Genes : un accélérateur systolique pour les réseaux de neurones

Autor(en): **Lehmann, Chrisitan**

Objekttyp: **Article**

Zeitschrift: **Bulletin des Schweizerischen Elektrotechnischen Vereins, des Verbandes Schweizerischer Elektrizitätsunternehmen = Bulletin de l'Association Suisse des Electriciens, de l'Association des Entreprises électriques suisses**

Band (Jahr): **83 (1992)**

Heft 5

PDF erstellt am: **22.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-902803

Genes – Un accélérateur systolique pour les réseaux de neurones

Christian Lehmann

La définition théorique d'une architecture d'ordinateur systolique mimant la structure des réseaux de neurones biologiques a conduit à la conception d'une famille de circuits intégrés à large échelle (VLSI). Une carte prototype contenant 256 de ces circuits est une première étape du projet Mantra dans lequel plusieurs cartes «neuronales» travailleront en parallèle.

Die theoretische Architektur-Definition eines systolischen Rechners, welcher die Netzwerkstruktur biologischer Neuronen nachahmt, hat zur Konzeption einer VLSI-Chips-Familie geführt. Mit der Herstellung eines Prototyp-Prints mit 256 VLSI-Bausteinen ist eine erste Etappe des Mantra-Projekts erreicht, in welchem mehrere «neuronale» Prints parallel zusammenarbeiten werden.

Adresse des Autors

Christian Lehmann, ing. dipl. EPFL, Groupe Neurone, LAMI-EPFL, 1015 Lausanne.

De nombreuses solutions technologiques ont été examinées afin d'implanter sur silicium des réseaux de neurones adaptatifs [1]. Des solutions systoliques mono-dimensionnelles ont déjà été étudiées par plusieurs chercheurs [2;3]. Dans ce qui suit, le développement du calcul du produit matrice-vecteur, principale opération du calcul neuronal, réalisé sur la matrice bi-dimensionnelle Genes, permettra au lecteur de se rendre compte des caractéristiques essentielles de cette technique. Ces caractéristiques, et notamment le parallélisme élevé qui en résulte, permettent d'envisager des performances surpassant celles des super-calculateurs scientifiques très coûteux. Plus loin, une méthode de modification des coefficients de la matrice synaptique, permettant l'apprentissage, sera brièvement expliquée. En effet, l'apport principal de la famille Genes est de proposer un système suffisamment général pour s'adapter à de nombreux algorithmes neuro-mimétiques.

Le calcul du produit matrice-vecteur

Les structures systoliques sont particulièrement adaptées aux calculs matriciels. Cet aspect a été étudié fondamentalement par plusieurs auteurs dont S.Y. Kung [4] et P. Quinton [5]. Dans le cas qui nous intéresse, les réseaux de neurones artificiels, l'opération à effectuer lors de la phase dite d'application est la suivante:

$$y_i = \sigma\left(\sum_{j=1}^{N} W_{ij}x_j\right) \qquad (1)$$

où i et j sont les indices, N le rang de la matrice «synaptique» $\boldsymbol{W}$, x_j les stimuli de l'espace d'entrée, y_i l'activité du neurone i et $\varsigma(\ )$ est une fonction nonlinéaire.

L'expression à déterminer correspond en fait, mis à part l'évaluation de la fonction $\sigma(\ )$, à calculer une succession de sommes pondérées.

$$p_i = W_{i1}x_1 + W_{i2}x_2 + \ldots + W_{iN}x_N \qquad (2)$$

Cette opération sera répétée pour chaque potentiel de neurone p_i, c'est-à-dire, pour chaque ligne i de la matrice. On peut écrire l'expression 2 sous sa forme récurrente en introduisant une variable supplémentaire, k:

$$\begin{aligned} p^{(0)}{}_i &= 0 \\ p^{(k)}{}_i &= p^{(k-1)}{}_i + W_{ik}x_k \\ p_i &= p^{(N)}{}_i \end{aligned} \qquad (3)$$

Les seuls calculs à effectuer à chaque pas de la récurrence sont une multiplication et une addition. On utilisera donc des cellules MAC (multiply

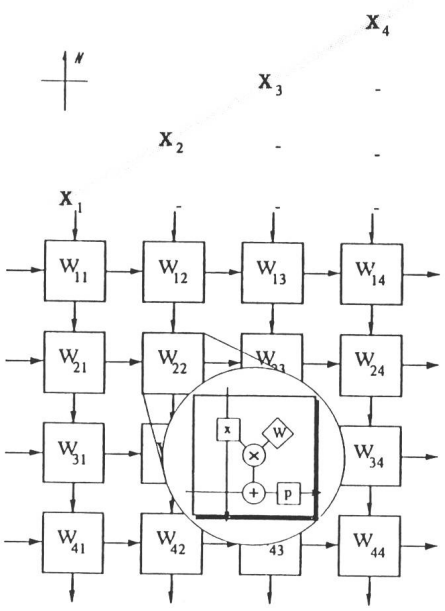


Figure 1 Multiplication matrice-vecteur sur réseau systolique bi-dimensionnel carré
$t = 0$, $N = 4$

and accumulate) qui réaliseront les opérations de calcul. Chaque cellule contient un coefficient de la matrice W (fig. 1).

L'algorithme systolique fonctionne de façon synchrone tel que décrit sur la figure 2.

– Les composantes x circulent inchangées du Nord au Sud du réseau. Elles sont disposées diagonalement de manière à ce que les produits partiels calculés à l'Ouest puissent être transférés dans la cellule voisine Est en même temps que la prochaine composante y parvient.

– Les sommes partielles des composantes p circulent d'Ouest en Est. Leur valeur initiale est nulle, et en passant à travers le réseau, chaque p_i accumule ses produits partiels.

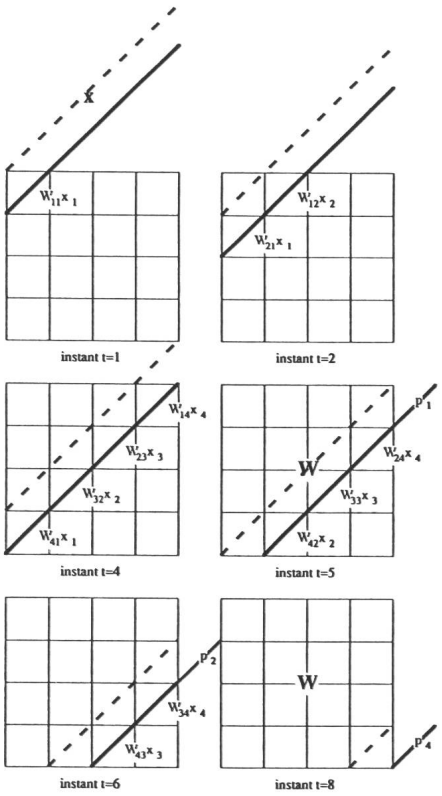


Figure 2 Multiplication matrice-vecteur sur réseau systolique bi-dimensionnel carré
$t = 1$ à 8, $N = 4$

Ainsi, au temps $t = 1$, le produit $W_{11}x_1$ est calculé dans la cellule (1,1). Au temps $t = 2$, les produits $W_{12}x_2$ et $W_{21}x_1$ sont calculés dans les cellules (1,2) et (2,1) respectivement. Le calcul de la cellule (1,2) est additionné au résultat fourni par la cellule (1,1) du temps $t = 1$. L'algorithme se poursuit ainsi jusqu'au temps $t = 4$ où le vecteur x se trouve sur la diagonale reliant la cellule (4,1) à la cellule (1,4). On remarque alors l'apparition des résultats p_1, p_2, p_3 et p_4 successivement élaboré à l'Est dès l'instant $t = 5$. Ainsi, il faut $2N = 8$ cycles pour réaliser le produit matrice-vecteur avec un tel réseau carré.

Un tel système parallèle n'est intéressant que s'il est possible d'utiliser avec une grande efficacité les processeurs le composant, ceci en réduisant au maximum les moments où ceux-ci n'effectuent pas d'opérations utiles. Le vecteur indiqué en trait-tillé sur la figure 2 suggère qu'il est possible d'enchaîner dans le temps des vecteurs qui sont présentés successivement au réseau. Il a été montré [6] qu'en utilisant un tel enchaînement de données (pipeline), le réseau systolique Genes a, en régime permanent, une efficacité de 100%.

Interface et calcul neuronal

Le passage de la structure systolique ci-dessus à une architecture adaptée au calcul neuronal amène certaines modifications dictées par la nécessité de simplifier la connectique de l'interface.

La première modification concerne l'adjonction de la fonction caractéristique des neurones σ( ) qui sera placée sur l'arête Est de la matrice systolique (fig. 3) fournissant ainsi la valeur y de l'activité des neurones selon la formule 1. Du fait que les stimuli x peuvent aussi bien provenir du réseau lui-même (réseaux récurrents) que de l'extérieur, il est intéressant de prévoir un chemin systolique permettant de reboucler les données de l'Est vers le Nord (trait-tillés sur la fig. 3). Les répercussions de ces modifications sur les cellules systoliques sont les suivantes (fig. 4):

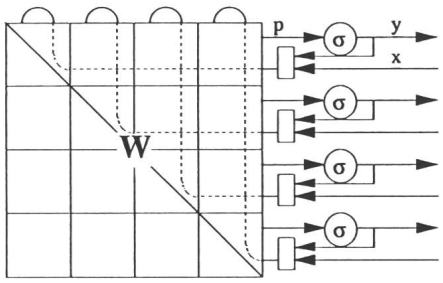


Figure 3 Réseau systolique et fonction neurone

- ajout d'un chemin menant de l'Est aux cellules diagonales (NV),
- spécialisation des cellules diagonales afin d'aiguiller les données provenant de l'Est sur le chemin Nord-Sud,
- ajout d'un chemin menant des diagonales au Nord (U).

Il est facile de constater que les modifications apportées ne seront utiles que dans la sous-matrice triangulaire supérieure. La partie triangulaire inférieure peut cependant être utilisée a bon escient dans certains modèles neuro-mimétiques en constatant les faits suivants:

– les vecteurs stimuli voyageant du Nord au Sud pendant le calcul peuvent rebondir sur l'arête Sud afin d'être renvoyés aux cellules diagonales, là, il est possible de les comparer (fonction

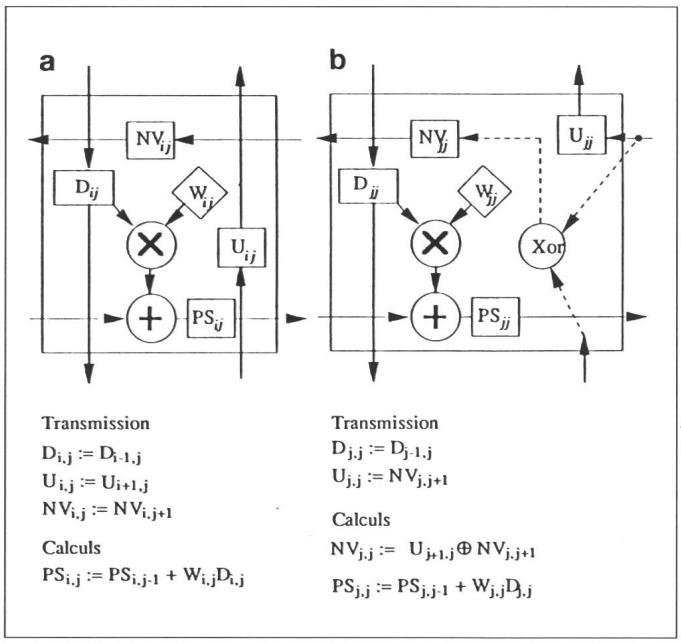


Figure 4
Description des cellules synaptiques systoliques:
a) non diagonale,
b) diagonale

XOR de la fig. 4), composante par composante, avec les activités des neurones ayant transité, à l'Est, par les fonctions neurones,
– les résultats de ces comparaisons peuvent transiter vers l'Ouest où ils seront récoltés dans un dispositif ad hoc (fig. 5).

Cette structure est maintenant tout à fait adaptée à la réalisation matérielle des réseaux de neurones artificielles. C'est ainsi que plusieurs circuits ont vu le jour.

– Le premier circuit, nommé Aplysie [7], transite des valeurs binaires et est ainsi réellement spécialisé dans la résolution du modèle de Hopfield.
– Le deuxième circuit, Genes HN8 [8], réalise les mêmes fonctions mais sur des valeurs représentées sur 8 bit, ouvrant ainsi le système vers plus de généralité. Ce circuit, contenant 4 processeurs fonctionne correctement à une fréquence d'horloge de 10 MHz. Il a permis la réalisation d'une première carte d'évaluation contenant 256 synapses. Cette carte sera décrite plus en détail dans la suite de cet article.
– Le troisième circuit, Genes HH8 [9], ajoute, aux fonctions décrites précédemment, la possibilité de modifier en parallèle les coefficients synaptiques selon la règle de Hebb. Contenant 16 synapses, ce circuit est encore en phase de test.
– Un dernier circuit, Genes VM16 [10], vient de revenir de cuisson. Plus sophistiqué, ce circuit (fig. 6 et 7) doit permettre l'apprentissage pour le réseau de Kohonen dont nous allons parler maintenant. Des possibilités supplémentaires permettent l'utilisation de matrices virtuelles étendant ainsi le système à des plus grands réseaux de neurones.

Réalisation du réseau de Kohonen

Le réseau auto-organisateur d'adaptation topologique (feature mapping) proposé par Teuvo Kohonen [11], nécessite l'application répétitive d'un algorithme particulier sur l'ensemble des neurones qui constituent le réseau. Les propriétés structurantes du réseau de Kohonen sont particulièrement adaptées au pré-traitement de données (voir exemples dans les articles suivants). Ces opérations nécessitent le calcul en temps réel et exigent l'utilisation d'accélérateurs appropriés.

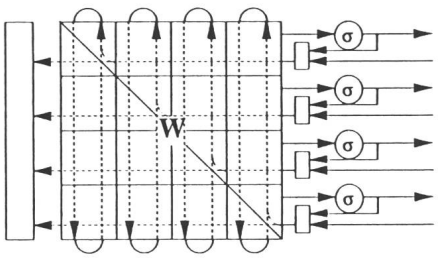


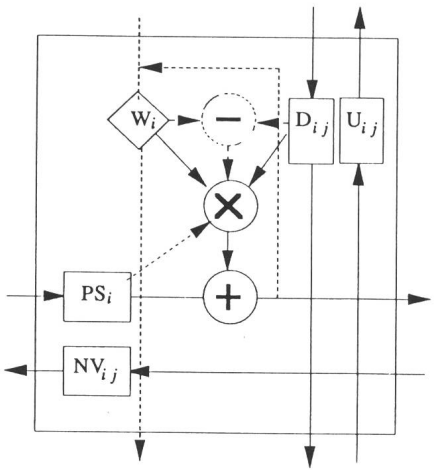
Figure 5 Solution Genes pour le calcul des réseaux de neurones

Le réseau Genes décrit plus haut sera utilisé pour implanter physiquement les synapses et les neurones du réseau de Kohonen. L'algorithme à implanter se divise en trois phases:

a) application des signaux d'entrée au travers du réseau,
b) localisation du groupe de neurones les plus sensibles (bubble, best match, etc.) et
c) mise à jour des connexions reliant ces neurones à l'espace d'entrée.

En analysant ces différentes étapes, on obtiendra les constatations suivantes qui ont une importance cruciale pour l'implantation systolique qui est proposée ici.

– La première phase consiste trivialement en l'application de l'opération matricielle décrite ci-dessus.
– La deuxième phase correspond également à l'application du produit matrice-vecteur. Dans ce cas cependant, la matrice de connexion ***W*** contient la définition du couplage latéral entre les neurones. Ce couplage est utilisé pour faire émerger la bulle des neurones «gagnants» au travers de la compétition.
– L'ordonnancement des opérations de la deuxième phase correspond à l'application du réseau de Hopfield.
– La dernière phase nécessite la mise à jour des coefficients de la matrice sur les lignes correspondant aux neurones qui ont gagné la compétition.



Transmissions

$D_{i,j} := D_{i-1,j}$ $\quad NV_{i,j} := NV_{i,j+1}$ $\quad U_{i,j} := U_{i+1,j}$

Calculs

$PS_{i,j} := PS_{i,j-1} + W_{i,j-1} D_{i,j-1}$ Phases a et b

$W_{i,j} := W_{i,j} + PS_{i,j} (D_{i,j} - W_{i,j})$ Phase c

Figure 6 Cellule synaptique Genes VM 16 non diagonale permettant l'adaptation selon la méthode de Kohonen

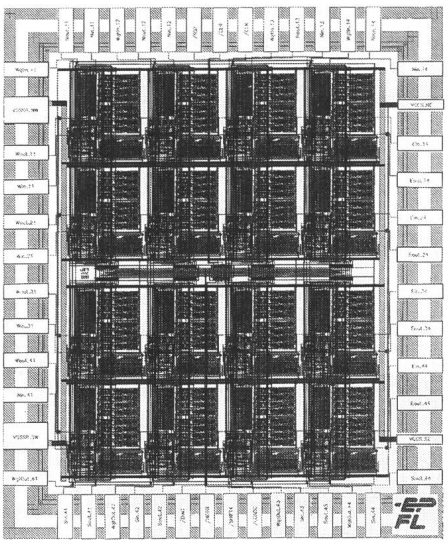


Figure 7 Circuit Genes VMI 6

Il est montré que l'algorithme ci-dessus peut être réalisé au moyen des cellules systoliques décrites dans la figure 6. On remarquera que les éléments utilisés sont identiques à ceux de la figure 4; seuls les chemins diffèrent. Ainsi on peut constater les possibilités d'extension du système Genes vers de nombreux algorithmes. Les cellules diagonales ne sont pas montrées sur la figure 6, elles comportent, en plus, le dispositif de comparaison indiqué plus haut.

Réalisation matérielle

Les réalisations VLSI de ce système systolique ont conduit à la réalisation d'une première carte d'évaluation (fig. 8) contenant 256 synapses reliant 16 neurones. Le rôle de l'interface entre un tel système parallèle et un ordinateur classique est de permettre à ce dernier d'utiliser au maximum, et avec le plus de facilité, la puissance de calcul potentielle du premier.

Les 256 processeurs fonctionnant à 10 MHz et composant le cœur de la carte peuvent théoriquement calculer 100 millions de connexions synaptiques par seconde (MCPS). Ceci implique que tous les processeurs soient utilisés. Cette hypothèse n'est évidemment valable qu'en régime permanent. Comme nous l'avons vu plus haut, cette hypothèse est tout à fait réaliste, encore faut-il que l'interface permette le transfert de données à une vitesse suffisante.

Le bus d'extension typique (VME, Nubus, SBUS) d'une station de travail permet le transfert vers des périphériques avec une vitesse de plusieurs dizaine de million d'octets par seconde (20 Mbyte/s pour Nubus par ex.). On démontre facilement que Genes est adaptable à la bande passante qu'offre un tel système hôte. En effet, il suffit, pour équilibrer la puissance de calcul, d'augmenter ou de réduire le nombre de processeurs (N^2) selon la formule suivante:

$$B = 2N\,f/m \qquad (4)$$

où B est la bande passante, f la fréquence des circuits Genes, N le nombre de neurones et m la longueur des résultats.

Pour garantir la facilité d'utilisation de l'accélérateur neuronal, il faut pouvoir travailler dans les formats de données traditionnellement utilisés par les ordinateurs classiques, c'est-à-dire des mots de 32 bits en représentation virgule flottante. Les circuits Genes travaillant sur des nombres en virgule fixe de 8 bits et calculant de plus en série, des circuits de conversions d'un format à l'autre doivent être contenus dans l'interface.

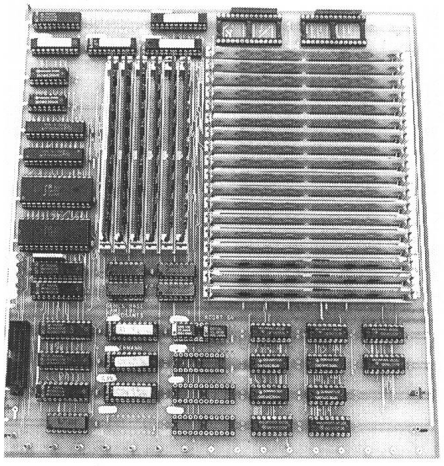


Figure 9 Carte Prototype Genes SY 1

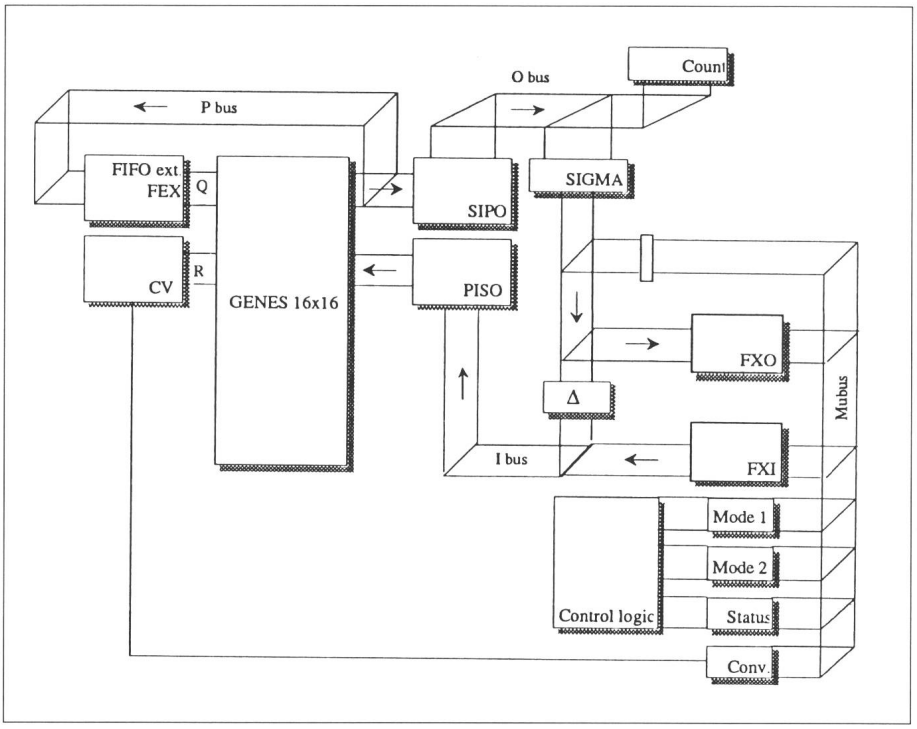


Figure 8 Interface Genes SY1
Les données traversent le fifo d'entrée (FXI) puis les sérialiseurs (piso) pour terminer dans la matrice Genes. Les résultats traversent alors les parallèliseurs (sipo) puis adressent une mémoire (sigma) pour terminer dans le fifo de sortie (FXO). Un fifo supplémentaire (FEX) permet de reboucler les résultats pour l'extension aux matrices virtuelles.

Les problèmes posés par l'interface sont théoriquement résolus et la carte prototype Genes SY1, que nous expérimentons actuellement (fig. 9), apporte les premiers éléments de réponses pratiques. Pour l'instant, cependant, la traduction virgule flottante – virgule fixe est laissée aux soins du programmeur.

Travaux futurs

La suite du projet Genes consistera essentiellement à concrétiser pratiquement les résultats théoriques qui sont maintenant bien établis. Ainsi, il s'agit de réaliser un système contenant plusieurs cartes du type de celle décrite plus haut ainsi que des cartes d'acquisition et de visualisations éventuelles. Ce système aura pour tâche de réaliser le réseau de Kohonen et d'offrir des puissances de traitement importantes aussi bien pendant la phase d'application que la phase d'apprentissage. Ce système sera la première concrétisation des efforts interdisciplinaires du groupe Mantra [12].

Bibliographie

[1] *P. Treleaven, M. Pacheco* et *M. Vellasco*: VLSI Architectures for Neural Networks, IEEE Micro, Decembre 1989, pp. 8–27.

[2] *A. Guérin*: Crasy: un calculateur de réseaux adaptatifs systolique. Application au calcul neuromimétique, Thèse DI, INPG-Enserg., Grenoble, 1987.

[3] *M. Weinfeld*: A Fully Digital Integrated CMOS Hopfield Network Including the Learning Algorithm, VLSI for Artificial Intelligence, H.G. Delgado-Frias et W.R. Moore Eds, Kluwer Academic Publishers, 1989.

[4] *S.Y. Kung*: VLSI Array Processors, Prentice-Hall, 1988.

[5] *P. Quinton* et *Y. Robert*: Algorithmes et Architectures systoliques, Masson, 1989.

[6] *F. Blayo*: Une implantation systolique des algorithmes connexionnistes, Thèse de Doctorat n° 904, EPFL, 1990.

[7] *P. Hurat*: Aplysie: Un circuit neuro-mimétique: réalisation et intégration sur tranche, Thèse de Doctorat, LGI, Grenoble, 1989.

[8] *C. Lehmann*: Conception VLSI: De la Parole au Geste, Rapport Interne, LAMI-EPFL, 1990.

[9] *J.T. Randriamalazarivo*: Conception du layout du circuit Genes–HH8–4 x 4, Rapport Interne, LAMI-EPFL, 1991.

[10] *C. Lehmann*: Genes VM16: 2e version du circuit systolique Genes en ES2, Rapport Interne, LAMI-EPFL, 1991.

[11] *T. Kohonen*: Self Organization and Associative Memory, Springer Verlag, Berlin, 1984.

[12] *M. Hasler*: Neuronale Netzwerke im Verbund, Das Projekt Mantra der ETH Lausanne, Bulletin ASE/UCS 82(1991)13, p. 24–26.