Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Neurocomputing 51 (2003) 225–235

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Cooperative supervised and unsupervised learning algorithm for phoneme recognition in continuous speech and speaker-independent context

Najet Arous*, Noureddine Ellouze

Unité de Recherche: Signal, Image, Reconnaissance de Formes, Groupe: Reconnaissance Vocale,
Ecole Nationale d'Ingénieurs de Tunis, BP-37 Campus Univesitaire 1002 Tunis, Tunisia

Received 5 July 2001; accepted 8 May 2002

Abstract

Neural networks have been traditionally considered as an alternative approach to pattern recognition in general, and speech recognition in particular. There have been much success in practical pattern recognition applications using neural networks including multi-layer perceptrons, radial basis functions, and self-organizing maps (SOMs).

In this paper, we propose a system of SOMs based on the association of some supervised and unsupervised learning algorithms inherited from the most popular neural network in the unsupervised learning category, SOM. The case study of the proposed system of SOMs is phoneme recognition in continuous speech and speaker independent context. Also, we propose a way to save more information during training phase of a Kohonen map in the objective to ameliorate speech recognition accuracy. The applied SOM variants serve as tools for developing intelligent systems and pursuing artificial intelligence applications.

Keywords: Neural network; Supervised learning; Unsupervised learning; Self-organizing map; Continuous speech recognition

1. Introduction

Among the most challenging areas of artificial intelligence and pattern recognition, automatic speech recognition is one of the different topics which has attracted particular

* Corresponding author.
E-mail addresses: najet.arous@enit.rnu.tn (N. Arous), n.ellouze@enit.rnu.tn (N. Ellouze).

PII: S0925-2312(02)00618-5

attention and research effort. In spite of this important progress and cumulated improvements of such systems, under natural and realistic conditions, they are still far from the level that humans attain under similar conditions. The proposed and developed systems are based on a large number of techniques and models [16,18,19,22]. Artificial neural networks have also been applied successfully in speech recognition applications including multi-layer perceptrons, radial basis functions, and self-organizing maps (SOMs) [3,7,8,21] among others [1,2,6].

Kohonen network (SOM), the most popular artificial neural network algorithm in the unsupervised learning category, is a structured non-hierarchical network of simple, laterally interconnected computing units. This means that all neurons are of the same kind and they are all connected to their nearest neighbors with respect to the structure of the network. The layout of these neurons is in most practical applications two-dimensional and the structure often determines the nearest neighbor set of each unit. All neurons act as input units and receive the same external N-dimensional input signal in parallel. Furthermore, they receive internal feedback signals through the lateral connections from other units. To each unit i there is associated a real valued weight vector w_i of the same dimension as the external input stimuli [11]. Training a Kohonen network to do a specific task is a matter of finding the right set of weight vectors (codebook vectors) in relation to the given set of input patterns. These codebook vectors, while being trained, also tend to approximate $p(x)$, the probability density function of a stochastic data vector x. One might then say that the SOM is a non-linear projection of the probability density function $p(x)$ of the high-dimensional input data vector onto the two-dimensional display. Vector x may be compared with all w_i in any metric, in many practical applications, the smallest of the euclidean distances $\|x - w_i\|$ can be made to define the best matching unit (BMU) [15,24].

Kohonen has done a large number of experiments [13] where he uses SOM to create phoneme maps of input vectors which are short time spectra of digitized speech. A phoneme mapping is a mapping from some representation of speech constituting an input domain of real valued vectors into some domain in which the points are interpreted as phoneme representations [14]. Thus, such a mapping is a phoneme classifier known as “ phonotopic map”.

In this paper, we propose a way to enrich information in a Kohonen map with the purpose to study speech signal classification by means of unsupervised learning rule and also to reach best phoneme recognition rates. We present also a cooperative system for continuous speech recognition composed of associated supervised and unsupervised learning algorithm based on the famous Kohonen learning rule. The goal of such system of SOMs is to create autonomous systems, the parts of which can learn also from each other. On the other hand, competitive nature of each SOM system participates in the decision in recognition phase.

In the following, we describe how to enrich information in a Kohonen map. Thereafter, we describe the combined systems of SOMs by detailing each of its competitive learning algorithms. Finally, we illustrate experimental results of the application of each isolated competitive system and associated SOM components to recognize vowels of the TIMIT database in the objective to be extended for continuous speech recognition.

2. How to enrich information in a Kohonen map ?

In the study of phoneme clustering by means of SOM [17], we have concluded that SOM is not able to decide optimally how many quasi-phonemes correspond to one phoneme and what the most probable phoneme is. For this reason, we propose a procedure to enrich information in a Kohonen map during training phase. In fact, SOM in its unsupervised traditional learning nature update BMU weight vector and its neighbor vectors based, on one hand on the minimal distance between a sample input vector and codebook vectors of a Kohonen map, on the other hand on a neighborhood radius.

Since TIMIT corpus is labelled, we will benefit from this information and we suggest to save more information inside each selected BMU unit.

During training phase, when a sample input vector is attributed to a BMU unit, we save its corresponding vector, its label and we update its frequency activation. By this way, each unit of a Kohonen map is characterized by

- A general centroid vector (GCV): determined by means of Kohonen update rule.
- Information relating to each phoneme class attributed to a unit :
- Mean vector (MV) of the phoneme class.
- Label of the phoneme class.
- Activation frequency of the phoneme class.

3. System of SOMs

In this paragraph, we present a cooperative system based on the association of supervised and unsupervised learning algorithms inherited from SOM (Fig. 1) which are:

- SOM based on a sequential learning.
- SOM based on an optimized sequential learning.
- SOM with locally adapting neighborhood radii:AdSOM.
- SOM with a conscience term:DeSieno.
- SOM based on a supervised sequential learning.

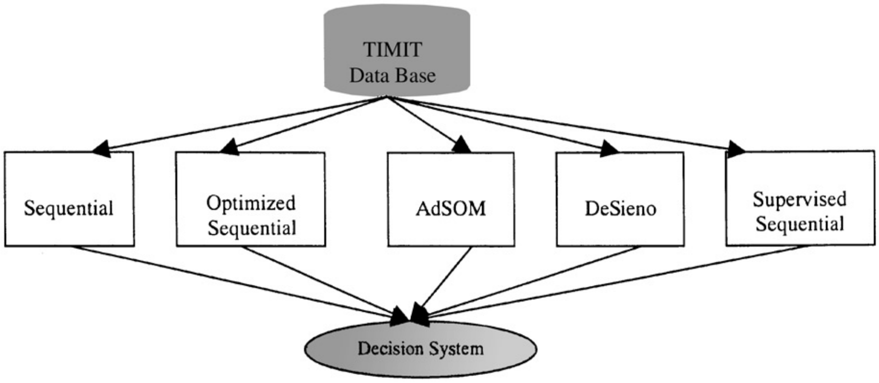

Fig. 1. Systems of SOMs.

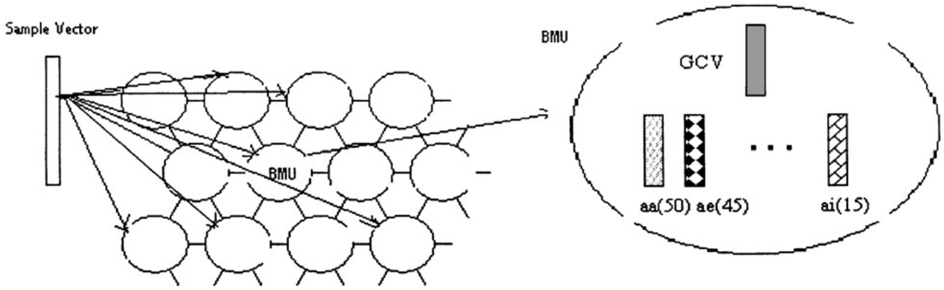

Fig. 2. Recognition schema.

The goal of such association is to create a cooperative system of different learning algorithms based on similar competition principle. Each isolated subsystem can be considered as an autonomous system. Moreover, each of them can learn from the other. The resulting of each subsystem is a sort of a phonotopic map.

During training phase, each subsystem learns the input space in its own competitive manner. We modify each SOM variant in order to enrich information of each associated Kohonen map as described in the previous paragraph. During recognition phase, a decision rule maximises proposed solutions of each SOM subsystem. Each subsystem presents to the decision system the phoneme label and its activation frequency by the following way.

Inside a SOM subsystem phoneme recognition is performed at frame level and performance is evaluated by comparing each recognized frame with the reference one. A recognition decision is operated in two steps (Fig. 2). At a first step, for a test sample vector presented to a modified SOM, we search for the BMU among all GCV of a map. Thereafter, inside the selected BMU unit, we search for the best mean vector (MV) of different classes of the selected unit, in terms of minimal euclidean distance. Finally, for the selected best mean vector we retain its label and its frequency activation. These informations (label and its frequency activation) will be presented thereafter to the decision system.

Example. Suppose that for a given two test sample vectors TSV1 and TSV2 presented to each isolated subsystem, we have the following responses:[1]

	TSV1	TSV2
Sequential learning	aa(50)	aa(50)
Optimized sequential learning	ae(49)	ae(49)
AdSOM	aa(60)	aa(60)
DeSieno	ae(40)	ae(40)
Supervised sequential learning	aa(30)	ao(39)

[1] A response correspond to: phoneme label(frequency activation) e.g. aa(50).

In the case of TSV1, a classification retained by the decision system is the phoneme class “aa” because the majority of subsystems (sequential learning, AdSOM and supervised sequential learning) have presented to the decision system the same class which is “aa”.

In the case of TSV2, a classification retained by the decision system is the phoneme class “aa” because the total frequency activation ($50 + 60 = 110$) presented by the subsystem based on sequential learning and AdSOM one is superior to that ($49+40=89$) presented by subsystem based on optimized sequential learning and DeSieno one.

In the following, we describe associated SOM variants.

3.1. Sequential learning

In sequential learning algorithm, the SOM is trained iteratively in time sequential manner. In each step, distances from the weight vectors of the current map and a randomly chosen input vector are calculated and BMU is found. After finding the BMU, his weight vector is updated so that the BMU is moved closer to the current input vector [20,23]. The topological neighbors of the BMU is also updated. This adaptation procedure stretches the BMU and its topological neighbors towards the input sample vector. The SOM update rule for the weight vector of the unit i in the neighborhood of the BMU is

$$\mathrm{m}_i(t+1) = \mathrm{m}_i(t) + \alpha(t)h_{ci}(t)[x(t) - \mathrm{m}_i(t)], \quad \forall i \in [1 \ldots n] \tag{1}$$

where $x(t)$ is the input vector randomly drawn from the input data set at time t, n is the number of neurons, $h_{ci}(t)$ is the neighborhood kernel around the winner unit c and $\alpha(t)$ is the learning rate at time t. The neighborhood kernel is a non-increasing function of time and of the distance of unit i from the winner unit c. It defines the region of influence that the input sample has on the SOM [4,13].

3.2. Optimized sequential learning

The traditional sequential learning algorithm is modified in such a way that each unit has its own specific learning rate α_i and neighborhood radius [9].

3.3. SOM with locally adapting neighborhood radii:AdSOM

In the AdSOM, each neuron i has its own neighborhood radius[2] parameter σ_i associated with it. When training of the AdSOM is started, the parameter σ_i is set to half the diameter of the lattice for all i. During training, sample vectors are presented to the AdSOM just as to the basic SOM, and the weights are adjusted according to the familiar update rule, with decreasing learning rate. However, the width of the neighborhood is determined by the σ_i value of the best matching unit.

[2] The neighborhood radius σ_i: determines the size of the topological neighborhood surrounding the BMU, where prototype vectors are updated. Usually, this parameter decreases in time, like the learning rate.

The parameter σ_i are determined by the local topographic errors. If for a sample vector x the two nearest weight vectors are w_j and w_k, and the corresponding best matching units n_j and n_k are not adjacent, there is local topographic error. Then, for units n_i that are near the two BMUs, a new value for the neighborhood radius σ_i is calculated [12]:

$$\sigma_i = \begin{cases} \|n_j - n_k\| & \text{if } \max\{\|n_i - n_j\|, \|n_i - n_k\|\} \leqslant \|n_j - n_k\|, \\ \|n_j - n_k\| - s & \text{when } s = \min\{\|n_i - n_j\|, \|n_i - n_k\|\} < \|n_j - n_k\|, \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

Thus, for units between the two BMUs, the new σ_i is equal to the distance on the map between the BMUs; outside that area, the σ_i decreases linearly to one.

3.4. SOM with a conscience term:DeSieno

The purpose of the proposed improvement to Kohonen learning is to form a better approximation of a probability density function of an input sample x. When an input vector is presented to the network, a competition is held to determine which processing element's weight vector is closest in any metric to the input vector called BMU. The winning processing element in the competition is not necessarily the element to have its weights reinforced. A bias is developed for each processing element based on the number of times it has won the competition. The weights of the processing element winning this biased competition are adjusted as follows [5]:

$$\mathrm{m}_i(t+1) = \mathrm{m}_i(t) + \alpha(t)h_{ci}(t)[x(t) - \mathrm{m}_i(t)] - c(1/n - p_i(t)), \quad \forall i \in [1 \ldots n], \quad (3)$$

where c is a positive scalar, n is the number of neurons and p_i is the likelihood of selection of unit i.

As a network is trained using this algorithm, and processing elements start winning their share of the competitions, the bias terms have less impact and the process tends to revert to a simple Kohonen learning rule. The term conscience arises because a processing element that wins too often begins to "feel guilty" and prevents itself from winning excessively.

3.5. Supervised SOM

The classification accuracy could be improved by a significant amount if information about the class identity could be taken into account in the learning phase. In order to make the SOM supervised, the input vectors were formed of two parts x_s and x_u, where x_s was a 12 mel cepstrum coefficients computed over 16 ms, and x_u corresponded to a unit vector with its components assigned to one of the 20 phonemic classes that were taken into account. The concatenated 32-dimensional vectors $x = [x_s^{\mathrm{T}}, x_u^{\mathrm{T}}]^{\mathrm{T}}$ were then used as inputs to the SOM. Supervised learning means that whereas the classification of each x_s in the training set is known, the corresponding x_u value must be used during training. During recognition of an unknown x, only its x_s part is compared with the corresponding part of the weight vectors, x_u was not considered [10].

4. Experimental results

We have implemented a cooperative system for continuous speech recognition based on systems of SOMs composed of three main components: a pre-processor for vowels sounds and producing mel cepstrum vectors. The sound input space is composed by 12 mel cepstrum coefficients each 16 ms frame. Three frames are selected at the middle of each phoneme to generate data vectors. The second component is a system of multi-SOMs. The third component is a phoneme recognition module.

The speech database used is the DARPA TIMIT acoustic-phonetic continuous speech corpus.

4.1. The DARPA TIMIT speech corpus

We have used the TIMIT corpus for the purpose of developing and evaluating the proposed multi-SOMs system for speaker independent continuous speech recognition. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. A speaker's dialect region is the geographical area of the US where they lived their childhood years. The data were recorded at a sample rate of 16 KHz and a resolution of 16 bits. In our experiments, we have used the New England dialect region (DR1) composed of 31 males and 18 females. The corpus contains 14 399 phonetic unit for training. Each phonetic unit is represented by three frames selected at the middle of each phoneme to generate data vectors. Training has been made on vowel phonemes, which are : 'iy', 'ih', 'eh', 'ey', 'ae', 'aa', 'aw', 'ay', 'ah', 'ao', 'oy', 'ow', 'uh', 'uw', 'ux', 'er', 'ax', 'ix', 'axr', and 'ax–h'; detailed number of frames are shown in Table 1.

We have implemented above cited SOM variants and described cooperative system of SOMs. We have compared each isolated subsystem and associated system of SOMs using a data set that consisted of 13 669 sample vectors. The neural lattice was bidimensional. Three experiments were conducted. In the first, the map size is 10×20, in the second, the map size is 20×20 and in the last experiment the map size is 40×40. All maps are trained for $10 \times$ length data set iterations. For all maps the learning rate decreases linearly from 0.99 to 0.02. The radius width also decreases linearly from half the diameter of the lattice to one. All maps of the same size have the same initial conditions (that is the same $m_i(0)$).

Table 1
Frame number of each phoneme

Phonemes	iy	ih	eh	ey	ae	aa	aw	ay	ah	ao
Frames	1552	1103	946	572	1038	762	180	600	580	665
Phonemes	oy	ow	uh	uw	ux	er	ax	ix	axr	ax-h
Frames	192	549	141	198	400	392	871	2103	739	86

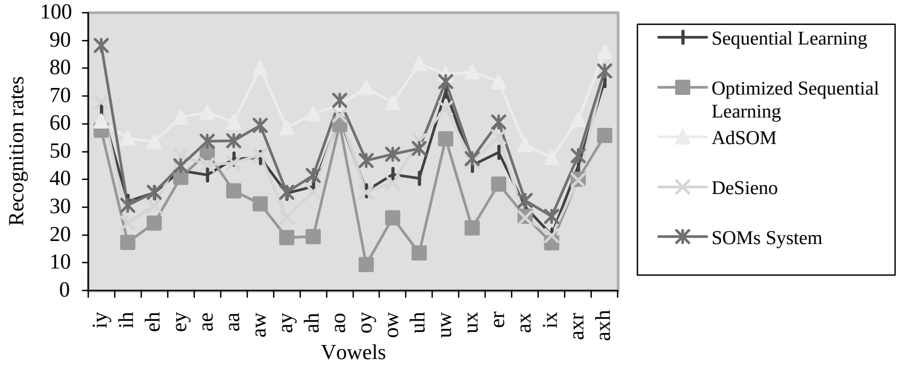

Fig. 3. Vowel recognition rates map size: 10 × 20.

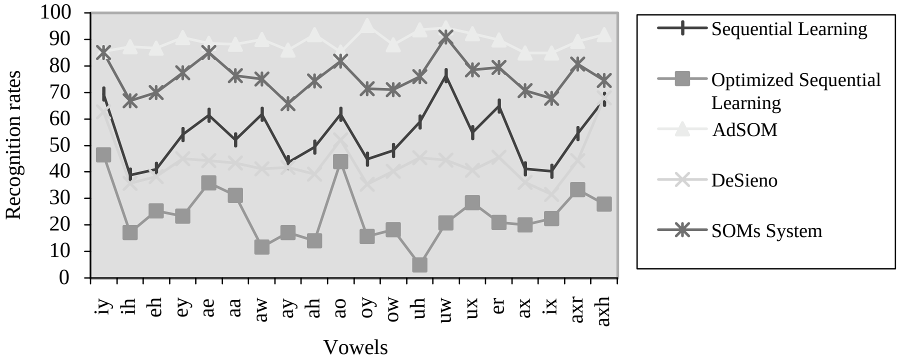

Fig. 4. Vowel recognition rates map size: 20 × 20.

Figs. 3–5 show a comparison of different recognition rates obtained by using, respectively, SOM based on a sequential learning, SOM based on an optimized sequential learning, SOM with locally adapting neighborhood radii:AdSOM, SOM with a conscience term:DeSieno and system of SOMs.

From Figs. 3–5 AdSOM provides best recognition rates accuracy in comparison with other variants. However, we should note that recognition rate depends on the number of phoneme frames presented during training phase. That is why, we propose to present phonemes to SOMs with the same probability (equiprobability). For example, for vowels “iy” and “axh” we remark that is well recognized even for isolated subsystems and also the best accuracy for the associated SOMs. We can conclude also that phonemes which resemble each other phonetically such as “ih” and “eh” do not provide even a mean recognition rate only AdSOM reach 87.31% (map size 40 × 40). That is, SOM in its different implemented variants is not able to separate “near” phonemes. In this sense, we suggest to hybridize SOM and genetic algorithm in order to reach

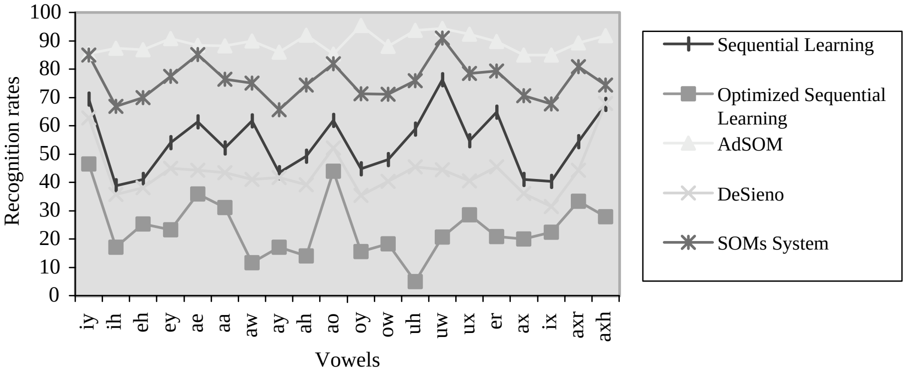

Fig. 5. Vowel recognition rates map size: 40 × 40.

such separation. On the other hand, we suggest to study SOM parameters for each variant.

5. Conclusion

In this paper, we have studied the learning quality of supervised and unsupervised learning algorithm of the most popular artificial neural network algorithm in the unsupervised learning category. The case study of such algorithms is phoneme recognition. We have proposed a procedure to enrich information in a Kohonen map. Also, we have proposed a cooperative system of SOMs for phoneme recognition.

The main results are as follows:

- Kohonen update rule is used to determine the GCV and to create ordered map and extra information saved during training serves as a tool to reach better recognition accuracy.
- AdSOM provides best recognition rates in comparison with other SOM variants.
- Map size 40 × 40 provides best recognition accuracy and particularly by applying AdSOM.
- System of SOMs provides best recognition rates if it is proved by all subsystems.
- All variants of SOMs are not able to separate phonetically "near" vowels.

We suggest to hybridize SOM and genetic algorithm on one hand to fine tune SOM parameters and on the other hand for training data set input in the objective not only to ameliorate recognition rates but also to create frontier between phonetically "near" phonemes.

Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable comments which improved the presentation and contents of this paper considerably.

References

[1] M. Beveridge, Using self-organizing maps for the objective assessment of /s/ misarticulations by patients with intra-oral cancers, Speech and Language Processing, MSc. Dissertation, Department of Linguistics, University of Edinburgh, September 14, 1993.

[2] J.C. Chappelier, RST: une architecture connexionniste pour la prise en compte de relations spatiales et temporelles, Thèse de doctorat de l'Ecole Nationale Supérieure des Télécommunications Spécialité:Informatique et Réseaux, Paris, January 11, 1996.

[3] A. Choppin, Unsupervised classification of high dimensional data by means of self-organizing neural networks, Master Thesis, Département d'Ingénierie Informatique, Université Catholique de Louvain Faculté des Sciences, Appliquées, 1997–1998.

[4] M. Cottrell, J.C. Fort, G. Pagès, Theoretical aspects of the SOM algorithm, Neurocomputing 21 (1–3) (1998) 119–138.

[5] D. DeSieno, Adding a conscience to competitive learning, IEEE Internat Conference on Neural Networks, New York (1988) pp. 117–124.

[6] T. Graepel, M. Burger, K. Obermayer, Self-organizing maps: generalizations and new optimization Techniques Neurocomputing 21 (1–3) (1998) 173–190.

[7] A. Hockstra, M.F.J. Drossaers, An extended Kohonen feature map for sentence recognition, Proceedings of the ICANN'93, 1993, pp. 404–407.

[8] O.B. Jensen, M. Olsen, T. Rohde, Automatic speech recognition & Neural networks, Thesis in Computer Science, Computer Science Department, Aarhus University, Denmark, April 25, 1991.

[9] J.A. Kangas, T. Kohonen, J. Laraksonen, Variants of self-organizing maps, IEEE Transactions on Neural Networks 1 (1) (1990) 93–99.

[10] J. Kangas, On the analysis of pattern sequences by self-organizing maps, Thesis for the degree of Doctor of Technology, Helsinki University of Technology, 6 May, 1994.

[11] S. Kaski, T. Honkela, K. Lagus, T. Kohonen, WEBSOM—self-organizing maps of document collections, Neurocomputing 21 (1–3) (1998) 101–117.

[12] K. Kiviluoto, Topology preservation in self-organizing maps, Proceedings of the International Conference on Neural Networks (ICNN), 1996, pp. 294–299.

[13] T. Kohonen, Self-Organizing Map, 3rd Edition, Springer, Berlin, 2001.

[14] T. Kohonen, P. Somervuo, Self-organizing maps for symbol strings, Neurocomputing 21 (1–3) (1998) 19–30.

[15] T. Kohonen, J. Hynnnien, J. Kangas, J. Laaksonen, SOM_PAK The self-organizing map program package, Laboratory of Computer and information Science, Helsinki University of Technology, April 7, 1995.

[16] C.G. Looney, Pattern Recognition Using Neural Networks, Oxford University Press, New York, Oxford, 1997.

[17] A. Najet, Speech clustering by means of self organizing maps, Tunisan–German Conference on Smart Systems and Devices, Hammamet-Tunisia, 27–30, March 2001.

[18] S. Nakagawa, K. Shikano, Y. Tohkura, Speech, Hearing and Neural Network Models, IOS Press, Amsterdam, Oxford, Tokyo, Washington, DC, 1995.

[19] S. Saito, Speech Science and Technology, IOS Press, Ohmsha, 1992.

[20] P. Samervuo, T. Kohonen, Self-organizing maps and learning vector quantization for feature sequences, Neural Process. Lett. 10 (1999) 151–159.

[21] V.D. Sanchez, A Neural networks based pattern recognition, in: S.K. Pal, et al., (Eds.), Pattern Recognition from Classical to Modern Approaches, World Scientific, Singapore, 2001, pp. 281–300.

[22] C. Tadj, Méthodes connexionnistes de quantification vectorielle à apprentissage compértitif. Application à la détection de mots clés, Thèse de doctorat de l'Ecole Nationale Supérieure des Télécommunications Spécialité :Signal et Images, Paris, April 3, 1995.

[23] J. Vesanto, Data mining techniques based on the self-organizing map, Thesis of the degree of Master of Science in Engineering, Helsinki University of Finland, 26 May, 1997.

[24] J. Vesanto, J. Himberg, E. Alhoniemi, The self-organizing map function package for Matlab 5, Laboratory of Computer and information Science, Helsinki University of Technology, 15, February 2000.

Najet Arous received computer science engineering degree from Ecole Nationale des Sciences d'Informatique, Tunis, Tunisia, the MS degree in electrical engineering (signal processing) from Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia, is currently working towards the Ph.D. degree in electrical engineering (signal processing) from ENIT. She is currently a computer science assistant in the computer science department at FSM, Tunisia. Her research interests include scheduling optimization, speech recognition and evolutionary neural networks.

Noureddine Ellouze received a Ph.D. degree in 1977 from l'Institut National Polytechnique at Paul Sabatier University (Toulouse-France), and Electronic Engineer Diploma from ENSEEIHT in 1968 at the same University. In 1978, Dr. Ellouze joined the Department of Electrical Engineering at the National School of Engineer of Tunis (ENIT—Tunisia), as assistant professor in statistic, electronic, signal processing and computer architecture. In 1990, he became Professor in signal processing, digital signal processing and stochastic process. He has also served as director of electrical department at ENIT from 1978 to 1983, General manager and President of the Research Institut on Informatics and Telecommunication IRSIT from 1987 to 1990, and President of the Institut during 1990–1994. He is now Director of Signal Processing Research Laboratory LSTS at ENIT, and is in charge of Control and Signal Processing Master degree at ENIT.Dr. Ellouze is IEEE fellow since 1987, he directed multiple Masters and Thesis and published over 200 scientific papers both in journals and proceedings. He is chief editor of the scientific journal Annales Maghrébines de l'Ingénieur. His research interests include neural networks and fuzzy classification, pattern recognition, signal processing and image processing applied in biomedical, multimedia, and man machine communication.