# Speech Recognition using Based-Phone, Syllables, Triphones and Words on a Digit Corpus in Spanish Language

José Luis Oropeza Rodríguez[1]
Sergio Suárez Guerra[2]

Computing Research Center,
Av. Juan de Dios Batiz esq. Miguel Othon de Mendizabal s/n
Col. Nueva Industrial Vallejo, Mexico, D. F.
[1]joropeza@cic.ipn.mx, [2]ssuarez@cic.ipn.mx,

**Abstract.** This paper reports results obtained in Automatic Speech Recognition (ASR) using linguistic structural like phonemes, syllables, triphones and words on a digit corpus in Spanish Language. During five decades approximately, the Speech Recognition has been studied by researchers in all the world, simple recognizers have been built, yielding credible performance. But it was soon found that the techniques used in these systems were not easily extensible to more sophisticated systems. For a long time the phoneme has been used because a little amount of them exists. The words also have been utilized but the great quantity of them is a problem when an ASR system is created. The syllable and triphones are the new approaches where this field is focused. In this paper we show the results obtained when we use the continuous and discontinuous speech recognition on a digit corpus, employing the HTK (Hidden Markov Model Toolkit) for that. HTK is a toolkit for building Hidden Markov Models (HMMs). However, HTK was designed for building HMM-based speech recognition tools, in particular recognizers. Thus, much of the infrastructure support in HTK is dedicated to this task. The HTK counts with two major processing stages involved: training tools and recognizer. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools. The results obtained show the capability and difference that exists between the structural linguistic employed for the corpus purposed. We found that the triphones achieve an error rate low in comparison with another structural linguistics proposed in this experiment.

## 1 Introduction

Automatic Speech Recognition is a recent technology employed in the computer science that allows a computer to identify the words that a person speaks into a microphone or telephone. Despite several decades of research in the area, accuracy greater than 90% is only attained when the task is constrained in some way. For

© L. Sánchez, O. Espinosa (Eds.)
Control, Virtual Instrumentation and Digital Systems.
Research in Computing Science 24, 2006, pp. 213-226

large-vocabulary speech recognition of different speakers over different channels, accuracy is no greater than 87%, and processing can take hundreds of times real-time.

In English language the use of the triphones has started to try to increment the accuracy and capabilities of the systems mentioned before like Rabiner mentioned in [Rabiner 86]. In the Spanish language, the triphones like a new approach are also interesting. The dominant technology in ASR since 80s is called the HMMs. In recent years, the syllabic structure has been realized incorporating a Knowledge-based system to the training stage. In the training phase an expert system uses ten rules for syllable splitting in Spanish. It receives the energy components STTEF (Short-Term Total Energy Function) and the ERO (Energy Function of the High Frequency) parameters extracted from the speech signal [Oropeza 2005].

In the ASR phoneme-based or another linguistics structure (syllable or triphone), each word in a vocabulary list is specified in terms of its component linguistic. A search procedure is used to determine the sequence of structural linguistic with the highest likelihood. This search is constrained to only look for structural linguistic sequences that correspond to words in the vocabulary list, and the structural linguistic sequence with the highest total likelihood is identified with the word that was spoken.

Thus, the triphone must be studied to know the capability that can offer to the ASR task in the Spanish language and its grammatical structure.

## 2 Characteristics and Generalities

Speech recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols. The ASR is constitutive by: training and recognition stages.

The frequency bandwidth of a speech signal is about 16 KHz. However, most of speech energy is under 7 or 7.5 KHz (woman or man voice can change the range mentioned before) dependently. Speech bandwidth is generally reduced in recording. A speech signal is called *ortophonic* if all the spectral components over 16 KHz are discarded. A telephonic lower quality signal is obtained whenever a signal does not have energy out of the band 300-3400 Hz. Therefore, digital speech processing is usually performed by a frequency sampling ranging between 8000 samples/sec and 3200 samples/sec. These values correspond to a bandwidth of 4 KHz and 16 KHz respectively.

Voice is a static procedure that can to have a duration time between 80-200 ms. a simple but effective mathematical model of the physiological voice production process is the excitation and vocal tract model.

The excitation signal is assumed periodic with a period equal to the pitch for vowels and other voiced sounds, while for unvoiced consonants, the excitation is assumed

white noise, i.e. a random signal without dominant frequencies. The excitation signal is subject to spectral modifications while it passes through the vocal tract that has an acoustic effect equivalent to linear time invariant filtering. The model is relevant because, for each type of excitation, a phoneme (or another structural linguistic) is identified mainly by considering the shape of the vocal tract. Therefore, the vocal tract configuration can be estimated by identifying the filtering performed by the tract vocal on the excitation. Introducing the power spectrum of the signal $P_x(\omega)$, of the excitation $P_v(\omega)$ and the spectrum of the vocal tract filter $P_h(\omega)$, we have:

$$P_x(\omega) = P_v(\omega)P_h(\omega) \qquad [1]$$

The speech signal (continuous, discontinuous or isolated) is first converted to a sequence of equally spaced discrete parameter vectors. This sequence of parameter vectors is assumed to form an exact representation of the speech waveform on the basis that for the duration covered by a single vector (typically 10-25 ms) the speech waveform can be regarded as being stationary. Although it is not strictly true, it is a reasonable approximation. Typical parametric representations in common use are smoothed spectra or linear predictive coefficients plus various other representations derived from these. The training stage is involved in the extraction of the parameters mentioned. In our experiments we used the following block diagram for the isolated speech recognition (fig. 1).

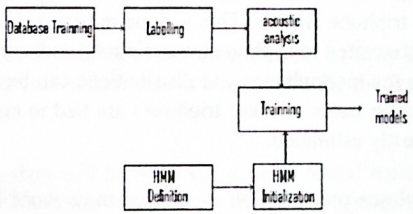


Fig. 1 Block diagram of training step for ASR for isolated words

The database employed consists of the ten digits (0-9) for the Spanish language. Many of the operations performed by HTK which involve speech data assumes that the speech is divided into segments and each segment has a name or label. The set of labels associated with the speech data will be the same as corresponding speech file but a different extension.

Often the final stage of data preparation is to parameterize the raw speech waveforms into sequences of features vectors. In this experiment, Mel Frequency Cepstral Coefficients (MFCCs), which are derived from FFT-based log spectral, will be used.

As figure 2 shows. In HTK is possible to create a set of files to define HMMs structures of each unit employed for the speech recognition. Often, the first step in

HMM training is to define a prototype model. The parameters of this model are not important; its purpose is to define the model topology.

When we employed different linguistic structural to word (phoneme, syllable or triphon), the block diagram changed and it is shown in figure 2.

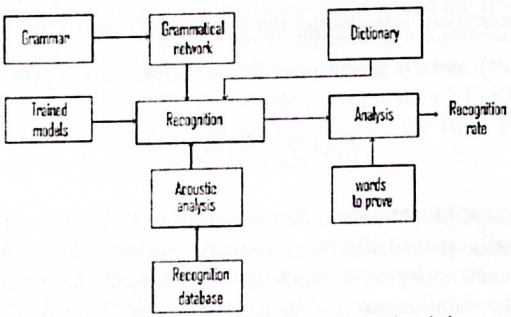


Fig. 2 Training step for ASR using structural linguistic

The new blocks added, in comparison with the block diagram showed in figure 1, represent the integration of the linguistic units mentioned above. Figure 2 shows blocks focused to the phonemes, if the linguistic unit is not that, only it will be changed for the linguistic unit selected.

Given a set of monophone HMMs, the final stage of model building is to create context-dependent triphone HMMs. This is done in two steps. Firstly, the monophone transcriptions are converted to triphone transcriptions and a set of triphone models are created by copying the monophones and distributions can be robustly estimated. Secondly, similar acoustic states of these triphones are tied to ensure that all state distributions can be robustly estimated.

The style of triphone transcription is referred to as word internal. Note that some biphones were generated as contexts at word boundaries were sometimes only include two phones. You can see the alignment data stage and the linking states showed in the figure 2.

In the recognition stage, the block diagram was the following (figure 3).

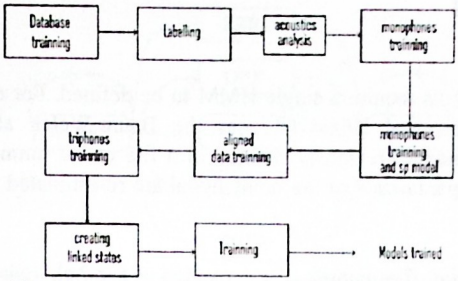


Fig. 3 Block diagram speech recognition used in HTK

As we know, the first stage or any recognizer development project is data preparation. Speech data is needed both for training and for testing. In this case all speech signals will be recorded from scratch and to do scripts are needed to prompt each sentence. In the case of the test data, these prompt scripts will also provided the reference transcriptions against which the recognizer's performance can be measured and a convenient way to create them is to use the task grammar as a random generator. In the case of the training data, the prompt scripts will be used in conjunction with a pronunciation dictionary to provide the initial phone level transcriptions needed to start the HMM training process.

## 3 Image Acquisition, processing and Segmentation

The database described in this paper counts with a set of forty three utterances for each word (0-9), in total 200 utterances. We used 100 for the training task and the rest for data test.

The training and test data will be recorded using a sound recorder provided with Microsoft Windows. HTK counts with HSLab, this is a program with graphical interface that permits a combined waveform recording and labeling a speech signal. The speech signal has the following characteristics:

- 8 bits per sample
- 11025 frequency sample
- PCM format
- High quality without noise integrated, samples providing from laboratory

HMMs implemented have the following structure:

- 6 states for isolated speech recognition
- 5 states for continuous speech recognition
- Left-right structure
- One Gaussian mixture per state

## 3.1 Trainning

Some HTK tools require a single HMM to be defined. For example, the isolated-unit re-estimation tool HRest (it uses the Baum-Welch algorithm for the re-estimation). A model previously defined in a file called hmmdef needed to be defined. Then the parameters of the input signal are re-estimated using the speech data files s1, s2, etc.

HMM definition files consist of a sequence of symbols representing the elements of a simple language (HTK uses HInit tool to initialize HMM [Baum 96]). HRest for isolated words and HERest for connected structures are the final tools in the set designed to manipulate isolated unit HMMs. Its operation is very similar to HInit (tool employed to predefine HMM). It expects the input HMM definition to have been initialized and it uses Baum-Welch re-estimation in place of Viterbi training. This involves finding the probability of being in each state at time frame using the Forward-Backward algorithm. This probability is then used to form weighted averages for the HMM parameters.

### 3.1.1 Dictionary and grammatical structures employed

Before to create a HMM, we must to create a set of word models and define a basic architecture for the recognizer. HTK provides a grammar definition language for specifying simple task grammars. It consists of a set variable definitions followed by a regular expression describing the words to recognize.

The first step in building a dictionary is to create a sorted list of the required words. For this experiment it is quite easy to create a list of required words by hand. For this experiment, the recognizer must handle digit strings only. Examples of typical inputs might be:

- Isolated words: one, zero, three, etc. (one per one not consecutively)
- Connected words: one one two four six seven; two four five zero nine eight, etc.

For the isolated words, a suitable grammar might be:

$digit = ZERO | ONE | TWO | THREE | FOUR | FIVE | SIX | SEVEN | EIGHT | NINE

(SENT-STAR ( sil <$digit> sil) SENT-END)

The grammatical network for the experiment is showed in the figure 4.

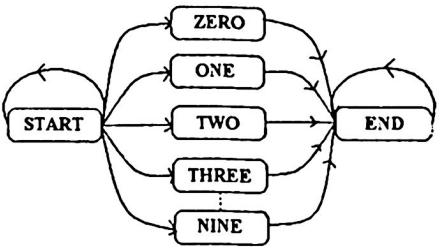


Fig. 4 Grammatical digit corpus structure for isolated speech recognition

When we employed connected digits, the grammatical network looking for such as is illustrated in the figure 5.

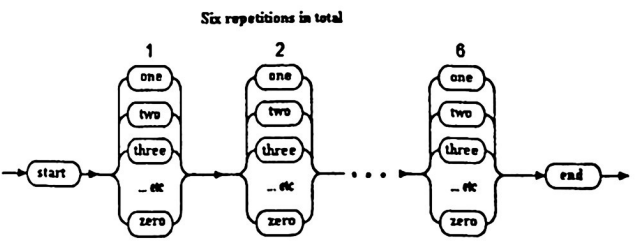


Fig. 5 Grammatical digit corpus structure for connected speech recognition

## 3.1.2 Trainning Hidden Markov Models

The training and test data can be recorded using the HTK tool HSLab. The training sentences employed in this experiment was obtained from a database of our laboratory, HSLab was used to labeling this training utterances. When HSLab is invoked, a window with a waveform display area in the upper half and a row of buttons appears. When the name of a normal file is given as argument, HSLab displays its contents. To train a set of HMMs, every file of training data must have an associated phone level transcription. In the following figure 6 we can see the labeling realized on a speech signal in the connected words experiment.

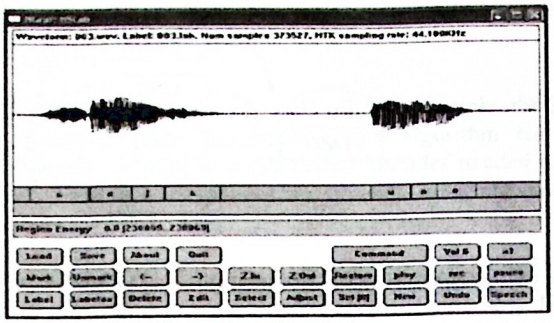

Fig. 6 Labeling realized with HSLab

### 3.1.3 Feature extraction of the speech signal

The final stage of data preparation is to parameterize the raw speech waveforms into sequences of feature vectors. HTK supports both FFT-based and LPC-based analysis. Here Mel Frequency Cepstral Coefficients (MFCCs), which are derived from FFT-based log spectra, were used. Also, we use the tool HCOPY to automatically convert its input into MFCC vectors.

We use $W_0$ as the energy component, the frame period is 10msec (HTK uses units of 100ns). The FFT used a Hamming Window and the signal had first order preemphasis applied using a coefficient of 0.97. The filter bank was constituted by 26 channels and 12 MFCC coefficients. Then, the system utilized 39 parameters calculated from the length of the parameterized static vector MFCC plus the delta coefficients plus the acceleration coefficients, to represent the speech signal in both: training and test data steps.

### 3.1.4 Definition of the Hidden Markov Models

The figure 7 shows definition of Hidden Markov Models employed during isolated word recognition. At left the HMM initialized and right the HMM re-estimated.

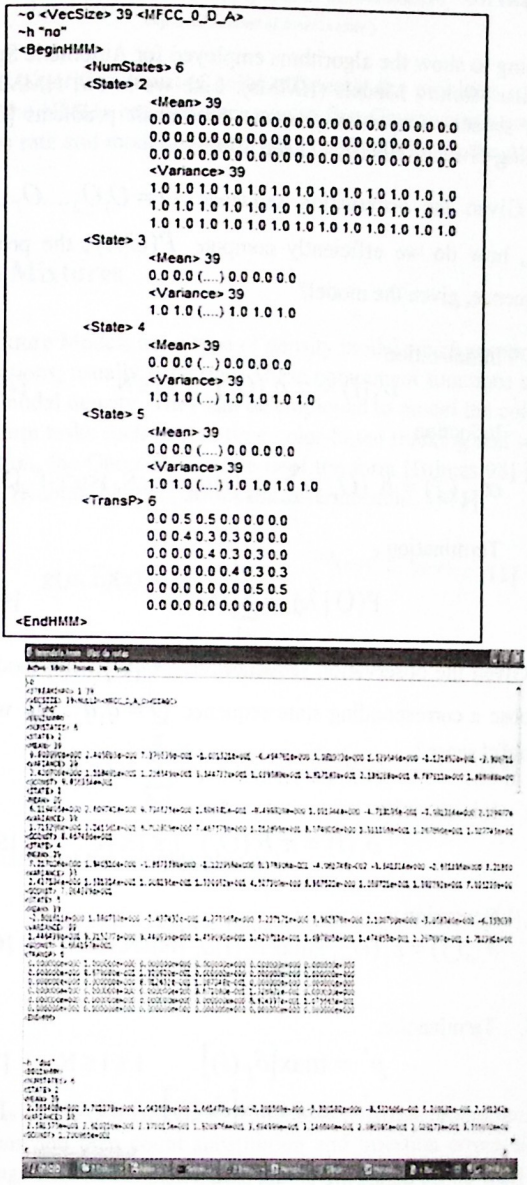


Fig 7 Definition and final results of a Hidden Markov Model training stage

For the connected word experiment, 6 words were pronounced by the speaker, it represents six digits connected. A raw of silence was interposed between each word pronounced to avoid overlap between them.

## 4 Hidden Markov Models

Now, we are going to show the algorithms employed for Automatic Speech Recognition using Hidden Markov Models (HMMs). Like we know, HMMs mathematical tool applied for speech recognition presents three basic problems [Oropeza, 2000] [Rabiner and Biing-Hwang, 1993] y [Zhang 1999]:

Problem 1. Given the observation sequence $O = O_1 O_2 .... O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O \lambda)$, the probability of the observation sequence, given the model?

1. Initialization

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N \qquad [2]$$

2. Induction

$$\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij} \quad 1<j\le N,\ 1\le t\le T-1 \qquad [3]$$

3. Termination

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad [4]$$

Problem 2. Given the observation sequence $O = O_1 O_2 .... O_T$ and the model $\lambda$, how do we choose a corresponding state sequence $Q = q_1 q_2 ... q_T$ which is optimal in some meaningful sense?

*1.* Inicialización

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N \qquad [5]$$

2. Recursión

$$\delta_{t+1}(j) = b_j(O_{t+1}) \left[ \max_{1\le i\le N} \delta_t(i) a_{ij} \right] \quad 1\le j\le N, 1\le t\le T-1 \qquad [6]$$

3. Terminación

$$p^* = \max[\delta_T(i)] \quad 1 \le i \le N \qquad [7]$$

$$q^* = \arg\max[\delta_T(i)] \quad 1 \le i \le N \qquad [8]$$

Problem 3. How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O \lambda)$?

$$a_{ij} = \frac{\text{exp ected number of times from state } s_i \text{ to state } s_j}{\text{exp ected number of transition s from state } s_i} \qquad [9]$$

$$b_{jk} = \frac{\text{expected number number of times in } s_j \text{ and observating } v_k}{\text{expected number of times in state } j} \qquad [10]$$

Then, HMMs algorithms must to solve efficiently the problems mentioned above. For each state, the HMMs can use since one to five Gaussian mixtures both to reach high recognition rate and modeling vocal tract configuration in the Automatic Speech Recognition.

## 5 Gaussian Mixtures

Gaussian Mixture Models are a type of density model which comprise a number of component functions, usually Gaussian. These component functions are combined to provide a multimodal density. They can be employed to model the colors of an object in order to perform tasks such as real-time color-based tracking and segmentation. In speech recognition, the Gaussian mixture is of the form [Bilmes 98] [Resch, 2001a], [Resch, 2001b], [Kamakshi et al., 2002] and [Mermelstein, 1975].

:

$$g(\mu,\Sigma)(x) = \frac{1}{\sqrt{2\pi^d}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \qquad [11]$$

Equation 12 shows a set of Gaussian mixtures:

$$gm(x) = \sum_{k=1}^{K} w_k * g(\mu_k, \Sigma_k)(x) \qquad [12]$$

In 12, the summarize of the weights give us

$$\sum_{i=1}^{K} w_i = 1 \quad \forall \quad i \in \{1,.....,K\} \ : w_i \geq 0 \qquad [13]$$

## 6 Results

HTK counts with the tool HResults, which uses dynamic programming to align the two transcriptions and then count substitution and insertion errors. Figure 8 shows that the percentage obtained was 95% for isolated words.

```
hresults -A -T 1 -e ??? sil lista_hmm.txt rec.mlf
====================== HTK Results Analysis =======================
  Date: Wed Jul 05 17:53:48 2006
  Ref : 
  Rec : rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=95.00 [H=95, S=5, N=100]
WORD: %Corr=95.00, Acc=95.00 [H=95, D=0, S=5, I=0, N=100]
===================================================================
```

Fig. 8 Digit isolated results using words

And for the connected words the results are shown in figure 9.

```
------------------------ Overall Results ------------------------
SENT: %Correct=60.87 [H=14, S=9, N=23]
WORD: %Corr=92.75, Acc=92.03 [H=128, D=1, S=9, I=1, N=138]
-----------------------------------------------------------------
```

Fig. 9 Digit Connected results using words, format HTK

The line starting with SENT; indicates that of the 23 test sentence utterances (with 6 digits each one, a total of 138), 14 (60.87%) were correctly recognized (discontinuous speech recognition). The following line starting with WORD: gives the world level statistics and indicates that the 138 words in total, 92.75% were recognized correctly. There were 1 deletion error (D), 9 substitution error (S) and 1 insertion error (I).

The use of phone-based (monophones) with the corpus data used, we obtained the following results:

```
hresults -A -T 1 -e ??? sil lista_hmm.txt rec.mlf
====================== HTK Results Analysis =======================
  Date: Wed Jul 20 11:10:58 2006
  Ref :
  Rec : rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=73.91 [H=17, S=6, N=23]
WORD: %Corr=97.03, Acc=97.03 [H=597, D=8, S=9, I=17, N=573]
===================================================================
```

Fig. 10 Connected digits application using phonemes as linguistic structure

Context-dependent triphones can be made by simple cloning monophones and then re-estimating using triphone transcriptions. HTK counts with HLEd tool to do that. Figure 11 shows the results:

```
hresults -A -T 1 -e ??? sil lista_hmm.txt rec.mlf
====================== HTK Results Analysis =======================
  Date: Wed Jul 17 12:13:48 2006
  Ref :
  Rec : rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=60.87 [H=14, S=9, N=23]
WORD: %Corr=96.76, Acc=93.52 [H=597, D=10, S=10, I=20, N=617]
===================================================================
```

Fig. 11 Connected digits application using triphones linguistic structure

From [Oropeza 2005] a 98% was reported using syllables. The results show the comparison between different linguistic structural in the corpus digit for Spanish language.

## 7 Conclusions and future works

The main feature that is used to characterize the complexity of speech recognition is weather the speech is connected or is spoken one word at time. In connected speech, it is difficult to determine where one word ends and another begins, and the

characteristics acoustic patterns of words exhibit much greater variability depending on the context. Isolated word recognition systems do not have these problems since words are separated by pauses.

HTK is a set of tools that permit to create an Automatic Speech Recognition (ASR) secure and efficient. It permits to use words, phoneme or triphone based on monophoneme definition as linguistic structures. Also, it contains with a graphical tool that permits to create labels for each word. The rest of tools are accessed via DOS shell prompt.

In this paper we show the results obtained in both Isolated and connected speech recognition systems using HTK for a digit corpus in Spanish language. At the same time, we employed three structural linguistic (words, phonemes and triphone) and we compared them with the results obtained when the syllables were used for the same corpus without employ HTK.

The differences found it between the accuracy recognition using different linguist structures is a consequence of so much things. Firstly, the labeling proposed here was effected using HSLab tool, which does not permit an automatic splitting of the word. Then, human errors are possible to appear and it affects a good response of the system. It explains because the error rate is lower than compared with the experiment using syllable structures, where a Knowledge based system was used for the splitting the word.

So, the result obtained shows an important accuracy to consider an adequate real implementation of them. The most important result extracted from the experiment is the integration of triphone structure and its repercussions for future works. The possibilities to use other linguistic structures in ASR increment the faculty to create real systems using new approaches. Also, incorporating the analysis previously realized in automatic speech recognition with noise can be another way to consolidate the study of linguistic structures.

# References

1. Pope A. R. "Model-Based Object Recognition. A survey of recent research", University of British Columbia, Vancouver, Canada, Technical Report 94-04, January 1994.

2. Singh S., Markou M., Haddon J., "Detection of new image objects in video sequences using neural networks", Proc. SPIE Vol. 3962, p. 204-213, Applications of Artificial Neural Networks in Image Processing V, Nasser M. Nasrabadi; Aggelos K. Katssagelos; Eds., 2000.

3. Fay R., Kaufmann U., Schwenker F., Palm G., "Learning objects recognition in a neurobotic system". In: Horst-Michael Grob, Klaus Debes, Hans-Joachim Böhme (Eds.) 3rd Workshop on Self Organization of AdaptiVE Behavior (SOAVE 2004). Fortschritt-Berichte VDI, Reihe 10 Informatik / Kummunikation, Nr. 743, pp. 198-209, VDI Verlag, Düsseldorf, 2004.

4. Wang W., Zhang A. and Song Y., "Identification of objects from image regions", IEEE International Conference on Multimedia and Exp (ICME 2003), Baltimore, July 6-9, 2003.
5. Vision Components, Technical Documentation VCM40, VCM50, Vision Components GmbH Ettlingen, Germany, Oktober 2003.
6. Abhijit S. Pandya, Robert B. Macy, Pattern Recognition with Neural Networks in C++, edit. CRC PRESS & IEEE PRESS. A CRC Cook Published in Cooperation with IEEE PRESS, 1996.
7. Felzenswalb P. and Huttenlocher D., "Efficiently computing a good segmentation". In IEEE Conference on Multimedia and Expo (ICME 2003), Baltimore, July 6-9, 2003.
8. Bishop C. M., Neural networks for Pattern Recognition, Oxford University Press, 1995.
9. Nicolas Amezquita Gomez and Rene Alquezar., "Object Recognition in Indoor Sequences by Classifying Image Segmetnation Regions Using Neural Networks, 10th Iberoamerican Congreso on Pattern Reognition, CIARP 2005, Havana, Cuba, November 2005, Proceedings.