Control of a nurse robot using voice commands and associative memories

Roberto A. Vázquez and Humberto Sossa

Centro de Investigación en Computación-IPN
Av. Juan de Dios Batíz, esq. Miguel Othón de Mendizábal.
Mexico City, 07738. Mexico
Contact: ravem@ipn.mx, hsossa@cic.ipn.mx

(Paper received on February 29, 2008, accepted on April 15, 2008)

Abstract. NR-Alpha is a prototype of a surgical instrument server nurse robot. NR-Alpha is designed to provide to the surgeon the demanded surgery instruments. NR-Alpha perceives the surgeon's voice, recognizes the name of the instrument demanded and starts to look for the demanded instrument on the working area; once localized the instrument, NR-Alpha grapes it and finally reaches the hand of the surgeon to give him the demanded instrument. NR-Alpha is composed by three main modules: Artificial Vision Module (AVM), Voice Recognition Module (VRM) and Control Module (CM). In this paper, we describe how the VRM module could be implemented. To recognize the name of the instrument pronounced by the surgeon, VRM uses a dynamic associative memory (DAM). This DAM stores associations between a voice signals that encode the name of a surgery instrument and images of the corresponding instrument. Once the associative memory is trained, we would expect that when the surgeon pronounces, for example, "Forceps" the associative memory would recall the image of a forceps. Subsequently, the image recalled by the DAM could be used to localize the instrument. In order to test the accuracy of the proposal, we firstly train the DAM with associations of the instrument we would like the DAM learned. We then use a benchmark composed by 1800 voice signals to test the performance of the proposal.

1 Introduction

Robotics technology is developing dramatically. In this sense, a robot system is an alternative since accuracy for sensors and control system is increasing and computer technology for robots is rapidly developing. Robots are been used in increasingly complex surgical procedures. However these robots are not autonomous machines that carry out simple, pre-programmed instructions.

NR-Alpha is a surgical instrument server nurse robot (under development) composed by three main modules: An Artificial Vision Module (AVM), for the details refer to [11], A Voice Recognition Module (VRM) and Control Module (CM). In this paper we particularly emphasize on the implementation of the VRM module by means of reported associative memories.

An associative memory is a particular kind of neural network specially designed to recall output patterns in terms of input patterns that might appear altered by some kind

Special Issue in Electronics and Biomedical Informatics,
Computer Science and Informatics
Research in Computing Science 35, 2008, pp. 77-85

of noise, refer for example to [1], [2], [3], [4], [6], [7] and [8]. Most of these associative models have several constraints that limit their applicability in real life problems. In order to achieve the best performance the input patterns have to satisfy several conditions. Recently in [6] and [7] a new dynamic associative model (DAM) was proposed. This model can be used to recall a set of images even if these images suffer affine transformations. This model also has been applied to different pattern recognition problems. Refer for example to [6], [9] and [10].

In this paper we describe how the proposed associative model was implemented into the VRM for recognizing the name of the instrument pronounced by the surgeon. This DAM stores associations between a voice signals that encode the names of a surgery instruments and images of them. Once trained the associative memory we expect that when the surgeon says for example "Forceps" the associative memory recall the image of a forceps. Subsequently, the image recalled by the DAM could be used to localize the instrument. In order to test the accuracy of the proposal, we firstly train the DAM with associations of the instrument we would like the DAM learned, then during recognition, we use a benchmark composed by 1800 voice signals.

2 Description of the dynamic associative model

This model is not an iterative model as Hopfield's model [1]. The model emerges as an improvement of the model proposed in [4] which is not an iterative model. Let $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} \in \mathbf{R}^m$ an input and output pattern, respectively. An association between input pattern $\mathbf{x}$ and output pattern $\mathbf{y}$ is denoted as $(\mathbf{x}^k, \mathbf{y}^k)$, where k is the corresponding association. Associative memory: $\mathbf{W}$ is represented by a matrix whose components w_{ij} can be seen as the synapses of the neural network. If $\mathbf{x}^k = \mathbf{y}^k \forall k = 1, \ldots, p$ then $\mathbf{W}$ is auto-associative, otherwise it is hetero-associative. A distorted version of a pattern $\mathbf{x}$ to be recalled will be denoted as $\tilde{\mathbf{x}}$. If an associative memory $\mathbf{W}$ is fed with a distorted version of $\mathbf{x}^k$ and the output obtained is exactly $\mathbf{y}^k$, we say that recalling is robust.

2.1 Building the associative memory

This model is bio-inspired in some biological ideas of human brain. Humans, in general, do not have problems recognizing patterns even if these are altered by noise. Several parts of the brain interact together in the process of learning and recalling a pattern. This model defines several interacting areas, one per association we would like the memory to learn. Also integrate the capability to adjust synapses in response to an input stimulus. Before an input pattern is learned or processed by the brain, it is hypothesized that it is transformed and codified by the brain. This process is simulated using the procedure introduced in [5].

This procedure allows computing *codified patterns* from input and output patterns denoted by $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ respectively; $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are *de-codifying patterns*. Codified and de-codifying patterns are allocated in different interacting areas and d defines of much these areas are separated. On the other hand, d determines the noise supported by our model. In addition a simplified version of $\mathbf{x}^k$ denoted by s_k is obtained as:

$$s_k = s\left(\mathbf{x}^k\right) = \mathbf{mid}\ \mathbf{x}^k \tag{1}$$

where **mid** operator is defined as $\mathbf{mid}\ \mathbf{x} = x_{(n+1)/2}$.

When the brain is stimulated by an input pattern, some regions of the brain (interacting areas) are stimulated and synapses belonging to those regions are modified. In this model, the most excited interacting area is call *active region* (AR) and could be estimated as follows:

$$ar = r(\mathbf{x}) = \arg\left(\min_{i=1}^{p} \left|s(\mathbf{x}) - s_i\right|\right) \tag{2}$$

Once computed the *codified patterns*, the *de-codifying patterns* and s_k we can compute the synapses of the associative memory as follows:

Let $\left\{\left(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k\right) | k = 1, \ldots, p\right\}, \overline{\mathbf{x}}^k \in \mathbf{R}^n, \overline{\mathbf{y}}^k \in \mathbf{R}^m$ a fundamental set of associations (codified patterns). Synapses of associative memory $\mathbf{W}$ are defined as:

$$w_{ij} = \overline{y}_i - \overline{x}_j \tag{3}$$

In short, building of the associative memory can be performed in three stages as:

1. Transform the fundamental set of association into codified and de-codifying patterns by means of previously described Procedure 1.
2. Compute simplified versions of input patterns by using equation 1.
3. Build $\mathbf{W}$ in terms of codified patterns by using equation 3.

2.2 Modifying synapses of the associative model

In this model, synapses could change in response to an input stimulus; but which synapses should be modified? There are synapses that can be drastically modified and they do not alter the behavior of the associative memory. In the contrary, there are synapses that only can be slightly modified to do not alter the behavior of the associative memory; we call this set of synapses *the kernel* of the associative memory and it is denoted by $\mathbf{K}_\mathbf{W}$. In this model are defined two types of synapses: synapses that can be modified and do not alter the behavior of the associative memory and synapses belonging to the kernel of the associative memory. These last synapses play an important role in recalling patterns altered by some kind of noise.

Let $\mathbf{K}_\mathbf{W} \in \mathbf{R}^n$ the kernel of an associative memory $\mathbf{W}$. A component of vector $\mathbf{K}_\mathbf{W}$ is defined as:

$$kw_i = \mathbf{mid}\left(w_{ij}\right), j = 1,\ldots,m \tag{4}$$

Synapses that belong to $\mathbf{K}_\mathrm{W}$ are modified as a response to an input stimulus. Input patterns stimulate some ARs, interact with these regions and then, according to those interactions, the corresponding synapses are modified. Synapses belonging to $\mathbf{K}_\mathrm{W}$ are modified according to the stimulus generated by the input pattern. This adjusting factor is denoted by Δw and can be computed as:

$$\Delta w = \Delta\left(\mathbf{x}\right) = s\left(\overline{\mathbf{x}}^{ar}\right) - s\left(\mathbf{x}\right) \tag{5}$$

where ar is the index of the AR.

Finally, synapses belonging to $\mathbf{K}_\mathrm{W}$ are modified as:

$$\mathbf{K}_\mathrm{W} = \mathbf{K}_\mathrm{W} \oplus \left(\Delta w - \Delta w_{old}\right) \tag{6}$$

where operator $\oplus$ is defined as $\mathbf{x} \oplus e = x_i + e \;\forall i = 1,\ldots,m$. As you can appreciate, modification of $\mathbf{K}_\mathrm{W}$ in equation 6 depends of the previous value of Δw denoted by Δw_{old} obtained with the previous input pattern. Once trained the **DAM**, when it is used by first time, the value of Δw_{old} is set to zero.

2.3 Recalling a pattern using the proposed model

Once synapses of the associative memory have been modified in response to an input pattern, every component of vector $\overline{\mathbf{y}}$ can be recalled by using its corresponding input vector $\overline{\mathbf{x}}$ as:

$$\overline{y}_i = \mathbf{mid}\left(w_{ij} + \overline{x}_j\right), j = 1,\ldots,n \tag{7}$$

In short, pattern $\overline{\mathbf{y}}$ can be recalled by using its corresponding key vector $\overline{\mathbf{x}}$ or $\tilde{\mathbf{x}}$ in six stages as follows:

1. Obtain index of the active region ar by using equation 2.
2. Transform $\mathbf{x}^k$ using de-codifying pattern $\hat{\mathbf{x}}^{ar}$ by applying the following transformation: $\tilde{\mathbf{x}}^k = \mathbf{x}^k + \hat{\mathbf{x}}^{ar}$.
3. Compute adjust factor $\Delta w = \Delta\left(\tilde{\mathbf{x}}\right)$ by using equation 5.
4. Modify synapses of associative memory $\mathbf{W}$ that belong to $\mathbf{K}_\mathrm{W}$ by using equation 6.
5. Recall pattern $\tilde{\mathbf{y}}^k$ by using equation 7.
6. Obtain $\mathbf{y}^k$ by transforming $\tilde{\mathbf{y}}^k$ using de-codifying pattern $\hat{\mathbf{y}}^{ar}$ by applying transformation: $\mathbf{y}^k = \tilde{\mathbf{y}}^k - \hat{\mathbf{y}}^{ar}$.

The formal set of prepositions that support the correct functioning of this dynamic model and the main advantages again other classical models can be found in [12].

3 Experimental results

Several experiments were performed in order to test the accuracy of the proposal when a person pronounces the name on the instrument he needs. Firstly, we recorded a collection of signal voices. We recorded the name of five different instruments (backcock forceps, sponge forceps, adson forceps, allis forceps and rat tooth forceps).

Each signal voice was recorded in a wav file (PCM format, 44.1 KHz, 16 bits and mono). In average the duration of each sample was of 450 ms. Some samples of the signal voices are shown in Fig. 1. In total, 1960 signals were analyzed through the experiments.

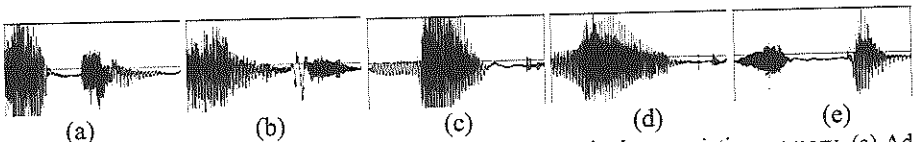

Fig. 1. Signal voice from the different instruments used to train the associative memory. (a) Adson. (b) Allis. (c) Backcock. (d) Rat tooth. (e) Sponge.

In addition, we proceeded to obtain an image of each instrument in order to associate it with its corresponding signal voice. Some images are shown in Fig 2.

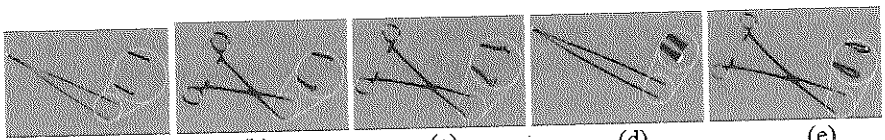

Fig. 2. Images from the different instruments used to train the associative memory. (a) Adson. (b) Allis. (c) Backcock. (d) Rat tooth. (e) Sponge.

In order to train the associative memory, we firstly transformed the signal sound of each instrument into a raw vector and the image of each instrument into a raw vector. Then each signal sound vector was associated with its corresponding image vector. Finally the associate memory was trained using the procedures described in section 2.

Experiment 1. In this experiment, we verified if the associative model was capable to recall the fundamental set of associations, in other words, if the DAM was able to recall the image associated to the signal voice used as input pattern. In average, the accuracy of the proposal in this experiment was of 100%, some examples are shown in Fig. 3.

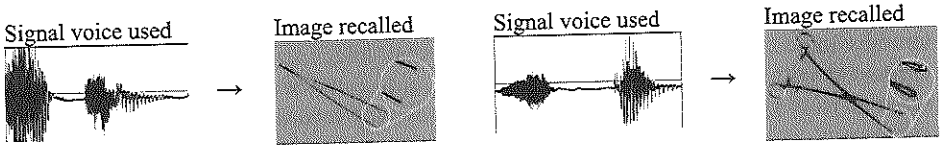

Fig. 3. Some examples of Images recalled using an specific signal voice.

Experiment 2. In this experiment, we verified if the DAM was able to recall the image associated to the signal voice used as input pattern, even if the signal voice is altered by additive noise (AN). To do this, each signal voice previously recorder was contaminated with additive noise altering from 2% until 90% of the information. 89 new samples were generated from each signal voice already recorder. This new set of signal voices was composed for 440 samples, some examples are shown in Fig. 4. In average, the accuracy of the proposal using this set of signal voices was of 71.3%.

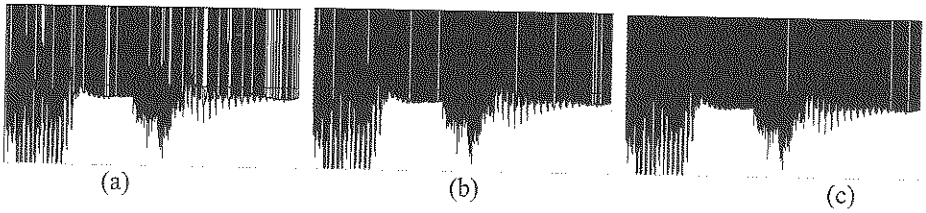

Fig. 4. (a-c) Signal voice of the adson instrument contaminated with additive noise.

Experiment 3. In this experiment, we verified if the DAM was able to recall the image associated to the signal voice used as input pattern, even if the signal voice is altered by subtractive noise (SN). To do this, each signal voice previously recorder was contaminated with subtractive noise altering from 2% until 90% of the information. 89 new samples were generated from each signal voice already recorder. This new set of signal voices was composed for 440 samples, some examples are shown in Fig. 5. In average, the accuracy of the proposal using this set of signal voices was of 71.6%.

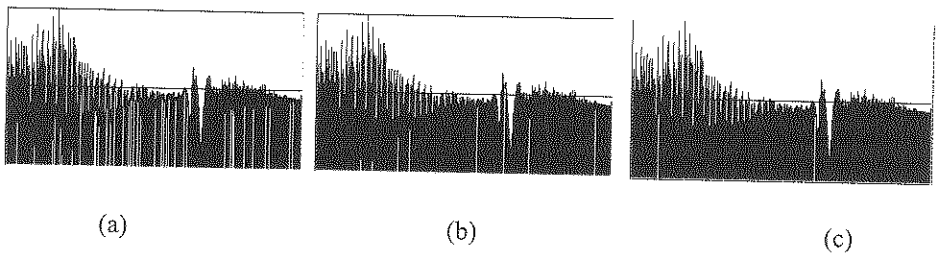

Fig. 5. (a-c) Signal voice of the allis instrument contaminated with subtractive noise.

Experiment 4. In this experiment, we verified if the DAM was able to recall the image associated to the signal voice used as input pattern, even if the signal voice is altered by mixed noise (MN). To do this, each signal voice previously recorder was contaminated with mixed noise altering from 2% until 90% of the information. 89 new samples were generated from each signal voice already recorder. This new set of signal voices was composed for 440 samples, some examples are shown in Fig. 6. In average, the accuracy of the proposal using this set of signal voices was of 79.55%.

Experiment 5. In this experiment, we verified if the DAM was able to recall the image associated to the signal voice used as input pattern, even if the signal voice is altered by Gaussian noise (GN). To do this, each signal voice previously recorder was contaminated with Gaussian noise altering from 2% until 90% of the information. 89 new samples were generated from each signal voice already recorder. This new set of signal voices was composed for 440 samples, some examples are shown in Fig. 7. In average, the accuracy of the proposal using this set of signal voices was of 74%.

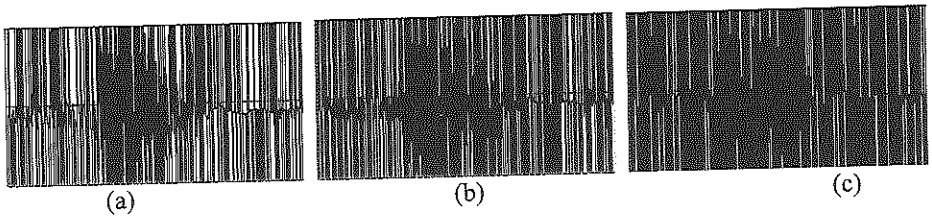

Fig. 6. Signal voice of the backcock instrument contaminated with mixed noise.

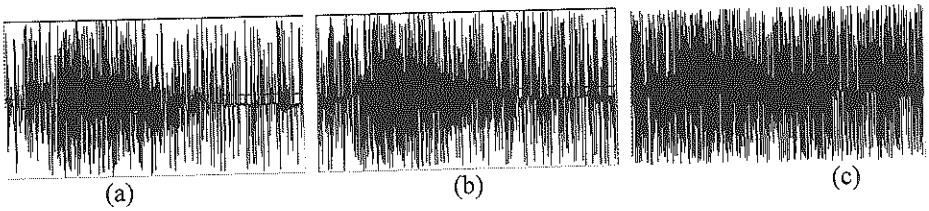

Fig. 7. (a-c) Signal voice of the rat tooth instrument contaminated with Gaussian noise.

Experiment 6. In this experiment, we verified if the DAM was able to recall the image associated to the signal voice used as input pattern, even if the signal voice is recorded at different tempo. To do this, each signal voice previously recorder was recorded 10 times. Ten new deformed samples (DEF) were recorded from each signal voice already recorder. This new set of signal voices was composed of 200 samples, some examples are shown in Fig. 8. In average, the accuracy of the proposal using this set of signal voices was of 32%.

Experiment 7. Despite of the low accuracy obtained, the results are encouraging because in more than the 50% of the samples, used on each experiment, a human is unable to perceive the name of the instrument. However, in order to increase the accuracy of the proposal we decided to apply the technique described in [10] for face recognition to the voice recognition problem. In [10], the authors suggest to compute a simplified version of the DAM model by using a random selection of stimulating points. For the details, refer to [10].

We tested again the accuracy of the proposal by using 1, then 2, them 3 until 50 stimulating points. In Fig. 9 it is shown the results obtained. As you can appreciate, the accuracy increases when the number of stimulating points increases. When we used more than 20 stimulating points the accuracy of the proposal was almost of 100% for the samples altered by additive, subtractive, mixed and Gaussian noise. Also we can

appreciate that the accuracy slightly increases to 50% for the temporal deformed voice patterns.

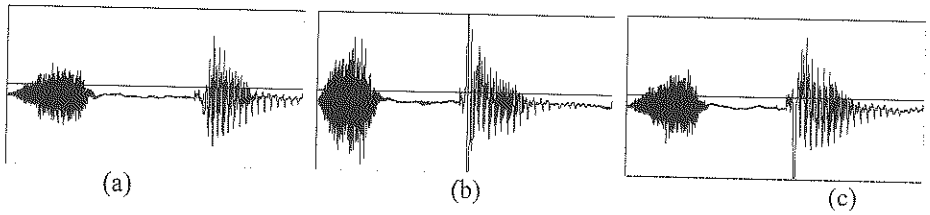

Fig. 8. (a-c) Deformed signal voice of the sponge instrument.

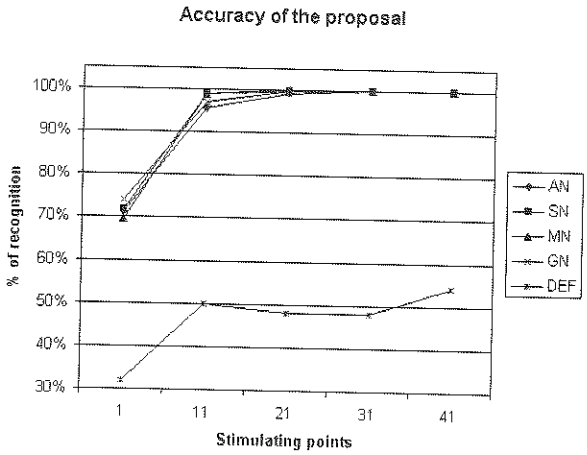

Fig. 9. Behavior of the proposal with the collections of altered signal voices using different number of stimulation points.

4 Conclusions

We have described how the VRM of NR-alpha prototype robot was implemented. We have demonstrated that associative memories, particularly dynamic associative memories can be used as powerful tools for voice recognition.

The accuracy of the proposed model was tested by using different sets of complex signal sounds and the result obtained supports the robustness of the proposals. We have studied the behavior of the model when a voice signal is contaminated with additive, subtractive, mixed, Gaussian noise and temporal deformations. The presented results are highly encouraged.

Nowadays we are applying some preprocessing techniques to increase the accuracy of the proposal when the voice signals are pronounced at different tempo by different people. Also we are integrating the VRM and AVM modules to localize an instrument using a voice command.

Acknowledgment. This work was economically supported by SIP-IPN under grants 20071438, 20082948 and CONACYT under grant 46805.

References

[1] Hopfield, J. J.: Neural networks and physical systems with emergent collective computational abilities. *Proc. of the Nat. Academy of Sciences*, 79: 2554-2558 (1982)

[2] Kohonen T.: Correlation matrix memories. IEEE Trans on Comp. 21(4):353-359 (1972)

[3] Ritter G. X., Sussner P., Diaz de Leon J. L.: Morphological associative memories. IEEE Trans Neural Networks 9(2):281–293 (1998)

[4] Sossa H., Barron R., Vazquez R. A.: New associative memories to recall real-valued patterns. In Alberto Sanfeliu, José Francisco Martínez Trinidad, Jesús Ariel Carrasco-Ochoa (Eds.) CIARP 2004. LNCS, vol. 3287, pp. 195-202, Springer (2004)

[5] Sossa H., Barrón R., Vázquez R. A.: Transforming Fundamental set of Patterns to a Canonical Form to Improve Pattern Recall. *LNAI* 3315:687-696 (2004)

[6] Vazquez R. A., Sossa H.: Associative Memories Applied to Image Categorization. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) CIARP 2006. LNCS, vol. 4225, pp. 549–558. Springer, Heidelberg (2006)

[7] Vazquez R. A., Sossa H., Garro B. A.: A New Bi-directional Associative Memory. In: Gelbukh, A., Reyes-Garcia, C.A. (eds.) MICAI 2006. LNCS (LNAI), vol. 4293, pp. 367–380. Springer, Heidelberg (2006)

[8] Sussner P., Valle M.: Gray-Scale Morphological Associative Memories. IEEE Trans. on Neural Networks, vol.17, No.3, pp.559-570 (2006)

[9] Vazquez R. A., Sossa H., Garro B. A.: *3D Object recognition based on low frequencies response and random feature selections*. In Alexander Gelbukh, A. F. Kuri Morales (Eds.): MICAI 2007. LNAI, vol. 4827, pp. 694-704 (2007)

[10] Vazquez R. A., Sossa H., Garro B. A.: *Low frequency responses and random feature selection applied to face recognition*. In Mohamed Kamel, Aurelio Campilho (Eds.): ICIAR 2007. LNCS, vol. 4633, pp. 818-830, Springer (2007)

[11] Sossa H., Vazquez R. A., Barron R.: *Reconocimiento y localización de instrumental medico usando análisis automatizado de imágenes*. Revista Mexicana de Ingeniería Biomédica, 26(2):75-85, SOMIB (2005)

[12] Vazquez R. A., Sossa, H.: A new associative memory with dynamical synapses (submitted to Neural Processing Letters, 2007)