HAL open science

Comparison between two spatio-temporal organization maps for speech recognition

Zouhour Neji Ben Salem, Laurent Bougrain, Frédéric Alexandre

To cite this version:

Zouhour Neji Ben Salem, Laurent Bougrain, Frédéric Alexandre. Comparison between two spatio-temporal organization maps for speech recognition. Second IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition - ANNPR 2006, Aug 2006, University of Ulm/Germany, pp.11-20, 10.1007/11829898_2 . inria-00104156

HAL Id: inria-00104156

https://inria.hal.science/inria-00104156v1

Submitted on 5 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparison Between Two Spatio-Temporal Organization Maps for Speech Recognition

Zouhour Neji Ben Salem[1], Laurent Bougrain[2], Frédéric Alexandre[2]

[1] AI Unit, CRISTAL Laboratory, National School of Computer Sciences, Manouba Campus, Tunisia
zouhourbensalem@hotmail.com

[2] Cortex Team, LORIA Laboratory, Nancy, France
{Laurent.bougrain,Frederic.Alexandre}@loria.fr

Abstract. In this paper, we compare two models biologically inspired and gathering spatio-temporal data coding, representation and processing. These models are based on Self-Organizing Map (SOM) yielding to a Spatio-Temporel Organization Map (STOM). More precisely, the map is trained using two different spatio-temporal algorithms taking their roots in biological researches: The ST-Kohonen and the Time-Organized Map (TOM). These algorithms use two kinds of spatio-temporal data coding. The first one is based on the domain of complex numbers, while the second is based on the ISI (Inter Spike Interval). STOM is experimented in the field of speech recognition in order to evaluate its performance for such time variable application and to prove that biological models are capable of giving good results as stochastic and hybrid ones.

1 Introduction

Spatio-temporal classical connectionist networks comprise an important class of neural networks that can deal with patterns distributed both in time and space. In the case of Automatic Speech Recognition (ASR), this class of models has been shown to yield good performance (sometimes better than Hidden Markov Models) on short isolated speech units. By their recurrent aspect and their implicit or explicit temporal memory they can perform some kind of integration over time. Yet till now spatio-temporal biologically inspired models are the most less used for ASR. This class of models is very relevant to be exanimate because the advance realized in neurophysiologic researches have yield to the emergence of neuromimetic models especially for explicitly processing of temporal information and we know that in ASR, there is a time dimension or a sequential dimension which is highly variable and difficult to handle directly in ANNs. These models present an alternative approach to ASR which might in the long term help to overcome restrictions of current speech recognition technology with regard to noise tolerance or speaker independence. Some of these biological models have demonstrated good performance for ASR, we can cite the work of Béroule [4]

concerning dynamic propagation or Durand [8] for super units map. In this paper we present STOM map extending the SOM map from the processing of purely spatial signals to the processing of spatio-temporal signals using biologically inspired coding approaches and algorithms. STOM model represents the time dimension, depending on the parameters used, either as the level of weights or map. The use of SOM in this paper is coming from the conviction of universal auto-organization concept covering the major part of human brain [19]. This concept could be able to explain all experimental findings concerning the plasticity of cortical topographies. A possible candidate for universal self-organization is the SOM map. STOM map will be trained using two different spatio-temporal algorithms. The first one is the ST-Kohonen algorithm proposed by Mozayyani [10]. It is an algorithm using spatio-temporal inputs which are represented by a spatio-temporal coding approach proposed by Vaucher [15]. The latter uses the domain of complex numbers to encode inputs. The choice of this domain derives from the fact that it is the only domain supplying two degree of freedom allowing to represent the two correlated data. The second algorithm is TOM (Time Organized Map) [19]. Its main additional idea comparing to SOM is the functionally reasonable transfer of temporal signal distances into spatial signals distances in topographic neural representations. This is achieved by neural dynamics of propagating waves. The inputs processed by this algorithm are encoded using an approach based on the ISI (Inter Spike Interval)[6],[14]. Each input is presented as a pair containing the spatial information of the stimuli, and the temporal distance that separate it from its successor.

STOM is experimented in the domain of speech recognition of isolated digits. Section two of this paper presents the encoding approach used for STOM inputs, while section three is devoted to describe ST-Kohonen training model and propose also an extension of TOM for two dimensions. In section four, we show the application of the two encoding approaches to the analysis of acoustic signal and we compare experimental results between the two models obtained in the domain of speech recognition. Section five concludes the paper and presents some perspectives for the model.

2 The Spatio-temporal encoding approaches

2.1 The Complex approach

The spatio-temporal technique takes its roots from the work undertaken by neurobiologists to model passive electric properties of the dendrites trees [16], in particular Rall's work [11]. It is based on a particular modeling of the Post-Synaptic Potentials (PSP) and of their mix. According to the formalization made by Agmon-Snir [1], which characterizes one PSP(t) by its moments of order k, only the first two moment are kept, each one of them being respectively associated to the norm and phase of a complex number. The use of complex domain is justified by the fact that it is the only domain offering numbers having two degree of freedom. This property allows encod-

ing the two correlated data at the same time. The Spatio-Temporal coding technique (**Fig.1**.) is introduced for the aim to provide the classical artificial neurons the capacity of processing sequences in asynchronous manner, leading to the emergence of STANN [15] (Spatio-Temporal Artificial Neural Network). It consists on adding a delay; at the level of input; to introduce temporal information. It improves the work of 'Integrate and Fire' neuron [7], [17] using the field of complex numbers. The spatio-temporal coding is done as following: Each input produces a train of impulses, for each impulse or spike I characterized by its amplitude μ_I and the temporal delay d_I which separates the current instant of time from the time at which the impulse has been occurred, a complex number z_I is assigned. z_I Contains μ_I as its module and φ_I as its phase as following:

$$\begin{aligned} z_I &= \mu_I e^{i\rho_I} \\ \rho_I &= \arctan(\mu_t d_I) \end{aligned} \quad \textbf{(1)}$$

μ_t is a constant inverse to time. To incorporate the attenuation of amplitude through time, the amplitude of the input could be calculated using the following formula:

$$\mu_I(t+\Delta t) = \mu_I(t) e^{-\mu_s \Delta t} \quad \textbf{(2)}$$

According to Baig [3], the couple (μ_s , μ_t) must satisfy the following condition in order to obtain a good complex encoding.

$$\mu_s = \mu_t = 1/T_w \quad \textbf{(3)}$$

T_W represents the temporal window used to encode all the coming inputs.

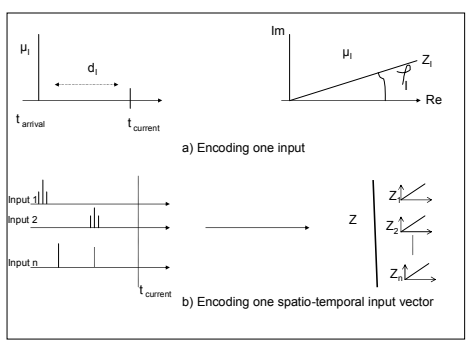

Fig. 1. The Spatio-Temporal Complex coding approach

2.2 The Temporal coding approach

Recent neurobiological findings experimented by Spengler et al [12] have demonstrated the importance of temporal interval between stimuli. In these experiments, tactile stimuli with different relations in space and time were applied to monkey. After months of training, the primary somatosensory cortex was mapped, showing that temporally separated stimuli (with an ISI of about 300ms) were segregated within cortical topography, while stimuli having an ISI under 100ms are integrated. Thus the dynamic of incoming stimuli reflects the stimuli's relatedness with regard to their functional meaning. Stimulus dynamic is therefore important for the learning of topography. Accordingly, many researchers think that the design of incoming flux of n stimuli S_i could be represented as sequence of couple containing the stimulus and ISI_i interval expressing the temporal proximity of consecutive stimuli.

$$(S_i, ISI_i) \qquad i = 1,..,n \tag{4}$$

3 Spatio-temporal training

3.1 The ST-Kohonen

ST-Kohonen map [10] is a SOM having complex neurons. These latter have input and weighting vectors defined in the complex domain as described in the above section. The ST-Kohonen algorithm works in the same manner as classical Kohonen one, however, the winner is chosen according to the Hermitien distance instead of the Euclidian one:

$$\delta(X, W_i) = \|X - W_i\| = \sqrt{{}^t(X - W_i)\overline{(X - W_i)}} \tag{5}$$

X designes the map input and W_i is the weighting vector of the neuron i. Both the input and the weight vectors are defined in the complex domain C^p .The adaptation rule for ST-Kohonen is the same as the one presented in Kohonen, yet we manage complex vectors instead of real. With ST-Kohonen we made a spatio-temporal classification which is very relevant in ASR because it allows classifying speech according to its feature extractors and their temporal occurrence or sequences. However, topology obtained by the map is spatial because weights are updated according to the distance of neurons in the map and not in the input space.

3.2 The Time Organized Map Algorithm (TOM)

The TOM algorithm was presented by Wiemer [19] for a better understanding of the self-organization and the geometric structure of cortical signal representations. TOM was proposed for one dimension SOM map, reader can refer to [19] for more details about the algorithm. In the present section, we present an extension of TOM for two dimensions.

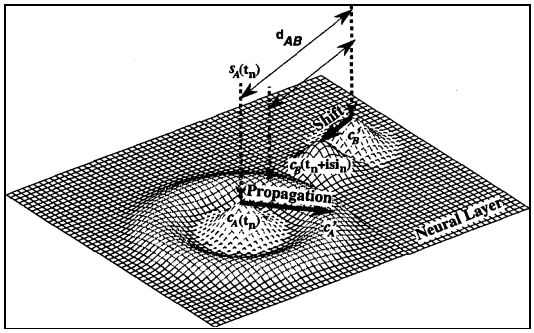

Fig. 2. Wave propagation and interaction in a two dimension TOM map

High-dimensionally coded stimuli are applied to TOM two dimensions (**Fig. 2**.) at discrete times. They are described by a sequence (S_n, ISI_n).The layer is composed of $N_1^c.N_2^c$ neurons ('c' for cortical). At time t_n, a stimulus $s_n = s_A(t_n)$ is presented to the map resulting in a feedforward activation obtained by multiplying the connection matrix with the stimulus: The activity $c_A(t_n)$ builds up from the current feedforward activation $c^{ff}(t_n)$, and the activity state of the map, that has evolved out of the earlier activity $c(t_{n-1})$ from time t_{n-1} to t_n. The evolution of map activity between two stimulations is of wave-like type; excitation propagates into its surround. This dynamic may result from local interactions [2],[13],[18] e.g. between excitatory and inhibitory neurons. The dynamic is fundamental in the sense of a general principal of locality. In other words, 'effects propagate from point to neighboring point. In case of a monomodal and rotation symmetric of map activity, an 'elementary wave' propagates as is shown by **Fig. 2**. Assuming linear superposition, the dynamic corresponding to general activity patterns can be reduced to the superposition of elementary waves. As an analogue, one may think of water waves that a raindrop generates when it falls down. During the isi_n, and assuming a constant propagation speed, the wave has propagated along a certain distance reaching neurons having position that verify the following formula:

$$c'_A(n) = k_{ff}(n) \pm v.ISI_n \quad \textbf{(6)}$$

Where $k_{ff}(n)$ is the neuron activated by the stimulus $s_A(t_n)$. The stimulus $s_B(t_n + isi_n)$ (applied at that time), generates also a feedforwrad activation resulting in the activity c'_B. This activity will propagates in the map yielding to an interaction zone between the two waves. Considering this interaction, the activity of the map must be shifted to towards this zone because it consists of neurons that are excited by the two stimuli. In a neural field model, and also in biological model [20] the length of the shift depends on the distances between the wave front of the stimulus $s_A(t_n)$, which is $c'_A(n)$, and $k_{ff}(n + isi_n)$, the neuron activated by $s_B(t_n + isi_n)$. This distance could be determined using the shortest path linking the two neurons $k_{ff}(n)$ and $k_{ff}(n + isi_n)$. Thus the two dimensional interaction could be reduced to one-dimensional interaction along this shortest path. This reduction allows us to apply one dimensional TOM algorithm along this path. The neurons that constitute this shortest path must have a coordinate (i_k, j_k) verifying the following formula:

$$j_k = round(A.i_k + b) \text{ and} \quad \textbf{(7)}$$

$$[\, i_{k_{ff}(n)} < i_k < i_{k_{ff}(n+isi_n)} \text{ or } i_{k_{ff}(n+isi_n)} < i_k < i_{k_{ff}(n)} \,] \text{ and}$$

$$[\, j_{k_{ff}(n)} < j_k < j_{k_{ff}(n+isi_n)} \text{ or } j_{k_{ff}(n+isi_n)} < j_k < j_{k_{ff}(n)} \,]$$

Where:

$$A = \frac{j_{n+isi_n} - j_n}{i_{n+isi_n} - i_n}, \text{ et } b = j_n - A.i_n \quad \textbf{(8)}$$

(i_n, j_n) and $(i_{n+isi_n}, j_{n+isi_n})$ are the coordinates of the neuron $k_{ff}(n)$ and $k_{ff}(n + isi_n)$. Once the path is determined, the different steps proposed for TOM one dimension algorithm could be applied. These steps are as follows:

Determination of the shift toward the interaction zone due to the propagation of waves

The interaction shift is calculated according to the distance that separates the current winner from the border of the propagation waves introduced by the earlier stimulus

$$\Delta_{\text{int}}(n) = f(c'_A(n) - k_{ff}(n)) \quad \textbf{(9)}$$

The interaction function f expresses the distance dependency of the shift. Wiermer proposed the following function

$$f(k) = K.\tanh(\frac{k}{K})\exp(-\frac{k^2}{2\sigma_K{}^2}) \quad \textbf{(10)}$$

K represents the strength of the interaction. The form of the interaction function is shown by (**Fig. 3**.) for different values of K and σ_K .

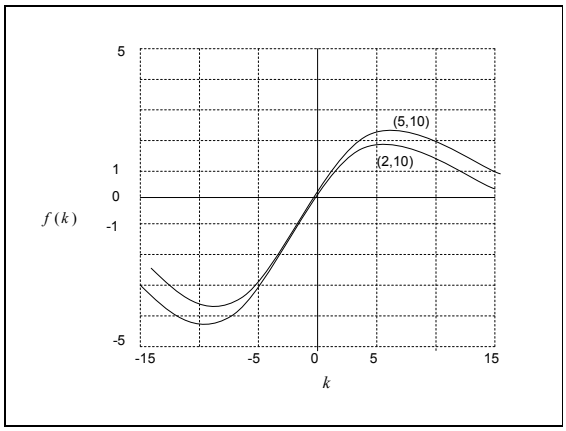

Fig. 3. The non linear interaction function for (K, σ_k) = (5,10) and (2,10)

Shift due to noise

Noise in the network is expressed by a noise term that is randomly drawn from a normal distribution with zero mean and standard deviation

$$\Delta_{noise} \rightarrow N(0, \sigma_{noise}(n)) \qquad \textbf{(11)}$$

The standard deviation decreases monotonically in time from its initial value σ_0 to its final value σ_f and remains constant to its final value for the remaining of the steps learning.

$$\sigma_{noise}(n) = \sigma_0 \left(\frac{\sigma_f}{\sigma_0}\right)^{n/n'_f} \text{ si } n \leq n'_f \qquad \textbf{(12)}$$

New winner position

The winner $k_{learn}(n)$ is determined as the integer closest to the position of the maximal feedforward activation shifted by interaction and noise.

$$k_{learn}(n) = round(k_{ff} + \Delta_{\text{int}} + \Delta_{noise}) \qquad \textbf{(13)}$$

New winner adaptation weights

Only winner weights are adapted by learning step. They are shifted toward the presented stimulus.

$$\Delta W_{k_{learn}(n)}(n) = \alpha.(S_n - W_{k_{learn}(n)}(n)) \quad \textbf{(14)}$$

α is the learning rate, it is constant during all the training.

4 Experimentation

4.1 The Data coding and model

The map is trained and tested with isolated words of TI-Digit database. The number of speakers contributing to the application is 110 distributed as 55 men and 55 women pronouncing the eleven vocal digits (0, 1, ..,9 and 'oh'). Each speaker pronounces one digit twice: one occurrence is used for the training and the other for the test. We use the Mel Frequency Cesptral Coefficients (MFCC) vectors [5] as feature extractor; it is composed of 12 coefficients. The speech signal were collected in quiet environment and digitized at 16 KHz. The window size for calculating each MFCC contains 512 points. Windows are overlapping and are separated with 10ms (160 points) between two consecutive MFCC. To permit the spatio-temporal coding with complex numbers, we suppose that all coefficients for one MFCC vector have the same time of occurrence. Thus, complex MFCC vectors will have the following form:

$$MFCC_{real} = \begin{Bmatrix} c_1 \\ . \\ .. \\ c_{12} \end{Bmatrix} \Rightarrow MFCC_{complex} = \begin{Bmatrix} c_1 e^{i\rho_1} \\ . \\ .. \\ c_{12} e^{i\rho_1} \end{Bmatrix} \quad \textbf{(15)}$$

The phase is calculated using the following formula:

$$\rho_i = \arctan(1 - 0.01(i-1)) \quad \textbf{(16)}$$

The temporal encoding approach is straight forward. In fact each input is couple of the following form:

$$(S_n, isi_n) = \left(\begin{Bmatrix} c_1 \\ ... \\ c_{12} \end{Bmatrix}, 10ms \right) \quad \textbf{(17)}$$

4.2 Experimental results

STOM tested by either ST-Kohonen algorithm or TOM is composed of 30 (5x6) neurons connected to a second map containing 11 neurons corresponding to the 11 digits to be recognized. The recognition rate obtained for ST-Kohonen is evaluated at 95,9%, while it reaches 97,5% for the TOM model. Thus the spatio-temporal interaction taken into account in the TOM model performs better than the spatio-temporal classification made by ST-Kohonen. This interaction can capture more the spatio-temporal invariance of the speech signal. However, the main drawback of TOM is the time of convergence. In fact, the algorithm has needed 10^9 loops to converge. This slowness could be interpreted by the fact that TOM update only the winner by each loop yielding to an increasing of time for the map to be organized. ST-Kohonen presents better result compared to other time-based neural networks [9]; this fact is explained by the use of coding approach gathering spatio-temporal data as inputs allowing a spatio-temporal classification.

5 Conclusion

The Spatio-temporal Organization Maps presented in this paper are an extension of SOM map to spatio-temporal domain. This extension is based on biologically inspired approaches related to coding and processing data. STOM is trained using two spatio-temporal processing algorithms. The results obtained by applying STOM in speech recognition are good and both algorithms are qualified for processing the fine spatio-temporal structure of speech signal. This might provide an insight into speech recognition using biologically models and could in the long run overcome the limitation of HMM-based speech technology or hybrid models regarding for example noise. Further research could be focused on testing the presented models using corpus collected in real world. It could be also possible to combine biologically model for speech perception, a model of human cochlea for example, with STOM because we think that it can be a fruitful area that needs to be explored.

References

1. Agmon-Snir, H. and Segev, I. :Signal Delay and iInput Synchronization in Passive Dendritic Structures. Journal of Neurophysiology, 70(5) (1973).
2. Amari,S.:Topographic organization of nerve fields. Bull Math Biology. 42 (1980) 339-364.
3. Baig, A. B. :Une approche methodologique de l'utilisation des STAN applique a la reconnaissance visuelle de la parole. PhD thesis, Suplec, campus universitaire de rennes (2000).
4. Béroulle, D. :Un modèle de mémoire adaptative, dynamique et associative, pour le traitement automatique de la parole. Thèse de l'université de Paris 11 (1985).
5. Calliope. La parole et son traitement automatique. *Masson*, Paris, Milan , Barcelone, (1989).
6. Cariani, P.:As If Time Really Mattered: Temporal Strategies for Neural Coding of Sensory Information'. CC-AI. 12(1-2) (1995).

7. Casti, A. R. R. Omurtag, A., Sornborger, A., aplan, E., Knight, B., Victor, J., Sirovich, L.:A population study of integrate and fire or burst neur*on*, Neural Computation**,** Volume 14 Issue 5 (2002).
8. Durand, S.:Learning speech as acoustic sequences with the unsupervised model TOM. In NEURAP, 8th international conference on neural networks and their applications. Marseille french 1995.
9. Laurence, S. Tsoi, A. C. Back, A. D. : The gamma MLP for speech phoneme recognition. Advances in Neural Information Processing System, 8 (1996) 785-791.
10. Mozayyani, N., Alanou, V., Derfus, J. and Vaucher, G.:A spatio-temporal data coding applied to kohonen maps. Inter conf on Artificial Neural Natwork (1995) 75-79.
11. Rall, W.:Core conductor theory and cable properties. In handbookof physiology:the nervous system, Americain physiology society (1977).
12. Spengler, F. Hilger, T. Wang, X. and Merzenich, M:Learning induced formation of cortical populations involved in tactile object recognition. Social Neurosciences. 22. (1999) 105-110.
13. Szentagothai, J.:The module concept in cerebral cortex architecture. Brain research (1995) 475-496.
14. Thorpe, S.:Spiking arrival times: Ahighly efficient coding scheme for neural networks. In Parallel Processing in Neural System, Elseiver Press (1990).
15. Vaucher, G. :A la recherche d'une algerbre neuronale spatio-temporal. P.hD thesis. Nancy University (1996).
16. Vaucher, G.:A Complex-Valued spiking machine, ICANN (2003) 967-976.
17. Vinh ho, H. :Un réseau de neurones a décharge pour la reconnaissance des processus spatio-temporels..PhD thesis, Genie Electric Department, Monreal University (1992).
18. Wiemer, J. Spengler, F. Joublin, F. and Wacquant, S.:Learning cortical topography from spatiotemporal stimuli. Biology cybernetic. 82 (2000) 173-187.
19. Wiemer, J. C.:The Time-Organized Map (TOM) algorithm: extending the self-organizing map (SOM) to spatiotemporal signals. Neural Networks, 15 (2003).
20. Wilson, H. Cowan, J.:A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Biology Cybernetic. 13 (1973) 55-80.