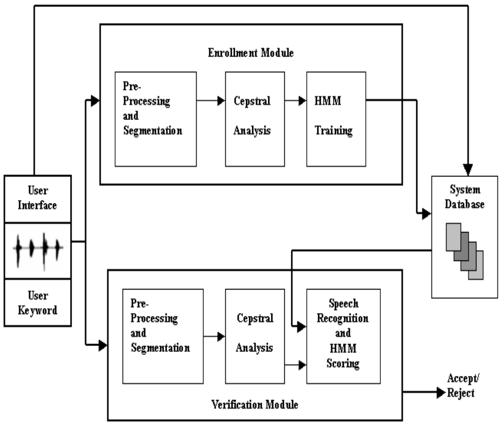


Fig. 2. Blocks diagram of a typical speaker recognition system.

III. NEURAL NETWORKS FOR VOICE RECOGNITION

We used the sound signals of 20 words in Spanish as training data for a supervised feedforward neural network with one hidden layer. We show in Table 1 the results for the experiments with this type of neural network.The results of Table I are for the Resilient Backpropagation training algorithm because this was the fastest learning algorithm found in all the experiment (required only 7% of the total time in the experiments). The comparison of the time performance with other training methods is shown in Figure 6. Table.1 shows the results of feed forward neural networks for 20 words in Spanish.

Stage	Time (min)	Num. of Words	No. Neuron s	Words Recog ized.
1a.	11	20	50	17
2a.	04	20	50	19
1a.	04	20	70	16
2a.	04	20	70	16
3a.	02	20	25	20
1a.	04	20	25	18

1a.	03	20	50	18	90%
2a.	04	20	70	20	100%
2a.	04	20	50	18	90%
1a.	07	20	100	19	95%
2a.	06	20	100	20	100%
1a.	09	20	50	10	50%
1a.	07	20	75	19	95%
1a.	07	20	50	19	95%
2a.	06	20	50	20	100%
1a.	29	20	50	16	80%
1a.	43	20	100	17	85%
2a.	10	20	40	16	80%
3a.	10	20	80	16	80%
1a.	45	20	50	11	55%
2^{a}	30	20	50	15	75%
3^{a}.	35	20	70	16	80%

Table 1. Results of Feedforward Neural Networks for 20 words in Spanish.

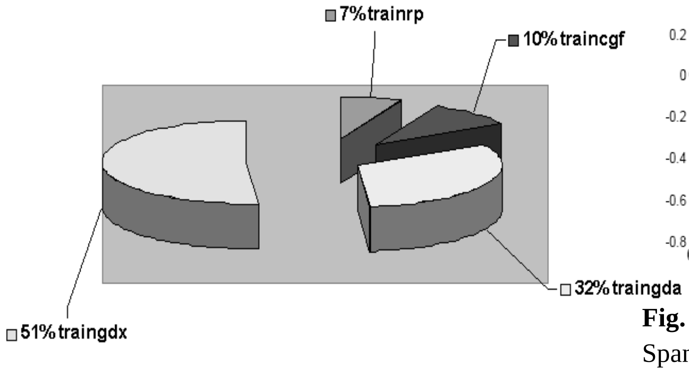


Fig. 6. Comparison of the time performance of several training algorithms.

We now show in Table 2 a comparison of the recognition ability achieved with the different training algorithms for the supervised neural networks. We are showing average values of experiments performed with all the training algorithms. We can appreciate from this table that the resilient backpropagation algorithm is also the most accurate method, with a 92% average recognition rate.

Method	Average Recognition	
trainrp	92%	
TRAINCGF-srchcha	85%	
traingda	90%	81%
traingdx	100%	70%

Table II. Comparison of Average Recognition of Four Training Algorithms.

We describe below some simulation results of our approach for speaker recognition using neural networks. First, in Figure 7 we have the sound signal of the word "example" in Spanish with noise. Next, in Fig. 8 we have the identification of the word "example" without noise. We also show in Fig. 9 the word "layer" in Spanish with noise. In Fig. 10, we show the identification of the correct word "layer" without noise.

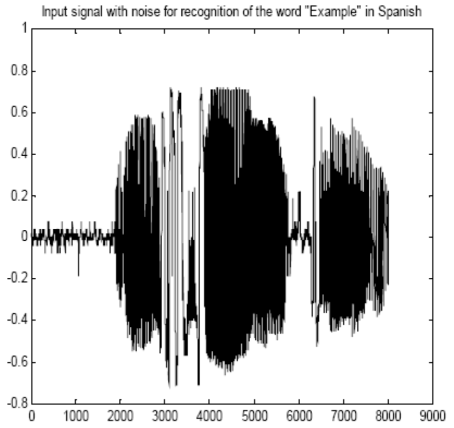


Fig. 7. Input signal of the word "example" in Spanish with noise.

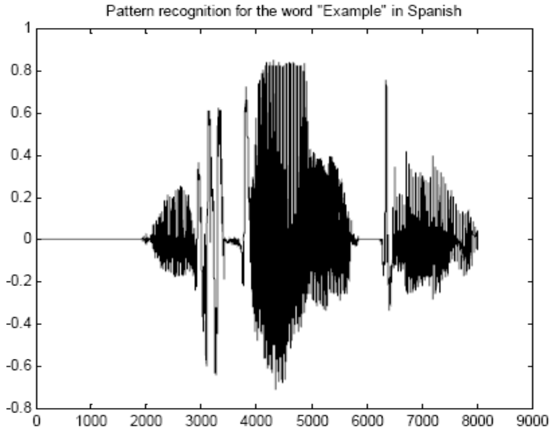


Fig. 8. Indentification of the word "example".

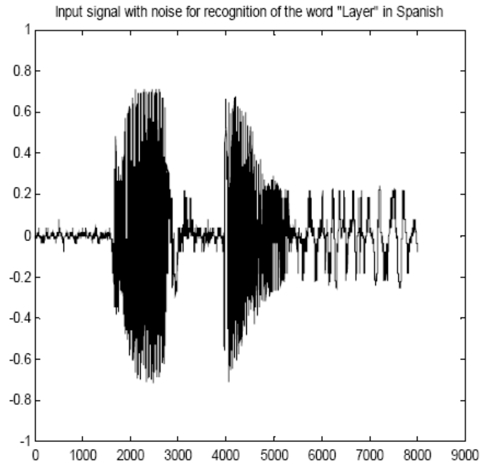


Fig. 9. Input signal of the word "layer" in Spanish with noise added.

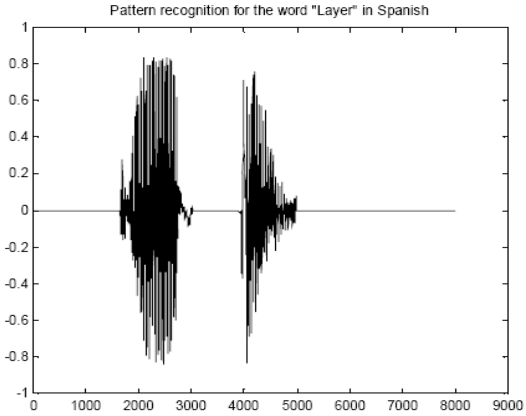


Fig. 10. Identification of the word "layer".

From the figures 7 to 10 it is clear that simple monolithic neural networks can be useful in voice recognition with a small number of words. It is obvious that words even with noise added can be identified, with at least 92% recognition rate (for 20 words). Of course, for a larger set of words the recognition rate goes down and also computation time increases. For these reasons it is necessary to consider better methods for voice recognition.

IV. Voice Recognition with Modular Neural Networks and Type-2

We can improve on the results obtained in the previous section by using modular neural networks because modularity enables us to divide the problem of recognition in simpler sub-problems, which can be more easily solved. We also use type-2 fuzzy logic [9] [16] to model the uncertainty in the results given by the neural networks from the same training data. We describe in this section our modular neural network approach with the use of type-2 fuzzy logic in the integration of results [1] [13].

We now show some examples to illustrate the hybrid approach. We use two modules with one neural network each in this modular architecture. Each module is trained with the same data, but results are somewhat different due to the uncertainty involved in the learning process. In all cases, we use neural networks with one hidden layer of 50 nodes and "trainrp" as learning algorithm. The difference in the results is then used to create a type-2 interval fuzzy set that represents the uncertainty in the classification of the word. The first example is of the word "example" in Spanish, which is shown in Fig. 11.

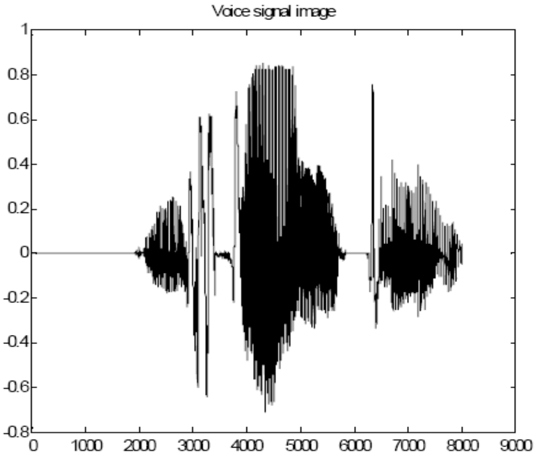


Fig. 11. Sound signal of the word "example" in Spanish.

Considering for now only 10 words in the training, we have that the first neural network will give the following results:
SSE = 4.17649e-005 (Sum of squared errors)
Output = [0.0023, 0.0001, 0.0000, 0.0020, 0.0113, 0.0053, 0.0065, 0.9901, 0.0007, 0.0001]

The output can be interpreted as giving us the membership values of the given sound signal to each of the 10 different words in the database. In this case, we can appreciate that the value of 0.9901 is the membership value to the word "example", which is very close to 1. But, if we now train a second neural network with the same architecture, due to the different random inicialization of the weights, the results will be different. We now give the results for the second neural network:

SSE = 0.0124899
Output = [0.0002, 0.0041, 0.0037, 0.0013, 0.0091, 0.0009, 0.0004, 0.9821, 0.0007, 0.0007]
We can note that now the membership value to the word "example" is of 0.9821. With the two different values of membership, we can define an interval [0.9821, 0.9901], which gives us the uncertainty in membership of the input signal belonging to the word "example" in the database. We have to use centroid deffuzification to obtain a single membership value. If we now repeat the same procedure for the whole database, we obtain the results shown in Table II. In this table, we can see the results for a sample of 6 different words.

Example		Daisy		Way	
M1	M2	M1	M2	M1	M2
0.0023 0.0001 0.0000 0.0020 0.0113 0.0053 0.0065 0.9901 0.0007 0.0001	0.0002 0.0041 0.0037 0.0013 0.0091 0.0009 0.0004 0.9821 0.0007 0.0007	0.0009 0.9957 0.0001 0.0080 0.0005 0.0035 0.0011 0.0000 0.0049 0.0132	0.0124 0.9528 0.1141 0.0352 0.0014 0.0000 0.0001 0.0021 0.0012 0.0448	0.0081 0.0047 0.0089 0.9797 0.0000 0.0074 0.0183 0.0001 0.0004 0.0338	0.0000 0.0240 0.0003 0.9397 0.0126 0.0002 0.0000 0.0069 0.0010 0.0007
Salina		Bed		Layer	
M1	M2	M1	M2	M1	M2
0.9894 0.0031 0.0019 0.0024 0.0001 0.0000 0.0006 0.0001 0.0067 0.0040	0.9780 0.0002 0.0046 0.0007 0.0017 0.0017 0.0000 0.0024 0.0051 0.0012	0.0028 0.0104 0.9949 0.0221 0.0003 0.0003 0.0032 0.0003 0.0094 0.0051	0.0014 0.0012 0.9259 0.0043 0.0025 0.0002 0.0002 0.0004 0.0013 0.0001	0.0009 0.0032 0.0000 0.0001 0.9820 0.0017 0.0070 0.0132 0.0003 0.0010	0.0858 0.0032 0.0005 0.01 0.92 0.00 0.00 0.00 0.00 0.00

Table II. Summary of Results for the Two Modules (M1 AND M2) for a Set of Words in "SPANISH".

The same modular neural network approach was extended to the previous 20 words (mentioned in the previous section) and the recognition rate was improved to 100%, which shows the advantage of modularity and also the utilization of type-2 fuzzy logic. We also have to say that computation time was also reduced slightly due to the use of modularity.We now describe the complete modular neural network architecture (Fig. 12) for voice recognition in which we now use three neural networks in each module. Also, each module only processes a part of the word, which is divided in three parts one for each module.

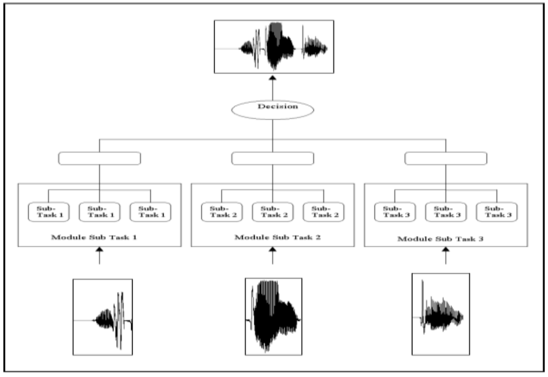


Fig. 12. Complete modular neural network architecture for voice recognition.

We have also experimented with using a genetic algorithm for optimizing the number of layers and nodes of the neural networks of the modules with very good results. The approach is very similar to the one described in the previous chapter. We show in Fig. 13 an example of the use of a genetic algorithm for optimizing the number of layers and nodes of one of the neural networks in the modular architecture. In this figure we can appreciate the minimization of the fitness function, which takes into account two objectives: sum of squared errors and the complexity of the neural network.

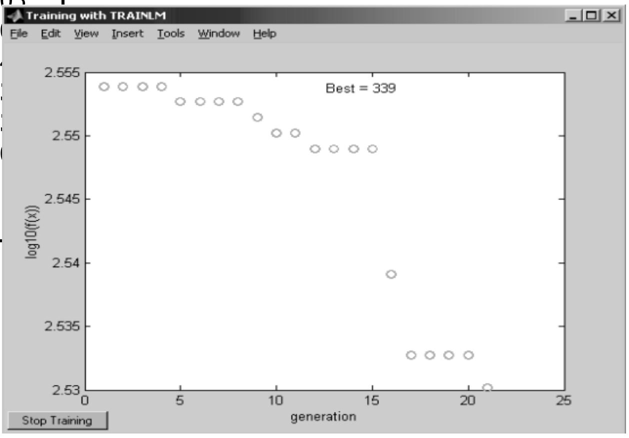


Fig. 13. Genetic algorithm showing the optimization of a neural network.

V. CONCLUSIONS

We have described in this paper an intelligent approach for pattern recognition for the case of speaker identification. We first described the use of monolithic neural networks for voice recognition. We then described a modular neural network approach with type-2 fuzzy logic. We have shown examples for words in Spanish in which a correct identification was achieved. We have performed tests with about 20 different words in Spanish, which were spoken by three different speakers. The results are very good for the monolithic neural network approach, and excellent for the modular neural network approach. We have considered increasing the database of words, and with the modular approach we have been able to achieve about 96% recognition rate on over 100 words. We still have to make more tests with different words and levels of noise.

REFERENCES

[1] O. Castillo, O. and P. Melin, "A New Approach for Plant Monitoring using Type-2 Fuzzy Logic and Fractal Theory", International Journal of General Systems, Taylor and Francis, Vol. 33, 2004, pp. 305-319.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Transactions on Acoustics, Speech and Signal Processing, 29(2), 1981, pp. 254-272.

[3] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques", Speech Communication, 5(2), 1986, pp. 183-197.

[4] S. Furui, "Speaker-independent isolated word recognition using dynamic features of the speech spectrum", IEEE Transactions on Acoustics, Speech and Signal Processing, 29(1), 1986, pp. 59-59.

[5] S. Furui, "Digital Speech Processing, Synthesis, and Recognition". Marcel Dekker, New York, 1989.

[6] S. Furui, "Speaker-dependent-feature extraction, recognition and processing techniques", Speech Communication, 10(5-6), 1991, pp. 505-520.

[7] S. Furui, "An overview of speaker recognition technology", Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994, pp. 1-9.

[8] A. L. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting", Digital Signal Processing, Vol. 1, 1991, pp. 89-106.

[9] N.N Karnik, and J.M. Mendel, "An Introduction to Type-2 Fuzzy Logic Systems", Technical Report, University of Southern California, 1998.

[10] T. Matsui, and S. Furui, "Concatenated phoneme models for text-variable speaker recognition", Proceedings of ICASSP'93, 1993, pp. 391-394.

[11] T. Matsui, and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability", Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994, pp. 59-62.

[12] P. Melin, M. L. Acosta, and C. Felix, "Pattern Recognition Using Fuzzy Logic and Neural Networks", Proceedings of IC-AI'03, Las Vegas, USA, 2003, pp. 221-227.

[13] P. Melin, and O. Castillo, "A New Method for Adaptive Control of Non-Linear Plants Using Type-2 Fuzzy Logic and Neural Networks", International Journal of General Systems, Taylor and Francis, Vol. 33, 2004, pp. 289-304.

[14] P. Melin, F. Gonzalez, and G. Martinez, "Pattern Recognition Using Modular Neural Networks and Genetic Algorithms", Proceedings of IC-AI'04, Las Vegas, USA, 2004, pp. 77-83.

[15] P. Melin, A. Mancilla, C. Gonzalez, and D. Bravo, "Modular Neural Networks with Fuzzy Sugeno Integral Response for Face and Fingerprint Recognition", Proceedings of IC-AI'04, Las Vegas, USA, 2004, pp. 91-97.