Universidad Nacional Autónoma de México

FACULTAD DE INGENIERIA

LAS REDES NEURONALES ARTIFICIALES CONTEMPORANEAS EN HARDWARE

T E S I S

QUE PARA OBTENER EL TITULO DE

INGENIERO MECANICO ELECTRICISTA

P R E S E N T A

CESAR VELEZ ANDRADE

Director: DR. JOSE ISMAEL ESPINOSA ESPINOSA

MEXICO, D. F. 1996

DEDICATORIA

A mis padres, César Vélez y de la R. y Haydeé Andrade D.:

> Por el infinito amor, confianza y dedicación que me han brindado. Por criarnos, a mis hermanas y a mi, bajo una insustituible atmósfera de amor, principios y dedicación. Para ustedes este trabajo, que sin duda, sólo es una minúscula representación de lo mucho que les quiero y les agradezco.

A mis hermanas Haydeé, Verónica, Rubria y Ana Luisa:

> Por estar siempre conmigo y ser parte de mi formación. Por haberme enseñado, con ejemplo vivo, el comportamiento de unión, honestidad y nobleza. Las quiero mucho.

A Gloria:

> Por toda la confianza, amor y dedicación que depositas cada día en mí y por hacer todo lo posible para siempre estar conmigo.

A la memoria de mi tío Héctor A. Vélez y de la R.

A mis tíos Karina Vélez y de la R. y Armando Andrade D.

> Por siempre apoyar a mis padres y estar pendientes de cada uno de nosotros, siempre con el único interés de nuestro bienestar. Los quiero mucho y gracias.

A la familia Espinosa Lara: Ismael, Ritaluz y Malors.

> Por su atención, cariño e invaluable amistad. Especialmente al Dr. José Ismael Espinosa E. por haberme concedido el orgullo de ser su amigo y el enorme placer de haberle aprendido tantas cosas.
> Ismael, mil gracias.

A todos mis amigos.

> Especialmente a: Margarita A., Carlitos S., Eduardo Z., J. Carlos E., Rodrigo P. y Wilphen V. Por haber utilizado el paso de cada día para consolidar y demostrar su gran amistad.

AGRADECIMIENTOS

Al Prof. Jorge López S.

Por enseñarme la disciplina de la esgrima y a asociarla a mi vida profesional.

A mis compañeros del Laboratorio de Cibernética:

Irma, Alberto, Fidel, Javier, Jorge Q., Juan Carlos, Juan Manuel, Luis, Pepe, Roberto y los Rubenes.

A mis compañeros de trabajo:

Alejandro G., Alejandro Z., David V., Fernando B., Irving H., J. Carlos Z. y René Ch. Gracias a todos y en especial a René por su constante motivación y apoyo.

A la U.N.A.M.:

Por haberme brindado educación y conciencia de la realidad nacional. Por haberme permitido ser parte de ese singular mosaico social que la conforma, que de alguna manera, induce la responsabilidad que un profesional debe tener hacia un país tan complejo como lo es nuestro México.

Por mi Raza Hablará el Espíritu.

ÍNDICE

CAPÍTULO 1: INTRODUCCIÓN, OBJETIVOS Y MÉTODOS

La creciente preocupación vivida actualmente por optimizar el tiempo abarca todo tipo de actividades: desde planear una ruta para cumplir cierto itinerario en el menor tiempo posible, hasta la generación de códigos de programación que demanden poco tiempo del procesador en la realización de operaciones. En todas las formas de vida humana se hace manifiesta la "variable reina", el tiempo. Es en función de ésta que se define que tan bien o mal van las cosas.

La economía, la política, la mecánica, la astronomía y cada una de las disciplinas que ha desarrollado el ser humano tienen como parámetro al tiempo. Actualmente la capacidad de las empresas prestadoras de servicios como las de computación, bancos, comunicaciones, salud, etc. se catalogan a partir del tiempo de respuesta que puedan ofrecer, mientras éste sea más reducido y evidentemente el resultado sea correcto, la empresa tendrá mayor reconocimiento en su rama.

La automatización es un hecho en nuestro tiempo, toda actividad que quiera rendir en operación tiende a ser automatizada, lo que significa adaptar a su mecánica laboral un sistema autónomo capaz de realizar en el menor tiempo posible el mayor número de tareas. Actualmente estas tareas han crecido notablemente tanto en número como en complejidad, lo que exige optimizar tiempo y, por lo tanto, las características de dichos sistemas han tenido que mejorarse continuamente. Esto es fácil de comprobar poniendo como ejemplo una computadora personal (PC): los requerimientos mínimos de hardware para poder utilizar paquetería de actualidad, que es la demandada por el ritmo de vida actual, rebasan hasta en un 200% lo que era el mínimo necesario hace 1 año. Si en algo tan simple como una PC para trabajar en casa se requiere de

mayores velocidades en el procesamiento, es fácil imaginarse como se incrementan los requerimientos de este tipo en instituciones o empresas que atienden a mas de 60 personas.

La capacidad de realizar un gran número de operaciones en períodos cortos de tiempo está relacionada con las características del procesador utilizado y la forma en la que se programa su operación. Aún cuando el mejor programador realice las rutinas de trabajo del procesador optimizando tiempos, existen infinidad de procesos que demandan mayor velocidad de cálculo que la que nos pueda entregar incluso, una supercomputadora. Para conocer este tipo de procesos no es necesario ser un experto en programación o en el uso de computadoras, basta con dedicarle un poco de tiempo a razonar cada una de la cosas que hacemos día con día. Imaginemos que leemos un libro u observamos un programa de T V, no resultaría difícil tomar una taza que esté junto y dar un sorbo de café sin perder detalle alguno de la actividad primaria, o bien podemos modificar por completo la posición de nuestro cuerpo mientras seguimos atendiendo la T V o la lectura. Esto, aún tratándose de una actividad intrascendente por su aparente "simplicidad" para llevarse a cabo, engloba una enorme cantidad de datos y procesamientos previos y durante cada movimiento, para que cada uno de éstos se pueda realizar al mismo tiempo, sin desatender la operación de los demás procesos. Si algo tan común como acomodarse mientras se lee un libro involucra un número de operaciones hasta ahora desconocido, imaginemos la complejidad en la realización de operaciones y la distribución de éstas para un bateador de base-ball que en un período de décimas de segundo debe identificar la bola del resto del campo visual, definir si abanica o no a una bola que se aproxima a una velocidad cercana a los 160 Km./hr, sincronizar brazos, piernas, cadera, muñecas y aplicar la máxima fuerza en un preciso instante. Qué se puede decir cuando la tarea es procesar información de naturaleza abstracta como la música o las

matemáticas, cuánta complejidad hay en estos casos en los que no se trata de coordinar el movimiento del cuerpo, sino de asociar, procesar y dar secuencia a ideas.

Si este tipo de procesos no se ha podido igualar en ningún sistema diferente al cuerpo humano, se debe a que la operación y leyes que rigen a la unidad encargada de asociar, almacenar y coordinar información que controla nuestro cuerpo o sea el cerebro, es prácticamente desconocido en lo que se refiere a la constelación de redes neuronales que lo constituyen.

Objetivos

Esta tesis tiene como objetivo desarrollar una aplicación en hardware de redes neuronales para clasificación de señales neuro-eléctricas y, en general, para reconocimiento de patrones generados por sensores.

En este trabajo no se pretende, de ninguna manera, repetir con tecnologías contemporáneas, modelos ya realizados en los años sesentas y setentas. En otras palabras, el objetivo no es desarrollar un modelo electrónico de neurona, ni siquiera un modelo electrónico de una red neuronal pequeña. Todo lo contrario, se trata de aplicar las propiedades emergentes de las redes neuronales artificiales que poseen masividad tanto en el número de procesadores (neuronas) como en el de conexiones. Tal masividad puede implicar cientos de procesadores y cientos de miles de conexiones.

Aún cuando no deja de ser interesante reproducir un modelo básico de sinapsis o de neurona con dispositivos actuales, desde el punto de vista de la ingeniería no resulta muy práctico debido a que en las aplicaciones contemporáneas se hacen manifiestas las propiedades emergentes que han tenido las RNA's en los últimos años, como por ejemplo, el procesamiento en paralelo,

esta característica no se contemplaba como representativa de las estructuras neuronales hace 20 o 30 años y actualmente es un comportamiento característico con enorme relevancia para los procesos en que se aplican.

Actualmente las aplicaciones de las RNA en hardware son procesos complejos que engloban diferentes disciplinas, es casi imposible encontrar un CI comercial con una estructura neuronal que opere únicamente como oscilador. Por esta razón, el reproducir modelos clásicos de sinapsis o de neuronas individuales no pone en evidencia las capacidades de un RNA, por el contrario, proyecta un sistema extremadamente complejo y con capacidades reducidas para atacar problemas contemporáneos. Como ejemplo representativo de esto se puede utilizar una aplicación muy conocida en el campo de las RNA´s: La asignación de rutas para el "vendedor viajero". En este problema se pretende obtener el mayor provecho en la visita de 6 ciudades, cada una con "X" prioridad. Para la solución de este problema a través de una RNA se necesitan al menos 36 neuronas y 360 interconexiones (Hopfield, 1987). Este ejemplo da una idea de la enorme cantidad de neuronas y sinapsis que debe contener un CI para una aplicación más compleja como reconocimiento de patrones, o reconocimiento de imágenes.

Como podrá verse, el hardware contemporáneo de redes neuronales no está desarrollado para ejecutar tareas simples, la justificación de recursos para el diseño y fabricación de un CI neuronal depende de la aplicación a la que se dirija, por lo que es de esperarse que las tareas para las cuales se desarrolle un CI con estructura neuronal sean bastante complejas. Desafortunadamente una industria muy interesada en el desarrollo de esta tecnología es la industria militar, el infortunio no sólo se refiere a lo funesto de las típicas aplicaciones de esta

industria, sino también a la dificultad para obtener información referente a las tecnologías aplicadas y a las tareas específicas donde las RNA's son utilizadas.

El estudio y aplicación de las redes neuronales en hardware es un tema desconocido pero de reconocida importancia debido a que sólo logrando desarrollar este tipo de estructuras se podrá obtener todo el beneficio que ofrece esta tecnología y hablar de la generación real o cuando menos de una aproximación cercana a un sistema biológico, lo que proporcionaría una sólida plataforma para el desarrollo científico.

Mostrando el potencial de estos dispositivos se busca también hacer manifiesta la necesidad de señalar la importancia y trascendencia del estudio interdisciplinario así como de impulsarlo para poder comprender, aportar y equilibrar el mundo que nos rodea.

Métodos

En primer lugar, se realizará una revisión de la literatura sobre el tema.

En segundo lugar, se hará la clasificación de una base de señales neuro-eléctricas (potenciales de acción que son pulsos con una cierta forma de onda y una duración aproximada de 2 ms) simulando con DynaMind v3.0 una red neuronal en cascada entrenada con el algoritmo de retropropagación.

En tercer lugar, utilizando DynaMind v3.0 y NeuroLink v2.0 se emularán el neurochip Intel 80170NX y la tarjeta multineurochip EMB de Intel.

Por ultimo, la red clasificadora de señales se transportará al emulador de la red de neurochips 80170NX y se hará una verificación de la capacidad de generalización de la red neuronal propuesta.

En las etapas en que se presenten dudas, se utilizará INTERNET y correo electrónico para consultar con INTEL y con quien sea necesario.

En este trabajo se presenta en el capítulo 2 una breve descripción de la operación del cerebro desde el punto de vista funcional y como sistema de control. En el capítulo 3 se aborda el tema de las redes neuronales artificiales estableciendo, primero, la definición de una RNA a partir de los tres posibles casos de relación que existen entre las neuronas, es decir: excitación, inhibición e independencia. En el capítulo 4 se habla acerca de las características de operación de una red neuronal artificial y acerca del algoritmo de retropropagación y se mencionan algunas aplicaciones de las RNA´s. El capítulo 5 trata algunas de las RNA´s desarrolladas en hardware. En el capítulo 6 se mencionan y comentan las técnicas de desarrollo de circuitos para RNA´s, modelos tanto de neuronas como de sinapsis. El capítulo 7 habla acerca de las diferentes tendencias en lo que a la elaboración de RNA´s en hardware se refiere, esto es, a partir de los modelos y tecnologías aplicadas en el desarrollo de RNA´s se han identificado diferentes modelos que obedecen a la aplicación de determinada tecnología en un punto determinado de la red (pesos o neuronas). El capítulo 8 hace referencia a los sistemas neuromórficos y se dedica también a dar una explicación del sistema, desarrollado en base a esta tecnología, llamado el "visor- auditivo". El capítulo 9 presenta el comportamiento, así como las técnicas aplicadas para el entrenamiento de una RNA en software, todo esto realizado con el sistema Dynamind. En el capítulo 10 se vierten las recomendaciones emanadas de este trabajo y se discute lo que realmente se pudo hacer con los recursos disponibles y de acuerdo con los objetivos propuestos. Finalmente, se presentan un glosario de abreviaciones utilizadas en el texto y una lista de referencias consultadas y mencionadas, así como una lista de URL's consultados por medio de NETSCAPE.

CAPÍTULO 2: EL CEREBRO

"El hombre debería saber que del cerebro, y no de otro lugar vienen las alegrías, los placeres, la risa y la broma, y también las tristezas, la aflicción, el abatimiento y los lamentos. Y con el mismo órgano, de una manera especial, adquirimos el juicio y el saber, la vista y el oído y sabemos lo que está bien y lo que está mal, lo que es trampa y lo que es justo, lo que es dulce y lo que es insípido, algunas de estas cosas las percibimos por costumbre y otras por su utilidad. Y a través del mismo órgano nos volvemos locos y deliramos, y el miedo y los terrores nos asaltan, algunos de noche y otros de día, así como los sueños y los delirios indeseables, las preocupaciones que no tienen razón de ser, la ignorancia de las circunstancias presentes, el desasosiego y la torpeza. Todas estas cosas las sufrimos desde el cerebro" (Hipócrates: Sobre la Enfermedad Sagrada) (Smith, 1970)

Siendo esta porción del cuerpo de apenas 1.350 Kg., en promedio, la región más misteriosa del ser humano parece increíble saber que desde hace aproximadamente 2,500 años ya se consideraba como un elemento excepcionalmente complejo según lo muestra Hipócrates en sus escritos.

Existen muchas razones y principios que nos permiten suponer que la unidad básica de la vida es la célula, esto puede comprobarse fácilmente echando un vistazo a través del microscopio a cualquier organismo o tejido vivo. Pero resulta mucho más sorprendente el pensar que no sólo los tejidos como la piel o los que conforman los diferentes órganos del ser humano están compuestos por células, sino que también el sistema nervioso central, el encargado de interconectar todo nuestro cuerpo y darnos conciencia de qué y quiénes somos, está compuesto

también por estas unidades. Las células que componen al Sistema Nervioso Central (S.N.C.) corresponden básicamente a dos tipos: las células nerviosas o neuronas y las células neurogliales o gliales. En lo que respecta a las neuronas, es a éstas a las que se les atribuye todo el funcionamiento del S.N.C: conducción de los impulsos nerviosos, elaboración de la información sensitiva, determinación de los patrones apropiados a estímulos específicos, etc. En lo que respecta a las células gliales que, aunque se imponen en número, sólo se consideraba que su función era de soporte en la operación de las neuronas y tal vez alguna relación con su nutrición, sin embargo hay evidencias recientes de que juegan otros papeles.

Las células neuronales o neurosensitivas pueden clasificarse en diferentes tipos dependiendo de su morfología la cual se determina a partir de la función en la que está involucrada. En otras palabras, podemos considerar que todas las células nerviosas son células neurosensoriales altamente especializadas. En los mamíferos las células neurosensoriales se encuentran en la mucosa nasal, en los músculos y en muchos órganos internos. Esto nos lleva a pensar que cada célula neurosensorial responde a un tipo específico de estímulo: presión, color, olor, etc. Pero a pesar de la diversidad de funciones que llevan a cabo este tipo de células y a la enorme variedad de estímulos a los que pueden responder, en cualquier caso podremos hablar de ciertas especializaciones anatómicas:

La Zona Dendrítica que es sensible ante la presencia de cualquier actividad extracelular.

El Cuerpo de la Célula o Pericarión.

El Axón, que emerge desde el pericarión y permanece sin dividirse hasta:

El Teledendrón, que es capaz de segregar sustancias químicas específicas.

La figura siguiente muestra diferentes clases de células nerviosas que van desde el tipo más primitivo de célula conductora hasta la motoneurona y la interneurona.

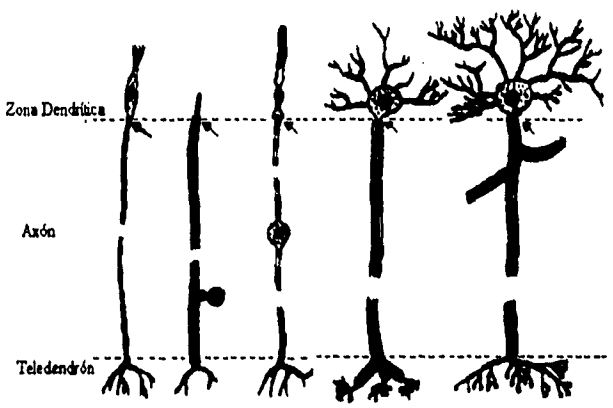


FIGURA 1.-. *Morfología de neuronas y células neurosensoriales*
(tomado de Smith, 1970 p. 82)

De esta forma se tiene que en los diferentes tipos de células nerviosas y neurosensitivas es posible ver que la terminología presentada puede generalizarse. Esta generalización fue propuesta en 1963 por D. Bodian (Tresguerres, 1992), quien lo estableció asignando nombres similares a regiones con funciones también similares.

Aún con la diferencia considerable que a simple vista se percibe en la morfología de estas células se tiene que en cualquier caso, la zona dendrítica responde a *alguna forma de estimulación* (esta estimulación es un impulso eléctrico producto del potencial producido por el intercambio de iones entre interior y exterior de la membrana de la célula, dicho potencial tendrá una intensidad determinada), la unión entre el pericarión y el axón (colina axónica) es el lugar donde se genera el *potencial de acción* o *impulso nervioso* (este potencial es el que indica la activación de una célula neuronal), el axón conduce *el impulso sin decremento* y el teledendrón o

botón sináptico segrega una determinada cantidad de sustancia neurotransmisora al llegar dicho impulso que genera la reacción correspondiente, excitación o inhibición dependiendo de la sustancia neurotransmisora involucrada, en la siguiente célula. Este proceso se lleva acabo a través de la llamada sinapsis.

El funcionamiento del cerebro ha sido a lo largo de la historia una de las grandes interrogantes del ser humano. Hasta hoy la investigación sólo ha logrado encontrar explicación de algunas características particulares y otras tantas generales. Si bien el conocimiento de la operación del cerebro humano a nivel anatómico está considerablemente avanzado, la investigación de esta porción del cuerpo humano en lo que se refiere a su operación en la generación de sentimientos, estados de ánimo, personalidad, conductas agresivas y en general como centro de asociación de ideas sigue teniendo grandes incógnitas.

Actualmente es sabido que cada movimiento del cuerpo humano encuentra su origen en el cerebro y es controlado por impulsos eléctricos, impulsos que varían en intensidad y tipo de reacción (excitación o inhibición) en la o las células incidentes. La razón por la cual un impulso eléctrico puede terminar siendo excitación o inhibición depende fundamentalmente de la sustancia neurotransmisora que se libere ante la presencia de éste. La ausencia o presencia no correcta de cualquiera de estas dos reacciones (excitación e inhibición) dan como resultado un comportamiento inadecuado del cuerpo humano o de alguna de sus regiones que le impedirá al individuo interactuar con toda su capacidad con el medio que le rodea.

2.1 El cerebro como un sistema complejo

Una de las grandes interrogantes que se tienen y que su respuesta sería de gran ayuda no sólo para la medicina sino para todos los campos que estudian al cerebro es conocer la función de la red de interconexión que existe, si no entre los 300 millones de células nerviosas que tiene el cerebro humano en su totalidad, sí al menos en alguna de las regiones que lo conforman.

La gran diversidad de tareas y funciones que puede realizar el cerebro, van desde la aparente sencillez del movimiento de un pequeño músculo o algún miembro del cuerpo hasta la conformación y manejo de conceptos tan complejos como lo es la personalidad. Los conocimientos, los sentimientos, la capacidad de razonar, la creatividad y todas las funciones que hacen al ser humano una especie intelectual muy superior a las demás no se podrían realizar si no existiera el arreglo de conectividad que tiene el cerebro humano, una neurona por si sola no podría servir para realizar ni la tarea más básica de un ser humano.

De hecho lo que somos y como somos está en función de la operación del cerebro, por ende, definido a través de la red que se forma entre las neuronas que componen al más complejo elemento del universo, el cerebro humano.

La red de comunicación existente en el cerebro de cada ser humano es precisamente la orquestadora de todos los elementos que integran al cuerpo humano, es decir gracias a esta podemos movernos, hablar, diferenciar sabores, formas, colores y olores. Aparte de permitirnos conocer el mundo, esta red hasta ahora desconocida es la generadora de todas y cada una de las herramientas que la tecnología pone a nuestro servicio y, creando un círculo, utilizamos estas para entender o aproximarnos a la operación de su creador, el cerebro.

Resulta que el organismo animal realmente comunica sus distintas partes por medios eléctricos, sin embargo el método adoptado difiere substancialmente del empleado en la transmisión de datos o de telefonía. Esto tiene que ser así porque los "hilos telegráficos" del organismo son conductores de electricidad del tipo cable submarino. Hodgkin calculó que un metro de axón humano normal tiene la resistencia eléctrica equivalente a la de 16^{10} kilómetros de hilo de cobre (Hodgkin, 1958 ; Smith 1970).

CAPÍTULO 3: REDES NEURONALES

Actualmente es claro que el comportamiento individual de la neurona no entrega información suficiente para pretender encontrar respuesta a interrogantes como la conducta, la memoria, los gustos, etc. Esto es sabido desde años atrás y de profundo interés no sólo para las áreas biológicas o neurofisiológicas, sino también para las matemáticas, la física y la ingeniería. Gracias a los experimentos en preparaciones biológicas del axón gigante del calamar se logró hacer por primera vez la inserción de electrodos en una fibra nerviosa y obtener así un potencial de membrana (Young, 1936 , Smith, 1970) esto demostró la presencia de niveles de electricidad en el impulso nervioso y permitió entrar a mas detalle en su investigación permitiendo que se desarrollaran modelos matemáticos para tratar de representar una neurona y posteriormente tratar de desarrollarla con componentes eléctricos.

3.1 La conectividad entre neuronas

Con la intención de aclarar un poco mas el concepto de red neuronal y la importancia de las características de su sinapsis en el comportamiento de ésta, se presentan a continuación los tres casos más importantes en una configuración de red simple que involucra un par de neuronas. Para ilustrar este ejemplo se utilizará el paquete de software *NEURORED* (Alcántara, 1992) desarrollado en el laboratorio de Cibernética. En este paquete es posible simular configuraciones de redes neuronales pequeñas y es una poderosa herramienta para el estudio de las mismas. La simulación de una red con *NEURORED* va desde la definición de la estructura de la red (numero de neuronas, interconexiones, tipo de interconexiones, etc.), la representación en el tiempo de la

ocurrencia de actividad de las neuronas (para facilitar dar una idea de la posible dependencia entre neuronas), hasta la obtención de un gráfico de correlación entre las neuronas que se deseen estudiar. La correlación cruzada es una herramienta altamente utilizada en la fisiología debido a que a través de esta técnica es posible determinar si existe dependencia temporal en la operación de células. Esta técnica consiste en un análisis de la actividad celular la cual es representada en un histograma bidimensional (para el caso de dos neuronas) en el que el eje de las abcisas representa el tiempo y el de las ordenadas el número de ocurrencias de intervalos temporales entre neuronas.

El interés de la fisiología en esta técnica se debe a que a través de ésta es posible conocer qué células están relacionadas con cuáles y qué tipo de relación guardan entre sí: excitatoria o inhibitoria. La correcta interpretación del histograma que entrega la correlación cruzada puede dar una imagen aproximada de la estructura que guarda el arreglo celular en estudio, es decir, permite conocer un poco más la conectividad funcional de la red celular en cuestión.

La interpretación del histograma consiste en determinar dentro de un intervalo de tiempo determinado la característica de dependencia o independencia en la activación de las neuronas a través de la forma de la gráfica que genera la actividad celular. Cuando el histograma tiene una forma plana, es decir, sin espigas que sobresalgan del resto de las activaciones, esto representa independencia de operación entre las neuronas involucradas. Si por el contrario, en la gráfica existe una espiga sobresaliente o bien una ausencia de disparos se habla de una relación excitatoria e inhibitoria, respectivamente (Espinosa, 1977). En las figuras de este capítulo se presentan gráficas de correlación en las cuales se aprecia y comprueba claramente el fenómeno mencionado.

3.1.1 Neuronas con actividad independiente

En primer lugar se presenta un ejemplo para un par de neuronas con actividad totalmente independiente:

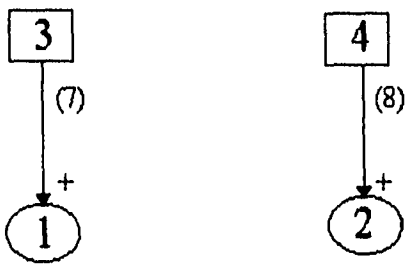


FIGURA 2.- *Configuración para dos neuronas independientes. Donde 1 y 2 son neuronas; 3 y 4 son fibras activadoras y la conexión va de 3 a 1 y de 4 a 2 para producir actividad "eléctrica" independiente en 1 y 2 .Los números 7 y 8 indican los pesos sinápticos.*

En el diagrama de la figura 2 se representan dos fibras (3 y 4) que excitan a las neuronas 1 y 2 respectivamente. La conexión entre las fibras y neuronas está representada por líneas que unen a los elementos involucrados, en esta línea está establecida la "intensidad o influencia" (peso) de esta conexión en la operación de las neuronas, en este caso particular los pesos son 7 para la conexión "3 a 1" y 8 para la conexión "4 a 2". Para este ejemplo no existe conexión entre neuronas con la idea de representar un comportamiento independiente entre éstas. En este caso en las fibras se estableció una probabilidad de disparo de 50 % para la 3 y 15 % para la 4. Esa probabilidad se verá reflejada en la actividad de las neuronas independientes tal como se muestra en la figura 3, donde la neurona 1 se dispara 53 veces y la neurona 2 sólo 25 veces.

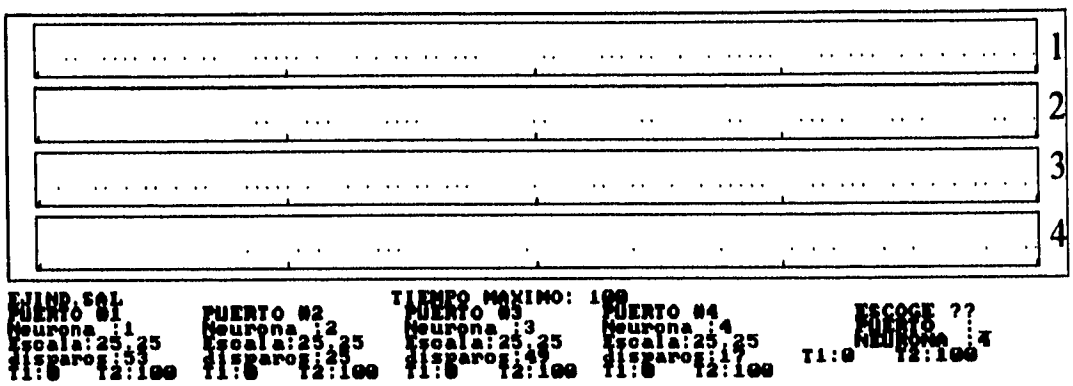


FIGURA 3.- *Diagrama de puntos de fibras y neuronas para el caso de neuronas independientes.*

Cada elemento, tanto neuronas como fibras, será identificado con el número que se le asigne en el diagrama inicial. En esta representación gráfica el tren de potenciales de acción o impulsos nerviosos de cada elemento se representan por puntos en función del tiempo (para este caso un período de 100 ms). Cada punto significa la activación (un disparo) del elemento correspondiente en el momento indicado, con este diagrama es posible darse una idea de la dependencia existente entre los diferentes elementos de la red. Este diagrama puede, en un momento dado, indicar que elementos están relacionados en lo que se refiere a su activación; en este caso se conoce de antemano qué elementos dependen de cuáles. Suponiendo que se desconoce el conexionamiento de la red es posible determinar esta dependencia funcional a través del histograma de correlación cruzada.(Espinosa, 1977)

En las cuatro gráficas siguientes se presenta la correlación cruzada entre los elementos de la red de la figura 2. El eje de las abcisas representa el tiempo y el de las ordenadas el número de ocurrencias. Con este tipo de histogramas es posible determinar si existe dependencia, y de qué tipo (excitación o inhibición), o no entre la activación de los elementos en estudio.

Para el caso de la figura 4 el histograma es prácticamente plano, es decir no existe una presencia o ausencia súbita de ocurrencias en determinado tiempo, lo que representa que no existe relación alguna entre la operación de ambas. Esto era de esperarse debido a que la actividad de cada fibra es completamente independiente una de la otra.

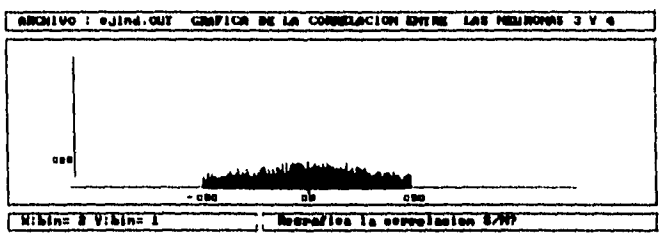


FIGURA 4.- *Gráfica de correlación para las fibras 3 y 4.*

En el caso de la figura 5 la gráfica presenta una espiga a la derecha del origen lo que representa un número mayor de ocurrencia en ese instante, es decir en determinado momento, a la activación de la fibra le sucede la activación de la neurona, esto es, que 3 excita a 1. Esto toma sentido al recordar que la fibra 3 está conectada directamente a la neurona 1 y esa sinapsis tiene características excitatorias.

FIGURA 5.- *Gráfica de correlación para la fibra 3 y la neurona 1.*

El comportamiento en la gráfica 6 presenta el mismo comportamiento que la inmediata anterior pero no es igual la densidad de disparos debido a que la actividad no es la misma, esta

diferencia se debe a que las fibras que excitan a las neuronas no tienen la misma probabilidad de disparo y que el peso en las conexiones es diferente.

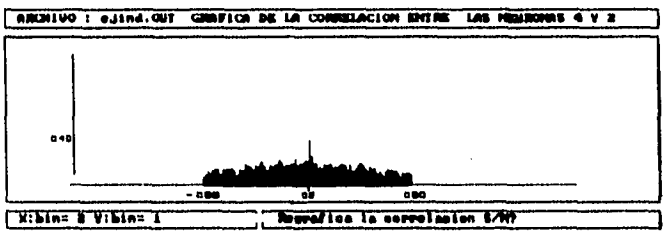


FIGURA 6.- *Gráfica de correlación para la fibra 4 y la neurona 2.*

Por último, en la gráfica de la figura 7 se presenta la correlación cruzada de las neuronas 1 y 2, como era de esperarse esta gráfica no arroja ninguna información que pueda indicar dependencia entre estas neuronas, al igual que en la figura 4 la correlación se hace entre elementos que operan de manera totalmente independiente por lo que el resultado es una gráfica casi paralela al eje de abcisas.

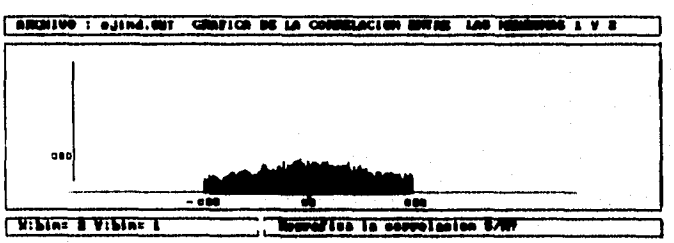


FIGURA 7.- *Gráfica de correlación para las neuronas 1 y 2.*

3.1.2 Neuronas con conexión directa: Excitación

Ahora se presenta una red con la misma morfología que la anterior pero con características excitatorias entre las neuronas 1 y 2.

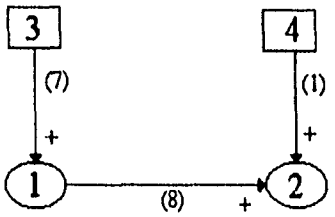


FIGURA 8.- *Configuración para dos neuronas dependientes con excitación de 1 a 2.*

Los valores de los pesos y de la probabilidad de disparo de las fibras se modificaron de acuerdo a lo indicado en la figura 8 con la idea de que las gráficas sean representativas del fenómeno, las probabilidades de disparo de las fibras 3 y 4 son 40% y 10%, respectivamente. Además, a diferencia del diagrama presentado en la figura 2, en este caso si existe una conexión directa entre las neuronas y se representa por la línea que va de 1 a 2, esta tiene características excitatorias con un peso de 8; las demás conexiones permanecen con las mismas características del ejemplo anterior. Una vez hecha la simulación se utilizan las herramientas de NEURORED (diagramas de puntos y correlación) para verificar el comportamiento de la red. Primeramente se obtiene el diagrama de puntos de la figura 9:

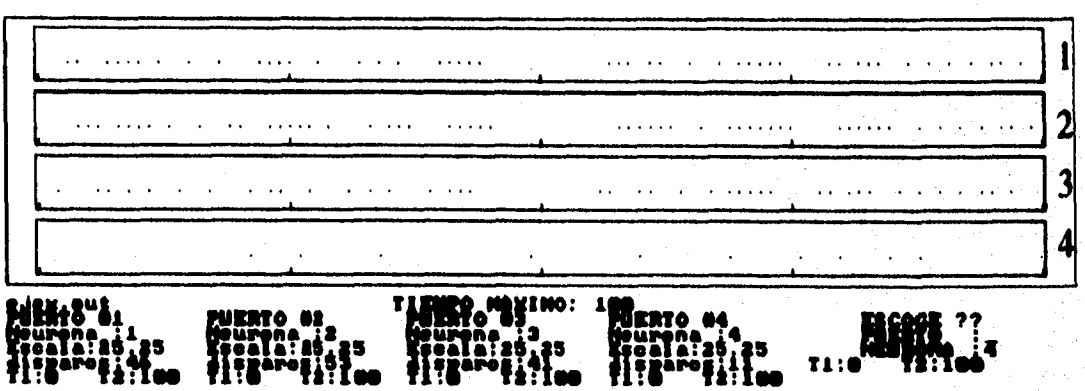


FIGURA 9.- *Diagrama de puntos de fibras y neuronas para el caso de neuronas con excitación.*

A diferencia de la figura 3 en este diagrama de puntos (fig. 9) se nota una activación diferente sobre todo en la neurona 2, esto debido a que ahora no sólo recibe excitación de la fibra 4 sino también de la neurona 1. Como se mencionó anteriormente, las características de las conexiones de la red (con excepción de la excitación) no han cambiado, por lo que el comportamiento entre los demás elementos permanece igual, por esta razón sólo se presenta la gráfica que permite comprobar el efecto excitatorio entre las neuronas.

FIGURA 10.- *Diagrama de correlación entre las neuronas 1 y 2.*

En la figura 10 es claro que existe una dependencia en la operación de las neuronas 1 y 2. La espiga, ligeramente corrida a la derecha del centro del histograma indica que existe una excitación de la neurona 1 hacia la neurona 2. En otras palabras, la neurona 1 intentará mantener activa a la neurona 2 independientemente de que algún otro elemento, en este caso la fibra 4, lo influya.

3.1.3 Neuronas con conexión directa: Inhibición

A continuación se presenta un tercer caso en el cual la relación entre las neuronas 1 y 2 es inhibitoria y las probabilidades de disparo para las fibras 3 y 4 son de 70% y 85%, respectivamente y con las intensidades indicadas en la figura 11.

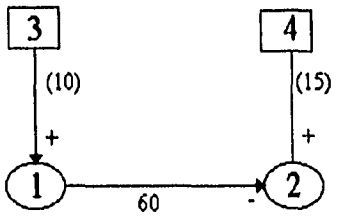


FIGURA 11.- *Configuración para dos neuronas dependientes con inhibición de 1 a 2.*

Es claro que la topología es la misma que para el caso de la figura 9. Sólo que ahora la sinapsis entre las neuronas 1 y 2 tiene características inhibitorias con un peso de -60 (el signo menos indica que es inhibición), esto se establece en la edición del archivo generado al dar de alta la red. A continuación se presenta el diagrama de puntos de esta simulación:

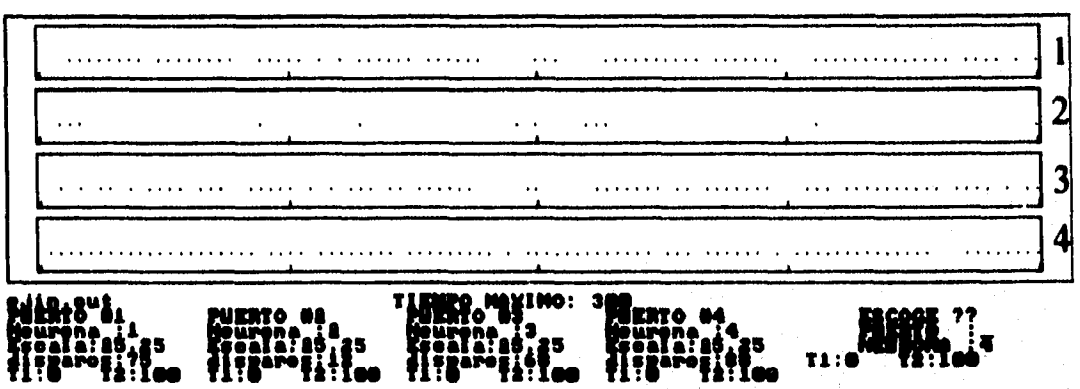


FIGURA 12.- *Diagrama de puntos de fibras y neuronas para el caso de neuronas con inhibición.*

La presencia de la activación de las neuronas debe cambiar obligadamente ya que a la neurona 2 no le llega el mismo tipo de información que en el caso de la excitación. Para comprobar la dependencia entre estas neuronas se muestra a continuación el diagrama de correlación entre 1 y 2.

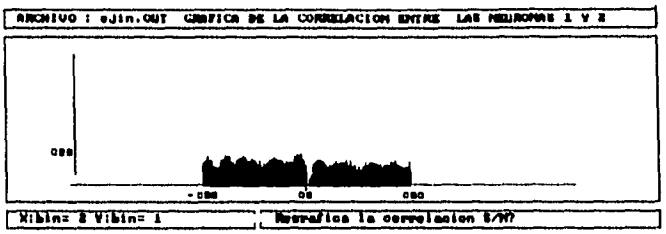


FIGURA 13.- *Diagrama de correlación entre las neuronas 1 y 2.*

En contraparte al ejemplo de la figura 10 ahora se presenta un hueco ligeramente corrido a la derecha del origen, este caso la correlación indica que existe una inhibición de 1 hacia 2, motivo por el cual la gráfica toma esta forma.

La trascendencia de los pulsos eléctricos y, por lo tanto, de la excitación e inhibición se debe no sólo a su presencia como tales, sino a la sincronización en su ocurrencia y la intensidad con que se presente cada una, es decir, que una excitación puede ser sucedida por otra excitación, un tren de excitaciones o bien por una inhibición o un tren de inhibiciones; en cada caso el impulso puede tener mayor, igual o menor intensidad. Esta secuencia, hasta ahora desconocida, varía para cada tarea que realiza el cuerpo humano, pero es un hecho que ocurre en la ejecución de cualquier tipo de operación desde una tarea común que demande el movimiento controlado de un brazo hasta una tarea abstracta como la realización de un modelo matemático. El equilibrio entre excitación e inhibición en regiones muy específicas del cerebro humano permite la operación ordenada del cuerpo. Por ejemplo, el sistema motor del cuerpo humano se comporta desordenadamente cuando no existen las inhibiciones debidas en el momento adecuado, este desorden se manifiesta por el movimiento involuntario de los músculos del cuerpo como ocurre con el mal de Parkinson o la epilepsia en el peor de los casos. Por el contrario, existen casos de parálisis en los que la movilidad de algún miembro o alguna región del cuerpo no existe aún

cuando se tenga la voluntad de hacerlo, esto debido a que no se generan o reciben la excitaciones en las regiones correspondientes. Así como es posible tener deficiencias motoras pueden ocurrir fallas en la percepción visual, espacial, asociación de ideas, etc. Todas producto de una incorrecta secuencia o intensidad en la excitación o la inhibición en una red neuronal.

Con los tres ejemplos mostrados anteriormente queda clara cual es la idea y la importancia de las características de la conectividad de una red neuronal. Evidentemente para redes neuronales complejas, las neuronas y fibras involucradas son mucho mas que dos, pero el principio de interacción es exactamente el mismo.

CAPÍTULO 4: LAS REDES NEURONALES ARTIFICIALES

En 1943 Mc Culloch y Pitts (Grey, 1961) publican algunos teoremas de modelos neuronales que establecen:

- La actividad de una neurona se define como un proceso "todo o nada".
- Cierto número de sinapsis se activan , durante un período de adición, con el fin de excitar una neurona. Este número de sinapsis es independiente de la actividad previa de dicha neurona.
- La estructura de las redes es invariante con el tiempo.

El resultado de estos teoremas fue el primer modelo matemático de neurona, que se llamó "Neurona Formal".

A partir de la aceptación de que la actividad de una neurona obligadamente involucra a otras, el concepto de *red* empieza a tomar fuerza. Este concepto se refiere a la conexión, con características muy específicas en la sinapsis que determinan el grado de dependencia y tipo de influencia que tiene la activación o inhibición de una neurona con la neurona vecina, de al menos un par de neuronas entre las cuales existe cierta interdependencia operacional. En 1949, Donald. O. Hebb (Hebb, 1949) define por primera vez un método para la asignación de esta "influencia" (que posteriormente llamará "peso sináptico") de acuerdo a la actividad neuronal. A este método se le conoce como "Sinapsis Hebbiana" y se rige por el siguiente postulado neurofisiológico:

"Cuando el axón de una neurona A está en contacto con las dendritas de una neurona B y la excita repetidamente, algún proceso metabólico toma lugar en ambas células. De tal manera

que la eficiencia de la neurona A como una de las neuronas que excita a la neurona B, se incrementa."

La consideración de elaborar sólo una aproximación del comportamiento de la sinapsis y de la neurona biológica en hardware a través de componentes electrónicos es obligada debido a la gran cantidad de factores y la compleja operación que existe entre ellos antes, durante y después de cada sinapsis. Para poder generar una RNA es necesario saber que elementos involucrados en el caso de la sinapsis biológica son imprescindibles y cuales no para la aproximación a la cual se pretende llegar. Es sabido que la estructura del cerebro se establece a partir de una muy alta densidad de interconexión (conexiones hechas a través de las dendritas), entre las neuronas que lo componen lo que significa que existe un gran número de canales de comunicación entre neuronas a través de los cuales viaja información.

Para definir el concepto de Red Neuronal Artificial (RNA) se retomará el esquema básico de red es decir, la conexión entre dos neuronas, sin importar ahora si se excitan o inhiben, como en el ejemplo mostrado en la figura 8. A partir de que existe una conexión entre dos neuronas se puede hablar de una Red Neuronal. En nuestro caso una RNA se obtiene al sustituir a la neurona y la sinapsis por componentes electrónicos configurados de tal modo que su comportamiento sea similar al de los casos mostrados en el capítulo 3, debido a que a partir de ahora la unidad básica de la red no es una neurona como tal, nos referiremos a estas como procesadores. Evidentemente la mayoría de las RNA contienen mucho más que dos procesadores con la idea de que el trabajo se distribuya entre todas en cuanto se le encomiende un proceso a la red. El número de neuronas, tipos de interconexión, algoritmo de operación y, en general, las características físicas de la red definen lo que se conoce como la arquitectura de la RNA. Cuando se diseña una RNA, definir que

arquitectura deberá tener va a depender fundamentalmente de dos factores: del algoritmo de aprendizaje y del tipo de tarea para la cual sea desarrollada la RNA, a partir de estos parámetros se definen casi automáticamente el resto de los elementos que definen completamente a la red.

4.1 El aprendizaje de una red

El aprendizaje, para el caso de una RNA, considerada como un sistema entrada-salida, es la capacidad de "***Asociar***" las características de una serie de datos presentados a la entrada de la red con una salida o respuesta específica. El aprendizaje de una RNA no comprende únicamente a la capacidad de identificar los patrones con los que se llevó a cabo el entrenamiento, sino también a la capacidad de clasificar una entrada con características diferentes a las que la red "ya conoce" en la categoría que le corresponda. Esto proceso se conoce como la capacidad de ***generalizar***.

Para entender cuando una RNA ha aprendido consideremos lo siguiente:

Sea un conjunto U_X de "n" vectores "P" con dimensión "E" donde cada vector representa un patrón de entrada: ($U_X = P_1, P_2, .. , P_{n-1}, P_n$) a cada patrón P_n le corresponde un vector V_n con dimensión "S" que indica el tipo o clase a la que P_n pertenece. Por ejemplo: Suponiendo que se tienen 4 caracteres diferentes representados en una matriz de 4x4 y nos interesa clasificarlos en 4 diferentes clases, una para cada uno, necesitaremos 2 bits para obtener estas 4 posibilidades de clasificación ($2^2 = 4$) y 16 bits para representar cada dígito. Por lo tanto los vectores Pn y Vn contienen 16 y 2 componentes respectivamente (E=16 y S=2).

Se dice que la RNA ha aprendido cuando al presentarle un patrón P_n, la red entrega a la salida la correspondiente clase V_n, es decir, cuando asocia correctamente P_n con V_n. Y que,

además generaliza, cuando al presentarle a la red un patrón P que no pertenezca al conjunto Ux, la RNA entregue una clase V_n que indique que el patrón de entrada tiene características de él o los patrones que corresponden a dicha clase, es decir, cuando se logre cumplir la generalización.

El proceso fundamental para conseguir que la red aprenda es la asignación del valor de cada uno de los pesos sinápticos de los procesadores que conforman la RNA. El proceso de búsqueda de estos valores es propiamente el "entrenamiento de la red" y consiste en ejecutar un determinado *algoritmo de aprendizaje*, este algoritmo busca mediante iteraciones el valor para cada peso tal que al iniciarse la operación de la RNA se activen o desactiven algunos procesadores de tal forma que la operación global de la RNA permita obtener la respuesta esperada, es decir, la asociación que define al aprendizaje. La decisión de qué algoritmo utilizar depende del tipo de datos que se procesen y del tipo de tarea que se quiera realizar.

Cuando a una RNA se le presentan por primera vez los patrones de entrenamiento y se ejecuta un algoritmo de aprendizaje, se dice que la red está en entrenamiento, es decir, el algoritmo empezará a operar en función de los datos de los patrones de entrenamiento hasta que se obtengan las condiciones establecidas por el algoritmo. En este momento el valor del peso para cada sinapsis estará asignado por lo que se considerará que la RNA está entrenada. A continuación se presenta una clasificación de redes hecha a partir de los parámetros mencionados (tipo de tarea y de datos), indicando en cada caso cual es el algoritmo adecuado para las RNA's en cuestión.

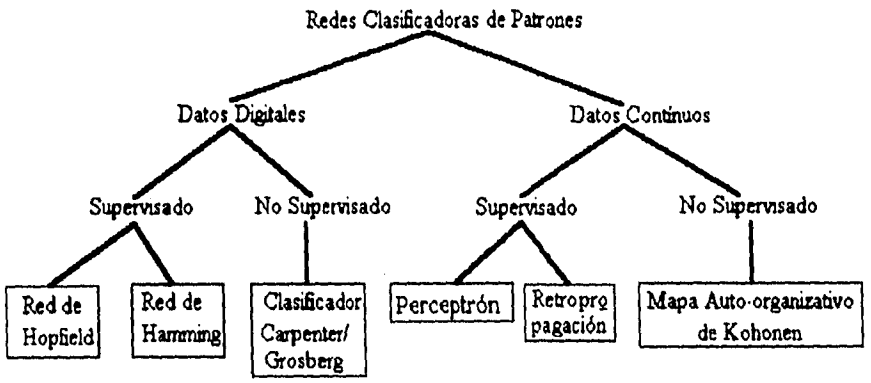


FIGURA 14.- *Agrupación de las seis redes clasificadoras más conocidas.* (Lippmann , 1987)

En la figura 14 se puede observar que la primera división se establece a partir del tipo de datos con que se trabaje, es decir, que los vectores de entrada que representan los patrones a estudiar pueden estar compuestos por valores digitales (binarios o bipolares) o analógicos, esto dependerá de la representación que se quiera dar al fenómeno. La segunda división se hace a partir del tipo de entrenamiento que recibe cada red, éste puede ser de dos formas: una en la cual la red asocia un conjunto de patrones de entrada a una salida que se conoce de antemano, conocido como aprendizaje supervisado. O bien cuando la red sólo recibe patrones de entrada sin que tengan una salida asociada, conocido como aprendizaje no supervisado.

Las redes de Hamming, Hopfield y Carpenter / Grossberg se han utilizado típicamente en casos en los cuales los fenómenos pueden representarse fielmente como entradas digitales (Lippmann , 1987) , por ejemplo imágenes en blanco y negro en donde la entrada es el valor de los pixeles o bien cuando se trata de textos ASCII en los que la entrada puede representarse por 8 bits. Estas redes resultan poco eficientes para procesos en los cuales el fenómeno se manifiesta con valores analógicos debido a que el patrón original debe ser convertido a digital. Por otra

parte, el resto de las redes mencionadas pueden operar satisfactoriamente con valores de entrada tanto digitales como analógicos (Lippmann, 1987), por ejemplo para procesar señales biológicas o reconocimientos de patrones auditivos entre otros.

4.2 Aprendizaje no supervisado

A diferencia del aprendizaje supervisado, en este caso no se tiene una salida predeterminada para el patrón de entrada a la red, la red recibe sólo las entradas e intenta agruparlas o clasificarlas por si sola. En las redes con este tipo de aprendizaje se lleva a cabo una competencia entre los procesadores de la última capa con el fin de que uno sólo sea el ganador. La adaptación de los pesos y otros parámetros de la red se hace en función de características simbólicas del conjunto de patrones de entrada para que así la red los clasifique en categorías similares. Las redes mas conocidas que operan con este tipo de aprendizaje son: Las basadas en la Teoría de Resonancia Adaptiva (ART) y los mapas organizativos de Kohonen.

4.3 Aprendizaje supervisado

Este tipo de aprendizaje depende de la asociación de una entrada a una salida que de antemano se conoce, es decir, para un patrón de entrada (P_n) debe existir un vector de salida (V_n) el cual representa su clase o tipo. Este aprendizaje se utiliza principalmente en redes que tendrán una aplicación de memoria asociativa o bien como clasificador de patrones.

Este proceso de aprendizaje opera, *grosso modo*, de la siguiente manera: El vector de entrada (P_n) es presentado a la RNA, ésta genera un vector de salida (V_S) el cual es comparado con el vector deseado (V_n) para el patrón de entrada, la diferencia entre estos vectores genera un

error determinado, los pesos de la red son modificados de acuerdo a un algoritmo específico buscando disminuir dicho error. Este procedimiento es repetitivo y el número de veces que se realiza es proporcional a la longitud del vector que representa el patrón de entrada y al número de patrones con que se entrena la RNA. El cálculo del error se hace hasta lograr que éste sea aceptable para cada uno de los patrones de entrada de la red.

Es inadecuado el tratar de determinar que tipo de RNA es mejor, la eficiencia de cada una depende del tipo de aplicación para el que se utiliza, de lo significativo de los datos con que una red se entrene y sobre todo de la interpretación que se tenga de los resultados que la red entrega. Para el caso de este proyecto el algoritmo de retropropagación resulta idóneo aparte de que se cuenta con las herramientas necesarias para hacer una buena representación de una RNA operando bajo esta arquitectura y algoritmo de entrenamiento.

4.3.1 Retropropagación

El algoritmo de retropropagación resulta una perfecta herramienta académica, además de sus cualidades en la práctica, para mostrar el proceso de aprendizaje de una red neuronal. La aplicación de este algoritmo se puede hacer en prácticamente cualquier problema que involucre un mapeo de patrones.

El algoritmo de retropropagación fue propuesto por Paul Werbos en 1974, en forma independiente por Y. Le Cun y D. Parker. La arquitectura de redes que utilizan retropropagación consiste de una capa de procesadores de entrada (con tantos procesadores como elementos tenga el patrón de entrada) , una o más "capas ocultas" y una capa de salida. Como se muestra en la figura 15.

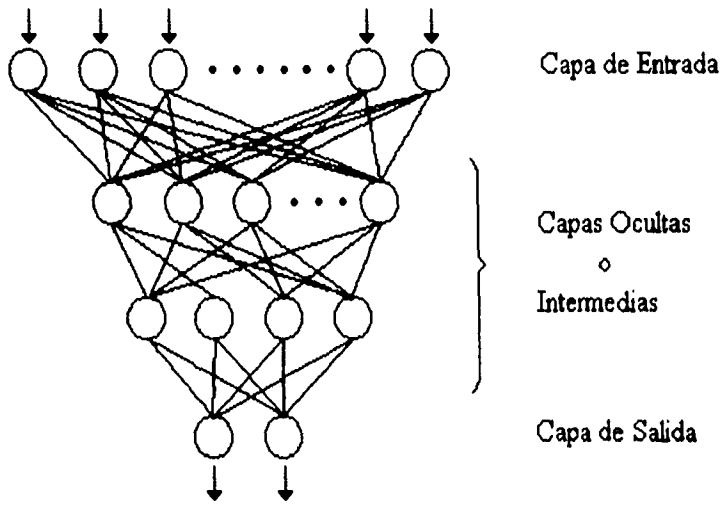


FIGURA 15.- Arquitectura *para una red con retropropagación.*

La conexión se hace sólo entre capas, no existen conexiones entre procesadores de un mismo nivel. Cada procesador debe estar conectado con cada uno de los procesadores del nivel siguiente y nunca se conectan con el nivel anterior, es decir tienen una conexión en cascada.

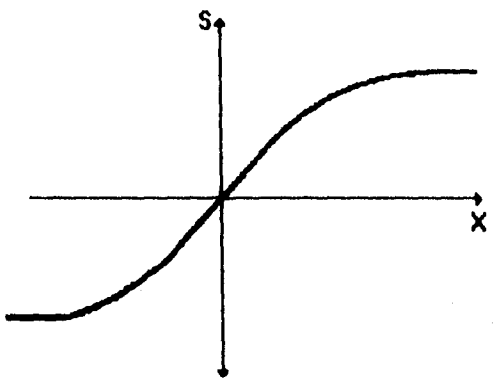
En las redes que utilizan retropropagación se consideran como niveles de procesadores las capas intermedias y la capa de salida; la capa de entrada queda excluida debido a que su operación no es propiamente la de un procesador, sino que sólo toma el valor del elemento del vector de entrada para entregarlo a la siguiente capa, por lo tanto el número total de procesadores "m" será igual a la suma de procesadores de las capas ocultas y las de la capa de salida.

El peso de la conexión entre las unidades *i* y la unidad *j* se denota como W*ij*. La operación del algoritmo se basa en el principio del gradiente decreciente y se logra de la siguiente manera: Primero se asignan valores arbitrarios a los pesos W*ij*, una vez establecidos estos valores se presenta a la RNA el patrón de entrada Pn, el cual es propagado por todas las capas

intermedias hasta la capa de salida para poder definir el nivel de activación de cada uno de los procesadores. Para lograr esto cada procesador realiza la suma de todas sus entradas y el resultado lo aplica a una función no lineal, típicamente en una sigmoide, aunque puede utilizarse cualquier no linealidad, para así generar el nivel de activación. La ecuación representativa de la función sigmoidal es la siguiente:

$$S = \frac{1}{1+e^{-x}}$$

La representación gráfica es:



Si denotamos como T*j* a la suma total de las entradas que tiene la unidad *j* entonces tendremos:

$$T_j = \sum_i U_i * W_{ij}$$

Donde U*i* es el nivel de activación de la unidad *i* y W*ij* representa el peso sináptico entre el procesador *i* y el *j*. La asignación del umbral de activación para el procesador *j* se obtiene entonces por:

$$U_j = \frac{1}{1+e^{-T_j}}$$

para lograr que la red converja en el menor tiempo se incorpora un procesador más en cada nivel el cual tiene por definición un umbral igual a uno, este parámetro es conocido como "umbral de bias".

Al presentar un patrón a la entrada de la red, ésta generará una salida Vs la cual será comparada con el valor deseado Vn y de esta comparación surgirá un error denotado por ε. Este proceso se hace para cada uno de los patrones de entrada y en forma repetitiva. El proceso de cálculo de error se representa esquemáticamente en la figura 16.

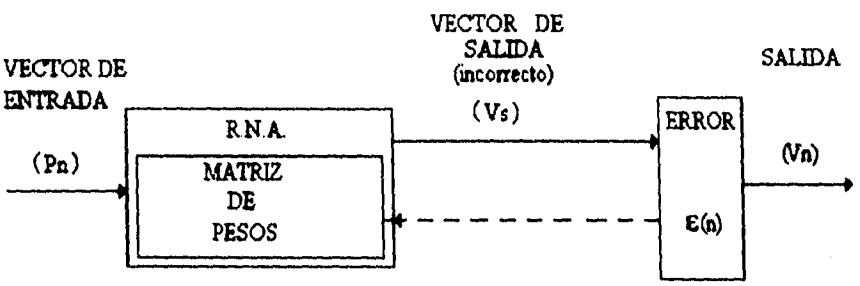


FIGURA 16.- *Esquema básico de aprendizaje por retropropagación.*

El valor para cada peso es calculado de la siguiente manera:

$$W(t+1) = Wij(t) + \alpha\delta jUi$$

donde $Wij(t)$ representa el peso de la sinapsis de la neurona i a la neurona j en el instante t. α es un factor de ganancia. δ es el error de j y Ui es el nivel de activación. El valor del peso converge mas rápidamente si se le agrega un término conocido como momentum, por lo que tendremos:

$$W(t+1) = Wij(t) + \alpha\delta jUi + \beta(Wij(t) - Wij(t-1))$$

donde $0 < \beta < 1$ (Lippmann, 1987).

El error para cada unidad de la capa de salida se calcula de la siguiente manera:

$$\varepsilon_S = (y_s - U_s) f'(T_s)$$

donde y_S es el valor del patrón deseado para la unidad S, U_S es el umbral de activación del procesador S y f '(x) es la derivada de la función no lineal. La cantidad de error la da la diferencia (y_S- U_S) en tanto que la derivada de la función sigmoidal escala el error para una corrección más significativa.

El error para cualquier unidad de las capas intermedias se calcula como:

$$\varepsilon_i = \left(\sum_j \varepsilon_j * W_{ij}\right) f'(T_i)$$

en este caso se realiza la suma de ε para todos los procesadores que reciben salida de la unidad i, con sus respectivos pesos. La derivada de la función cumple el mismo propósito que en la capa de salida.

La evaluación de la red se hace realizando el cálculo del error cuadrático, el cual tiene una tendencia a cero, aplicando siguiente método:

$$\varepsilon = \sum_{t=1}^{n} \varepsilon(t) = \sum_{t=1}^{n} \sum_{i=1}^{m} \left(\frac{1}{2}\right) (V_i(s) - V_i(c))^2$$

que resulta ser un caso particular del método de mínimos cuadrados donde se realiza la sumatoria del error de cada uno de los procesadores para cada uno de los patrones.

La figura 17 indica en que dirección se realizan los procesos mencionados dando una justificación clara del porque recibe este método el nombre de retropropagación.

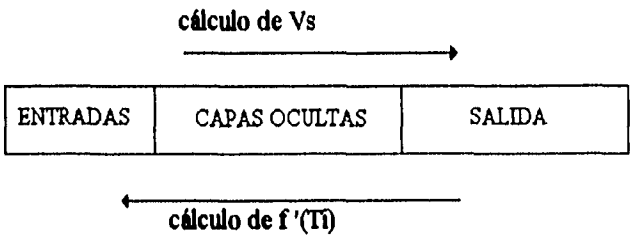


FIGURA 17.- *Flujo de información en una red con retropropagación*

Investigaciones acerca de las iteraciones del cálculo del error (Werbos, 1990) establecen que si el número de patrones de entrenamiento excede el número de pesos en la red es muy probable que el error oscile y no tienda a converger en cero.

4.4 Aplicaciones de las RNA

Las RNA han adquirido mucha importancia debido a la gran capacidad de procesamiento que prometen operando en paralelo, esto implica que la velocidad en los procesos encargados a una RNA, aunque sean muy complejos, es muy alta.

La investigación contemporánea sobre RNA se ha enfocado principalmente al desarrollo de algoritmos y simulaciones de redes a nivel teórico, mostrando una gran eficiencia en tareas como clasificación de patrones o memorias asociativas, por ejemplo.

Las simulaciones en software de procesos a muy alta velocidad está obligada a realizarse por pasos regidos por ciclos de reloj, lo que hace de este proceso algo demasiado lento para cumplir con los objetivos. Además de que para implementar un modelo neuronal es necesario trabajar en un gran sistema de ecuaciones paralelas y este modelo es simulado en procesos seriales, esto provoca que la ejecución sea demasiado lenta cuando se compara con estructuras

elaboradas directamente en hardware. Se han realizado simulaciones de grandes redes interconectando computadoras convencionales obteniendo como resultado un proceso ineficiente debido a la gran lentitud de procesamiento (Graf, 1988).

Evidentemente la manifestación del gran potencial que tienen las redes sólo se logrará cuando se implementen los modelos neurales desarrollados teóricamente en un hardware específico y concretamente en Circuitos Integrados (C.I's) especializados para este tipo de arreglos.

Este comportamiento de las redes lleva a pensar en el desarrollo necesario de nuevas arquitecturas de sistemas computacionales.

Actualmente existen aceleradores digitales que ayudarían a disminuir el tiempo de procesamiento pero aún con estos elementos no se logra acelerar significativamente. Varias investigaciones (Graf, 1988) han demostrado que implementando estos modelos con componentes discretos es posible estudiar ciertas características importantes de estos arreglos como por ejemplo su comportamiento dinámico, pero la ejecución con esta técnica arroja un sistema voluminoso y difícilmente útil para un aplicación real.

Los llamados sistemas neuromórficos son el resultado de una red neuronal inspirada en formas biológicas y que contienen software para realizar integración y clasificación visual, síntesis de lenguaje, reconocimiento de caracteres y análisis de riesgos. La elaboración en hardware de sistemas neuromórficos es una gran necesidad para aplicaciones en tiempo real como robótica, reconocimiento de patrones y diferentes tareas de control y aplicaciones de procesamiento de señales. Esta fabricación debe ser ideada de tal forma que cubra tanto la velocidad en la generación de los sistemas de ecuaciones y la utilización de arquitecturas paralelas.

La alta interconectividad que requiere una red neuronal permite pensar en utilizar medios ópticos de conexión para poder realizarla ya que así se pueden hacer un bloque tridimensional. Esto presenta una gran ventaja sobre las conexiones en dos dimensiones a las que orilla una superficie común para C.I. , desgraciadamente la tecnología óptica está aún en una etapa muy temprana de desarrollo por lo que no puede pensarse en una aplicación inmediata con estas características (Graf, 1988).

¿Por qué es tan importante la velocidad en los procesos que involucran redes neuronales? La respuesta la ofrecen los sistemas en los que se fundamenta esta filosofía de desarrollo, los sistemas biológicos. Como ejemplo podemos tomar la información visual. Esta información tarda en ser procesada unos cuantos milisegundos. La velocidad de esta respuesta tiene ventajas obvias en cuanto a las posibilidades de adaptación y supervivencia que exige un ambiente dinámico. La técnica que más se aproxima a la realización de estos modelos es la fabricación de dispositivos de propósito especial con una muy alta escala de integración (un Chip V.L.S.I.). Los sistemas neuronales sintéticos (SNS) basados en la tecnología VLSI, pueden entregar tiempos de respuesta similares a los mencionados, lo que los hace útiles para este tipo de aplicación. Por lo tanto sólo con el desarrollo de un hardware específico podremos obtener las velocidades de respuesta suficientes para implementar SNS.

En un sistema biológico, el proceso de aprendizaje y memorización de las células incluye modificaciones continuas en las características de comunicación entre ellas (sinapsis) atribuibles a la liberación y detección de sustancias conocidas como neurotransmisores, a la modificación de la morfología del sistema (síntesis de proteínas), y a cambios de la membrana celular que permiten intercambios selectivos de sustancias con el exterior, entre otros procesos. Algunos de los

mecanismos celulares involucrados en el proceso de aprendizaje mencionados anteriormente no sugieren una elaboración de estos con dispositivos electrónicos convencionales, sin embargo pueden realizarse sistemas, no tan robustos, que logren retener determinada información.

La necesidad de tomar como parámetro a los cerebros biológicos para el desarrollo de sistemas neuronales artificiales es obligada debido a que todo arreglo robusto que involucre procesamiento paralelo está inspirado en ellos. Esta comparación permite conocer datos que si bien indican lo lejano que se está de crear sistemas similares al cerebro, también motiva a implementar mejoras substanciales en las técnicas de generación de SNS. Por ejemplo, en términos de disipación de energía, la comunicación sináptica que es un proceso esencial del Sistema Nervioso Central (SNC), es aproximadamente 100 veces mas efectivo que cualquier computadora actual (Faggin and Mead, 1990). Esta diferencia se atribuye principalmente a que los arreglos operacionales del cerebro minimizan la longitud de las conexiones por sinapsis. Aunque numerosos axones conectan regiones distantes del cerebro, al final de cada axón existe una "ramificación axonal" utilizada para distribuir señales axonales a miles de sinapsis. La longitud de cada axón atribuible a cada sinapsis es entonces reducida por este factor de "fan-out". Esta división de "cable" y la minimización de cables locales dedicados es posible debido a que la mayoría de operaciones son concentradas en zonas cercanas.

CAPÍTULO 5: REDES NEURONALES DESARROLLADAS EN HARDWARE

Gracias al gran avance que ha existido en las últimas décadas de dispositivos y técnicas de encapsulado, la elaboración de hardware para RNA tiene un campo muy extenso. Estas pueden ir desde la elaboración de un circuito simple que genere una función no lineal (como se considera la operación de una neurona) hasta el diseño o fabricación de una RNA con arquitectura específica y con posibilidad de modificar su estructura. El objetivo en el trabajo de las RNA en hardware puede ser tan ambicioso como el tratar de generar un circuito capaz de reproducir alguna tarea básica del SNC con todos los factores externos que en este se involucran. La gran utilidad de una RNA no depende directamente de la fidelidad con la que represente al sistema biológico en el que esté inspirado, si así fuera aún no existiría aplicación alguna ya que en esta rama el SNC es prácticamente desconocido, sino de lograr obtener aproximaciones a las estructuras ya trabajadas en las que está visto que una correcta manipulación de la información permite optimizar infinidad de tareas.

La atención en este caso se centra en buscar configuraciones que permitan ser diseñadas y posteriormente aplicadas en un circuito integrado a gran escala VLSI. Ya que sólo este nivel, o uno superior de integración, pueden ofrecer una alta velocidad de procesamiento, bajo consumo de potencia, alta densidad de conexión y una aceptable aproximación al sistema biológico. O sea que haría de una RNA un elemento con características que lo colocan como una perfecta herramienta de aplicación y desarrollo.

Existen dos diferentes tendencias en lo que a la elaboración de un SNS se refiere, la primera que busca la ejecución de las ecuaciones de un sistema neuronal y la segunda que persigue la obtención del comportamiento de un sistema biológico en el hardware. La definición

de cual es la tendencia correcta depende de la rama de investigación para la cual se quiera elaborar la RNA, evidentemente para algún neurofisiólogo sería mucho mejor tener una representación de un sistema biológico para modificarlo a placer sin preocuparse por entender o comprobar el cumplimiento de ciertos algoritmos.

El principal problema de diseñar en VLSI un sistema biológico es que el mismo diseño limita en aspectos como la capacidad de conectar elementos a la salida y a la entrada de los dispositivos (fan-in /fan-out). Los sistemas biológicos imponen un alto requerimiento de conectividad que para tan sólo aproximarlos se debería pensar en una técnica de ultra escala de integración (ULSI).

Uno de los retos más grandes en el diseño de hardware de redes neuronales de propósito general es, como ya se ha mencionado, el nivel de interconexión requerido. La interconexión presenta problemas aún cuando se implementa en software. La limitante en este caso es la memoria disponible y la velocidad requerida para hacer tangible el beneficio de estos algoritmos.

Debido al panorama que se presenta con los requerimientos para el diseño de hardware es importante definir la técnica de acoplamiento del sistema en VLSI. Un factor fundamental en el diseño de dispositivos y de sistemas es que exista la capacidad de aislar cada dispositivo del ambiente de los demás componentes excepto de aquellos con los que esté prevista la ocurrencia de determinados efectos planeados a través de la interconexión de la red. Este problema llega a ser serio en el diseño de microsistemas en los que los componentes se empacan demasiado juntos.

La elaboración en VLSI se busca por tres principales razones: Primero, el tamaño de los dispositivos puede ser reducido a través de las técnicas de fabricación y con métodos de litografía. Segundo, el área necesaria para implementar una RNA con esta técnica es mínima. Tercero, la

"habilidad" del circuito puede reducir el número de componentes necesario para el diseño. Existe como limitante el antecedente de que el número de niveles de interconexión sólo se ha elevado un poco en los últimos 28 años, esto es, ha tenido un incremento de dos a cinco veces, lo que representa un gran compromiso para la aceleración en el estudio de técnicas de encapsulado.

La aceptación de que exista un grado de error o tolerancia en la elaboración del diseño de sistemas VLSI es una buena consideración para poder diseñar e implementar este tipo de sistemas. En circuitos que operan como memorias existen tolerancias para el ahorro de renglones y columnas y aun así son extensamente utilizados, lo mismo ocurre con microprocesadores y controladores.

El problema de utilizar arreglos de dispositivos con alto nivel de integración y acoplarlos con otros no radica en la elaboración de los arreglos sino en introducir un control jerárquico que logre ejecutar completamente los algoritmos necesarios, esto lo podremos ver en algunos ejemplos de fabricación.

En 1982 Hopfield publica una serie de fascinantes características y posibilidades del diseño y arquitectura en VLSI, muchas de sus ideas y aproximaciones habían sido ya expuestas por diferentes investigadores, pero Hopfield fue el primero en presentar esta información con tendencias claras hacia su ejecución en un chip (Hopfield, 1987) . Este chip compuesto de elementos analógicos no lineales altamente interconectados podría generar procesos muy eficientes, pero no necesariamente óptimos, en la solución de problemas computacionales complejos.

Existen diferentes métodos para lograr un acoplamiento entre los dispositivos que componen a un C.I , uno de los más promisorios es el autómata celular. En este arreglo cada

elemento tiene una conexión bidireccional con el resto de los elementos vecinos. Cuando se aplica un estímulo al arreglo, el estado de todos los elementos afectados evoluciona hasta lograr estabilizarse. La respuesta del autómata celular a una entrada dada puede ser predicha a través del acoplamiento entre los "vecinos" más cercanos y así obtener el comportamiento de una memoria asociativa, similar a una red neuronal. La diferencia entre un autómata celular y una red neuronal es que la segunda está densamente interconectada en distancias largas, lo que implica un serio problema para la disposición de espacio dentro del chip.

Actualmente los C.I. digitales disipan cerca de 10^8 veces mas energía por operación que lo que utiliza el cerebro (10^{-7} con 10^{-15} J por operación) (Zornetzer, 1990) . De esta energía un factor de 100 se atribuye a la energía irreducible para la operación de la compuerta de los dispositivos, un factor de 10^4 al número de dispositivos involucrados durante una operación y un factor de 100 optimizando el uso de cable. Es posible disminuir en un factor de 100 la potencia disipada por operación utilizando tecnología de silicio para reducir los niveles de alimentación y compactando las dimensiones de los dispositivos. También se puede mejorar el factor de cableado adoptando patrones de conectividad local en arquitecturas de VLSI. Esto se puede lograr a través de la técnica "escala-oblea". El diámetro de las obleas en producción actual son de aproximadamente 14.5 cm. En un futuro cercano el diámetro de estas obleas podrá ser de 25 a 30 cm. Durante este mismo periodo de tiempo se podrá disminuir el grosor de las líneas de conexión que actualmente van de 1.0 a 1.2 μm podrá ser de 0.3 a 0.4 μm. Estos avances implican que el nivel de integración pueda crecer de 500 millones de componentes por oblea (con 200 μm^2 por componente) a 10 mil millones de componentes en una oblea de 25 cm. con un diámetro de líneas de conexión de 0.35 μm.

Es claro que la gran importancia que tiene el realizar redes con componentes analógicos radica principalmente en dos cosas: en la velocidad y capacidad de almacenamiento y en la fidelidad en la representación de una red biológica en comparación con lo que puede entregar un circuito puramente digital. También es claro que las ventajas en el estudio de características dinámicas es mas factible en sistemas con componentes digitales ya que para implementar estas se tienen más herramientas

Es muy importante entender que el realizar un método analógico de resolución de problemas no necesariamente implica una técnica de diseño analógica. Es posible tener un diseño de bloques digitales (pueden ser elementos procesadores) para realizar operaciones elementales como sumas, multiplicaciones e integraciones y entonces conectarlo a un bloque analógico capaz de resolver ecuaciones diferenciales.

CAPÍTULO 6: TÉCNICAS DE DISEÑO DE REDES NEURONALES EN HARDWARE

Gracias a la interconexión y al procesamiento espacio-temporal que se hace en el cerebro al total de la información que recibe cada neurona a través de cada una de sus dendritas, se logra que esta información (visual, auditiva, sensorial, etc.) se procese bajo un esquema distribuido obteniendo así la inmediata capacidad de respuesta, hasta ahora inigualable por cualquier sistema artificial, que ofrece el cerebro en condiciones normales.

Una aproximación del procesamiento descrito anteriormente que, aunque es muy limitada en relación al comportamiento biológico, es aceptable para las disciplinas que se ocupan del estudio y diseño de redes neuronales y ha sido de mucha utilidad tanto para la investigación como para la industria se presenta en la figura 18.

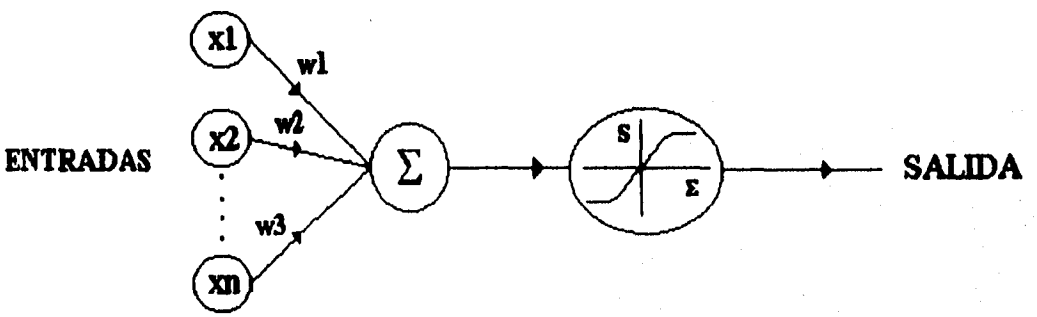


FIGURA 18.- *Diagrama del modelo artificial de neurona.*

En la figura 18 las entradas a la neurona están representadas por $x_1, x_2, \ldots x_n$, en el caso de los sistemas neuronales artificiales el valor de cada entrada se multiplica por el valor del peso sináptico Wn que corresponde a esa unión, la consideración del total de la información que recibe la neurona se hace a través de la sumatoria de cada uno de los productos y este resultado se

somete a la no linealidad de la función de transferencia que se aplique a la red; sigmoidal, binario o de umbral (cap. 4), una vez aplicada la no linealidad se define si la neurona se activa o no con excitación o inhibición.

A partir de este modelo de neurona es que la gran mayoría, si no es que el total, de grupos interesados en el desarrollo de RNA's, tanto de software como de hardware, han trabajado.

La figura 19 muestra el esquema de una neurona biológica y una neurona artificial, la célula neuronal típicamente puede tener hasta 10^4 conexiones de entrada y salida (Denker *et al.*, 1988), evidentemente esta densidad de conexión resulta una gran limitante para los sistemas artificiales..

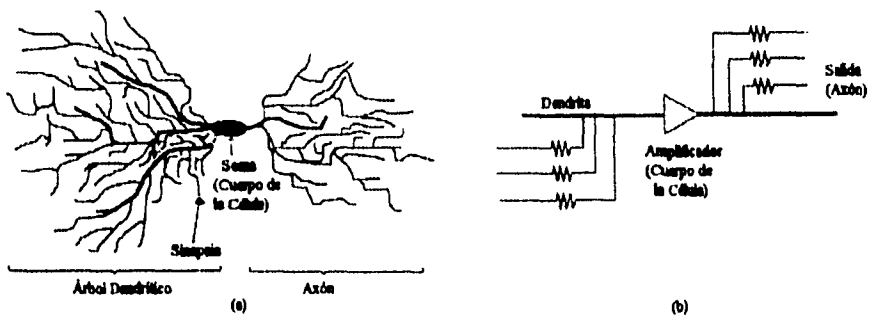


FIGURA 19.- *Estructura de la neurona: a) Modelo biológico. b)Diagrama esquemático del modelo electrónico.*

Debido a que el objetivo central es obtener una operación aproximada a la que se presenta en el caso biológico es importante describir, aunque no detalladamente, los procesos que se pretenden desarrollar en hardware. La aceptación del diagrama de la figura 19b como equivalente al de la figura 19a parte de las siguientes consideraciones: El proceso de sinapsis se considerará como el mecanismo de transmisión de información del axón de una neurona a las dendritas de aquellas con las que esté interconectada. En la sinapsis intervienen, neurotransmisores, iones y síntesis proteica entre otros elementos y procesos que al operar en concierto pueden producir un potencial de

acción, que transporta el axón sin decremento y se convierte en pequeños pulsos eléctricos positivos y negativos distribuidos sobre las dendritas que están conectadas al axón. Este proceso se lleva a cabo en el botón sináptico de las neuronas, el cual se representa esquemáticamente en la figura 20.

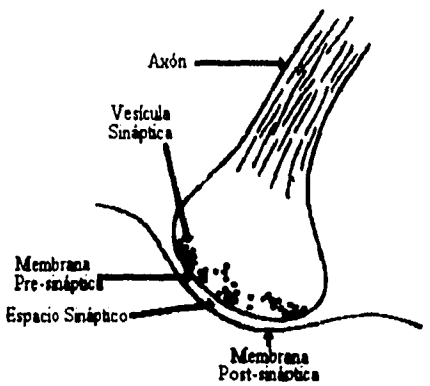


FIGURA 20.- *Diagrama esquemático de la sinapsis biológica.*

Por otra parte, cuando se habla de la activación o inhibición de una neurona, se refiere a la respuesta que tiene esta célula al generar los potenciales eléctricos producto de la sinapsis de cada dendrita que contenga la neurona, este potencial tiene características pasivas graduadas en cada una de las dendritas pudiendo ser positivo o negativo. La neurona de alguna manera obtiene el resultado global de todas estas señales y define a partir de esto si es excitada, inhibida o permanece sin cambio.

En el diagrama de la figura 19b el valor del pulso de la neurona es substituido por el valor de voltaje del amplificador y las conexiones sinápticas, con su respectivo peso, son representadas por conductancias. De esta manera un voltaje de entrada genera una corriente a través del cable (dendrita) en proporción al producto del voltaje de entrada y el valor de la conductancia

(sinapsis), por lo que amplificador deberá aplicar la no linealidad a la suma del total de corrientes que le inciden. Analíticamente esto es:

$$V_{out_j} = f(\Sigma I_j)$$

$$V_{out_j} = f(\Sigma (V_{out^i} - V_{in^j}) T_{ij})$$

Donde Vin y Vout son la entrada y salida de voltajes de un amplificador; I_j es la corriente que fluye a través de un resistor. T_{ij} es la conductancia del resistor que une al amplificador i con el j; y f es la función de transferencia. Esta ecuación es realizable utilizando componentes digitales utilizando un circuito multiplicador-sumador en la entrada de cada neurona resultando un circuito muy grande para una red con pocas neuronas.

La función no lineal que se aplica en una RNA puede ser principalmente de tres formas: Binaria, de Límites con Umbral o Sigmoidal. Aparentemente el sistema biológico opera con algo muy similar a la sigmoide, por lo cual se adoptará esta señal como la no linealidad aún cuando se dice que la elección de la no linealidad está en función del tipo de red con que se pretende trabajar (Foo, *et al.*, 1990).

La tendencia en lo que a la fabricación de RNA en VLSI se refiere es la utilización de componentes analógicos, la razón de elegir este tipo de componentes sobre los digitales se fundamenta en tres principales causas: La primera es la velocidad de procesamiento, que es muy superior en un dispositivo analógico que en uno digital. La segunda corresponde al amplio intervalo de operación por dispositivo que se puede obtener y por último el espacio requerido en un encapsulado para generar arreglos con dispositivos analógicos es menor que el que demandan arreglos digitales.

6.1 *Circuitos equivalentes de neuronas*

Existen varias aproximaciones en lo que a la generación de neuronas analógicas se refiere, pero todas tienen el mismo principio: La operación de la neurona se asume como la aplicación de una no linealidad al resultado de la suma de los potenciales sinápticos. La actividad de una neurona va a depender de la entrada que esta reciba, esta entrada va a esta compuesta por tantos valores como sinapsis haya en el árbol dendrítico, cada sinapsis a través del peso va a determinar la intensidad y tipo del pulso (excitatorio o inhibitorio) que aporte. Por lo tanto la entrada a cada neurona está en función de una suma de pesos, esto es:

$$Y_i = \sum_j W_{ij} V_j \qquad \text{---------- (1)}$$

Donde Y*i* es la entrada a la neurona *i*, W*ij* es el peso de la sinapsis *ij* y V*j* es el nivel de activación de la neurona *j*.

Generalmente el modelo de neurona pretende representar dicha suma como la suma de cada una de las corrientes que inciden en el nodo de entrada de la neurona y posteriormente someter este resultado a la función no lineal que determinará si la neurona se activa o no.

Un circuito que se utiliza como modelo electrónico de la neurona es el amplificador de transconductancia, cuyo diagrama esquemático, símbolo y operación se presentan en la figura 21.

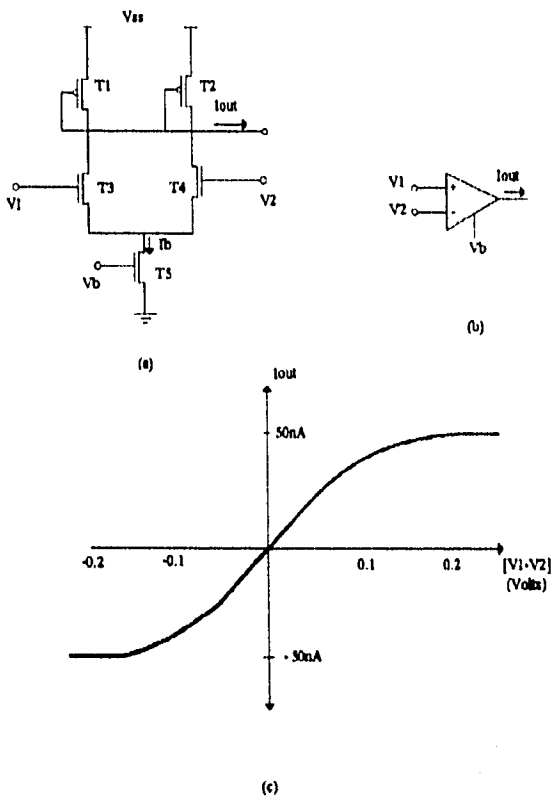


FIGURA 21.- ***Amplificador de transconductancia. (a) Diagrama esquemático del amplificador de transconductancia. (b) Simbolo del amplificador. (c) Comportamiento de la corriente de salida como función del voltaje de entrada***

En el circuito de la figura 21a los amplificadores T1 y T2 son *PMOS* y los restantes son *NMOS*. La corriente de bias I_b es activada por el voltaje V_b en T5 que a su vez controla la ganancia, V_b es menor que el voltaje de umbral de T5, lo que indica que dicho transistor opera bajo umbral, esto implica que I_b es una función de V_b. El comportamiento de la corriente de salida queda representado en la siguiente ecuación:

$$Io = I_b \tanh (V_1 - V_2)/2 = I_b \tanh V_{in}/2$$

La corriente de salida Io, es igual a la corriente de bias multplicada por la tangente hiperbólica de un medio de la diferencia de voltajes que aparece en sus entradas.

Este valor de corriente será transmitido a otra sinapsis que le aplicará un peso determinado y será parte de la entrada a otra neurona. La gráfica de la figura 21c muestra que sólo para diferencias entre V1 y V2 del orden de los 0.2 V ó menores es que el amplificador presenta una operación lineal, para diferencias de voltaje mayores la corriente se satura. Estos valores dan una idea de la magnitud que deben tener las señales con las que opera una RNA bajo esta estructura de neurona. La operación de T5 por debajo del umbral indica que el requerimiento de potencia de este dispositivo es mínimo, lo cual lo hace atractivo para su aplicación en RNA de VLSI.

Otra variante en lo que a la generación de neuronas vía hardware se refiere es la representación de la actividad de una neurona a través de la generación de pulsos, la frecuencia de los pulsos codifica el valor de la señal que representa al potencial de acción. Algunos autores consideran que esta operación es la que mas se aproxima al procesamiento biológico (Card, *et al.*, 1992). La neurona entrega una secuencia de pulsos en la cual está codificado el valor de dicha señal. En 1990 un grupo formado por Tomlinson, Walker y Silvioti demostró que la suma y determinación de la activación de la neurona que recibe los pulsos puede hacerse a través de compuertas OR. (Silvioti, *et al.*, 1990). Para el desarrollo analógico de esta neurona resulta imposible aplicar compuertas OR debido a que la presencia asíncrona de pulsos analógicos causaría problemas en la operación de las compuertas. Debido a esto y al problema que significa el limitado número de conexiones posibles a una compuerta (fan-in), la secuencia de pulsos analógicos es almacenada en forma de carga en un capacitor. Después, la operación de una neurona puede ser obtenida a través de circuitos integradores o bien por arreglos de

amplificadores de transconductancia, En cualquier caso, la secuencia de pulsos es convertida en información analógica a través de corriente eléctrica cuya magnitud y sentido son controlados por los pesos. La corriente que recibe la neurona es procesada y posteriormente convertida a voltaje para operar un VCO que entregará a la salida una secuencia de pulsos con una frecuencia determinada que indica la intensidad de la respuesta de esta neurona.

Un circuito que ha tenido mucha aceptación para aplicarlo en la fabricación de VLSI es el "Multiplicador Gilbert de intervalo amplio", el diagrama de este circuito se presenta en la figura 22.

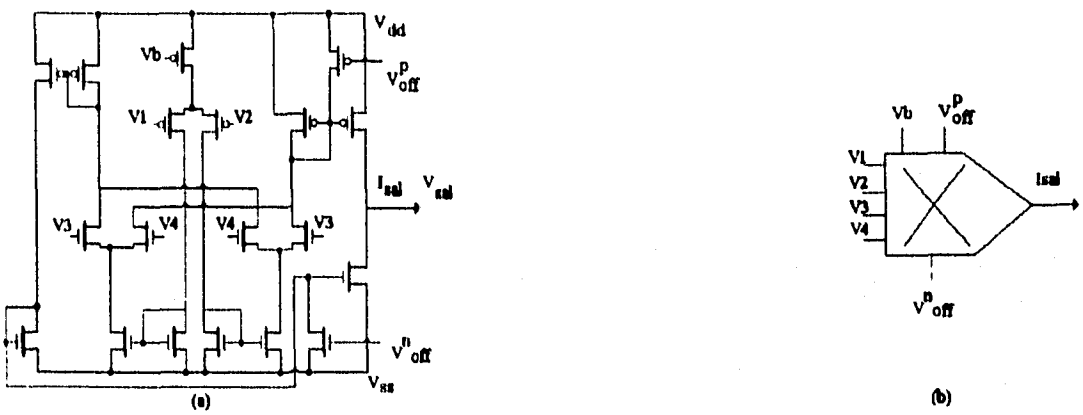


FIGURA 22.- *Circuito CMOS del multiplicador Gilbert de intervalo amplio. a) Diagrama esquemático. b) Símbolo.* **(Card, *et al.*, 1991).**

Aunque este multiplicador excede en número de componentes al amplificador de transconductancia y por lo tanto demanda mas espacio en su fabricación resulta una excelente representación en lo que a neuronas se refiere debido a que la corriente de salida que entrega es una excelente aproximación a la función no lineal tanh, las neuronas resultantes aparte de tener una ganancia ajustable representan una buena fuente de corriente, estas propiedades se muestran en la figura 23. Es importante notar que el intervalo de estas características es lo suficientemente

amplio para poder operar los transistores sobre el nivel de umbral, a diferencia del amplificador de transconductancia, aún cuando la potencia se incrementa, los requerimientos siguen siendo aceptables para un chip.

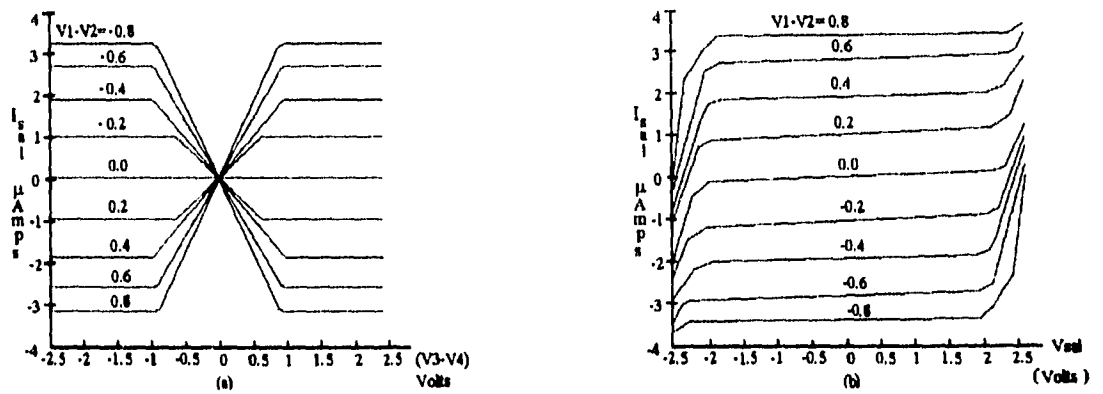


FIGURA 23.- *Gráficas del comportamiento del multiplicador Gilbert. a) Corriente de salida I_{sal} como función de V3-V4 para V1-V2 de -0.8 a 0.8 Volts. b) Corriente de salida I_{sal} Vs. voltaje de salida V_{sal} para V3-V4 = -2.5, V1-V2 de -0.8 a 0.8 Volts y Vdd=2.5V, Vss= -2.5V, Vb=0.9V*

6.2 *Circuitos equivalentes de sinapsis (pesos sinápticos)*

Para entender la operación de los circuitos equivalentes a los pesos sinápticos es importante saber como se representa analíticamente su operación. A partir del concepto de una RNA es posible determinar cual es la función de los pesos, debe quedar claro que esta representación analítica es sólo una muy simple aproximación de la operación biológica, pero al tratarse de una sistema artificial en el que las variables de interés son el voltaje y la corriente, esta es aceptable. La principal variable del aprendizaje de una RNA es el peso, por lo cual es preciso tratar de representar su comportamiento en forma analítica. El modelo de variación de peso que se pretende representar es el de la sinapsis Hebbiana, la cual establece una correlación entre

actividades presinápticas y postsinápticas, tratando de representar esto en variables eléctricas se tiene:

$$\Delta W_{ij} = \varepsilon X_j V_i \text{ ---------- (2)}$$

donde W*ij* representa el peso, ε es un patrón de aprendizaje y X*j* y V*i* son los valores pre y postsinápticos. Otra representación analítica es la conocida como regla delta generalizada que representa la relación de la actividad pre y postsináptica:

$$\Delta W_{ij} = -\varepsilon X_j \delta V_i \text{ ---------- (3)}$$

donde se agrega al final de la expresión el término δV*i* que representa el error.

El peso sináptico para una RNA pueden generarse en hardware a partir de tres diferentes principios: A) Pesos de valor fijo impuesto durante la fabricación. B) Pesos programables cuyo valor es establecido a través de la descarga de un registro a través de una RAM que puede estar dentro o fuera del chip. C) Pesos adaptivos, que son aquellos que se ajustan por el entrenamiento.

En el caso de las redes con sinapsis tipo (A) la aplicación de estas queda altamente restringida a un uso y caso determinado. El interés que tienen las redes compuestas con pesos fijos radica en que al realizar un chip en VLSI con estas características es posible estudiar el comportamiento dinámico de la RNA, principalmente cuando se trata de una con alta densidad de interconexión como lo es una de tipo Hopfield. Durante la fabricación de un chip con esta técnica la densidad de pesos sinápticos en una oblea puede llegar a los $4x10^8/cm^2$ (Graf and Jackel, 1989).

El resistor (T_{ij}) que establecerá el peso a la sinapsis de *i* a *j* no se incorpora durante el proceso inicial de fabricación del chip. Un orificio denominado "orificio vía" en la matriz de interconexión permite el acceso para conectar el resistor entre una línea de aluminio y una de

silicio. El resistor se incorpora esparciendo una capa amorfa de silicio sobre estos espacios vía,, la resistividad de esta capa está caracterizada a través del dopaje del Silicio. En la figura 24 se muestra la retícula de pesos para una red de pesos fijos, en esta red una matriz de conexión de 22x22 resistores ocupa una área de sólo 88x88 micrómetros.

FIGURA 24.- *Fotografía con microscopio electrónico de una red neuronal integrada.* (Tomada de Denker, 1987).

La complejidad para generar sinapsis y el espacio necesario para obtener cada elemento típicamente se incrementa de las de tipo (A) a las de tipo (C). El poder modificar el valor de los pesos de una RNA ya sea por programación o durante el proceso de entrenamiento hace que esta red sea más robusta en lo que a su campo de aplicación se refiere. En la figura 25 se presenta un diagrama esquemático de las sinapsis tipo (B) y (C).

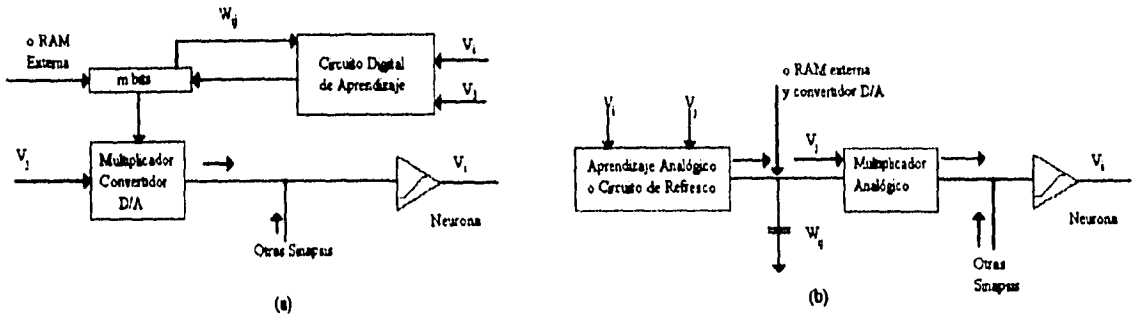


FIGURA 25.- *Diagrama de bloques para circuitos de control de peso.*
a) Peso programable. b) Peso adaptivo.

Considerando el principio de operación de los componentes, es de esperarse que un dispositivo analógico sea mucho mas sensible a cambios en el medio en que opera que uno digital (temperatura, voltaje de alimentación, humedad, etc.). Debido a esta consideración gran parte de los circuitos de control sináptico contienen elementos digitales aún cuando en algún momento se deba llevar a cabo la conversión D/A. Una buena justificación para aplicar circuitos digitales se tiene cuando la aplicación demanda una alta resolución en el valor de los pesos.

En las sinapsis tipo (B) el valor del peso es almacenado en registros binarios de la sinapsis, posteriormente este registro es multiplicado (a través de un multiplicador convertidor D/A que entrega una corriente analógica) por el valor de activación neuronal obteniendo así el término de la derecha de la ecuación 1 (pág. 48). Los multiplicadores convertidores emplean un número de transistores que se incrementa en proporción al número de bits significativos que contenga el registro, la contribución de cada transistor se va incrementando al doble en función del bit al que represente, esto hace que el área del chip se incremente exponencialmente con el incremento de la

resolución de los pesos. El diagrama esquemático de un multiplicador convertidor D/A se muestra en la figura 26.

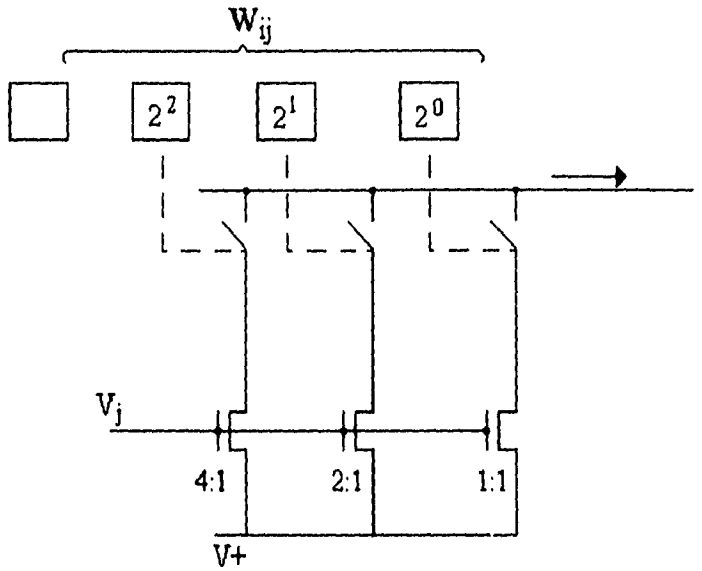


FIGURA 26.- *Multiplicador convertidor D/A de pesos positivos Wij a valores analógicos de activación determinados por Vj.*

Una alternativa al método de los transistores es utilizar un convertidor D/A que genere voltajes que serán almacenados en capacitores (caso de sinapsis tipo (C)). Esta variación se muestra esquemáticamente en la figura 25b. Con este método es posible minimizar considerablemente el área requerida para la sinápsis además de que utiliza un circuito analógico básico para su operación, lo que permite compactar más la elaboración de una RNA analógica. El principal problema de utilizar esta tecnología es que el valor de los capacitores cambia con el tiempo debido a las pequeñas fugas de corriente propias de los capacitores. Una solución a este problema es un circuito de refresco propuesto por Hochet (Abdo, *et al.*, 1991). Este circuito emplea un comparador de nivel para tener una lectura continua del voltaje del capacitor y en caso

de modificarse, activar un detector de fase para restablecer el voltaje. Otra solución es la presentación repetitiva de los patrones de entrenamiento, lo cual puede resultar en algunos casos poco práctico.

El problema de la variación de voltaje en un capacitor puede evitarse empleando elementos de almacenamiento no volátiles. Si una información analógica necesita ser almacenada por un periodo de tiempo corto (menor a 1 seg. p.ej.), un simple capacitor podría auxiliarnos en esta tarea. Generalmente la necesidad de almacenar información en este tipo de problemas requiere tiempos mucho mayores a un segundo, para este caso el capacitor no serviría debido a que la carga almacenada debería estarse recargando periódicamente. El proceso de recarga es relativamente sencillo cuando se trata simplemente de cargas binarias pero la complejidad se incrementa cuando los niveles de almacenamiento empiezan a crecer. Una solución a este problema puede ser almacenar la información en forma binaria para posteriormente realizar un conversión digital-analógica cuando sea necesario, de hecho esta técnica fue utilizada en una RNA experimental por Berger y colaboradores (Berger, *et al.*, 1987).

El único mecanismo fisico disponible en tecnología convencional VLSI con la propiedad de retener información durante un largo periodo de tiempo (hasta cientos de años) es la compuerta flotante de polisilicio. Una compuerta flotante es una pieza de polisilicio que no esta fisicamente conectada a nada y todo su alrededor esta cubierto por una capa de óxido térmico (SiO_2). Debido a que este óxido es un muy buen aislador cualquier carga que se encuentre en la compuerta puede permanecer por tiempo indefinido (aprox. 300 años a 50° C). Si la compuerta flotante se utiliza también como control de un transistor MOS, entonces podemos conocer la cantidad de carga almacenada en la compuerta a través de la corriente que fluye por el

dispositivo, así obtenemos un método de lectura de información. Para almacenar información es necesario aplicar electrones a la compuerta flotante atravesando el óxido; existen dos mecanismos a través de los cuales se puede lograr este objetivo: Inyección de electrones y aplicando el efecto túnel (tunelizando). En la figura 27 se muestra la estructura de la compuerta flotante.

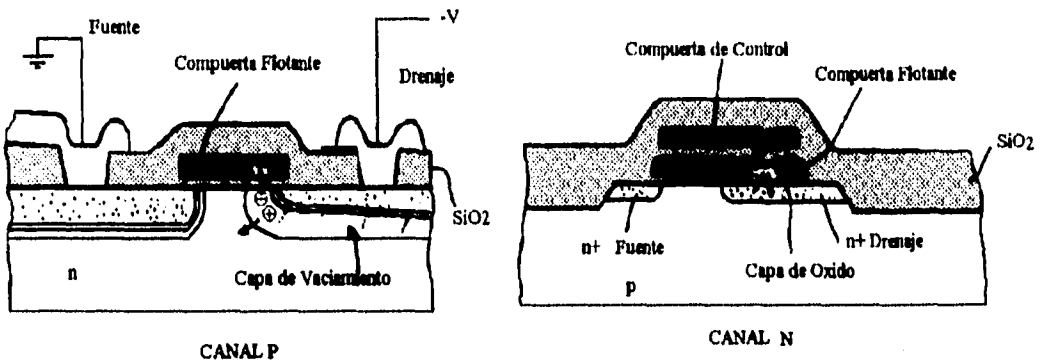


FIGURA 27 .- *MOS canal P* *MOS canal N.*

Para un dispositivo MOS canal P aplicamos el primer caso, este se logra generando electrones con alta energía cinética (llamados "electrones calientes"), para generar estos electrones es necesario aplicar un alto voltaje a la unión del drenaje, suficiente para generar una "condición de detenimiento" cuando se tiene este estado en esta unión, un pequeño porcentaje de portadores generan una región de vaciamiento, cerca de la superficie del dispositivo están los "electrones calientes", estos electrones tienen la suficiente energía para poder penetrar en la barrera de óxido y posteriormente encontrar la compuerta flotante cargando ésta negativamente. Es posible lograr la descarga de esta compuerta sometiendo al dispositivo a un campo electromagnético con suficiente energía (espectro ultravioleta) para acarrear a los electrones a través de la banda de conducción del polisilicio a la banda de conducción del óxido.

Para un dispositivo MOS canal N se utiliza el efecto de tunelizado para transportar electrones de y hacia la compuerta flotante. El dispositivo cuenta con una región muy delgada (de 50 a 100 A°) entre la compuerta flotante y el drenaje. Si se aplica un voltaje suficiente en la compuerta y el drenaje el campo eléctrico puede crecer lo suficiente para que se presente el efecto túnel y estos electrones atraviesen la barrera del óxido. Invirtiendo la polaridad del campo generado podremos obtener el efecto inverso, es decir sustraer electrones de la compuerta hacia el drenaje.

Las EEPROM's han tenido típicamente una aplicación de dispositivos de almacenamiento binario, pero ya se han empleado como registros de almacenamiento analógico en RNA (Holler, *et al.*, 1989). Una EEPROM emplea una compuerta flotante para poder almacenar carga que no sea volátil. La carga de esta compuerta es típicamente modificada a través de "corrientes tunelizadas" o bien puede establecerce el valor del peso a través de iluminación ultravioleta. La variación del valor almacenado es considerablemente menos probable que en el caso de un capacitor. Otra ventaja de utilizar dispositivos EEPROM en RNA's es la posibilidad de ajustar los voltajes de umbral en ambos modos de operación de transistor "vaciamiento y acrecentamiento" en el mismo circuito, esto permite una ejecución mas precisa de la multiplicación lineal requerida con sólo dos dispositivos EEPROM. Estos dispositivos (comp. flotante) pueden realizar tres funciones críticas esenciales para la fabricación de RNA's analógicas en VLSI. La primera es la autocompensación, la habilidad del circuito de continua y automáticamente compensar la inherente falta de precisión de sus elementos constitutivos. La segunda función es el aprendizaje, el aprendizaje no sólo implica leer, escribir y almacenar permanentemente una variable física analógica, sino que también involucra la habilidad de efectivamente modificar la carga en tiempo

real y de acuerdo con lo establecido por el algoritmo de aprendizaje. La ultima característica es la de poder cambiar la topología de la red. Las limitantes para esta tecnología son las características de programación del dispositivo EEPROM, los cuales requieren voltajes relativamente altos para su programación.

Al revisar la ecuaciones (1), (2) y (3) es evidente que tanto la activación neuronal como los pesos sinápticos contienen como operación fundamental la multiplicación analógica. Por esta razón existen configuraciones como el multiplicador de Gilbert que pueden aplicarse tanto en sinápsis como en neuronas.

A finales de la década de los 80´s diferentes grupos de investigación como Graf, Jackel y Hubbard de la Bell Laboratories 1986, Silvilotti, Emerling y Mead de Caltech en 1986 y Akers, Walker, Ferry y Grondin de Arizona en 1988 han diseñado chips con interconexiones programables. Esto es posible de implementar a través de la aplicación de transistores de cuatro pasos operando en la región óhmica (Akers, *et al.*, 1990).

CAPITULO 7: EJEMPLOS DE REDES NEURONALES DESARROLLADAS EN HARDWARE

En la fabricación de RNA existe una enorme separación entre los esquemas de investigación y los comerciales. Típicamente la fabricación de una RNA puede dividirse en tres tipos: analógica, digital e híbrida. Aún cuando en el capítulo anterior ha quedado claro que difícilmente se pueden conseguir circuitos puramente analógicos o puramente digitales es importante que una RNA especifique cuál de sus partes opera con qué tipo de señales.

En diferentes universidades y laboratorios industriales de Inglaterra, E.U.A. y Japón principalmente, se han desarrollado prototipos RNA en VLSI los cuales han tenido buen desempeño según se manifiesta en los artículos de divulgación. Desgraciadamente por tratarse de un tema de vanguardia con aplicaciones prometedoras que se ha manejado mucho en el campo militar y que muchos de los modelos desarrollados manejan patentes, es difícil encontrar una buena descripción y justificación de la mayoría de los circuitos desarrollados.

7.1 *Especificaciones de hardware*

La especificación básica de una RNA incluye la arquitectura de la red (multicapa, propagación hacia adelante, RBF, etc.), número de conexiones externas de entrada-salida I/O, número de neuronas, número de sinapsis por neurona, número de niveles, etc. Para el desarrollo de una RNA en hardware las especificaciones deben incluir datos más específicos, estos son: la tecnología utilizada (analógica, digital o híbrida), la precisión (en número de bits) en datos de I/O, de los pesos, de los acumuladores, etc. Existen varias formas de evaluar el desempeño del hardware, la forma más común de hacerlo es evaluando en número de conexiones por segundo

que realiza (CPS), número de conexiones actualizadas por segundo (CUPS) que es el valor que indica el número de pesos modificados durante el aprendizaje (en el caso de aprendizaje continuo). Debido a que existe una gran variedad en arquitecturas y técnicas de desarrollo de RNA en hardware hay casos en los que dos parámetros no dan una representación clara de la red, por esta razón se ha establecido la normalización del valor CPS con el número de pesos en el chip (CPSPW, o CPS por peso) valores que dan una visión mas general del la capacidad del chip, existen mas evaluaciones definidas a partir del parámetro CUPS. Normalmente el parámetro CUPS es aplicado para redes con aprendizaje por retropropagación aunque actualmente se ha aplicado para la evaluación de otro tipo de algoritmos, incluso existen redes como las RBF en las que los valores de CPS y CUPS no hablan mucho de las características del chip, en este caso los parámetros relevantes de evaluación se refiere al número de patrones presentados por segundo (P/seg).

A continuación se presentan algunas RNA's en VLSI que se han desarrollado, muchas de éstas han sido fabricadas por grupos de investigación de universidades y otras por grupos de empresas. En los casos que fue posible se explica el principio de operación general, costo o algunas otras características.

7.2 *Red con pesos fijos*

Inspirados en el trabajo de Hopfield un equipo de la Bell Laboratories diseñó, desarrollando y probando SNS que se componen de 22 y 256 neuronas. Ambos diseños utilizan resistores como elementos de interconexión los cuales no pueden ser modificados, el valor se establece en el proceso de fabricación. Ambos tienen disponibles salidas invertidas y no invertidas

para permitir la presencia de pesos positivos y negativos. El C.I. de 256 neuronas fue diseñado con dispositivos CMOS de 2.5 μm y contiene 512 amplificadores colocados en la periferia del encapsulado; el centro del integrado está reservado para la matriz de interconexión, a esta región le corresponde la mayoría del área del chip. Debido a que las 256 líneas de entrada-salida (I/O) necesarias resultan imposibles de incorporar al chip, éstas son multiplexadas con una longitud de 16 bits por palabra que va siendo almacenada en un buffer. El chip completo contiene alrededor de 25,000 transistores y mas de 130,000 sitios para resistor. El área del encapsulado es de 5.7 mm^2 (Graf, *et al.*, 1988). Si bien este chip no permite realizar modificaciones en los pesos sinápticos lo que podría verse, en algunos casos, como una desventaja, resulta una muy buena herramienta para el estudio de las características dinámicas de la RNA en el encapsulado.

7.3 *Redes con pesos variables*

Como se mencionó anteriormente, la generación de redes con pesos variables ha captado el interés debido a la grandes posibilidades que existen a través de esta técnica, de elaborar sistemas adaptivos en tiempo real.

A continuación se presentan dos tablas, la primera contiene RNA's con aprendizaje integrado en el C.I, con excepción de las últimas tres, desarrolladas por diferentes grupos de investigación. La segunda tabla contiene RNA's comerciales con diferentes configuraciones y tipos de aprendizaje.

Autores	Tecnologia	Área Sináptica	Aprendizaje	Tamaño	Capacidad
Allen, *et al.*	1.2 μm CMOS Pesos Digitales (5 bits)	10^5 μm^2	Boltzman y (CHL)	32 Neuronas 992 Sinapsis	10^8 CPS, CUPS
Fujita, *et al.*	0.8 μm CMOS Pesos Digitales (8 bits)	---	Hebbiano y Retropropaga-ción	24 Neuronas 576 Sinapsis	-----
Arima *et al.*	1.0 μm CMOS Pesos Capacitivos	5 x 10^3 μm^2	Aproximación binaria a Boltzman	336 Neuronas 28K Sinapsis	10^{12} CPS 3 x 10^{10} CUPS
Arima *et al.*	0.8 μm CMOS Pesos Capacitivos	32 x 10^3 μm^2	Aproximación binaria a Boltzman	400 Neuronas 40K Sinapsis	10^{12} CPS 10^{11} CUPS
Card, 1991	1.2 μm CMOS Pesos Capacitivos	40 x 10^3 μm^2	Hebbiano Analógico y CHL	Grupo pequeño de Neuronas	-----
Hochet *et al.*	2.0 μm CMOS Pesos Capacitivos con Refresco	30 x 10^3 μm^2	Kohonen completamente Analógico	Grupo pequeño de Neuronas	-----
Murray *et al.*	2.0 μm CMOS Ráfaga de Pulsos analógicos	13 x 10^3 μm^2	Externo	Grupo pequeño de Neuronas	-----
Holler *et al.*	1.0 μm CMOS EEPROM analógico	2 x 10^3 μm^2	Externo	64 Neuronas 8K Sinapsis	2 x 10^9 CPS
Chiang *et al.*	2-3 μm CCD	44 x 10^3 μm^2	Externo	144 Neuronas 2K Sinapsis	10^9 CPS

TABLA 1.- Prototipos de RNA's

Tipo	Nombre	Arquitectura	Aprendizaje	Precisión	Neuronas	Sinapsis	Velocidad
ANA-	Intel ETANN	FdFwd, ML	No	6b x 6b	64	10280	2 GCPS
LOGI- CAS	Synapics Silicon Retina	Neuromórfica	No	na	48x 48	red resistiva	na
	NeuraLogix NLX-420	FdFwd, ML	No	1-16b	16	Off-Chip	300CPS
	HNC 100-NAP	GP, SIMD, FP	Programado	32b	100PE	512K Off-chip	250MCPS 64MCUPS
	Hitachi WSI	Oblea, SIMD	Hopfield	9b x 8b	576	32K	138MCPS
	Hitachi WSI	Oblea, SIMD	Retro-Prop	9b x 8b	144	na	300MCUPS
	Inova N64000	GP, SIMD, int	Programado	1-16b	64 PE	128 K	870MCPS 220MCPUS
DIGI-	IBM ZISCO36	RBF	ROI	8b	36	64 X 36	250K pat/seg
TA-	MCE MT19003	FdFwd, ML	NO	13b	8	Off-chip	32MCPS
LES	Micro-Devices MD-1220	FdFwd, ML	NO	1b X 16b	1 PE	8	8.9MCPS
	Nestor /Intel NI1000	RBF	RCE, PNN	5b	1 PE	256x1024	40K pat/seg
	Philips Lneuro-1	FdFwd, ML	NO	1-16b	16 PE	64	26MCPS
	Siemens MA-16	Matrix ops	NO	16b	16 PE	16 x 16	400MCPS
	AT&T ANNA	FdFwd, ML	NO	3b x 6b	16-256	4096	2.1GCPS
	Belcore CLNN-32	FCR	Boltzmann	6b x 5b	32	992	100MCPS 100MCUPS
HIBRI	Mesa Research Neuroclasi-ficador	FdFwd, ML	NO	6b x 5b	6	426	21GCPS
DAS	Ricoh RN-200	FdFwd, ML	BP	na	16	256	3.0GCPS

TABLA 2.- Algunas RNA´s Comerciales

7.4 *Tecnología digital*

La tecnología digital en lo que a fabricación de RNA en hardware se refiere ha sido muy aplicada, para el diseñador la tecnología digital representa una serie de ventajas sobre cualquier otra, durante el proceso de fabricación, el almacenamiento de pesos en RAM, la ejecución exacta de operaciones aritméticas, la conservación de la precisión por el número de bits y los acumuladores son algunas de las herramientas digitales que hacen muy atractivo el diseño de redes a través de esta técnica. Desde el punto de vista del usuario una RNA digital encaja de manera relativamente fácil en la aplicación que así lo demande. La principal desventaja que tiene el desarrollar una red con tecnología digital es que la velocidad en la ejecución de una operación con técnicas digitales es menor que la obtenida con un circuito analógico, por otra parte, a cualquier operación digital le antecede una conversión A/D lo que implica involucrar más dispositivos y mayor tiempo. La eficiencia de la RNA digital en lo que al tiempo requerido para realizar un proceso se refiere, lo determina la aplicación, por lo cual el decir que un red es lenta o no, no depende únicamente de la tecnología con la cual fue desarrollada.

La tecnología digital en este campo se divide principalmente en tres sub-categorías que comprenden arquitectura en capas, SIMD y dispositivos en arreglo sistólico, y arquitecturas RBF, principalmente.

7.4.1 *Arquitectura de capas*

Siguiendo el concepto de capas de los procesadores digitales convencionales, el desarrollo de un chip a través de esta arquitectura permite construir bloques con el tamaño y la precisión

requerida, obteniendo velocidades de procesamiento moderadas y con aprendizaje "fuera del chip", esta técnica no representa un alto costo en la fabricación.

El C.I. MD-220 de Micro Devices fue realizado con esta técnica y se puede considerar como el primer chip de RNA comercial. Cada chip contiene 8 neuronas que operan con una función de transferencia (mencionada en el cap. 4) con umbral de límites binarios y 8 sinapsis de 16 bits con 1 bit de entrada. Utilizando multiplicadores en las sinapsis, el chip puede ejecutar aproximadamente 9MCPS. Es posible elaborar redes más grandes con mayor capacidad de bits de entrada utilizando un arreglo de chips.

Philips desarrolló el Lneuro 1.0, el cual cuenta con un procesamiento de 16 bits en el cual cada valor puede ser interpretado por las neuronas como 8-2 bit, 4-4 bit, etc. Este C.I. tiene la gran ventaja de contar con un caché interno de 1Kbite dentro del chip en el cual puede almacenarse el valor de los pesos. La función de transferencia se proporciona fuera del chip lo que permite obtener un producto sinapsis-entrada que facilite la interconexión de integrados.

7.4.2 *C.I's con multiprocesadores*

La tecnología de multiprocesadores adquiere su nombre por agrupar un gran número de elementos que realizan operaciones simples dentro de un encapsulado. Existen dos arquitecturas que dominan este tipo de diseño: Instrucción Única con Datos Múltiples (SIMD, de sus siglas en inglés) y arreglos sistólicos. Para el diseño SIMD cada procesador ejecuta la misma operación en paralelo pero para un dato diferente. En arreglos sistólicos un procesador ejecuta un paso del total del cálculo (siempre el mismo paso) el resultado de este cálculo pasa al siguiente procesador a la manera de "pipelined". La tecnología SIMD incluye chips como el N64000 de Inova y el 100-

NAP de Hecht-Nielson Computers (HNC). Existen sistemas que utilizan estos chips para elaborar un SIMD, un ejemplo es el sistema CNAPS de Adaptive Solutions que aplica el N64000 para este propósito. Este C.I. contiene 64 PE's y cada PE contiene un multiplicador entero de 9x16 bits, un acumulador de 32 bits y 4Kb de memoria (dentro del chip) para el almacenamiento de pesos. Todos los chips realizan la misma función y a través de control y buses de datos comunes es posible combinar la operación de varios C.I's.

El procesador con arreglo neuronal 100-NAP de HNC únicamente contiene 4 PE's pero puede realizar operaciones aritméticas de punto flotante de hasta 32 bits. Los pesos son almacenados en una memoria externa y permite la conexión en cascada para una red mas grande.

En cuanto a la elaboración de arreglos sistólicos, Siemens desarrolló el MA-16, este chip es una excelente herramienta para ejecutar rápidamente operaciones entre matrices (suma, multiplicación o resta) de 4x4 con elementos de hasta 16 bits, todas las salidas y acumuladores tienen una precisión de 48 bits. Los pesos y la función de transferencia se manejan externamente y el chip soporta la conexión en cascada.

7.4.3 *Arquitectura RBF*

Las redes con funciones de base radial (RBF) tienen una alta velocidad en el proceso de aprendizaje y la interpretación de su salida es simple. La mecánica de esta tecnología es muy similar tanto en arquitectura como en funcionamiento a la de retropropagación, la diferencia está en el tipo de operaciones matemáticas que aplica. Una red RBF es una red multiniveles, en los niveles intermedios se almacenan datos de vectores patrón o prototipo que serán comparados con los vectores presentados en el nivel de entrada de la RNA. La comparación entre estos vectores

arroja información que permite elaborar cálculos de distancia entre ellos y aplicando ésta para determinar la clase correspondiente del vector en cuestión.

Actualmente existen dos chips comerciales que operan con esta técnica: el ZISC036 (Zero Instruction Set Computer) de IBM y el Nestor Ni1000. En un área de 15.8 x 13.7 cm^2 el Ni100 alberga 3.75 millones de transistores, desarrollado con tecnología FLASH, EPROM CMOS, maneja una arquitectura con 256 elementos de entrada con una precisión de 5 bits cada uno, 1024 memorias de prototipo, una clasificación de 1 de 64 posibles categorías. Este encapsulado permite un entrenamiento dentro del chip utilizando tres posibles algoritmos: Restricted Coulomb Energy (RCE), Probabilistic Restricted Coulomb Energy (PRCE) y Probabilistic Neural Network (PNN). Aparte de poder operar con otros algoritmos pero en forma externa. El desempeño de este dispositivo es de 33,000 Patrones/Seg. y realiza hasta 16.5 GOPS (operaciones Por Segundo). La velocidad de operación es muy superior aún a la de una supercomputadora actual, para tener una idea más clara de la velocidad de operación se presenta la siguiente tabla:

COMPONENTE	GOPS
Ni1000	16.5
ETANN	4.2
PENTIUM	0.2
CRAY Y-MP/832	3.0
CRAY X-MP/416	1.5
CRAY X-MP/14-SE	0.4

TABLA 3.- Comparación de velocidad en giga operaciones por segundo de algunos C.I's. (Nestor Inc. 1995)

7.5 *Tecnología analógica*

La fabricación de un RNA en VLSI con componentes analógicos no resulta tan accesible como en el caso digital , una de las principales complicaciones en este tipo de diseños es obtener la obligada compensación que demandan los circuitos analógicos debido a las variaciones en el proceso de fabricación, de temperatura, etc. La generación de chips con esta tecnología ha evolucionado lentamente en lo que se refiere a su fabricación en serie, pero ha tenido un avance significativo en la investigación y desarrollo de componentes debido a que la obtención de un sistema con capacidad adaptiva (como el SNC) demanda la utilización de circuitos analógicos. El primer neurochip analógico comercial fue desarrollado por Intel, el 80170NX ETANN (Electrically Trainable Artificial Neural Network). Este C.I. es totalmente analógico, utiliza memorias EEPROM para el manejo de pesos. El valor de éstos puede actualizarse a una velocidad de 100K/Seg. El ETANN está diseñado para operar con retropropagación aunque no tiene aprendizaje dentro del C.I. El 80170 utiliza amplificadores Gilbert de intervalo amplio y cuenta con tres niveles de 64 neuronas cada uno totalmente interconectados, el C.I. puede operar con 128 entradas y 64 salidas o viceversa.

7.6 *Tecnología híbrida*

La tendencia del diseño híbrido es conjuntar las ventajas de un circuito analógico y de uno digital. De hecho la naturaleza de las aplicaciones de gran interés como reconocimiento de imágenes, con todo lo que esto implica, exige una integración de técnicas que permita optimizar la operación global del C.I. Un claro ejemplo de la necesidad de equilibrar arquitecturas analógicas y digitales es el problema del cableado que requiere una RNA, este problema puede solucionarse

con la aplicación de algún tipo de multiplexaje con técnicas digitales. Por otra parte, si el intercambio de información debe hacerse a grandes distancias la tecnología digital supera en mucho a las técnicas analógicas, aplicando multiplexaje por división de tiempo (TDM) se podría optimizar la operación de un circuito de esta naturaleza. Un sistema híbrido por lo tanto tendrá una composición de técnicas analógicas y digitales, en este chip las operaciones pueden estar realizadas por la etapa analógica mientras que la comunicación puede ejecutarse con mayor facilidad por la etapa digital. Esta técnica es prácticamente la que marca la pauta en el avance de la aplicación de las RNA.

En el campo de la investigación no comercial se han desarrollado un número muy interesante de prototipos con resultados alentadores, uno de estos casos bien documentado tuvo lugar en marzo de 1988 en la AT&T Bell Laboratories (Graf, *et al.*, 1988), donde desarrollaron un modelo de red conexionista tipo Hopfield. Las redes conexionistas se caracterizan por que cada procesador realiza una operación mínima, típicamente la detección de umbral a través del total de sus entradas, el resultado de la operación de la red dependerá de la interconexión entre los procesadores. La RNA consiste en un arreglo de 54 procesadores simples completamente interconectados con una matriz de conexión programable. La red está construida utilizando tecnología CMOS analógica y digital. El diagrama esquemático de esta red se muestra en la figura 28.

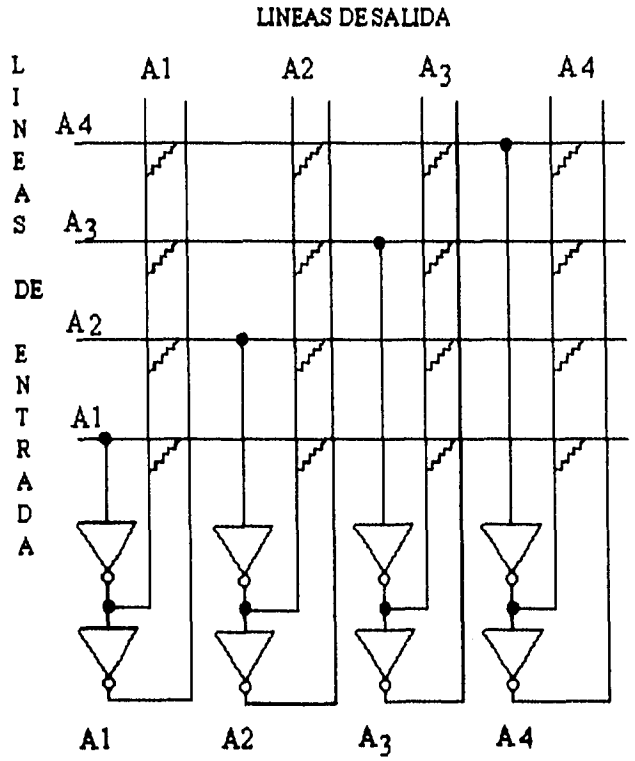


FIGURA 28.- *Diagrama esquemático de la RNA, los resistores representan la conexiones sinápticas entre las neuronas.*

En cada punto de intersección entre entrada y salida es colocada una sinapsis, cada una es programable. Dos amplificadores inversores unitarios son conectados en serie para poder obtener señales invertidas y no invertidas, de esta manera es posible controlar las conexiones excitatorias e inhibitorias. La operación de esta red puede hacerse en diferentes configuraciones programando las interconexiones entre los procesadores, para lograr esto se desarrolló el circuito que se muestra en la figura 29, el cual sustituye a las resistencias del diagrama esquemático:

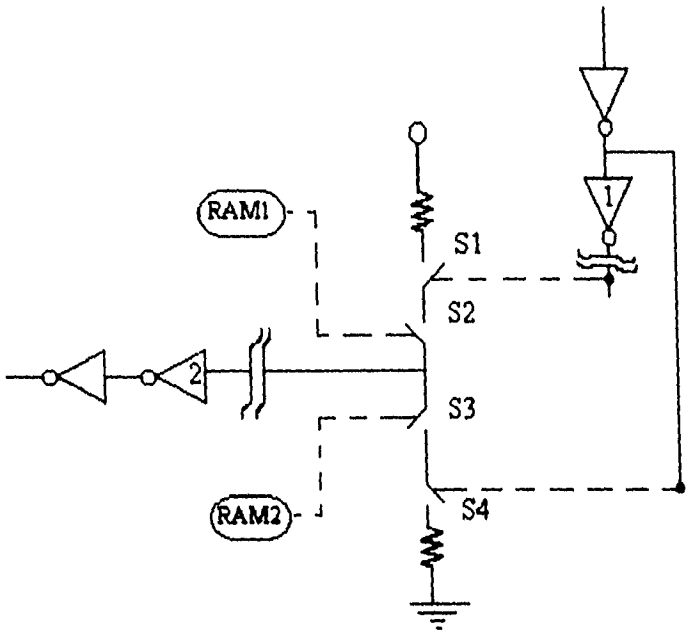


FIGURA 29.- *Diagrama esquemático del circuito para establecer el control de peso. Los interruptores S1, S2, S3 y S4 controlan el tipo de sinapsis, dos celdas RAM determinan el tipo de sinapsis.*

En este circuito la salida del amplificador 1 no se conecta a la línea de entrada del amplificador 2 sino que esta salida controla los interruptores S1 y S4 reduciendo así la carga del amplificador a la capacitancia de la línea de salida. Por cada conexión entre dos amplificadores se colocan dos celdas de memoria RAM que controlan los interruptores S2 y S3, el contenido de esta memoria determina el tipo de conexión. Una de tres posibles conexiones puede seleccionarse: Si un 1 es almacenado en RAM1, S2 está cerrado y la conexión es excitatoria. Si un 1 es almacenado en la RAM2, S3 está cerrado y la conexión es inhibitoria. Si ambas RAM almacenan un 0 no hay corriente en ninguna de las interconexiones.

El voltaje de entrada a un amplificador está determinado por la suma de corrientes debidas a todos los demás amplificadores que están en ese nodo. Por lo tanto el voltaje Vin*j* es un valor

analógico resultado de la contribución de cada amplificador conectado al nodo *j*. Este voltaje se ajusta al valor en el cual la corriente es cero. Debido a que la impedancia de entrada de un amplificador es muy alta se tiene:

$$\sum_{i=0}^{i=N} I_{ij} = \sum_{i=0}^{i=N} \frac{\Delta V_{ij}}{R_{ij}} = 0$$

donde I*ij* es la corriente que fluye a través de la resistencia del elemento de acoplamiento controlado por el amplificador *i*. V*ij* es el voltaje a través del resistor (Vin*j* - V_{DD}, Vin*j* - V_{SS}); R*ij* es la resistencia (R_+, R_-). Por lo tanto el voltaje Vin*j* es la suma analógica de la contribución de todos los amplificadores conectados al nodo de entrada *j*.

Los cálculos analógicos se realizan sólo en la matriz de conexión, las entradas, salidas y señales de control son digitales. Los datos de entrada y salida se transfieren a un registro donde una celda de memoria es conectada a cada amplificador. El dato de entrada primero se carga en l registro y después puede ser cargado en las celdas de memoria de la matriz de conexión o puede ser utilizada para inicializar el circuito. La inicialización del circuito se hace cargando los niveles de voltaje correspondientes al vector de entrada. Durante este proceso los amplificadores están apagados, cuando inician las operaciones se encienden y la red evoluciona hasta llegar a un estado estable sin ningún control externo o sincronización entre los amplificadores. Después de que se ha logrado un estado estable la salida de voltaje de cada amplificador es almacenada en el registro, el cual puede entonces ser leído.

Comprobaron el comportamiento de este chip utilizándolo en el reconocimiento de caracteres. El chip se conecta a una microcomputadora, la transferencia de datos puede ser hecha directamente por la memoria de la microcomputadora hacia el chip a una tasa que oscila entre

uno y dos mega bits por segundo. Esta tasa es limitada por la interfase y por la microcomputadora, no por el C.I. Un ciclo completo de procesamiento el cual incluye la carga del vector de entrada, la realización de los cálculos y la salida del resultado hacia la computadora requiere de aproximadamente 25 ciclos de reloj esto es de 25 a 50 microsegundos. La mayoría de este tiempo es utilizado en la lectura de datos de entrada y salida, el procesamiento en el circuito requiere de sólo un ciclo de reloj.

El proceso de reconocimiento opera con la siguiente secuencia: Una imagen se captura con una cámara digital que la normaliza a un tamaño de 128 X 128 pixeles. Luego la imagen se comprime a un tamaño de 16 X 16 pixeles en imagen binaria, después de esto la imagen es *"adelgazada"* - el ancho de las líneas es reducido a un pixel- y esta imagen adelgazada es identificada a través de un número representativo de sus características geométricas. La posición de estas características es comparada con las de entrenamiento y se determinará a cuál de éstas es mas parecida. De todo este proceso el adelgazamiento de la línea y la comparación de las características con las de entrenamiento es realizado por el C.I, el resto de las operaciones son realizadas por la computadora. La figura 30 muestra un ejemplo del resultado de la operación de adelgazamiento de línea.

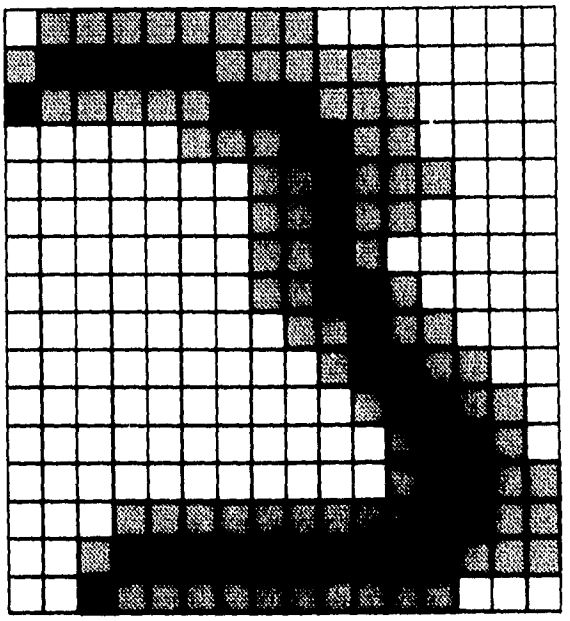


FIGURA 30.- *Resultado de la operación de adelgazamiento de línea en la escritura de un "3". Las líneas grises representan el carácter original y el área negra determina el resultado después de tres procesos de adelgazamiento* (Graf, *et al.* 1988).

El circuito contiene 75,000 transistores y 2916 elementos de acoplamiento en un área de 6.6 X 6.7 mm. Aproximadamente el 90% del total del área del chip es ocupada por la matriz de interconexión. Las pruebas de operación que hicieron fueron en tareas de clasificación de patrones y memoria asociativa. Los resultados que obtuvieron alientan a utilizar este chip como interfase en una PC y utilizarlo a este como un coprocesador en experimentos de reconocimiento de patrones.

7.7 *Otras RNA's*

Durante el desarrollo de este trabajo se llevó a cabo una búsqueda de RNA's en hardware, en algunos casos fue posible conocer el nombre de la compañía y de neurochip desarrollado. Como

complemento a los datos de la tabla 2 se presentan a continuación los nombres de los neurochips que se han desarrollado comercialmente. Estos chips no se incluyeron en ninguna de las tablas anteriores debido a que no se conocen características específicas, algunos de los neurochips presentados en este capítulo ya no son comercializados.

CNAPS-1016 y 1064	de Adaptive Solutions Inc.
FPIC	de Aptix Corp.
NLX 112 y 113	de American Neuralogix Inc.
ICMC (Intelligent Convolution Memory Chip)	de Oxford Computer Inc.
IPRMM (Intelligent Pattern Recognition Memory Module)	de Oxford Computer Inc.
MB4442	de Fujitsu Laboratories Ltd.
NISP (Neural Instruction Set Processor)	de Neural Technologies
NU32	de Neural Semiconductor
PVP16	de Meridian Parallel Systems Ltd.
RSC-164	de Sensory Circuits Inc.
SU32/32	de Neural Semiconductor
Neuro Chip	de Lockheed Missile and Space Corp.

En el capítulo final se comenta a cerca de las condiciones actuales de algunos de estos chips y compañías en lo que al campo de las RNA's se refiere.

CAPÍTULO 8: SISTEMAS NEUROMÓRFICOS

Si bien las técnicas de fabricación de SNS tienen líneas bien definidas de ejecución, la necesidad de obtener una RNA robusta encapsulada en un dispositivo ha obligado a grupos muy completos de investigación a desarrollar modelos innovadores, tal es el caso del grupo de Carver Mead en CALTEC (Mead, 1989), quien desarrolló un sistema integrado con un principio de operación neuromórfico. Un sistema neuromórfico consta de una distribución de elementos procesadores inspirados en los de un sistema neuronal biológico como es el caso del arreglo de conos y bastones en la retina del ojo. El arreglo de elementos procesadores se estructura de tal forma que cada elemento ejecuta una operación básica. La aproximación inicia con la identificación de varios niveles estructurales en el sistema nervioso con la tentativa de encontrar los principios organizativos de estas estructuras. La línea común de esta investigación es el uso de elementos adaptivos no exactamente como análogos a una sinapsis, sino como elementos de compensación de variaciones de dispositivo a dispositivo o de circuito a circuito, permitiendo lograr cierto grado de modificación o reconfiguración del hardware. Usando una aproximación neuromórfica es posible obtener tantos comportamientos como los logrados con técnicas digitales y analógicas.

El desarrollo de sistemas a partir de los modelos neuromórficos y basados en el principio de operación del concentrador seguidor han tenido asombrosos resultados, existen sistemas como la Retina de Silicio, el Censor Óptico de Movimiento , la Cóclea Electrónica o el sistema Visual Auditivo sistemas desarrollados por M.A. Mahowald, John Tanner, Richard F. Lyon, Lars Nielsen y Carver Mead. Más adelante se presenta la descripción general del sistema Visual

Auditivo el cual deja en claro la necesidad de desarrollar sistemas con procesamiento de señales en paralelo tal como ocurre en los sistemas neuronales.

8.1 *El concentrador - seguidor*

A continuación se presenta un ejemplo de esta nueva técnica en la que se utilizan amplificadores operacionales de transconductancia CMOS, descritos en el capítulo anterior, operando en un nivel inferior al del umbral para el diseño de esquemas neuronales en hardware. En este amplificador para la corriente de salida I_{out}, el voltaje es medido en unidades de KT /(qc), donde K es la constante de Boltzman, T es la temperatura absoluta, q es la carga del electrón y c es una constante dependiente del proceso. Para procesos la mayoría de los casos, KT /(qc) es cercana a 40 mv a temperatura ambiente. Para señales pequeñas el amplificador se comporta como una transconductancia con un valor, Gm, proporcional a la corriente de bias:

$$G_m = \delta I_{out} / \delta V_{in} = I_b\, cq / (2KT)$$

Para señales pequeñas , la salida de corriente está dada por:

$$Iout = G_m (V_1 - V_2)$$

Considerando que se tienen N amplificadores de transconductancia como se muestra a continuación:

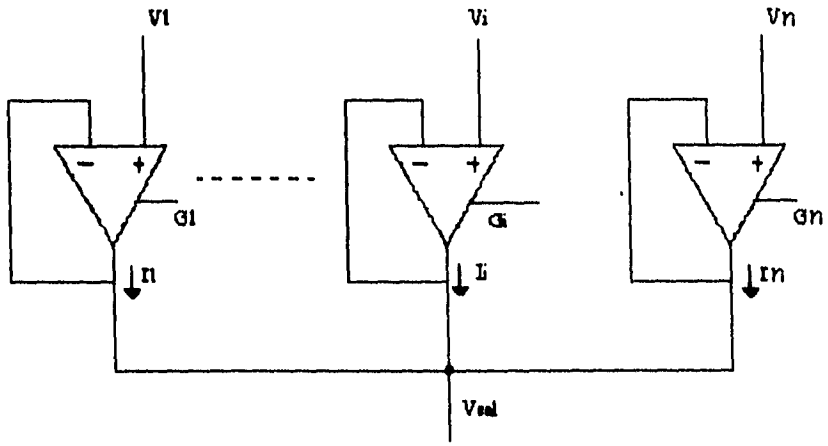


FIGURA 31 .- *Diagrama esquemático de un circuito concentrador-seguidor. Cada amplificador contribuye con una corriente I_i, proporcional a la diferencia de potencial de las entradas del amplificador V_i y el cálculo del voltaje de salida, V_o. Para un determinado (V_i - V_o), la corriente es proporcional a la transconductancia G_i del amplificador.*

Cada amplificador es conectado como un seguidor, llamándole seguidor porque la salida de voltaje es (sólo por algunos cuantos KT /(qc)) el voltaje de entrada.

Si el i^{th} amplificador tiene una transconductancia G_i y un voltaje de entrada V_i, entonces aplicando la ley de Kirchoff en el nodo de salida tenemos:

$$\Sigma_{i=1} G_i (V_i - V_{out}) = 0$$

Rearreglando la expresión para obtener el voltaje de salida:

$$V_{out} = \frac{\sum_{i=1}^{n} G_i V_i}{\sum_{i=1}^{n} G_i}$$

Este concentrador - seguidor calcula el promedio de los pesos de los voltajes de entrada $V_1, V_2, \ldots\ldots V_n$. La expresión inmediata anterior sólo es válida para cada amplificador que opera

en el régimen lineal. Se puede ver claramente que la transconductancia de un amplificador tiene como función de transferencia una "tanh", esto implica que tiene un comportamiento no lineal (una sigmoide) en el cual la corriente de salida está estrictamente limitada.

Por lo tanto, los voltajes de entrada muy diferentes del V_{out} actual, no tendrán gran impacto en el V_{out} más que sólo unos cuantos KT /(qc) diferentes a V_{out}. De esta manera, si se aplican muchas entradas, el valor de V_{out} ignorará a aquellas que no se aproximan mucho a este valor. El circuito concentrador seguidor provee un ejemplo muy simple de un circuito robusto: un circuito cuyas salidas no son afectadas por algunos datos equívocos o por alguna pequeña imprecisión de algún dispositivo. En base a este circuito se puede ilustrar el principio de operación de estructuras más robustas.

La transconductancia de cada amplificador en el sistema puede verse como la confidencia asignada (capacidad de modificar el valor de la salida, confianza en que se modifique dicho valor) a la entrada de este amplificador, por lo tanto la influencia de la salida de ese amplificador deberá tener consecuencias del proceso colectivo. Cualquier señal que no cambie en un largo periodo de tiempo no podrá acarrear información por lo que no se podrá asignar como una señal con alta confidencia. Si en el circuito mostrado se le asigna una confidencia a un amplificador en particular relacionado a la proporción de cambio de voltaje de entrada de este amplificador. Al instante en que se envía una señal errónea, este cambio podrá no ser diferenciado del dato real. Después de algún tiempo la influencia de esta señal decae, permitiendo el cálculo de otras entradas que estarán cambiando con el tiempo.

8.2 *La retina neuromórfica*

Un ejemplo de un diseño mas sofisticado utilizando aproximaciones neuromórficas es la retina de silicio en C.I. que realiza fotodetección y procesamiento de imagen en el mismo C.I., (Allen, *et al.*, 1988). El procesamiento de la imagen consiste en primero generar un voltaje proporcional al logaritmo de la intensidad de luz en la posición de cada pixel. Esta operación es entonces seguida del cálculo del centro y puntos vecinos, seguido del calculo de la orientación de la intensidad del gradiente local. Para cada posición de pixel cuatro valores analógicos son procesados: la intensidad de la luz, corregida por el calculo del centro y área vecina y la proyección en tres ejes del gradiente de intensidad local.

Este chip contiene aproximadamente 5000 pixeles y un número equivalente de procesadores. Este puede también realizar cálculos del orden de 1000 a 10,000 imágenes por segundo con un consumo de potencia de apenas 100 mW. Una aproximación del tiempo requerido para realizar esta operación utilizando técnicas de cómputo convencionales seria del orden de 5×10^8 a 5×10^9 operaciones por segundo.

Un aspecto esencial en los procesos neuromórficos es el uso de estructuras adaptivas capaces de compensar niveles de offset o incluso algunas fallas en el comportamiento de los dispositivos. Con estas características es posible obtener circuitos que cuenten con robustez, precisión y posibilidades de fabricación.

8.3 *El sistema visual - auditivo*

El sistema "Visual-Auditivo" está diseñado para proporcionar una representación del ambiente real a personas invidentes. El elemento principal del sistema es un chip analógico en el

cual las imágenes son recibidas por lentes. La función del sistema es realizar un mapeo de las señales visuales de elementos móviles en señales auditivas que puedan ser percibidas por el individuo a través de audífonos. Al generar sonidos a partir del movimiento de objetos, se pretende generar una sensación similar a la que se tiene al ubicar sucesos a través del sonido que emiten. El objetivo de los diseñadores de este chip es que una persona que carezca de visión pueda tener una percepción más completa de la realidad a través del sistema auditivo creando un modelo interno de su entorno, es decir, que pueda oír no sólo los sonidos de los objetos o sucesos que los causan, sino que también puedan ubicar elementos móviles asociando su desplazamiento con un sonido característico.

La información procesada por el chip fundamentalmente consiste en:

1.- Codificar la intensidad y posición de la fuente de luz en una proyección de arreglo bidimensional tipo retina.

2.- Procesar las señales eléctricas representativas del valor de intensidad y posición asociándolas a cambios temporales.

3.- Sintetizar un sonido obedeciendo las características psicofisiológicas que permitan determinar la ubicación de quién o lo qué lo originó.

Para poder obtener una respuesta adecuada del sistema a desarrollar es muy importante no sólo conocer el comportamiento que se desea obtener del sistema en sí, sino también conocer los principios de operación de los subsistemas en los que está inspirada su fabricación.

8.3.1 *Psicofisiología de la audición*

La investigación sobre la audición en los mamíferos ha logrado determinar características importantes que permiten entender como el cerebro puede procesar sonidos de diferentes fuentes para determinar su ubicación específica.

El trabajo de Bloom y Kendall (Bloom, 1977) (Kendall and Martens, 1984), ha logrado diferenciar exitosamente las señales que utiliza el hombre para la localización de sonidos. El sonido que será censado por ambos oídos es propagado por el aire y alrededor de la cabeza, las ondas sonoras son reflejadas por el pabellón auricular del oído externo para posteriormente entrar en el canal auditivo y llegar al tímpano en el oído interno. Las modificaciones de las ondas sonoras en esta trayectoria, tanto en el oído izquierdo como en el derecho, son las que nos permiten definir la ubicación de la fuente de sonido.

La percepción auditiva se hace en dos dimensiones, la localización vertical y la localización horizontal. Dentro de la localización horizontal existen dos características importantes de percepción, ambas son resultado de la interacción de los oídos. El primer tipo de percepción horizontal se lleva a cabo cuando el sonido incide en forma horizontal al individuo tal como se muestra en la figura 32.

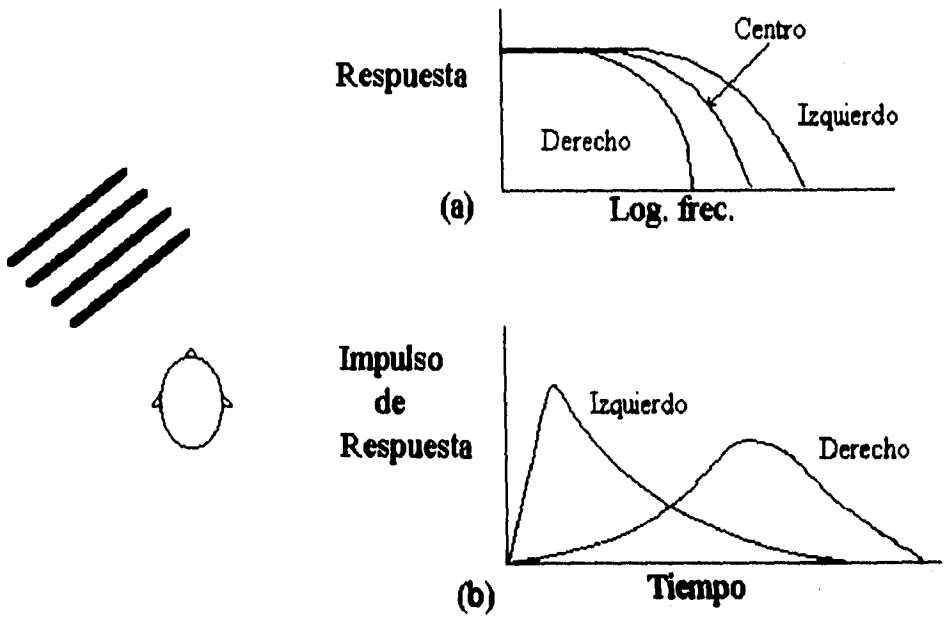


FIGURA 32.- *Representación de percepción de ondas sonoras horizontales. (a) Respuesta de cada oído. (b) Respuesta con retraso de tiempo para cada oído.*

La fuente de sonido está ubicada 45 ° a la izquierda del centro de la cabeza del individuo, las altas frecuencias de esta señal son atenuadas al pasar a través de la cabeza. La percepción de esta señal va a tener características diferentes de intensidad en el oído derecho (contralateral) y en el izquierdo debido a la atenuación de las frecuencias altas. Esta atenuación es conocida como Sombra Cefálica Acústica, su comportamiento se presenta en la figura 32(a). El segundo tipo de localización horizontal depende de la diferencia, producto del retraso, en la llegada de la onda sonora entre los dos oídos (retraso interauditivo). Un sonido que viene de frente al individuo encontrará la misma distancia entre ambos oídos y por lo tanto no existirá retraso interauditivo alguno. El máximo retraso se da cuando el sonido incide directamente sobre uno de los oídos, el valor de dicho retraso depende de las características anatómicas del cada individuo, típicamente el valor del retraso interauditivo va de 350 a 650 microsegundos. La respuesta de cada oído, para

el mismo caso de la percepción no frontal, se presenta en la figura 32(b). La respuesta del oído derecho presenta un retraso con respecto a la del oído izquierdo.

La localización de un sonido vertical es posible gracias al pabellón auricular. La llegada de una señal sonora al canal auditivo puede tener dos trayectorias: El primer caso se presenta cuando la señal llega directamente al canal auditivo, en el segundo caso la señal sonora rebota en las regiones del pabellón auricular. Estas formas de recepción se muestran gráficamente en la figura 33.

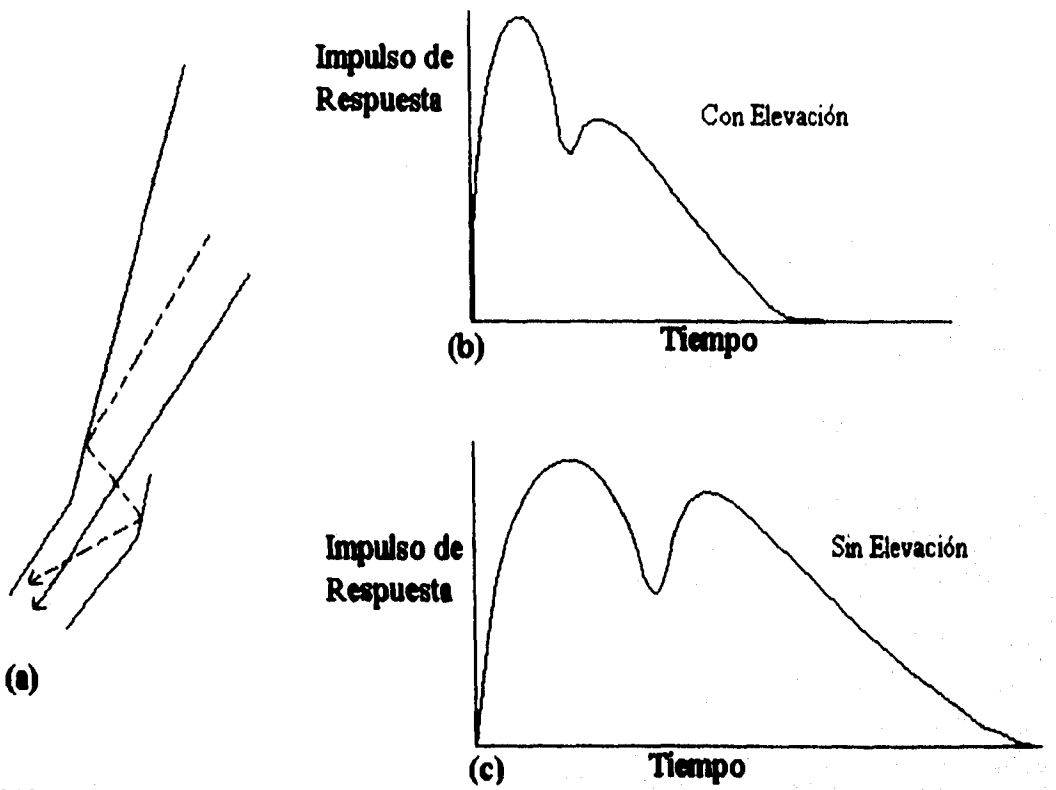


FIGURA 33.- *(a) Representación de las posibles trayectorias de llegada de una señal audible al canal auditivo en vista transversal. (b) Respuesta para niveles altos de elevación. (c) Respuesta para niveles bajos de elevación.*

Al igual que en el caso de la percepción horizontal, en este caso la señal también sufre modificaciones dependiendo del tipo de trayectoria que haya tenido, estas variaciones son las que ayudan a determinar la ubicación del sonido en posición vertical. En el caso de una recepción indirecta (trayectoria punteada) la señal rebota en el pabellón auricular, lo que genera un retraso mayor y amplitud menor que para el caso de la recepción directa (trayectoria continua). La diferencia entre el retraso de las dos trayectorias es una función lineal de la elevación, es corto para señales con gran elevación (figura 33(b)) y largo para señales con elevación pequeña (figura 33(c)). Para el ser humano el valor típico de retraso de 35 a 80 microsegundos.

8.3.2 *Sistema visual biológico*

El procedimiento para generar una representación visual es uno de los mas complejos y de los que más acaparan la atención por el enorme campo de aplicación que tiene. Pero es posible dar una descripción generalizada de cómo funciona la visión en los sistemas biológicos de mamíferos.

El primer paso en el procesamiento de la información visual es realizado en la retina en la cual la imagen se recibe en "lentes" donde la luz y sus características son censadas en forma bidimensional a través de una retícula de foto-receptores, cada uno de ellos genera un potencial analógico proporcional al logaritmo de la intensidad de luz que en ese punto tiene la imagen, el valor logarítmico provee un muy amplio rango de respuesta a la recepción, lo que permite establecer diferencias entre las salidas de los receptores en forma independiente a la iluminación global de la imagen. La ubicación de un foto-receptor en la retina involucra necesariamente la colocación de ese elemento en el espacio real. Esta información es transmitida y procesada a

través de capas múltiples de neuronas. El arreglo bidimensional de neuronas en la retina transmite su salida, producto del mapeo, a través del nervio óptico (formado por los axones de las neuronas ganglionares). Toda esta información es distribuida y procesada hacia y por las más de 250 regiones en el cerebro, hasta ahora conocidas, involucradas en el procesamiento de la visión. Este procesamiento construye un modelo tridimensional basado en los patrones espaciotemporales de la imagen que recibe la retina. En todos los animales el movimiento de las señales es una parte importante del proceso de reconstrucción. Un gran número de vertebrados generan su profundidad visual exclusivamente a través de los movimientos relativos de la imagen en la retina como resultado de los movimientos de su cuerpo (Lorenz, 1981). Aunque los seres humanos ocupamos la percepción binocular para detallar la información de profundidad en objetos cercano (de 1 a 2 metros), el paralaje que induce el movimiento de la cabeza y el cuerpo es un método efectivo para determinar la profundidad en la percepción aún con sólo un ojo (Richards, 1975). Para largas distancias el cambio de paralaje es la única forma para poder determinar la profundidad en la imagen. El cambio de paralaje es un fenómeno puramente geométrico, este no depende de la interacción binocular. Se pueden presentar diferentes casos de este fenómeno dependiendo de la variación entre la imagen de interés y el ojo y la línea de vista. Un buen ejemplo se presenta cuando ambos ojos se fijan en el infinito y la cabeza se mueve, la aparente velocidad de los objetos es una función lineal de la distancia a la que se encuentran. Los objetos cercanos se moverán rápidamente en tanto que mientras más lejanos se encuentren la velocidad aparente será menor. Los objetos en el infinito permanecen estáticos.

La función y tendencia de operación del sistema es clara, una vez contemplados los principios biológicos en los que se inspira el diseño pueden establecerse objetivos no tan

generales: Las señales que representan un evento visual deben ser codificadas de tal manera que al sintetizarlas en una señal acústica esta tenga las características adecuadas para asociarla a una ubicación específica.

8.3.3 *Diseño del sistema visual*

La parte visual de este chip es muy similar a la de un vertebrado. Un grupo de lentes mapea una imagen en un arreglo bidimensional de pixeles, cada pixel está asociado a un fotosensor. La respuesta de un pixel proviene de un punto específico del espacio real, por lo que la ubicación de un pixel en el arreglo bidimensional corresponde a la localización de un suceso en el espacio real. Al igual que en el caso biológico, este sistema genera señales logarítmicas en función de la intensidad de luz que recibe el pixel y, posteriormente, realiza cálculos de derivada en función del tiempo, manteniendo la información direccional de la señal. Este sistema cuenta con dos similitudes significativas con el sistema biológico: La primera es el arreglo del elemento receptor de imágenes, en ambos casos se ejecuta un mapeo bidimensional. La segunda es que las señales procesadas son de carácter analógico, esto en el sistema representa la garantía de no perder información al procesarlas y que contienen toda la información para poder generar una señal sonora completa.

8.3.4 *Diseño del sistema auditivo*

El sistema auditivo del chip es capaz de generar el comportamiento presentado en las figuras 32 (a) y (b). El principio de operación de un sistema capaz de detectar un suceso auditivo

horizontal con las características de comportamiento antes mencionadas se presenta en la figura 34.

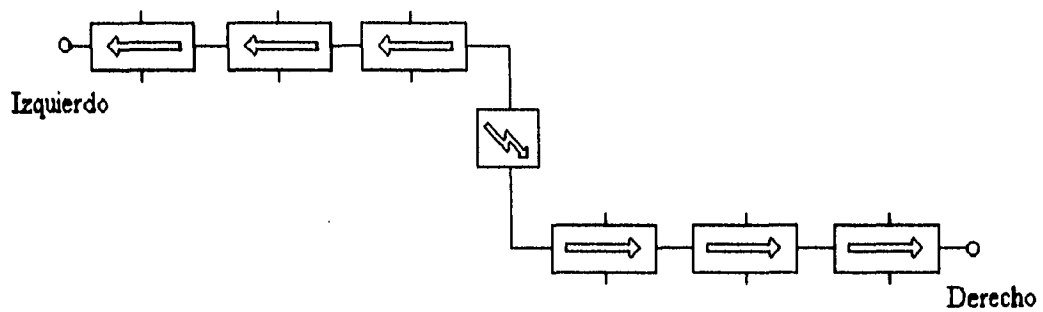


FIGURA 34.- *Diagrama de bloques para la detección horizontal de una señal auditiva. (el pixel está representado por un bloque con una flecha en zig-zag. El bloque de retraso se presenta con una flecha recta indicando con la punta la dirección de la onda sonora).*

La salida de un pixel es conectada a dos líneas de retraso, una para el oído izquierdo y otra para el oído derecho, cada línea genera un retraso en la señal de entrada, el valor del retraso está en función de la longitud de la línea y de la variable de control de cada bloque de retraso. El pixel genera una salida en función del nivel de luz que recibe del exterior, esta señal es transmitida en dos direcciones, en ambas se aplica un retraso. Si la línea izquierda produce un retraso menor que la derecha, esta señal llegará primero al oído izquierdo. De esta manera, el sistema es capaz de determinar la dirección de la cual proviene el evento sonoro.

Cada bloque de retraso contiene también un filtro de altas frecuencias, por lo que en tanto más viaje la señal a lo largo de una línea más atenuadas resultarán las frecuencias altas, con este comportamiento queda representado también el fenómeno de la Sombra Cefálica Acústica. Por tal razón la determinación de la dirección horizontal de un sonido está en función de la diferencia de longitudes entre las líneas de retraso del lado izquierdo y derecho.

La capacidad de procesamiento de señales del sistema auditivo debe ser mayor a una sola fuente sonora. El arreglo para el procesamiento múltiple de señales se presenta en la figura 35.

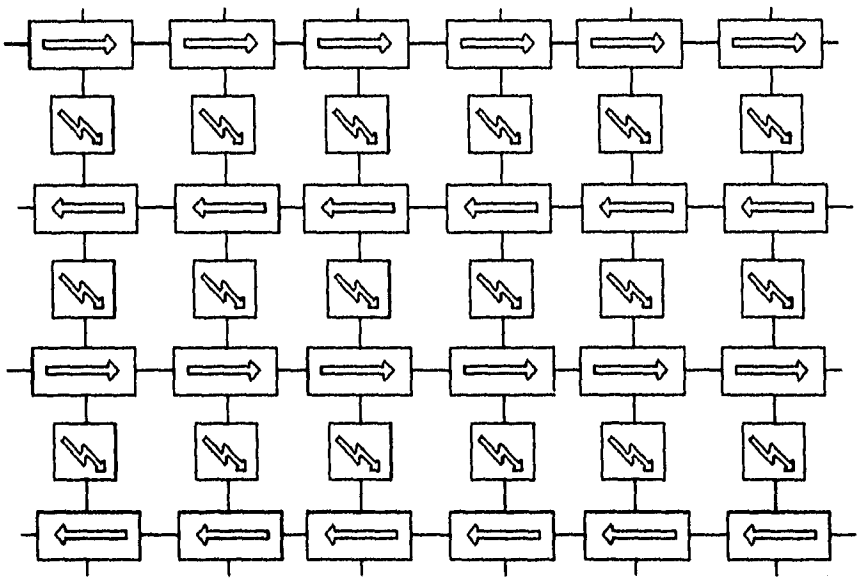


FIGURA 35.- *Arreglo de procesamiento múltiple para percepción sonora horizontal.*

El sistema cuenta con un arreglo bidimensional de dispositivos de entrada (en este caso los pixeles). Cada dispositivo de entrada es acoplado a las dos líneas adyacentes a él, una correspondiente al lado izquierdo y otra al lado derecho. La disposición de cada elemento de entrada en cada una de las filas representa la ubicación horizontal específica. Por tal razón el circuito mostrado en la figura 35 no sólo es capaz de procesar más de una señal de entrada, sino que gracias al carácter analógico de las señales lo puede hacer en forma simultánea.

Cada par de líneas de retraso comparte la misma entrada y recibe más de una señal, tantas como pixeles contenga la fila. La recepción múltiple de señales por cada línea es posible gracias al manejo de señales analógicas. El efecto de la superposición de señales emitidas por los pixeles en las líneas de retraso es similar al que sufre una señal sonora que en el medio se encuentra con

señales diferentes, la recepción en el oído es el resultado de la superposición de todas estas señales.

La ubicación vertical de eventos sonoros es codificada a través de la ubicación en el arreglo de la línea de retraso correspondiente. El modelo del pabellón auricular se muestra en la figura 36.

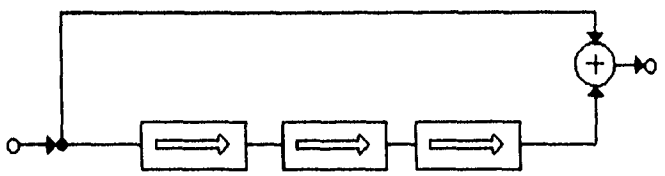


FIGURA 36.- *Modelo electrónico de pabellón auricular.*

En la figura 36 se presenta el caso con dos posibles vías a través de las cuales el sonido puede llegar al canal auditivo del individuo. Este principio permite representar el comportamiento del pabellón auricular si se coloca este circuito en los extremos de cada una de las líneas de retraso representadas en la figura 35.

La salida del circuito presentado en la figura 36 es la suma de la señal de entrada y una versión retrasada de esta misma señal, este comportamiento es equivalente al del canal auditivo biológico el cual recibe dos tipos de señales, las que inciden directamente y las que llegan reflejadas. El tamaño del retraso está determinado por la longitud de la vía, de tal forma que la percepción de señales a través de ambas vías permiten percibir si existe retraso o no de una señal con respecto a otra (vía directa o con retraso), lo que ayuda a determinar elevación la señal. La operación conjunta de los sistemas presentados tienen un comportamiento con las características de las gráficas mostradas en las figuras 33 y 34.

8.3.5 *La operación global*

La representación esquemática del C.I. Visual-Auditivo se muestra en la figura 37.

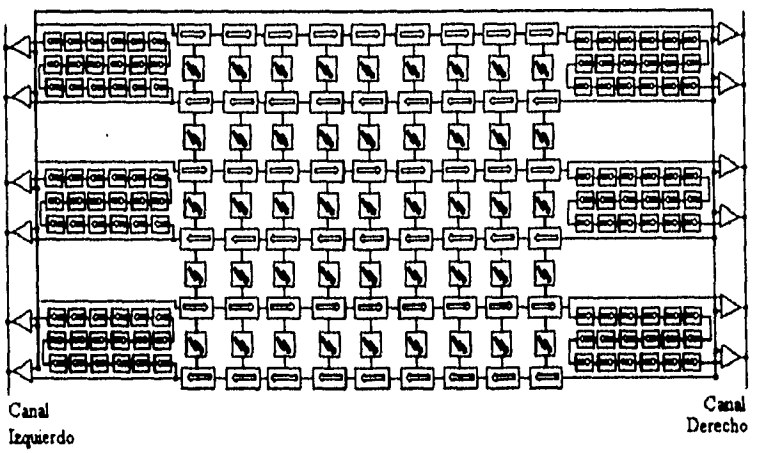


FIGURA 37.- *Diagrama esquemático del C.I. Visual-Auditivo.*

El arreglo de pixeles está establecido de tal forma que la señal de salida tenga una trayectoria horizontal a través de las líneas de retraso. La salida de un pixel actúa como entrada en las dos líneas adyacentes del sistema auditivo. La señal proveniente de los pixeles sufre un retraso y un filtrado de altas frecuencias en forma proporcional a la distancia que debe recorrer la señal desde el pixel hasta los extremos del C.I. Este filtrado y retraso le da a las señales las características requeridas para hacer la localización horizontal del movimiento. Para el caso de las señales verticales es importante notar que cada línea de pixeles corresponde a un ángulo de elevación específico en el espacio real. El modelo del pabellón auricular ubicado al final de cada línea es ajustado en sus valor de ganancia en cada caso para tener el retraso correspondiente al nivel de elevación que se encuentra. Para obtener la representación de la ubicación vertical en audífonos se utiliza la corriente que entrega como salida el modelo del pabellón auricular. La

salida para cada uno de los canales, izquierdo y derecho, es sumada en forma independiente. La superposición lineal de la información que tiene cada canal se obtiene automáticamente y se puede determinar a través de las leyes de Kirchoff.

La superposición de señales en el sistema auditivo permite realizar operaciones el paralelo a través de las cuales se realiza la asociación de una serie de señales auditivas con eventos en el espacio real en forma simultánea y codificar esta información en tan sólo dos canales de salida. La percepción de una imagen a través de múltiples canales visuales es transformada en una representación acústica variante en el tiempo que genera una percepción apropiada por el sistema auditivo humano. A continuación se muestra la integración de los tres bloques esenciales del sistema: pixel, modelo de sombra cefálica y modelo del pabellón auricular.

8.3.6 *Modelo de la retina*

El modelo de la retina está compuesto por un arreglo bidimensional de pixeles. Cada pixel está conformado por un elemento receptor (transconductor de luz), un circuito logarítmico y un diferenciador.

Cada fotorreceptor entrega una corriente de salida proporcional a la intensidad de luz que recibe el transductor, esta corriente es la entrada para el arreglo vertical de transistores bipolares el cual genera un voltaje de salida que es función logarítmica de la corriente entregada por el fotorreceptor. El rango de captura de intensidades es comparable al de los conos en el sistema visual humano. El diagrama esquemático del pixel se presenta en la figura 38.

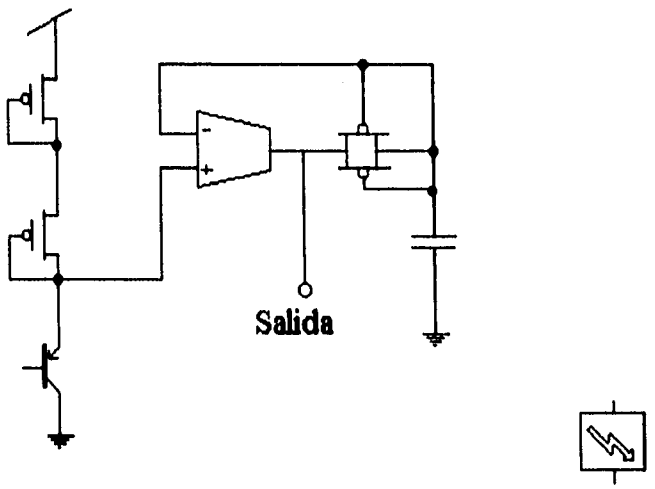


FIGURA 38.- *Diagrama esquemático de un pixel.*

Como se mencionó anteriormente, la salida del fotorreceptor es acoplada a la entrada del arreglo vertical de transistores. La salida de este circuito es la entrada para el derivador con histéresis, el cual como su nombre lo indica realiza la derivada en función del tiempo de su señal de entrada. La salida de este circuito es la entrada a las líneas de retraso del sistema auditivo.

La razón por la cual se utilizan dos transistores en serie conectados en "modo diodo" es debido a que un sólo transistor en estas condiciones de operación tiene un rango de salida considerablemente menor en comparación con el que opera un amplificador de transconductancia o uno de amplio rango, que en circuitos neuromórficos son los que típicamente se utilizan como etapa siguiente. Al conectar en serie dos transistores es posible incrementar de manera satisfactoria el rango de salida de este circuito. La variación en el voltaje de salida para este circuito es de 275 milivolts por un década de cambio en la intensidad de luz. La derivada de una señal en función del tiempo puede obtenerse a partir de un circuito simple RC o bien con amplificadores operacionales, en estos casos la aproximación a la derivada de la señal de entrada es buena pero se presentan problemas significativos al desarrollarlos en hardware el principal de

estos es el nivel de offset que genera el arreglo de operacionales. Para nulificar la presencia de este offset es necesario incorporar más elementos que incrementan la complejidad del circuito y de su encapsulado.

Partiendo del concepto fundamental de derivada (ecuación), esta función puede considerarse como la comparación de la señal con ella misma pero en un instante previo de tiempo, es decir un valor que no varia en función del tiempo no aplica para este concepto. Por tal razón el interés de realizar la derivada es amplificar las características variantes en el tiempo y no las estáticas.

Este comportamiento es posible a partir de un amplificador de amplio rango incorporándole realimentación a través de un elemento no lineal. En este caso el elemento no lineal es un capacitor controlado por un par de transistores, la carga del capacitor se hace a través del transistor canal P en tanto que la descarga de este se hace a través del tipo N. Este circuito genera ligeros cambios en su voltaje de salida cuando la derivada del voltaje de entrada con respecto al tiempo cambia de signo, por tanto, la máxima salida ocurrirá cuando la retina perciba un cambio brusco de intensidad, que bien puede ser producto de un cambio de escena o una variación en distancia de la misma. El rango de operación que se obtiene al manejar la derivada genera mayor oportunidad para obtener información a través del movimiento del cuerpo, hecho que aproxima más el comportamiento del sistema artificial al biológico.

8.3.7 *Modelo auditivo*

Cada línea de retraso está compuesta por una cadena de circuitos seguidor-integrador generados a partir del amplificador de transconductancia. Cada sección de esta línea retrasa y

filtra la señal de entrada. El retraso que sufre al viajar una señal de un extremo a otro de la línea esta en función del número de bloques que la integren y de la constante de tiempo de cada sección. El diagrama de este bloque se muestra en la figura 39.

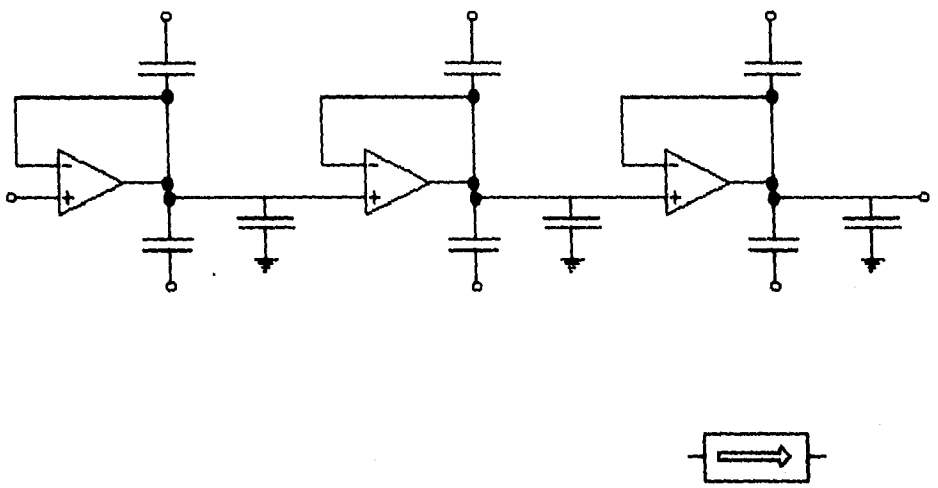


FIGURA 39.- *Diagrama de linea de retraso auditivo.*

Cada sección es acoplada a través de capacitores al pixel superior e inferior, el valor del capacitor que se encuentra a la salida de cada amplificador debe ser tal que a partir de que el voltaje de salida de un pixel tenga una variación de 1.5 Volts la entrada a la línea sea de aproximadamente 200 mv. Todas las líneas de retraso están conectadas en su parte inicial a un voltaje de referencia V_{ref} proporcionado fuera del chip.

El modelo del pabellón auricular esta compuesto por 18 bloques de retraso que actúan al final de cada línea, la variación en el retraso de cada modelo está en función de la elevación en la cual se encuentre colocado y es un valor independiente del retraso generado por la sección de la sombra cefálica.

Debido a que las líneas de retraso horizontal son conectadas a un V_{ref} el total de las señales acumuladas en una línea esta representado por la diferencia entre el voltaje de la línea y

V_{ref}. Esta diferencia es convertida a corriente a través de un amplificador de transconductancia conectado en la salida de la línea de retraso horizontal, esto modela una señal que incide de manera directa al canal auditivo, en tanto que colocando un amplificador al final del modelo del pabellón auricular representa la percepción de una señal indirecta. Estos amplificadores operan bajo el nivel de umbral, en la región lineal lo cual es suficiente para cubrir el rango completo de las señales que se manejan.

El C.I. original contiene 32 filas con 36 pixeles cada una, el tiempo de retraso establecido en el modelo de sombra cefálica es de 4 milisegundos, la diferencia en tiempos de percepción del modelo del pabellón auricular es de 3 milisegundos de la fila inicial a la final. El resultado obtenido es una excelente aproximación al comportamiento visual biológico. Diferentes pruebas con variación en la posición de recepción del haz de luz en el arreglo de pixeles han demostrado que las señales eléctricas son procesadas correctamente (Nielson, *et al.*, 1987).

CAPÍTULO 9: DISEÑO, ENTRENAMIENTO Y PRUEBAS EN SOFTWARE CON DYNAMIND

Como se mencionó en el capítulo 4 el desarrollo de RNA's se ha enfocado preferentemente al software, esto genera dos situaciones: La primera, y más común, es centrar la atención exclusivamente en el desarrollo de software tratando de demostrar que es la mejor manera de desarrollar y trabajar con RNA's, consecuentemente los espacios de difusión, y más aún en un país como el nuestro con limitada información científica, están reducidos a este tipo de información lo que de alguna manera impide la correcta difusión de información en lo que al campo de las RNA's corresponde. En segundo lugar está el desarrollo de software enfocado como una herramienta de evaluación en lo que se refiere a la aplicación de una RNA en una tarea determinada y también como una parte de la RNA misma, no como el total de ésta. La ventaja de ver el software de esta manera es que el concepto de red neuronal no se reduce simplemente a la programación o a la interconexión de componentes sino que se concibe como un sistema integral con los alcances ya mencionados.

El propósito de diseñar y entrenar una RNA a través de este paquete de software (Dynamind) es con el fin de proyectar las características de la red en hardware. Una vez que la red ha sido probada en software, las características que adquiere la RNA pueden, en principio, transladarse al neurochip 80170NX ETANN de Intel. Los valores de los pesos se escriben en memorias EPROM, la función de transferencia es proporcionada por el 80170NX y el arreglo de neuronas es configurable ya sea con uno a más C.I's. La programación del C.I. se hace a través del iNNTS (Intel Neural Network Training System). A pesar de que tanto el chip como el sistema de entrenamiento ya no se comercializan, se consiguió el valor del iNNTS durante el

período que estuvo disponible. El costo era de $11,800.00 dlls conteniendo dos CI´s junto con la tarjeta y software requeridos. Evidentemente el costo es muy elevado pero en muchos casos la aplicación lo justificaría.

El diseño, entrenamiento y prueba de una RNA se hizo a través del paquete de diseño y simulación de RNA´s "Dynamind" que además permite la emulación del neurochip 80170NX ETANN.

La RNA se diseño para poder obtener la clasificación de 72 señales biológicas (potenciales de acción) en cuatro diferentes clases aplicando durante el entrenamiento el algoritmo de retropropagación.

El proceso para obtener el clasificador de señales a partir de una RNA consta fundamentalmente de 6 pasos: 1° Asociación del fenómeno a estudiar con el proceso de una RNA; 2° Captura de datos; 3° Normalización de base de datos; 4° Determinación de la arquitectura de la RNA; 5° Creación de la RNA; 6° Entrenamiento y 7° Resultados y Pruebas.

Asociación del fenómeno con la operación de RNA

El fenómeno general para el que se propone la aplicación de una RNA es la clasificación de señales biológicas.

Estas señales patrón son presentadas a la RNA como señales de entrenamiento para que "aprenda sus características" y las agrupe en clases o categorías. Posteriormente se presenta a la red una nueva señal, diferente al menos a simple vista, a cualquiera de las señales de entrenamiento. Una vez que la señal es presentada a la red, la red es activada e inicia el proceso de comparación, generalización y clasificación. A la salida de la RNA se presenta una señal que

representa la categoría con la cual está asociada la nueva señal y, debido a la naturaleza de los datos, la red puede indicar también el porcentaje de pertenencia de esa señal con cada una de las categorías definidas.

Base de Datos

Las señales de entrenamiento y prueba son una colección de 72 señales de origen biológico (corteza auditiva cerebral de gato) obtenidas en el laboratorio de Moshe Abeles en Israel y facilitadas por George L. Gerstein. Estas señales están digitalizadas en 128 muestras cada una. La colección de señales digitalizadas se muestra en la figura 40.

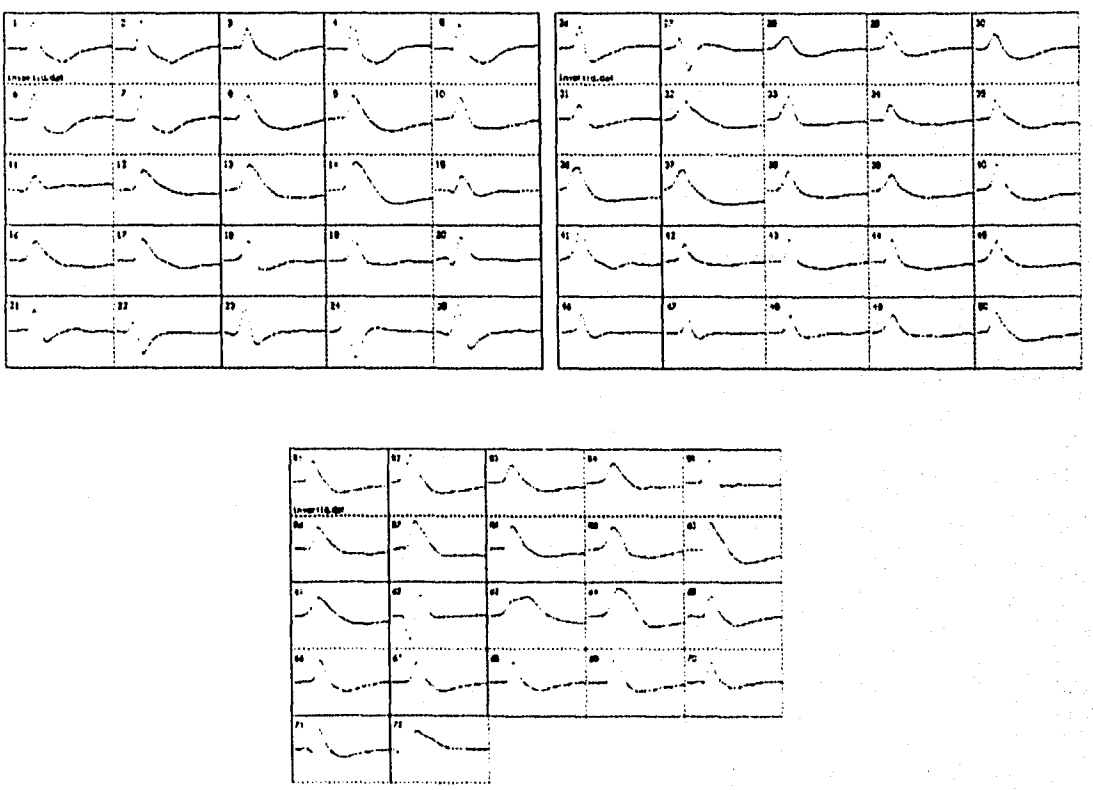


FIGURA 40.- ***Señales biológicas utilizadas para entrenamiento y pruebas de la RNA.***

Normalización de Base de Datos

La normalización es un proceso fundamental para poder acoplar los datos a las características matemáticas de operación de la RNA. La red en este caso opera con procesadores que aplican una señal sigmoidal como no linealidad con límites superior e inferior de ±1. Esto delimita los valores máximo y mínimo con que la red operará interiormente, lo que implica que cualquier valor superior o inferior a los casos de frontera (±1) no será considerado con su verdadero valor. Por ejemplo, si en una serie de muestras existe un valor de 1.5 al presentarlo a la red como tal, este valor a lo más podrá valer 1.0 perdiendo proporción con respecto a las demás muestras, lo que puede y de hecho altera la representación del fenómeno. En fenómenos naturales y más aún en el censado biológico es prácticamente imposible tener mediciones en las que el valor máximo y mínimo coincidan con los valores antes mencionados (los valores en censado neuronal alcanzan sólo algunos cuantos milivolts), para poder manipular y operar estos valores es necesario modificar sus escalas pero sin alterar la proporción que guardan, es decir normalizar. Al normalizar prácticamente se convierte el valor máximo de las muestras en +1 y el mínimo en -1 (para este caso).

Es importante mencionar que la representación del fenómeno es la que determina de que manera se manejan los datos para operarlos en la red, no necesariamente siempre tendrán que normalizarse con esta mecánica, pueden existir casos en los que las muestras puedan dividirse en "Si" y "No" y para este caso el manejo de datos puede sustituirse directamente por 1´s y 0´s sin mayor complicación.

Para el poder manejar las señales biológicas cada dato de las señales es normalizado de la siguiente manera:

$$X_S = \left[\frac{(X_d - d_{MIN})}{d_{MAX} - d_{MIN}} (TF_{MAX} - TF_{MIN}) \right] + TF_{MIN}$$ (Dynamind, 1991)

donde X_s es el valor normalizado, X_d es el valor bruto (valor original, antes de la normalización), d_{max} es el valor bruto máximo, d_{min} es el valor bruto mínimo TF_{max} es el valor máximo de la frontera de normalización y Tf_{min} es el valor mínimo de la frontera de normalización.

Una vez aplicada la normalización la señal conserva sus características pero "a escala" de tal manera que la red puede operar sin riesgo de no considerar algún dato con la proporción debida.

Determinación de arquitectura

Cada muestra cuenta con 128 datos, de las 72 señales que se tienen se diferenciaron 16 clases de acuerdo a la magnitud y proporción de valores positivos y negativos en la señal que a fin de cuentas determinan la forma.

Una vez que se tienen identificadas las características de las posibles entradas y las salidas asociadas es posible definir la arquitectura de la red. El número de elementos de entrada de la RNA deben coincidir con el número de muestras que compone a cada señal por lo que se tendrán 128 unidades de entrada. La salida debe contar con elementos suficientes para representar cada una de las clases establecidas, en este caso son 16 clases, para lograr una representación de 16 clases diferentes basta con tener 4 procesadores de salida donde cada uno puede tener valores entre 0 y 1, de esta manera las 16 clases pueden representarse con combinaciones binarias de estos 4 procesadores de salida.

Para definir el número de capas ocultas y número de neuronas en cada una de estas capas no existe un metodología determinada. La definición de cuántos niveles deben componer la RNA se hace de manera empírica, determinar cuántos niveles ocultos con cuántas neuronas cada uno sólo se puede establecer comparando el desempeño de la red en cada uno de los casos que se estudien, la evaluación del desempeño comprende desde el tiempo requerido del entrenamiento hasta la capacidad de reconocimiento y generalización de patrones. La determinación del número de procesadores en las capas intermedias no tiene alguna regla que lo determine pero comúnmente opera satisfactoriamente un nivel con tantas neuronas como el resultado de la media geométrica del número de procesadores de entrada y de salida.

La elección del algoritmo de entrenamiento se hizo también a partir de las posibilidades de operación de esta red en hardware, en este caso se aplicó el algoritmo de retropropagación ya que el ETANN 80170NX esta diseñado para una operación bajo la arquitectura de cascada.

Creación de la RNA

Una vez determinada la arquitectura de la RNA es necesario establecer el valor de cada uno de los parámetros involucrados en el proceso de aprendizaje como son: Tipo de función de transferencia, ganancia de la función de transferencia, intervalo de valores para los pesos y patrón de aprendizaje.

La función de transferencia a utilizar es de tipo sigmoidal y los valores máximo, mínimo y ganancia son configurables. La función de los valores máximo y mínimo se ha comentado anteriormente en el capítulo 5, el valor de este parámetro está en función de los límites que se quieran manejar en los valores de salida de la RNA. El parámetro de ganancia determina que tan

"extendida o comprimida" está la función sigmoidal. La definición del valor de este parámetro está en función del tipo de resultado que se quiera obtener, es decir valores binarios o analógicos. Para un valor de ganancia alto la red entregará típicamente valores saturados o en nivel cero, en tanto que para valores de ganancia pequeños (menor o igual a 1) la salida de la RNA entregará valores analógicos. El valor inicial de los pesos es un parámetro importante en tanto que a partir de este valor es que se inician los cálculos para buscar la convergencia, este parámetro es configurable pero el ETANN maneja limites de $\pm$ 2.5. La función del patrón de aprendizaje se explicó en el capítulo 5, el intervalo en el cual puede ser seleccionado el patrón de aprendizaje es [0.0 - 10.0] debido a que este parámetro determina en parte el tiempo necesario de entrenamiento el intervalo sugerido está comprendido entre [0.0 - 1.0]. Una vez que están establecidos todos los parámetros de la RNA se genera el archivo en el cual estará indicado el valor de cada uno de los parámetros representativos de la RNA, este archivo utiliza la extensión .NET.

Entrenamiento de la red

Una vez que la red está creada es necesario presentar los datos de entrenamiento que incluyen en este caso las asociaciones de las 16 señales representativas de las clases definidas anteriormente con la señal de salida que se desea obtener de la RNA. La forma de presentar estos datos a la red es a través del archivo conocido como archivo de entrada-salida (I/O), en el cual está toda la información que la red debe "aprender," cabe recordar que al tratarse de una red con retropropagación tiene un aprendizaje supervisado, por lo que hay que indicar la asociación entrada-salida deseada para cada caso. La creación de este archivo se puede hacer a partir de un archivo en ASCII, procesador de textos o una hoja de cálculo en los cuales estén los datos tanto

de entrada como de salida sin importar cual es el orden de presentación (entrada seguida de salida o viceversa). Para generar el archivo I/O existen tres métodos diferentes, la elección de qué método aplicar está en función de la distribución de valores que tenga la base de datos y de la aplicación del archivo, de entrenamiento o de pruebas (para detalles consultar el manual de Dynamind).

Una vez creado el archivo I/O se carga la red y se le asocia la base de datos correspondiente. Antes de iniciar el entrenamiento es necesario fijar las condiciones para las cuales el entrenamiento se considerará completo; el parámetro de terminación puede ser el valor del error global de la red o bien el número de épocas de entrenamiento, en este caso se fijó el número épocas en 2000.

Se generaron algunas redes, todas con la misma arquitectura pero con diferente valor en el factor de aprendizaje (F.A.) pero dentro del intervalo recomendado.

Resultados y Pruebas

Parámetros descriptivos de la red desarrollada:

Numero de Niveles: Dos

Unidades de Entrada: 128

Neuronas Ocultas (nivel 1): 23

Neuronas de Salida (nivel 2): 4

Algoritmo de Aprendizaje: Retropropagación

Función de Transferencia: Sigmoidal

Limites Superior e Inferior: +1 y -1 respectivamente

Ganancia:	1
Factor de Aprendizaje:	0.2, 0.35, 0.5, 0.7
Limites de Peso:	+2.5 Sup. y -2.5 Inf.

A continuación se muestran los tiempos necesarios para el entrenamiento y el error cuadrático de cada red diseñada. El entrenamiento de las redes se hizo en una P.C. pentium a 100 MHz. Debido a la velocidad de la computadora, la diferencia en tiempos de ejecución entre los cuatro casos es prácticamente nula. Se hizo el entrenamiento de la misma red en una P.C. 386 a 33 MHz y el tiempo requerido fue de 1:17:42. La diferencia entre procesadores es abismal y al incrementar el factor de aprendizaje de 0.2 a 0.35 y buscar el mismo valor del error el proceso tardó 1:38:27. Es claro que en este tipo de procesadores el valor del F.A. si se proyecta de manera notoria en el tiempo requerido de entrenamiento.

Red1.- F.A.= 0.20	Tiempo de entrenamiento =3:50 .
	Error Cuadrático = 2.8×10^{-4}
Red2.- F.A.= 0.35	Tiempo de entrenamiento =3:52.
	Error Cuadrático = 0.0016
Red3.- F.A.= 0.50	Tiempo de entrenamiento =3:53.
	Error Cuadrático = 0.0083
Red4.- F.A.= 0.70	Tiempo de entrenamiento =3:52.
	Error Cuadrático = 0.0014

Después de estos resultados es claro que la variación en tiempo y error cuadrático es mínima entre los diferentes valores del F. A.

Se trabajó también en dos redes más con patrones de aprendizaje de 1.5 y 2.5, el tiempo requerido para el cumplimiento de las 200 épocas fue el mismo que en los casos anteriores, pero el error tuvo variaciones considerables siendo de 0.398 y 0.418 respectivamente. Con esto se comprueba la validez de la recomendación de no asignar valores muy altos a este parámetro. La realización de las 200 épocas se ejecutó en el mismo tiempo que en los 4 casos anteriores pero el valor del error es considerablemente mayor, para una red con estas características (F.A. alto) es recomendable considerar como parámetro de entrenamiento el valor del error no el número de épocas calculadas.

En los capítulos anteriores se hizo patente la trascendencia del valor de los pesos de una RNA. A continuación se presentan en las figuras 41 a 44 las gráficas de los pesos de los dos niveles que componen la red1. En estas gráficas se considera el eje vertical como el valor de los pesos; los ejes diagonales representan el número de neuronas o elementos de entrada (para el caso del nivel de entrada), en la figura 41 el eje izquierdo representa el nivel uno y el eje derecho la capa de entrada.

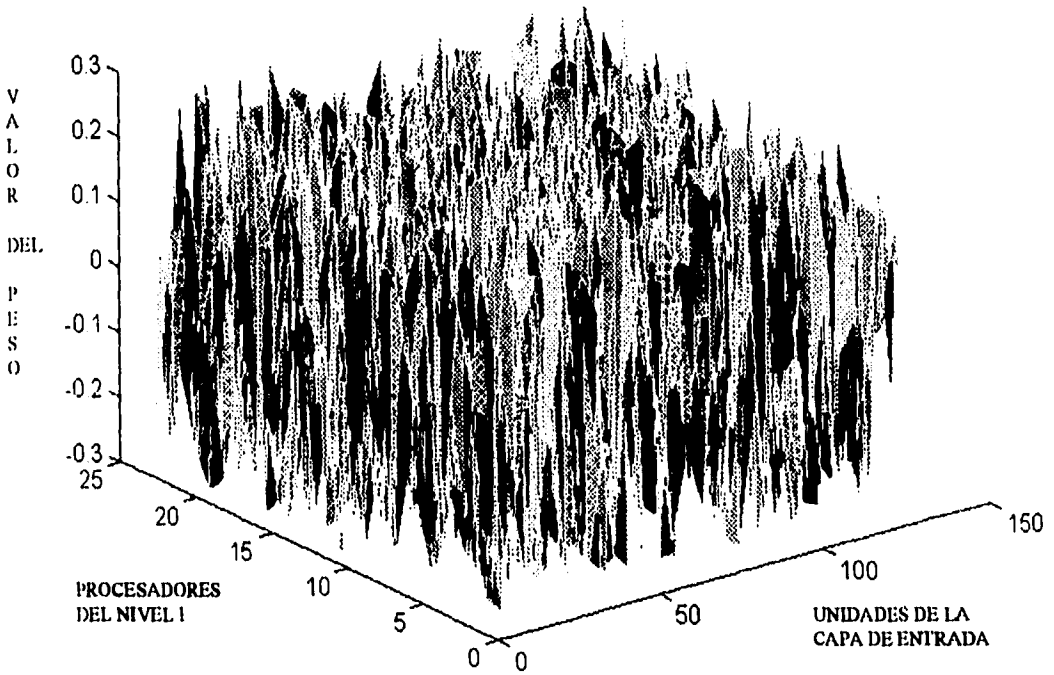


FIGURA 41.- *Gráfica de la matriz de pesos de la capa de entrada al nivel 1 de la red 1 sin entrenamiento.*

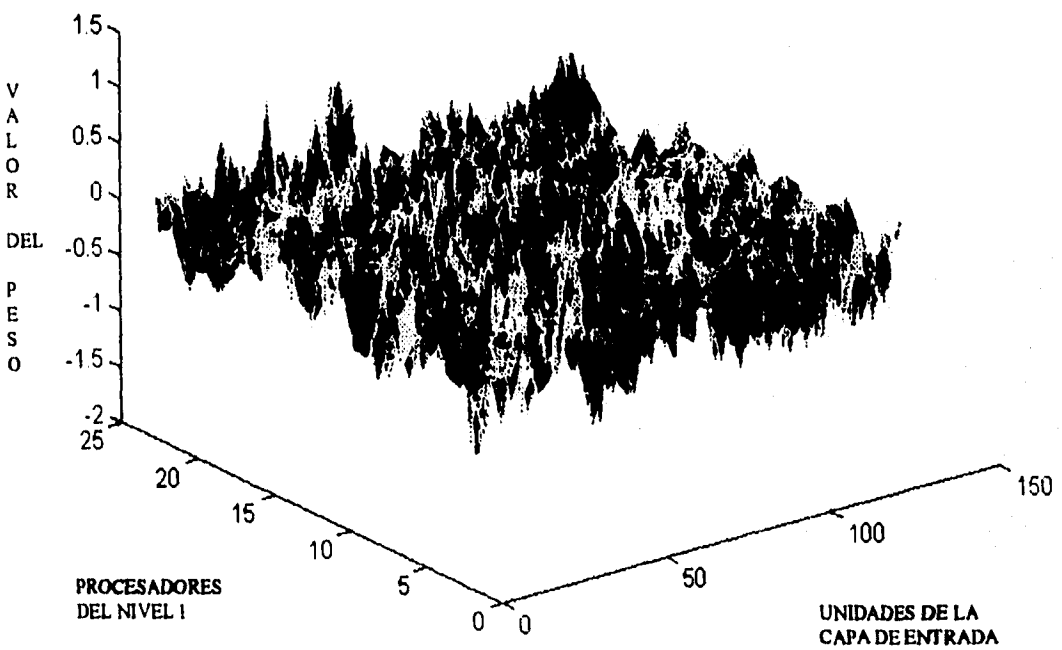


FIGURA 42.- *Gráfica de la matriz de pesos de la capa de entrada al nivel 1 de la red 1 entrenada.*

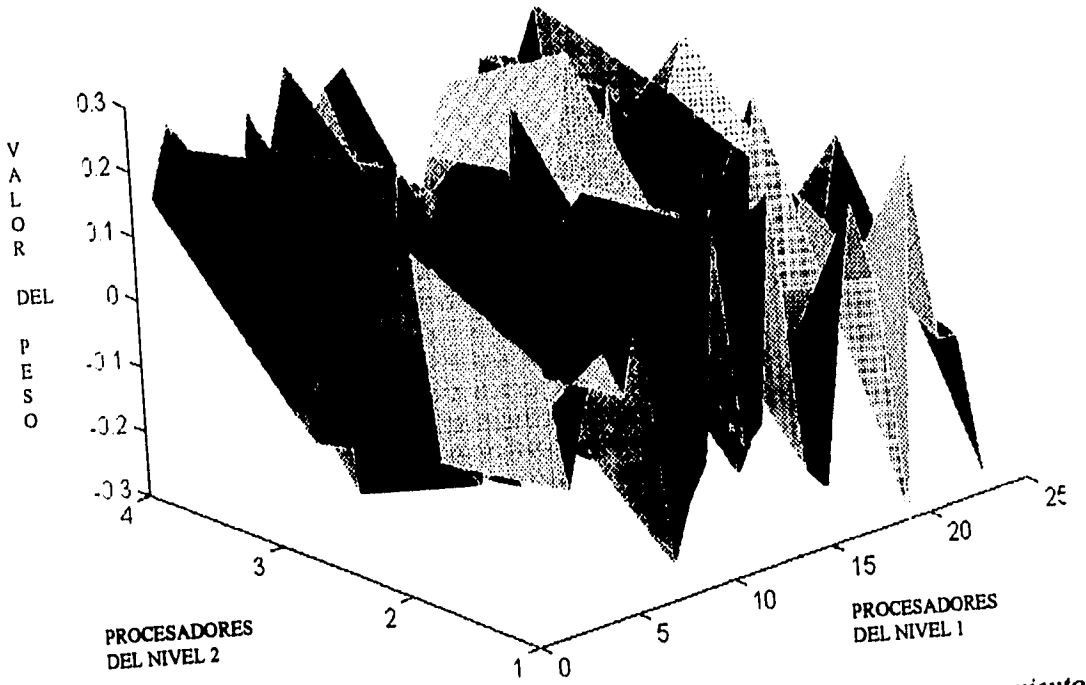


FIGURA 43.- *Gráfica de la matriz de pesos del nivel 1 al nivel 2 de la red 1 sin entrenamiento.*

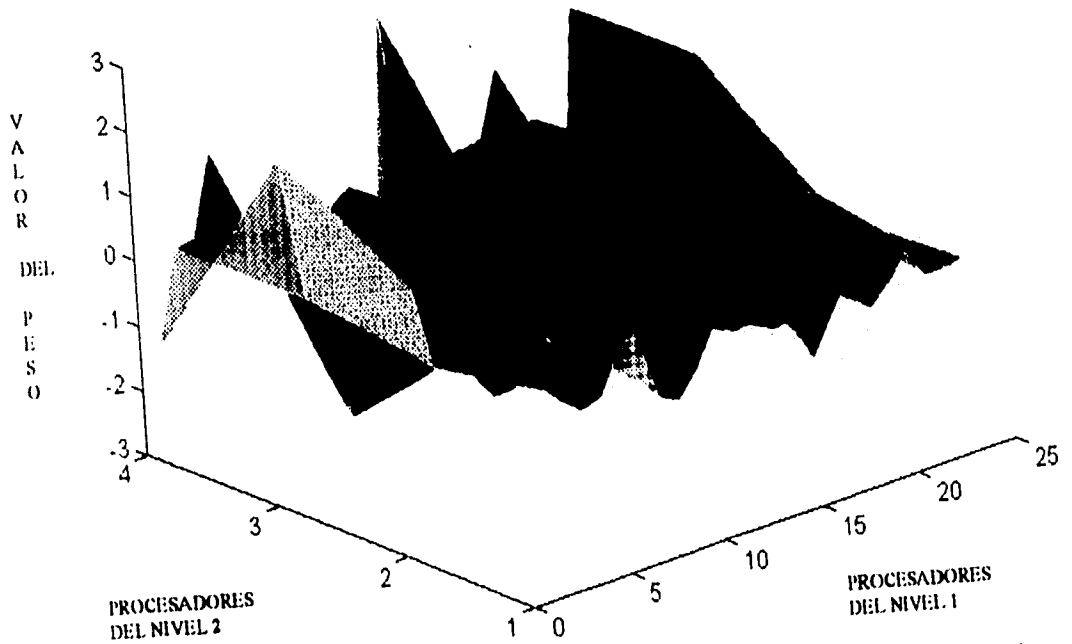


FIGURA 44.- *Gráfica de la matriz de pesos del nivel 1 al nivel 2 de la red 1 entrenada.*

En las gráficas sin entrenamiento es claro que el valor de los pesos no están acotado, la diferencia entre los valores de cada uno de los pesos es considerable, en este caso los pesos están establecidos al azar, esta es la condición inicial de la red. Una vez que la RNA ha sido entrenada los valores de los pesos son acotados de acuerdo a las condiciones establecidas en el diseño (a ± 2.5), el valor límite se establece en función del hardware en el que la red se quiera desarrollar. En esta gráfica cada peso tiene el valor que satisface simultáneamente las condiciones de cada uno de los patrones. El significado real de esta superficie se desconoce pero se decidió presentarla porque resulta una perfecta herramienta para visualizar el valor de los pesos en forma global, concepto comúnmente confuso en el campo de las RNA's.

La realización de pruebas se hizo en dos partes: La primera consistió en generar un archivo I/O para pruebas y cargarlo a la red ya entrenada. En este archivo sólo se indica el vector de entrada y el número de procesadores a la salida de la red sin asignarles valor alguno. El resultado de esta prueba fue muy bueno, todos los patrones de entrada fueron asignados a la clase indicada durante el entrenamiento. La segunda prueba fue la generalización, para realizarla fue necesario generar y cargar otro archivo I/O el cual contiene las 72 señales digitalizadas, estos valores son presentados como entrada a la red y únicamente se indica el número de procesadores de salida sin asignarles valor.

De las señales de la figura 40 fueron seleccionadas como patrón de entrada las indicadas en la columna izquierda de la tabla que se presenta a continuación, su asociación correspondiente

(establecida arbitrariamente) se presenta en la columna derecha. Estas columnas constituyen las asociaciones entrada-salida para el entrenamiento supervisado por retropropagación.

Señal Número	Clasificación
1	0000
4	0001
9	0010
10	0011
12	0100
15	0101
17	0110
19	0111
22	1000
23	1001
29	1010
34	1011
62	1100
65	1101
68	1110
70	1111

Los resultados de la generalización arrojaron información interesante que a simple vista es imperceptible. La clasificación de señales establecida fue hecha, como ya se mencionó, a partir de la forma que presenta cada señal. A simple vista sólo un pequeño grupo de señales presentaban valores muy negativos a su inicio, lo que en las gráficas las hacía diferentes a todas. La señal que presentaba esta característica con mayor énfasis fue la seleccionada como patrón de entrenamiento con la idea de que cualquier otra señal con estas características fuera agrupada con ésta.

Al realizar las pruebas de generalización se encontró que muchas más señales de las que se preveía que entraran en esa categoría fueron seleccionadas, la razón fue que aunque el inicio de la señal era plano, el valor en estos puntos era negativo. Esta característica es considerablemente difícil de determinar a partir de los registros biológicos, primeramente por los bajos niveles de registro y por la difícil ubicación de la referencia en estos procesos.

CAPÍTULO 10: DISCUSIÓN Y CONCLUSIONES

De acuerdo con el primer objetivo de esta tesis, se realizó una revisión de la literatura cuyos resultados se presentan en los capítulos del 2 al 8. No se pretende que la revisión incluya todo lo existente sobre el tema, pero si se cubren las áreas más importantes. Hasta donde sabemos, este es el primer reporte sobre el tema que se presenta en Español. Por otra parte, en el capítulo 9 se presentan los resultados de utilizar el simulador DynaMind y que era la segunda parte de los objetivos. En cuanto al resto de los objetivos de este trabajo que era iniciar un proyecto de redes neuronales en hardware, encontrar la manera de hacerlo no fue sencilla y la recomendación que ofrece esta tesis no es necesariamente única, pero también es cierto que no hay muchas alternativas en este momento, aunque este campo está evolucionando rápidamente y es posible que en un año o dos habrá otras alternativas. El estudio que se hizo nos llevó a concluir que lo mejor en este momento es el neurochip y sistema de desarrollo de NESTOR. Para esta adquisición el Laboratorio de Cibernética deberá presentar un proyecto a CONACYT, ya que con los presupuestos normales de las facultades, es imposible adquirir equipo y componentes que valen miles de dólares.

A continuación hacemos una discusión resumida del trabajo de investigación que se realizó para llegar a la recomendación mencionada anteriormente.

La evolución de las RNA se ha dado a pasos agigantados, las diferencias entre un hardware que demuestre el comportamiento básico de una red simple (un oscilador p. ej.) y uno que pueda aplicarse a una tarea más elaborada (generalización de vectores p. ej.) es abismal. Esto debido a que la investigación y el desarrollo en este campo ha tenido lugar exclusivamente en laboratorios

de universidades y en muy pocos laboratorios de algunas empresas. A diferencia del hardware de una PC que si bien se diferencia mucho entre las primeras versiones y las actuales, se cuenta con un gran número de versiones intermedias que reflejan un avance paulatino. El hardware para RNA es tajante, tanto en precio como en complejidad de los circuitos, en este caso no se tiene la fortuna de contar con versiones o modelos de desarrollo intermedio entre lo elemental y lo contemporáneo, por lo que la idea de conseguir, generar u operar una RNA en hardware no resulta trivial.

Debido a que el software a través del cual se realizó el diseño y entrenamiento de la RNA está inspirado en un CI particular, la búsqueda inicial fue sobre este chip, el neurochip ETANN 80170NX de Intel. Esto se discutió ampliamente en los capítulos 1 y 9.

La búsqueda de este neurochip fue algo muy representativo de lo que implica trabajar con hardware de RNA´s.

El contacto inicial con Intel fue vía fax solicitando información general del 80170NX y del sistema de desarrollo y la tarjeta multineurochip EMB necesarios para su programación. La carta fue enviada a Intel Corporation en Sta. Clara California E.U.A. Desgraciadamente nunca se obtuvo respuesta, en seguida se habló a Intel México, directamente con el gerente de comercialización encargado del mercado en Latino América, de igual manera se solicitó información general como cotizaciones, sistemas adicionales requeridos, etc. La respuesta, que no tuvo nada que ver con la esperada, se obtuvo tres y medio meses después vía telefónica con noticias nada innovadoras.

Durante este periodo se buscó información por otros medios, un herramienta importante fue Internet. A través de esta red se accesó la página de Intel, en la cual no se encontró absolutamente

nada referente a hardware de RNA's, este hecho fue un tanto desconcertante debido a que en la bibliografía del software utilizado para la simulación (capítulos 1 y 9), que está desarrollado en 1994, se comenta que este CI es el más avanzado, con un dato así es de esperarse que el CI al menos existe.

De la información encontrada en Internet acerca de RNA's un porcentaje muy alto de información corresponde exclusivamente a software (cerca del 95 %), esto complicó la búsqueda debido a que había que filtrar una enorme cantidad de información para conseguir referencias de hardware. Al ver que Intel no proporcionaba información por ningún medio sobre el 80170NX, se estableció contacto con la empresa que desarrolló el paquete de diseño y entrenamiento Dynamind, NeuroDynamx Inc. en Boulder Colorado EUA, este paquete emula la operación de dicho CI y sirve como herramienta de programación del mismo. Una vez localizada la página de esta compañía en la red, se encontró que ya existía una nueva versión de Dynamind que corre bajo Windows y, al igual que la primera, emula la operación del 80170NX, esta versión fue desarrollada en Noviembre de 1995, hecho que aumentaba las posibilidades de encontrar el CI. El contacto con NeuroDinamx se hizo de inmediato solicitando cotizaciones e información tanto de la nueva versión de Dynamind como del 80170NX. La respuesta que se obtuvo fue algo que posteriormente se volvió común para la mayoría de las empresas contactadas: "El giro de esta empresa es trabajar y desarrollar software, no hardware."

Debido a estas situaciones, la búsqueda se reinició bajo un contexto más general, es decir, no sólo referida al 80170NX. Para conseguir información acerca del hardware existente y disponible de RNA's se estableció contacto con gente del grupo del Dr. Edgar Sánchez Sinencio en la Universidad de Texas, concretamente con Glen Spencer y Han GunHe ambos alumnos del

Doctor Sánchez, la información proporcionada por este grupo fue importante, indicaron algunas direcciones electrónicas en las que se hablaba sobre este tema. En una de estas páginas se encontró un dato muy importante con información acerca del 80170NX. La nota, nada sobresaliente en forma con respecto a las demás, pero si en contenido, informa que Intel se retira del campo de las RNA's. La razón aparente es la canalización total de recursos al área de desarrollo de procesadores pentium. Un lote de aproximadamente 600 80170NX fue asignado a la compañía California Scientific para su comercialización.

Se estableció el contacto con California Scientific. Esta empresa es conocida en el campo de las RNA's, pero en el área de software, la respuesta al primer mensaje fue inmediata, pero no contenía información de ningún tipo referente al 80170NX, únicamente recomendaban visitar su página en la red en la cual mencionaban todos los productos disponibles, ninguno de estos es hardware. Después de insistir en esta compañía comentando el antecedente del lote de CI's de Intel, la respuesta fue que ya no está disponible y por lo tanto tampoco el sistema iNNTS, sin indicar motivos ni costos, que el chip es de Intel y que ellos son quienes podrían proporcionar información.

El hermetismo en esta información fue desconcertante, más aún cuando existen páginas actualizadas en Marzo de 1996 que mencionan al 80170NX como el CI analógico más rápido y aún en existencia. La búsqueda se enfocó a localizar los CI's existentes para conocer la tecnología que se está aplicando en su fabricación, las arquitecturas de redes utilizadas, la velocidad, las aplicaciones típicas y el costo .

Las compañías que de alguna manera se encuentran asociadas con hardware de RNA aún cuando en algunos casos no ofrecen ningún producto ni información respectiva a neurochips son:

Adaptive Solutions Inc. (Antes Aptix Corp. hasta 1992).

American Neuralogix Inc.

AT&T

California Scientific Software Inc.

Fujitsu Laboratories Ltd.

IBM

Intel Co.

Lockheed Missile and Space Corp.

Meridian Parallel Systems Ltd.

Nestor Inc.

Neural Semiconductors

Neural Technologies

NeuroDynamX Inc.

Oxford Computer Inc.

Philips

Ricoh Co. Ltd.

Sensory Circuits Inc.

Synaptics Inc.

De todas estas compañías sólo 9 cuentan con página electrónica, lo que significa una enorme ayuda para conocer datos de la empresa y poder establecer comunicación de inmediato. De las compañías restantes sólo de algunas se presentan, en páginas de información general, datos como dirección y en el mejor de los casos teléfono. Se enviaron correos electrónicos a cada una de las empresas que tenían esa posibilidad, sólo tres contestaron y de esas tres sólo una aportó información sobre su neurochip.

La información que en las mismas empresas se tiene referente a hardware de RNA's es prácticamente nula. Por ejemplo, la respuesta de gente de IBM fue que no tienen conocimiento de

algún componente relacionado con RNA desarrollado por esta compañía, aún cuando existe una página completa dedicada al ZISCO de IBM.

Las posibilidades *inmediatas* de aplicación de un neurochip que existen en el Laboratorio de Cibernética son principalmente dos: a) Aplicarlo como clasificador en tiempo real de potenciales de acción obtenido de electrofisiología celular en el hipocampo de rata. b) Aplicarlo como identificador de patrones de imágenes para operar como auxiliar del sistema visual y la fusión de sensores del robot "NOMAD-200". Como se ha mencionado, estas aplicaciones se han empezado a desarrollar y por lo tanto la aplicación del neurochip es de carácter inmediato.

Una vez evaluada la información obtenida, con la idea de determinar cual es la mejor opción para la aplicación de una RNA en hardware en una tarea específica, considerando disponibilidad del distribuidor para entregar equipo, herramientas de trabajo, soporte técnico, costos y aplicaciones se determinó que existen dos opciones para adquirir una RNA en hardware. La primera es comprar el neurochip Ni1000 de Nestor Inc. Este CI de 168 pines está desarrollado con tecnología de FLASH EPROM CMOS y puede aplicarse como clasificador operando con las siguientes características:

Cálculo de distancia entre vectores	**16 GOPS**
Velocidad del procesador matemático	**132 MFLOPS**
Microcontrolador	**16.5 MIPS**
Tamaño del vector	**256 Dim X 5 bit**
Número posible de vectores	**1024 (256 Dim) y 8192 (32 Dim)**
Número de transistores en el encapsulado	**3.75 Millones**
Dimensiones	**15.8 mm X 13.7 mm**

Este CI opera con un algoritmo propietario aparentemente similar a un mapa de Kohonen, aunque se indica que con este algoritmo, como tal, puede operarse en el modo "off-chip" al igual que el LVQ. Los algoritmos "on-chip" con los que el Ni1000 trabaja son: PNN, RCE y PRCE.

La disposición de la compañía Nestor Inc. es excelente en todo lo que se refiere a venta, cursos y asesoría en la operación del Nestor Ni1000. Aparentemente Nestor Inc. está a punto de sacar a la venta otro neurochip, más económico pero menos manejable, basado en una arquitectura de IBM, lo que deja pensar en que se trate del ZISCO.

La segunda opción puede implicar mayores capacidades de desarrollo tanto en el Laboratorio de Cibernética como en cualquier otro sitio interesado en aplicar redes neuronales en hardware, pero está enfocada a un proceso de largo plazo (aproximadamente 2 años) lo que de ninguna manera lo hace menos atractivo. Consiste en trabajar en forma conjunta con la compañía IBERCHIP, la cual se especializa en el diseño y elaboración de sistemas integrados en tecnología VLSI. Una tendencia interesante es elaborar un C.I. multifunción, pero desarrollar neurochips para aplicaciones específicas no es menos interesante y, de hecho, esta filosofía es la que mayores logros ha tenido en el desarrollo de sistemas neuromórficos, sistemas que en muchos sentidos son considerados como los más aproximados a operar como un sistema biológico. Lo complejo de la operación de las redes neuronales que contiene el cerebro indica, de alguna manera, que al menos hasta ahora es imposible desarrollar un sistema que se le asemeje en capacidades de operación. Pero la neurofisiología ha podido determinar, de manera aproximada, los mecanismos de operación de diferentes redes que comprenden el SNC, como la visión por ejemplo, estos avances

son significativos para poder pensar en desarrollar una RNA que pueda al menos cumplir con estas aproximaciones.

Existe la posibilidad de generar pequeños circuitos como osciladores o alguna memoria asociativa a partir de dispositivos electrónicos comunes pero el objetivo no sería demostrar que a partir de estos elementos se puede generar una RNA, sino en principio, desarrollar y aplicar una RNA desarrollada en VLSI para evaluar su operación bajo diferentes condiciones tanto de los parámetros de operación de la red como de las características de los datos de entrada, esto con la idea de establecer las ventajas y desventajas que puede implicar el utilizar redes neuronales en hardware en diferentes disciplinas. La intención de trabajar con una RNA en VLSI es penetrar en una vertiente de estudio prácticamente desamparada en nuestro país y, claro, tratar de aplicar RNA's contemporáneas en aplicaciones científicas y tecnológicas complejas. La distancia entre lo simple y lo actual es tan grande que se debe tomar una determinación realista, considerando las limitantes administrativas y económicas de las cuales estamos rodeados, que defina a partir de dónde es posible iniciar el estudio de las RNA contemporáneas dejando atrás el intento de reproducir experimentos ya ejecutados hace décadas, pero con los dispositivos actuales. La demostración de la eficiencia de un CI con arquitectura neuronal sólo se podrá llevar a cabo cuando se involucre en una aplicación de interés actual, en la cual sea de trascendencia el tiempo de procesamiento o la capacidad de diferenciar elementos ruidosos o incompletos, por ejemplo, no únicamente reproduciendo configuraciones ya estudiadas, que si bien representan una muy poderosa herramienta para el entendimiento y concepción de una RNA en hardware, no representan una sólida plataforma de desarrollo previa a la ejecución de un proyecto de aplicación contemporánea en la neurofisiología, medicina, comunicaciones, robótica y control, por ejemplo.

Proyectos que realmente se hagan con objetivos que demanden las bondades emergentes de una RNA no que sean recurrentes, es decir, proyectos con visión ingenieril sólo podrán plantearse y ejecutarse con un sólido apoyo económico-administrativo y evidentemente con la disposición intelectual de un grupo de gente interesada y convencida de los objetivos.

GLOSARIO

A / D.- Analógico / Digital.

ART.- Teoria de resonancia adaptiva.

CI.- Circuito integrado.

CMOS.- Superficie complementaria de Metal-Oxido.

CPS.- Conexiones por segundo.

CPSPW.- Conexiones por segundo por peso.

CUPS.- Conexiones actualizadas por segundo.

D/A.- Digital / Analógico.

EPROM.- Memoria de sólo lectura programable y borrable.

FA.- Factor de aprendizaje de la red para el algoritmo de retropropagación.

Fan Out.- Número límite de dispositivos conectados a la salida de un dispositivo electrónico.

GOPS.- Giga operaciones por segundo.

I/O.- Entrada / Salida.

LVQ.- Medición del vector de aprendizaje. (algoritmo de aprendizaje supervisado)

MOPS.- Mega operaciones por segundo.

MOS.- Superficie de Metal-Oxido.

Neuro Chip.- Circuito integrado que contiene un esquema de operación neuronal.

NMOS.- Superficie de Metal-Oxido tipo "N".

PC.- Computadora personal.

PE.- Elemento procesador.

PMOS.- Superficie de Metal-Oxido tipo "P".

PNN .- Red neuronal probabilística.

PRCE.- Restricción de probabilística por energía de Coulomb.

RAM.- Memoria de acceso aleatorio.

RBF.- Base de función radial.

RCE.- Restricción de energía de Coulomb.

RN.- Red neuronal.

RNA.- Red neuronal artificial.

RNA´s.- Redes neuronales artificiales.

S.N.C.- Sistema nervioso central.

SIMD.- Datos multiples con una instrucción.

SNS.- Sistemas neuronales sintéticos.

TDM.- Multiplexaje por division de tiempo.

ULSI.- Ultra larga escala de integración.

VCO.- Oscilador controlado por voltaje.

VLSI.- Muy larga escala de integración.

W_{xy}.- Peso sináptico de "x" a "y".

REFERENCIAS

Abdo, S., Declercq, J., Hochet, B., and Peiris, V., "*Implementation of a Learning Kohonen Neuron Based on a New Multilevel Storage Technique*", IEEE J. Solid St. Ccts. 26, pp. 267-267, 1991.

Akers, L. A., Ferry, D. K., and Grondin, R. O., "*Synthetic Neural Systems in VLSI*" , en "An Introduction to Neural and Electronic Networks", Zornetzer, D. L. (Ed.), Academic Press, 1990.

Alcántara, M., "Simulador y Analizador de Redes Neuronales Artificiales: Neurored", Tesis de Licenciatura para Ingeniería en Computación, Facultad de Ingeniería, UNAM 1992.

Allen, R. B., Alspector, J., Jayakumar, A., Zeppenfeld, T., and Meir, R., "*Relaxation Networks for Large Supervised Learning Problems*", in Advances in Neural Information Processing Systems 3, pp. 1015-1021, 1991.

Arima, Y., Mashiko, K, Okada, K., Yamada, T., Maeda, A., Notai, H., Kondoh, H., and Kayano, S., "*A 336-neuron, 28K synapse, self-learning neural network chip with branch-neuron-init architecture*", IEEE Solid St. Ccts.26, pp. 1637-1644, 1991.

Arima, Y., Mashiko, K., Okada, K., Yamada, T., Maeda, A., Notai, H., Kondoh, H., and Kayano, S., "*A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40 K Synapses*", in Proc IEEE Int. Solid State Cct. Conf. (ISSCC-92), Paper TP8.1 (Sn. Francisco, 1992).

Baxter, D. J., Murray, A. F., and Reekie, H. M., " *Fully Cascadeable Analogue Synapses Using Distributed Feed-back*", in VLSI for AI and Neural Networks, eds. J. G. Delgado-Frias and W. R. Moore (Plenum Press, New York), pp. 205-213, 1991.

Berger, R., Gilbert, S., Mann, J., Soares, A., and Raffel, J., "*A Generic Architecture for Wafer-Scale Neuromorphic Systems*", IEEE 1st International Conference on NN 3, pp. 501-513. 1987.

Bloom, P. J., "*Creating source elevation illusions by spectral manipulation*", Journal of the Audio Engineering Society, pp. 25-270, 1977.

Card, H. C. and Schneider C. R, "*CMOS Implementation of Analog Hebbian Synaptic Learning Circuits*", Proc. Intl. Joint Conf. Neural Networks 1, pp. 437-442, 1991.

Card, H. C., Schneider C. R, and Schneider, R. S, "*Learning Capacitive Weights in Analog CMOS Neural Networks*", J. VLSI Signal Proc. (in press).

Card, H.C. and Schneider, C.R., *"Analog CMOS Neural Circuits - In Situ Learning"*, World Scientific Publishing Company, International Journal of Neural Systems, Vol. 3, No. 2 , pp. 103-124, 1992.

Card, H.C. and Moore, W.R., *"VLSI Devices and Circuits for Neural Networks"*, International Journal of Neural Systems, Vol. 1, pp. 149-164, 1989.

Carterette, E.C. and Friedman, M.P., *"Handbook of Perception"*, Vol. 5. New York: Academic Press, p. 351, 1975.

Chiang, A. M. and Chuang, M. L., *"A CCD Programmable Image Processor and its Neural Networks Applications"*, IEEE J. Solid St. Cct. 26, pp. 1894-1901, 1991

Denker, J. S., Graf, H. P., Howard, R. E., Hubbard W. E., Jackel, L. D., Schwartz, D., Straughn, B., and Tennant, D. M., *"VLSI Implementation of a Neural Network Memory with Several Hundreds of Neurons"*, AIP Conf. Proc., pp.151- 182, 1986.

Denker, J. S., Graf, H. P., Henderson D., Howard, R. E., Hubbard W. E. and Jackel, L. D., *"Neural Network Chips"*, IEEE Engineering in Medicine & Biology Society 10th Annual International Conference, pp. 1495-1496, 1988.

Denker, J. S., Epworth, R.W., Graf, H.C., Howard, R.E., Hubbard W. E., Jackel, L. D., Schwartz, D., Straughn, B.L., and Tennant, D.M., *"An Asssociative Memory Based on An Electronic Neural Network Architecture"*, IEEE Trans. ED, 34, 1553, 1987.

Espinosa, E. I., *"Métodos Estadisticos Utilizados en la Neurofisiologia"*, Revista Ingenieria Vol. XLVII NUM. 1, pp. 101-109, Enero-Marzo 1977.

Faggin , F. and Mead, C. *"VLSI Implementation of Neural Networks"*, en "An Introduction to Neural and Electronic Networks", Zornetzer, D. L. (Ed.), Academic Press, 1990.

Foo, S. Y. , Anderson, L. R., and Takefuji, Y.,"Analog Components for the VLSI of Neural Networks", IEEE Ccts. and Devices Magazine, pp. 18-25, 1990.

Fujita, Y., Iida, T., Itakura, T., Kamatani, Y., Kimura, T., and Shima, T., *"Neuro Chip with on-chip Backprop and/or Hebbian Learning"*, in Proc. IEEE Int. Solid State Cct. Conf. (ISSCC-92), Paper TP8.4 (San Francisco, 1992).

Graf, H. P. and Jackel, L. D., "*Analog Electronic Neural Network Circuits*", IEEE Circuits and Devices Mag. 5:4, pp. 44-55, 1989.

Graf, H. P., Hubbard W. E., and Jackel, L. D., "*VLSI Implementation of a Neural Network Model*", IEEE Computer Magazine, pp. 41-19, March 1988.

Grey, W.,W., "El Cerebro Viviente", Fondo de Cultura Económica,1961.

Hebb, D. O., "The Organization of Behavior", New York: Wiley, 1949.

Hilburn, J.L. and Johnson D. E., "Manual of Active Filter Design", McGraw- Hill Book Company, 1973 .

Hodgkin, A.L., "*The Croonian Lecture: Ionic movements and electrical activity in giant nerve fibres*", Proc. Roy. Soc. 13, 125, pp. 1-37, 1958.

Holler, M., Tam, S., Castro, H., and Benson, R., "*An Electrically Trainable Artificial Neural Network With 1024 Floating Gate Synapses*", in Proc. IJCNN-89 Part II, pp. 191-196, 1989.

Hopfield, J. and Tank, D., "*Collective Computation in Neuronlike Circuits*", Scientific American, pp. 62-70, December 1987.

Kendall, G. S. and Martens, W. L., "*Simulating the cues of Spatial Hearing in Natural Environments*", Proceedings of the International Computer Music Conference, Paris: IRCAM, 1984.

Leonetti, G., "Simulador de Redes Neuronales Artificiales: Cibernet", Tesis de Licenciatura para Ingeniería en Computación, Facultad de Ingeniería, UNAM 1992

Lippmann, R.P., "*An Introduction to Computing with Neural Nets*", IEEE ASSP Magazine, pp. 4-22, April, 1987.

Lorenz, K.Z., "*The foundations of Ethology*", New York: Springer-Verlag, 1981.

Mead, C., "Analog VLSI Neural Systems", Addison-Wesley Publishing Company, 1989.

Murray, A. F., Del Corso, D., and Tarassenko, L., "*Pulse Stream VLSI Neural Network Mixing Analog and Digital Techniques*", IEEE Trans. Neural Networks 2, pp. 193-204, 1991.

Nestor Incorporation, "*Demo: N1-1000*", Nestor Inc. 1995.

NeuroDynamx, "Dynamind Developer User's Guide", Boulder, Co. U.S.A. 1991

Nielson, L., Mahowald, M., and Mead, C., "*SeeHear 1987*", International Association for Pattern Recognition, 5th Scandinavian Conference on Image Analysis, 1987.

Richards, W., "*Visual Space Perception*", In Carterette, E. C. and Friedman, M.P. (eds), Handbook of Perception, Vol. 5. New York: Academic Press, 1975.

Robinson, D. A., "*The electrical Properties of Microelectrodes*", Proceedings of the IEEE, Volume 56 Number 6, pp.1065-1071, June 1968.

Smith, C.U.M., "The Brain: Towards an Understanding", Alianza Ed.,1970

Silviotti. M., Tomlinson, M., and Walker, D. , "*A Digital Neural Network Architecture for VLSI*", in Int. Joint Conf. Neural Networks Vol. II, San Diego, pp. 545-550, 1990.

Tresguerres, J.A.F., "Fisiología Humana", McGraw Hill., 1992.

Werbos, P., "*Backpropagation Through Time: What It Does and How to do It*", Proceedings of the IEEE, Vol. 78, pp 1550-1560, October 1990.

Zornetzer D. L., "An Introduction to Neural and Electronic Networks", Ed. Academic Press, 1990.

BIBLIOGRAFÍA DE RED

http://crg.eee.kcl.ac.uk/clarkson/pram.html

http://msia02.msi.se/~lindsey/elba2html/elba2html.html

http://www.adaptivelogic.com/

http://www.asi.com/asi/manuals.html

http://www.calsci.com/home.htm

http://www.curtech.com/mm32k/index.html#sumary

http://www.mosaic-industries.com/

http://www.nestor.com/

http://www.neuralware.com/resource/reprint.html

http://www.neuroptics.com/

http://www.teleport.com/~cognizer/PRODUCTS/BYTECH1.HTM

http://www.wadinc.com/pgs_idx/ic256.shtml