

La détection de nouveauté basée sur le temps de stabilisation d'un réseau de neurones: application possible en reconnaissance de parole?

Jean ROUAT¹, Tuong Vinh HO^{1,2}

¹ ERMETIS, DSA, Université du Québec à Chicoutimi
555 boul. de l'Université, CHICOUTIMI, Québec, CANADA, G7H 2B1
Tél.: 1 418 545 5011 x5642 - Fax: 1 418 545 5012

² Ecole Polytechnique de Montréal, Canada
e-mail: Jean_Rouat@uqac.quebec.ca, vho@uqac.quebec.ca

ABSTRACT

We propose a spiking neural network model inspired from a simulated cortex model. Also, a new paradigm for pattern recognition by neural networks with complex dynamics is presented. The 'relaxation' time of the network is used as a criterion for novelty detection. We compare the proposed neural network with Hopfield and backpropagation networks for a noisy digit recognition task. It is shown that the proposed network is more robust. We also design a limited experiment based on the recognition of temporal sequences of vowels and we show that the network is able to perform the recognition with a rate of 100% (sequences of 5 and 11 vowels). Regarding speech and pattern recognition tasks, the proposed spiking network seems to have a strong potential.

1. INTRODUCTION

Il existe des systèmes de reconnaissance de parole indépendants du locuteur pour des vocabulaires limités et qui offrent des performances intéressantes à condition que l'environnement ne soit pas trop corrompu par le bruit et les interférences [Dup97]. Toutefois, l'interrogation de bases de données à distance, (téléphone, radio, etc.), la transcription des nouvelles télévisées ou radio par exemple, nécessitent de concevoir des systèmes polyvalents et peu sensibles aux conditions et aux environnements d'opérations ainsi qu'aux bruits ambiants (téléphone cellulaire, cabine téléphonique, voitures, avions, camions, etc.) [Ezz97][Rob97]. Or le traitement de la parole en milieu bruyant est loin d'être résolu et les systèmes de reconnaissance actuels ne peuvent fonctionner dans de tels environnements sans observer de dégradation significative des performances [Gau97][Woo97][Bak97].

La mise en place de système de reconnaissance versatile fait souvent appel à des analyses perceptives inspirées de modèles du système auditif [Pat95]. Ces analyses codent l'information liée à la parole dans un espace à représentation spatio-temporelle complexe et difficile à caractériser via les algorithmes couramment utilisés en reconnaissance de parole. En effet, ces représentations spatio-temporelles sont structurées à la fois spectralement et dans le temps. L'information est en général non stationnaire. Il y a donc lieu de mettre au point des paramètres et des algorithmes de reconnaissance de formes capables d'exploiter cette information temporelle. Or une ma-

jorité des algorithmes contemporains assument une certaine stationnarité du signal et des paramètres d'analyse. De plus, ces systèmes requièrent un apprentissage supervisé souvent long et fastidieux. Nous pensons qu'il y a lieu d'étudier l'élaboration de nouvelles techniques d'analyse et de reconnaissance adaptées à l'exploitation de l'information temporelle fine telle que générée par exemple via des analyses d'inspiration perceptive. Une information temporelle fine serait par exemple celle qui est liée à l'enveloppe des signaux à la sortie d'un banc de filtres cochléaires [Rou97].

2. MOTIVATION

Il apparaît intéressant d'élaborer de nouveaux outils aptes à traiter une information dynamique et non stationnaire. Ce type d'outil devrait pouvoir être utilisé en reconnaissance de formes afin de traiter une information codée dans le temps (parole, par exemple). Nous proposons un système qui devrait être en mesure de remplir ces critères.

Il s'agit d'un système de détection de la nouveauté basé sur le temps de stabilisation d'un réseau de neurones (dit bio-inspiré) dont l'architecture est inspirée de la couche IV du cortex. L'originalité du travail réside principalement dans la définition d'une règle de modification des connexions synaptiques et dans la création du critère de temps de stabilisation pour la reconnaissance. En effet, la majorité des techniques de reconnaissance des formes utilisant les réseaux de neurones (souvent formels) se base sur des critères relativement statiques (minimisation d'une fonction d'erreur, maximisation d'une probabilité, etc.) et codent l'information temporelle de façon statique via la structure spatiale des entrées [Hay94]. Par ailleurs, l'approche adoptée ici ne nécessite pas de supervision du réseau. Celui-ci est en mesure de détecter la nouveauté et de s'adapter de façon autonome. De plus, il n'y a pas de différence entre apprentissage et reconnaissance.

Dans un premier temps, il est important de tester et de valider ce type d'approche vis-à-vis des systèmes plus classiques afin de s'assurer que le réseau proposé puisse réaliser au minimum des tâches similaires à celles exécutées par les réseaux dits formels. Nous présentons donc une série de tests préliminaires effectués en reconnaissance de chiffres bruités ainsi qu'en reconnaissance de séquences de voyelles puis nous discuterons du potentiel de ce travail en reconnaissance de parole.

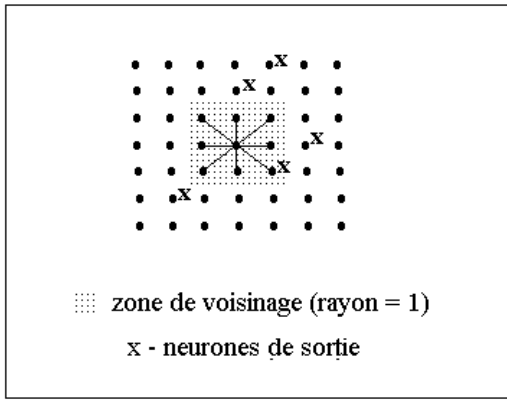


Figure 1: Architecture du réseau de neurone.

3. LE RÉSEAU DE NEURONES

3.1. Modèle du neurone

Le neurone est de type intégration et décharge ('Integrate and Fire') avec période réfractaire et potentiel post-synaptique décroissant. L'état du neurone à l'instant t est caractérisé par son potentiel interne U . La réponse $S_i(t)$ du neurone i est donnée par:

$$S_i(t) = \begin{cases} 0 & \text{pour } (t - t_{spike}) < \rho, \\ \mathcal{H}[U_i(t) - \theta] & \text{pour } (t - t_{spike}) \geq \rho \end{cases} \quad (1)$$

\mathcal{H} est la fonction d'Heaviside. Le potentiel de contrôle $U_i(t + 1)$ pour la cellule i est l'intégration de toutes les réponses afférentes à l'instant t et tient compte de la fréquence instantanée f_i de décharge de la cellule i .

$$U_i(t + 1) = \sum_j C_{ij}(t) S_j(t) + U_i(t) + e_i(t) + f_i(t) \quad (2)$$

$C_{ij}(t)$ est la connexion de la cellule j vers la cellule i à l'instant t . $e_i(t)$ est le signal externe (entrée du stimulus) pour la cellule i considérée.

3.2. Architecture et apprentissage

L'architecture est inspirée du modèle de la couche IV du cortex tel que proposé par Hill & Villa [Hil97]. Ils définissent une couche à 2 dimensions comprenant des neurones inhibiteurs et excitateurs avec récursivité. Chaque neurone est interconnecté à ceux qui appartiennent à son ensemble de voisinage. Cet ensemble est défini comme étant carré, centré autour du neurone et comprend une taille différente selon le type de connexion (inhibitrice ou excitatrice). Nous choisissons les neurones de sortie du réseau de façon aléatoire à partir de neurones de type excitateurs (figure 1). Le signal de sortie est caractérisé par l'état de décharge de l'ensemble des neurones de sortie.

L'apprentissage est inspiré du travail de Stassinopoulos et Bak [Sta94]. Pour chaque signal d'entrée, l'action du réseau est considérée comme étant un succès si au moins

Table 1: Taux d'erreurs pour un apprentissage sur les chiffres propres [0-4] et reconnaissance sur les chiffres [0-9] propres et bruités

Type de réseau	proposé	Hopfield	multicouches
Taux d'erreur (%)	10	21	27

Table 2: Taux d'erreurs pour un apprentissage sur les chiffres propres [5-9] et reconnaissance sur les chiffres [0-9] propres et bruités

Type de réseau	proposé	Hopfield	Multicouches
Taux d'erreur (%)	30	40	30

un neurone appartenant à l'ensemble des neurones de sortie se décharge. Dans ce cas, les connexions entre les neurones actifs sont renforcées. Si l'action est sans succès, les connexions entre les neurones actifs sont affaiblies. La mise à jour des connexions se fait selon la règle d'Hebb. Soit h , le signal de feedback ($h = +1$ pour le renforcement ou -1 pour affaiblir). La mise à jour des poids entre deux neurones est donnée par:

$$C_{ij}(t + 1) = C_{ij}(t) + \alpha C_{ij}(1 - C_{ij}) S_i(t) S_j(t) h(t) \quad (3)$$

avec α le taux d'apprentissage.

3.3. Critère de reconnaissance

Il n'y a pas de différence entre apprentissage et reconnaissance. La mise à jour des connexions synaptiques est toujours en cours sauf lorsque le réseau est stable. Après présentation d'un stimulus, le réseau a un comportement complexe qui finit par se stabiliser au cours du temps. Lorsque les changements de poids sont inférieurs à un seuil préétabli, on considère que le réseau est stable et que l'apprentissage ou la reconnaissance sont terminés. Le temps nécessaire T pour atteindre cet état stable est utilisé comme critère de détection de nouveauté. Ce temps T permet de caractériser le signal d'entrée. Un temps T très court (de l'ordre de 11 itérations) implique que le réseau a probablement déjà 'vu' ce signal au préalable. Cette notion permet de créer un système de détection de la nouveauté ayant un bon degré de robustesse vis-à-vis du bruit.

4. APPLICATION À LA RECONNAISSANCE DE CHIFFRES BRUITÉS

Des expériences de reconnaissance ont été réalisées à partir des chiffres 0 à 9 codés sur une matrice binaire de 7 par 5. Le réseau apprend uniquement sur les données non bruitées de 0 à 4 ou de 5 à 9 (colonne 1 ou 4 de la figure 2). Les expériences de reconnaissance portent sur les versions bruitées et propres des chiffres de 0 à 9 (toutes les colonnes de la figure 2). Le bruit est de type uniforme avec 20% des pixels modifiés. Le réseau est composé pour 70% de cellules excitatrices et pour 30% de



Figure 2: Images des chiffres propres (colonnes 1 et 4) et bruités (20% de bruit, colonnes 2,3,5 et 6).

cellules inhibitrices. Il comprend 49 cellules (7x7). Les chiffres sont codés sur une matrice de pixels 7x5. Chaque pixel est présenté sur l'entrée e_i d'une cellule choisie au hasard (i.e. 35 cellules sur les 49). Les tables 1 et 2 permettent de comparer le réseau proposé avec un réseau de Hopfield [Dem96] et un réseau multicouches utilisant la règle d'apprentissage généralisée avec moments [Ebe92]. Après un apprentissage à partir des chiffres propres de 0 à 4 (table 1) ou des chiffres propres de 5 à 9 (table 2), on présente l'ensemble des chiffres tel qu'illustrés à la figure 2. Le système doit ensuite différencier les chiffres qu'il connaît (ceux qui étaient représentés dans la séquence d'apprentissage) de ceux pour lesquels il n'a jamais 'vu' de représentant. Pour la série apprise à partir de [0-4], le réseau proposé est nettement supérieur (table 1). Pour la série apprise à partir de [5-9], les performances sont similaires à celles du réseau multicouches.

5. RECONNAISSANCE DE SÉQUENCES DE VOYELLES

Le but de ces expériences est d'illustrer la possibilité d'identifier des séquences de voyelles déjà apprises et de les isoler des autres séquences pour lesquelles on retrouve les mêmes voyelles dans un ordre différent.

5.1. Les données de parole et la taille du réseau

Les données de parole sont disponibles sur le site de [Mer98]. Nous utilisons les voyelles prononcées par un seul locuteur pour onze mots anglais: heed(i), hid(I), head(E), had (A), hard (a:) hud (Y), hod(O), hoard(C:), hood (U), who'd(u:) et heard(3:). Le signal est filtré passe-bas à 4.7 kHz puis il est échantillonné à 10 kHz et quantifié sur 12 bits. Une analyse LPC d'ordre 12 est réalisée afin d'extraire 10 coefficients par fenêtre de signal. Une fenêtre de Hamming de 512 points est centrée au préalable sur la zone stable de chaque voyelle. On utilise donc 10 paramètres LPC par voyelle. Le réseau comprend 10 cellules (matrice de 5x2). Chaque paramètre est présenté sur l'entrée e_i d'une cellule.

5.2. Méthodologie

Une première séquence de voyelles est utilisée comme séquence d'apprentissage. Chaque voyelle de cette

séquence est présentée au réseau de façon séquentielle pendant un intervalle de temps choisi T . Par exemple, la première voyelle est présentée au réseau pendant T itérations. Ensuite, la deuxième voyelle est présentée pour la même durée T . Il n'y a pas d'initialisation du réseau entre les présentations successives. On procède ainsi pour l'ensemble des voyelles de la séquence.

Après apprentissage, le comportement du réseau permet de savoir si celui-ci a identifié la séquence déjà apprise parmi les séquences qui lui sont présentées. Nous utilisons le temps de relaxation du réseau comme étant caractéristique du comportement. Chaque voyelle est présentée de façon séquentielle au réseau. Lorsque la première voyelle est présentée, le réseau oscille et se dirige vers un état d'équilibre. Ensuite, la deuxième voyelle est présentée, le réseau oscille à nouveau et atteint un autre état d'équilibre. On procède ainsi pour toutes les voyelles de la séquence. Après présentation de la séquence de voyelles, on obtient une série de temps de relaxation correspondant à la séquence de voyelles. En d'autres termes, nous avons effectué une 'projection' de la séquence de voyelles sur une séquence de temps de relaxation tout en réduisant la dimension des paramètres. L'analyse de cette séquence de temps permet d'indiquer si la série de voyelles a déjà été vue par le réseau.

Afin de faciliter l'analyse des séquences de temps, nous utilisons une méthode de codage très simple. Chaque intervalle de temps est codé par un chiffre. Par exemple, si le temps T se situe dans l'intervalle de 0 à 100 itérations ($0 < T \leq 100$), le chiffre associé est 0. De même, si le temps T se situe dans l'intervalle de 100 à 200 itérations ($100 < T \leq 200$), le code associé sera de 1, etc. De cette façon, on peut coder une séquence de temps de relaxation en une séquence de nombres entiers. En comparant les séquences de nombres correspondant aux séquences de voyelles de test avec la séquence déjà apprise, on peut savoir si une séquence de voyelles a déjà été 'vue' par le réseau. Les expériences suivantes illustrent ce principe.

5.3. Expérience 1: reconnaissance de séquences de 5 voyelles

Une séquence de 5 voyelles des mots anglais heed(i), hid(I), head(E), had (A) et hard (a:) est utilisée comme séquence d'apprentissage. Lors de l'apprentissage, chaque voyelle est présentée au réseau pour une durée de 200 itérations. Ensuite, cette même séquence est présentée au réseau afin de connaître le comportement de celui-ci face aux données déjà 'vues'. La séquence de temps de relaxation du réseau est de: 461, 11, 11, 245 et 145 itérations. En appliquant le codage précédent, on obtient la séquence de nombres suivante: 4, 0, 0, 2, 1. Pour les tests, nous utilisons un ensemble de séries de voyelles qui est composé de toutes les combinaisons possibles (120 combinaisons) des 5 voyelles déjà présentées au réseau lors de l'apprentissage. A titre d'exemple, la séquence de chiffres obtenue pour la série de voyelles (had (A), hid(I), head(E), heed(i), et hard (a:)) est 0, 0, 0, 5, 1. En comparant cette séquence à celle de l'apprentissage on peut conclure qu'elle n'avait jamais été vue par le réseau. Le taux d'erreur de reconnaissance des séquences est de 0% (au-

cune séquence en erreur).

5.4. *Expérience 2: reconnaissance de séquences de 11 voyelles*

Une séquence de 11 voyelles a été obtenue à partir des mots anglais heed(i), hid(I), head(E), had (A), hard (a:) hud (Y), hod(O), hoard(C:), hood (U), who'd(u:) et heard(3:)). Elle est utilisée comme séquence d'apprentissage. Chaque voyelle est présentée au réseau pendant une durée de 200 itérations. Lors des tests, nous utilisons un ensemble de séries de voyelles qui comprend 100 combinaisons des 11 voyelles d'apprentissage. Le taux d'erreur est de 0% (aucune séquence en erreur).

6. DISCUSSION ET CONCLUSION

Les expériences de reconnaissance de données statiques (les chiffres corrompus) indiquent que le réseau est en mesure de faire un travail aussi bon et souvent meilleur qu'un réseau de Hopfield et un réseau à rétropropagation.

Les expériences portant sur les séquences de voyelles permettent de mettre en valeur l'aptitude du réseau à traiter des séquences. En effet, la réponse du réseau (temps de stabilisation) à une voyelle dépend de ses états précédents. Pour une même voyelle, les caractéristiques dynamiques sont liées aux présentations antérieures.

Les expériences présentées sont relativement limitées et il y a lieu d'approfondir le travail afin de bien connaître les limites du réseau et son potentiel en reconnaissance de parole. Toutefois, sa robustesse au bruit ainsi que son aptitude à traiter de l'information structurée dans le temps, laissent à penser que ce système est un bon candidat pour la mise au point de techniques de reconnaissance de parole.

REMERCIEMENTS

Ce travail a été financé par le Conseil National de la Recherche en Sciences Naturelles et en Génie du Canada (CRSNG) ainsi que par la fondation de l'Université du Québec à Chicoutimi. Un merci tout particulier à Alessandro Villa pour les discussions stimulantes et enrichissantes en regard de ce travail.

BIBLIOGRAPHIE

- [Bak97] R. Bakis, S.Schen, P. Gopalakrishnan, R. Gopinath, S. Maes and L. Polymenakos (1997). Transcription of broadcast news - system robustness issues and adaptation techniques. *Proc. of the IEEE-ICASSP*, Vol. 2, 711-714.
- [Dem96] Demuth H., Beale M. (1996). Neural Network Toolbox for Use with MATLAB, *The Math Works Inc.*
- [Dup97] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine and J.M. Boite (1997). Hybrid HMM/ANN systems for training independant tasks: Experiments on Phonebook and related improvements. *Proc. of the IEEE-ICASSP*, Vol. 3, 1767-1770.
- [Ebe92] R. C. Eberhart, R. W. Dobbins. (1992). Neural Networks PC Tools: a practical guide. Academic Press. San Diego, USA.
- [Ezz97] H. Ezzaidi, I. Bourmeyster and J. Rouat (1997). A new algorithm for double talk detection and separation in the context of digital mobile radio telephone. *Proc. of the IEEE-ICASSP*, Vol. 3, 1897-1900.
- [Gau97] J.L. Gauvain, G. Adda, L. Lamel and M. Adda-Decker (1997). Transcribing Broadcast News Shows. *Proc. of the IEEE-ICASSP*, Vol. 2, 715-718.
- [Hay94] S. Haykin (1994). Neural Networks – A Comprehensive Foundation. *IEEE Computer Society Press and Macmillan College Publishing Company, Inc.*, 1994.
- [Hil97] Hill S., Villa A. (1997). Dynamic transitions in global network activity influenced by the balance of excitation and inhibition. *Network: Computation in Neural Systems*, UK, Vol. 8, 2, 165-184.
- [Ho97] Ho T.V., Rouat J. (1997). A Novelty Detector Using a Network of Integrate and Fire Neurons. *7th Int. Conf. on Artificial Neural Networks*, Lausanne, Switzerland, 8-10 Oct. 1997, Lecture Note on Computer Science (1327), Springer, pp. 103-108.
- [Ho98] Ho T.V., Rouat J. (1998). Novelty Detection Based on Relaxation Time of a Network of Integrate-and-Fire Neurons. *International Joint Conference on Neural Networks*, Alaska, May 1998.
- [Mer98] Merz, C.J., Murphy, P.M. (1998). UCI Repository of machine learning databases. [<http://www.ics.uci.edu/mllearn/MLRepository.html>]. Irvine, CA: Univ. of California, Dep. of Information and Computer Science.
- [Pat95] R.D. Patterson and M.H. Allerhand (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoust. Soc. Amer.*, Vol. 98 (4), pp.1890-1894.
- [Rob97] Robust speech recognition for unknown communication channels. *ESCA-NATO Tutorial and Research Workshop*, Pont-à-Mousson, France, 17-18 avril, 1997, éditions ESCA.
- [Rou97] Spatio-temporal Pattern Recognition with Neural Networks: Application to Speech. *Artificial Neural Networks-ICANN'97, Lecture Notes in Comp.Sci.*,1327, Springer, 43-48.
- [Rub95] A. J. Rubio and J. M. López (eds.) (1995). Speech Recognition and Coding, New Advances and Trends. *NATO ASI Series*, Springer.
- [Sta94] Stassinopoulos D., Bak P. (1994). Self-organization in a Simple Brain Model. *Proc. of WCNN'94*, San Diego, Jun, Vol. 1, pp. 4-26.
- [Woo97] P. C. Woodland, M. J.F. Gales, D. Pye and S. J. Young (1997). Broadcast News Transcription Using HTK. *Proc. of the IEEE-ICASSP*, Vol. 2, 719-722.