Evolving Spiking Neural Networks for Adaptive Audiovisual Pattern Recognition

Simei Gomes Wysoski

A thesis submitted to
Auckland University of Technology
in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

2008

Knowledge Engineering & Discovery Research Institute
Faculty of Design and Creative Technologies

Primary Supervisor: Nikola Kasabov
Secondary Supervisor: Lubica Benuskova

Table of Contents

“I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.”

Simei Gomes Wysoski

This dissertation is dedicated to my parents

Eleozina Gomes Wysoski

and

Paulo Silas Wysoski *(in memoriam)*

Of all the things I've learnt, what you have taught me is still by far the most important

Acknowledgements

Firstly, I would like to thank my primary supervisor Prof. Nikola Kasabov for providing ideas, inspiration and the necessary support to make this research possible. Special thanks go to my secondary supervisor, Dr. Lubica Benuskova, who devoted a lot of time teaching me the basic principles of neuroscience and for always having her office door open for "a quick discussion". Undoubtedly, it was from these discussions that most of the ideas for this work came and most of the implementations were designed. Through Dr. Benuskova's guidance, I was introduced to the modelling of brain functions, for which I am also truly indebted.

I would also like to thank the past and present members of KEDRI - Knowledge Engineering and Discovery Research Institute at the Auckland University of Technology, who directly or indirectly provided me with the conditions for the completion of this work. Dr. Ilkka Havukkalla deserves special acknowledgment for always bringing me challenging research questions. Peter Hwang for always being around for discussions, for providing diligent technical support, and teaching me how to use NeuCom. "Obrigado" to Dr. Paulo C. M. Gottgtroy for introducing me to the ontological way of describing the world and for teaching me to look beyond the spike domain. Dr. Michael D. Platel for many discussions over a cup of coffee. Vishal Jain, David Zhang, Akbar Ghobakhlou, Vishwa Shukla, Dougal Greer, Richard Walton, Helene Wall, Yoshitaka Inoue also contributed in different ways to this work. From KEDRI, I am indebted to Joyce D'Mello for always encouraging students at the times we were thinking of giving up and for being ready to help in whatever situation we faced. With her assistance, getting a PhD became a much easier task.

Thanks to Snjezana Soltic for the opportunity to share my thoughts about this research and for returning many intriguing questions that resulted in a deeper understanding of the topic. I am grateful to Dr. Marcia Gottgtroy, for kindly reading the draft of this dissertation and particularly, helping me to clarify the research design.

Thanks also to the Tertiary Education Commission of New Zealand, through the Top Achiever Scholarship, for the financial support for this research.

Finally, I would especially like to thank my mother Eleozina Gomes Wysoski and my wife Marisa van Arragon. Without their love and encouragement this work would probably have not even started.

Abstract

This dissertation presents new modular and integrative information methods and systems inspired by the way the brain performs information processing, in particular, pattern recognition. The proposed artificial systems use spiking neurons as basic elements, which are the key components of spiking neural networks. Of particular interest to this research are various spiking neural network architectures and learning procedures that permit different pattern recognition problems to be solved in an evolvable and adaptive way.

Spiking neural networks are used to model human visual and auditory pathways and are trained to perform the specific task of person authentication. The systems are individually tuned and trained to recognize facial information and to analyze sound signals from spoken sentences. The modelling of the integration of different sources of information (multisensory integration) using spiking neural networks is also a subject of investigation. A network architecture is proposed and a model for audiovisual pattern recognition is designed as an example.

The main original contributions of this thesis are:

a) Evaluation and further extension of adaptive learning procedures to perform visual pattern recognition. A new learning procedure that enables the system to change its structure, creating/merging neuronal maps of spiking neurons is presented and evaluated on a face recognition problem.

b) Design of two new spiking neural network architectures to perform person authentication through the processing of speech signals.

c) Design and evaluation of a new architecture that integrates sensory modalities based on spiking neurons. The integrative architecture combines opinions from individual modalities within a supramodal layer, which contains neurons sensitive to multiple sensory information. An additional feature that increases biological relevance is the crossmodal coupling of modalities, which effectively enables a given sensory modality to exert direct influence upon the processing areas typically related to other modalities.

The contributions were published in one journal paper and in four refereed international conference proceedings. The proposed system designs were implemented and, through computer simulations, demonstrated comparable performance with traditional benchmarking methods. The systems have some promising features: they can be naturally optimized in respect to different criteria: accuracy (when very accurate results are expected), energy efficiency (when management of resources play an important role), and speed (when a decision needs to be made within a limited time).

In this thesis, most of the parameters have been exhaustively optimized by hand or by using simple heuristics. As a direction for future work, there is an opportunity to include automated, specially tailored parameters optimization procedures or even general-purpose optimization algorithms, e.g., Genetic Algorithms and Particle Swarm Optimization.

Overall, the results obtained in this thesis clearly indicate that it is indeed possible to have fast and accurate adaptive pattern recognition systems scalable for multiple modalities computing with simple models of spiking neurons. However, it is important to advance the theory of spiking neurons to take advantage of its biological relevance to reach similar or better performance when compared to the human brain, for instance, exploring new neuron models, information coding schemes and network connectivity.

Chapter 1 - Introduction

1.1 Definition and motivation

Evolving spiking neural networks belong to the class of *Brain-like Evolving Connectionist Systems* (*Brain-like* ECOS). To fully understand the meaning of the term *Brain-like* ECOS, it is best to define each term individually.

Connectionist systems are systems composed of simple processing units connected to each other to form a network. Depending on how the configuration of these simple processing units is set up as a network, complex behaviours can be achieved.

In this context, the term *evolving* stands for the continuous and automatic adaptation of the processing units and the way these units connect in structured networks, triggered by internal and/or external causes. The evolving property enables the structures to perform suitable information acquisition (to separate "wheat from the chaff"), store knowledge, execute learning (learn/forget a specific task), or even only organize themselves internally to reach more stable states (Kasabov, 2007).

It is important to distinguish the concept of evolving systems with the definition of evolutionary algorithms (Back, 1996). Evolutionary algorithms basically "*models the collective learning process within a population of individuals*" using the biologically inspired processes of recombination, mutation and selection (Back, 1996). *Evolving* in this dissertation relates to connectionist systems that seek to model evolving processes that "*develop, change over time in a continuous manner*" (Kasabov, 2007).

Brain-like ECOS can be considered a special case of ECOS which, besides the ECOS requirements, are also concerned in giving a biological interpretation to the systems. Section 2.3.3 discuss the classification of COS, ECOS and *Brain-like* ECOS in more detail.

In this work, the biologically inspired processing units are implemented as spiking neurons, using the theory of spiking neural networks (Gerstner and Kistler,

2002) (Maass and Bishop, 2001). In particular, this research explores the use of evolving spiking neural networks, i.e., processing units grouped in ensembles that function, learn and evolve in a biologically inspired way, to solve audiovisual pattern recognition problems.

Pattern recognition is a broad area of artificial intelligence concerned with finding the correspondence of a certain stimulus with the information previously stored in memory. While some pattern recognition problems are easy to process automatically with artificial systems, e.g., searching for a string of characters in a document, other groups of patterns have been exhaustively studied and can only be recognized once many constraints are put into place. Classical examples in this group are the recognition of visual objects, sounds, smells, and taste patterns. Interestingly, some types of patterns that are difficult to handle by artificial systems are processed very accurately by human brains, most likely due to thousands of years of evolution and adaptation to environmental conditions (Darwin, 1859). The impressive performance of the brain has motivated many scientific studies. However, despite the tremendous effort put into deciphering the mechanisms that regulate the brain's activity, there are still many parts of the puzzle that remain unsolved.

The motivation of this research is twofold. First, the investigation of brain-like connectionist systems attempts to enhance the performance of current pattern recognition systems and solve engineering problems. In this direction, the aim is to understand the principles that rule biological brains and to simulate some properties of biological networks that can improve the behaviour of artificial systems. Second, this research can contribute to the emergence of new theories and knowledge which explain how the brain processes information and develops cognitive activities. The development and consequent simulation and analysis of new network architectures can be used as a tool to test and verify new hypotheses and discoveries.

This research intends to close the gap between artificial and biological models used for pattern recognition. In this attempt, this research follows the direction described in (Gerstner and Kistler, 2002) that states: "*If you want to avoid any prior assumption regarding the way the brain processes information, we need to simulate it using field potentials*", i.e. spiking models. Thus, all the new achievements of this research, namely, new network architectures and learning procedures, use the theory of field potential with information transmitted via pulses. As will be shown in the course of this dissertation, the use of processing units with field potentials for pattern recognition is very recent (Hopfield, 1995) and still has a fertile and open terrain to be explored. Currently, there are many research groups trying to elucidate whether computation using the field potential theory can indeed be considered the next generation of connectionist systems and whether they can substantially enhance performance on pattern recognition tasks (Maass and Bishop, 2001). While by the end of this research it may remain an open question, this work expects to contribute towards facilitating the answer.

1.2 Objective of the research

Inspired by the processing ability of the human brain, the general objective of this research can be defined as:

> *"To design new artificial systems to execute complex pattern recognition tasks that apply brain-like principles to information processing"*

1.2.1 Specific objectives

The general objective can be divided into three more concrete and specific objectives. They are:

1) to propose a new way of achieving adaptive/evolving learning in a spiking neural network model for pattern recognition;

2) to design new evolving spiking neural network architectures to perform audio and visual pattern recognition;

3) to present a new spiking neural network model to combine different sources of information coherent with brain-like principles.

1.2.2 Research design

In order to achieve the objectives proposed in the previous section, the research process is organised using the following steps:

a) review and evaluation of the existent methods;
b) conceptual design and software implementation;
c) experimentation and quantitative evaluation.

Literature review

The first step is to review and evaluate the state-of-the-art systems described in the literature that use traditional methods of audio, visual, and audiovisual information processing, i.e., methods that do not have biological representation but provide state-of-the-art performance in terms of pattern recognition.

In addition, biologically realistic ensembles of neurons and learning methods are evaluated, from low level processing realized in sensory systems up to the emergence of a higher level of cognitive activities. Emphasis is given to the computational models that emulate brain activity on different levels of information processing, utilizing adaptive connectionist ways of modelling.

Finally, how the human brain processes multimodal information and combines different sources of information in a coherent learning process is investigated. Some of the key points evaluated at this stage include: early or late integration of modalities, correlations between modes, response when partial information is missing and processing speed.

Conceptual design and implementation

The design of new SNN architectures and procedures to learn audio, visual, and audiovisual patterns follows the literature review. In this part, this research breaks new ground. The new designs as well as their dynamic behaviour and the tuning of their parameters are described to facilitate reproducibility. The new models are implemented in C++ or C# language and compiled in a general-purpose computer platform. The implementation is also a subject of analysis.

Test and validation

The design and implementation of the new systems is followed by an evaluation of performance. The systems are tested to process complex images and speech streams of benchmarking datasets. Parameters of the new systems are tuned to optimize accuracy. For the audio and visual systems alone, accuracy is then compared with benchmark systems under the same dataset setup. The new integrated audiovisual approach, on the other hand, is compared with individual modalities, also under the same dataset conditions.

The new methods are also analysed in respect to biological relevance and their usefulness to processing data to solve engineering problems. Finally, and equally importantly, directions for further advancement in the development of new evolving spiking neural networks are offered.

1.3 Scientific contribution

The primary contribution of this work is to use biologically realistic neural networks to solve pattern recognition problems. In this direction, a new computer model that performs visual pattern recognition using spiking neurons is designed and implemented (Wysoski *et al*, 2006) (Wysoski *et al*, 2008). The model includes a new learning procedure that adapts to new data through structural adaptation. An additional layer that accumulates opinions over several image frames proves useful in enhancing the decision-making process.

A new network architecture that resembles the human auditory system specifically able to recognize speakers is also presented (Wysoski *et al*, 2007a). In this architecture, a novel learning procedure is proposed based on mathematical models that perform normalization in the *similarity domain* (normalization of similarity scores), together with a new adaptive model to create/merge neurons.

Finally, the integration of audiovisual sensory information using the theory of spiking neurons is presented (Wysoski *et al*, 2007). A spike-based technique to perform *a posteriori* integration implements the combination of individual modalities where each modality can receive crossmodal influence. Thus, each modality is linked to other modalities before individual decisions are made.

All these new methods proposed are evaluated through various perspectives: accuracy, biological relevance, limitations and computational cost. As an overall conclusion, the methods present comparable results with some traditional pattern recognition algorithms.

1.4 Organisation of the dissertation

This dissertation is organised into six chapters. Chapter 2 presents a literature review covering the different ways of modelling brain activity, moving towards the modelling of audiovisual sensory stimuli. Chapter 3 focuses on the visual system and the corresponding modelling tools that perform visual information processing, in particular pattern recognition. Chapter 4 covers the auditory system and shows how patterns of auditory information can be processed using several artificial models. In Chapter 5, the integration of sensory modalities is presented. Chapters 3, 4 and 5 contain the core of this research, and it is where these new methods are described in detail. In each of these three chapters, a specific conclusion is given as well as further necessary evaluations and suggestions for future development. Chapter 6 summarises the achievements of the work and contains the overall conclusion. Hardware implementation and commercial applications are also briefly discussed, followed by future directions, which conclude the dissertation.

Chapter 2 - Modelling the biological brain and brain-like computation - A review

This chapter reviews what is known about the way the brain processes information, from neurons as the basic processing unit, up to the emergence of cognitive processes when ensembles of neurons are interconnected in a complex way. Then, a classification of neural-based information processing models is proposed. The last part of the chapter is devoted to introducing the theory of spiking neural networks, in particular, their use as a tool for artificial intelligence.

The review starts with the description of several artificial neuron models, followed by the integration of neurons into networks that resemble the brain's way of functioning and the brain's structure. As there is no single model that incorporates all the brain's capabilities, the most important properties of each model is highlighted and the differences among them are explained. The artificial models are presented in order of biological relevance, from the less biologically relevant to the most accurate models from a biological viewpoint.

From the pattern recognition perspective, this review proceeds with a brief explanation on the functional structure and information pathways that form two sensory mechanisms relevant to this research: the visual and auditory system. The integration of pathways and the way the information is integrated in these two systems are also described.

A classification of artificial models that use ensembles of neurons for information processing is introduced: *Connectionist Systems* (COS), *Evolving Connectionist Systems* (ECOS) and *Brain-like* ECOS. This research concentrates on *Brain-like* ECOS. As a special case of ECOS, *Brain-like* ECOS seeks to incorporate discoveries from neuroscience into the design of new artificial systems. Specific aspects to be considered in *Brain-like* ECOS are presented and discussed.

In the final part of the chapter, the theory of spiking neural networks (SNN) as having more biologically realistic neuronal processing units is discussed. An emphasis is placed on applying SNN to pattern recognition tasks.

Overall, this chapter gives the support and biological background to the main contributions presented in Chapter 3, Chapter 4 and Chapter 5, where existent and new artificial models for pattern recognition are presented.

2.1 Brain-like information processing units and their organization

2.1.1 Neurons

The brain is basically composed of neurons and glial cells. The role of glial cells is to provide support to the neurons, helping to increase conduction speed, forming a blood-brain barrier, holding neural structures together and defending the neurons against infection (Stein and Stoodley, 2006). Despite the fact that there are 10 to 50 times more glial cells than neurons, the role of information processing is undertaken exclusively by the neurons. For this reason, most artificial networks that model the processing of information simply do not take into account the glial cells, restricting their component blocks to neurons (neural networks).

Neurons are composed of three main parts: dendrites, cell body (or soma) and axons. Dendrites receive messages from other neurons, axons transmit messages to other neurons and the cell body contains the nucleus of the cells and the cell's genetic information. Axons from one cell and dendrites from another cell are connected through synapses. Neurons differ in type according to their main functionality. There are the sensory neurons, motor neurons, local interneurons, projection interneurons and neuroendocrine cells (Kandel, 2000). Yet, independent of type, a neuron is constituted of four functional parts: *input*, *trigger*, *conduction* and *output*. Figure 2.1 shows a graphical representation of a generic neuronal model.

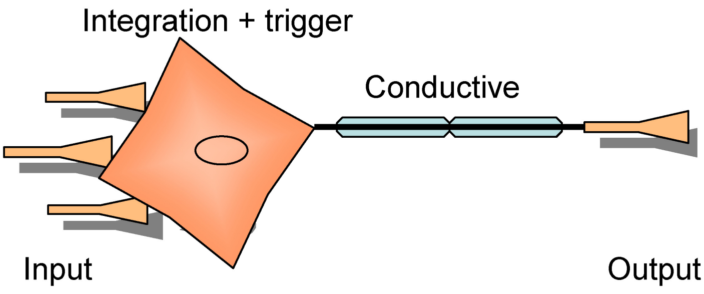


Fig. 2.1. Functional parts of biological neurons.

Over recent years, research about brain information processing has acquired enough evidence to suggest that biological neurons transmit information using short electrical discharges generated by electrochemical activity. The input and output signals in a single neuron have a short duration (generally 1ms) and a constant amplitude (typically 100 mV). Due to the short duration of the input/output signals they are commonly referred as pulses or spikes.

The typical behaviour of a neuronal unit can be roughly described as follows: an incoming pulse (received by a dendrite) increases the inner potential of a neuron (in the soma), which is called postsynaptic potential (PSP). When the inner potential of a neuronal cell (PSP) reaches a certain threshold, the neuron outputs a spike (through the axon). Figure 2.2 illustrates this process.

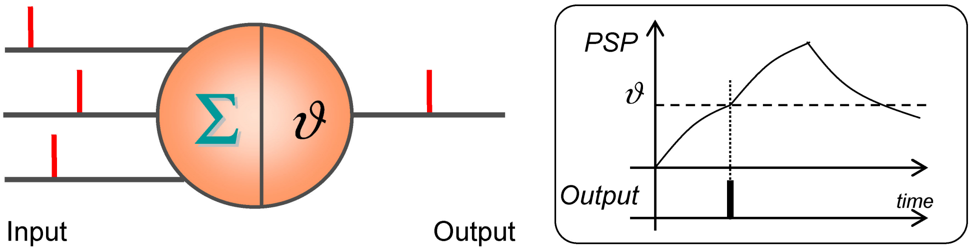


Fig. 2.2. Representation of a neuron and its basic functional behaviour. Incoming pulses increase the inner potential (PSP). Output pulses are released when the PSP reaches a certain threshold.

A wide range of models describing the functional behaviour of a single neuron has been proposed, e.g., integrate-and-fire, integrate-and-fire with leakage, spike response model, Izhikevich model (Gerstner and Kistler, 2002) (Izhikevich, 2003). In the majority of models, a neuron is represented with three parts: dendrites, responsible for collecting signals from other neurons; soma, the processing unit, and; axon, from which signals are released. Most of these attempts model information processing at a neuronal level and consider the spiking characteristic as a means of communication between neurons (Gerstner and Kistler, 2002).

Before describing how the neurons organize into ensembles, the main properties of the most popular models of neurons are succinctly presented. The models are presented in order of biological relevance. The connectivity of neurons in networks and, in particular, information encoding is described in the Section 2.1.2 (The organization of neurons).

1) McCulloch-Pitts neuron (1943)

McCulloch-Pitts neuron (Figure 2.3) (McCulloch and Pitts, 1943), also called the threshold logic unit, basically consists of a processing unit with binary inputs/outputs {0 or 1} and the threshold represented with real numbers. There is no notion of weights in its input connections. When an input receives a 1 signal, the inner state of the neuron is added by one unit (Threshold step function). This elementary processing unit was able to process AND and OR functions, but not XOR (exclusive OR) and NXOR (not exclusive OR).

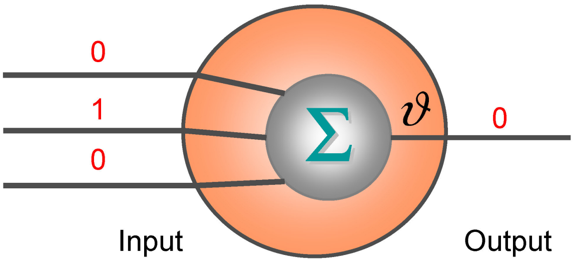


Fig. 2.3. McCulloch-Pitts neuron.

2) Perceptron (1949)

With Hebb's discovery in relation to changes in synaptic efficiency according to neuronal activity, further proposing that this mechanism could be the basis of learning and memory (Hebb, 1949), the McCulloch-Pitts neuron needed to be upgraded. Rosenblatt (Rosenblatt, 1962) evaluated the Perceptron model, with inputs having weights (to be trained) and a binary activation function. Rosenblatt's work also introduced the Delta rule for training. Differentiable activation functions (sigmoids, piecewise linear, Gaussian, etc) were introduced much later in 1986 by Rumelhart (Rumelhart, 1986), to train a multi-layer perceptron (MLP) structure through back-propagation of errors. Figure 2.4 shows an illustration of the perceptron.

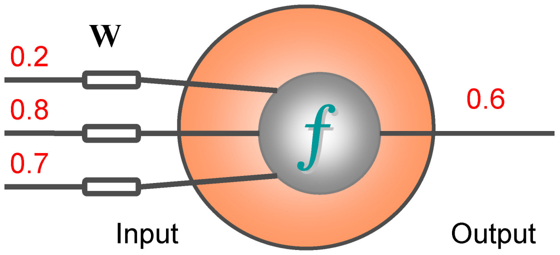


Fig. 2.4. Perceptron model. Note the innovations proposed over the years: changeable input weights, different activation functions, and inputs/outputs with real numbers.

3) Integrate-and-fire

The integrate-and-fire model was the first type of neuron that used the theory of spikes over time to convey information (See (Gerstner and Kistler, 2002) for an introduction). As the input spikes arrive in time, the inner potential of a neuron (postsynaptic potential - PSP) is increased/decreased with strength proportional to the connection weights. If the weights are positive, the connections are excitatory and an incoming stimulus acts to increase the PSP. Connections with negative weights are inhibitory as stimulus passing through them act to decrease the PSP. When the postsynaptic potential reaches a threshold, an output spike is released. In the event of an output spike, the PSP is reset to its resting potential, as can be seen in Figure 2.5.

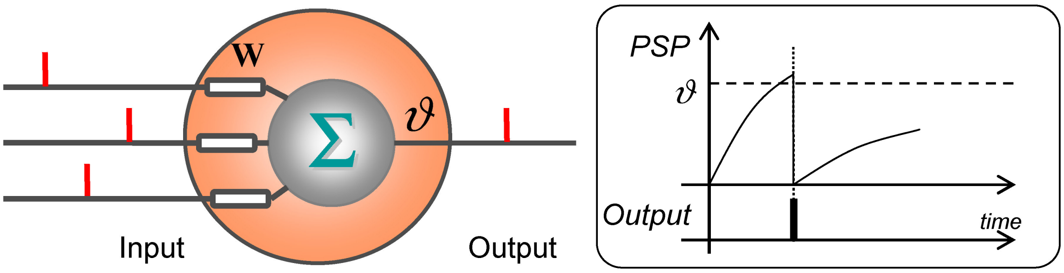


Fig. 2.5. Integrate-and-fire neuron.

4) Integrate-and-fire with leakage

Adding complexity to the integrate-and-fire neuron, the leaky integrate-and-fire neuron has its PSP ruled by a decay term, which decreases the magnitude of PSP over time (See (Gerstner and Kistler, 2002) for an introduction). Thus, when neurons cease to receive input excitation, the PSP gradually decreases and after some time reaches its resting potential. This mechanism, in its simplest form, can be associated with an RC electrical circuit where each neuron is composed of resistors and capacitors. Consequently, neuronal activity can be analyzed using the theory of electric circuits. The dynamics of a leaky neuron can be expressed by the change in the PSP (excitation/inhibition) upon spike arrival as:

$$PSP(t) = PSP(t) \pm A_{\max}\left(1 - \exp\left(-\frac{t - t_{ini}}{\tau_{rise}}\right)\right) \qquad \textbf{(2.1)}$$

where $A_{\max}$ is the maximum activation caused by a single spike, t_{ini} is the time of the incoming spike, and τ_{rise} is the excitatory/inhibitory time constant of the neuron. Some simplified models do not consider the exponential term in Equation 2.1. As a result, upon the arrival of a spike, the PSP is simply added to by the constant $A_{\max}$.

The PSP decay term, on the other hand, is described as:

$$PSP(t) = \pm A_{\max}\left(\exp\left(-\frac{t - t_{ini}}{\tau_{decay}}\right)\right) \qquad \textbf{(2.2)}$$

where $A_{\max}$ is the maximum activation caused by a single spike, t_{ini} is the time of the incoming spike, and τ_{decay} is the time constant for PSP decay in the neuron. Figure 2.6 shows the dynamics of a leaky integrate-and-fire neuron. Experiments have demonstrated that leaky integrate-and-fire neurons can very realistically reproduce the behaviour of biological neurons.

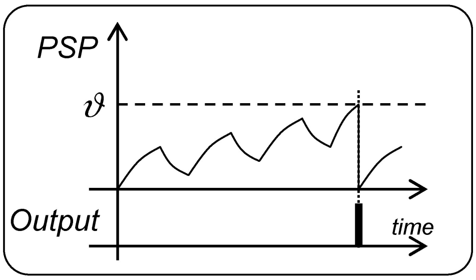


Fig. 2.6. Typical behaviour of a leaky integrate-and-fire neuron (rise and decay terms).

5) Spike response model

The spike response model (SRM) reproduces the electrical activity of the neuronal unit using kernels. The kernel representation gives a more general mathematical description of the neuron, enabling easy modification of the shape of the neuron's output curves by simply changing kernels. Gerstner and Kistler (Gerstner and Kistler, 2002) point out that leaky integrate-and-fire neurons can be considered a special case of the spike response model. The spike response model represents the PSP of a neuron according to two key events:

a) when a neuron fires: there is a strong depression of the PSP represented by the kernel $\eta(t - t^{(fire)})$, and

b) when a spike arrives: there is a excitation/inhibition of the PSP by an amount $w\varepsilon(t - t^{(IncomingSpike)})$. The variable w is the weight of the incoming connection and $\varepsilon(t - t^{(IncomingSpike)})$ is the kernel that represents the variation in the PSP when a spike is received.

The general equation that describes the PSP of a neuron i at time t is denoted by the integration of all events:

$$PSP_i(t) = \sum_{t_i^f \in OutputSpike} \eta_i(t - t_i^{(fire)}) + \sum_{j \in InputSpikes} w_{ij}\varepsilon_{ij}(t - t_j^{(IncomingSpike)}) \qquad \textbf{(2.3)}$$

A typical kernel equation for $\eta(t - t^{(fire)})$ is

$$\eta_i(t - t_i^{(fire)}) = C_{\max}\exp(-\frac{(t - t_i^{(fire)})}{\tau}) \quad \text{for} \quad t - t_i^{(fire)} > 0 \qquad \textbf{(2.4)}$$

where $C_{\max}$ is the maximum amplitude of the kernel and τ is the time constant.

Kernel $\varepsilon(t - t^{(IncomingSpike)})$ commonly has the form

$$\varepsilon_{ij}(t - t_j^{(IncomingSpike)}) = \exp(-\frac{t - t_j^{(IncomingSpike)}}{\tau_m}) - \exp(-\frac{t - t_j^{(IncomingSpike)}}{\tau_s}) \qquad \textbf{(2.5)}$$

for $t - t_j^{(IncomingSpike)} > 0$, where τ_m and τ_s are time constants.

The spike response model, similar to the integrate-and-fire neurons, does not consider the spatial structure of the neurons nor the mechanisms that rule the neuron dynamics at the ionic level (Gerstner and Kistler, 2002).

6) Neurogenetic model

Genetic studies show that human electroencephalogram has a strong genetic basis (Beijsterveldt and van Baal, 2002) (Buzsaki and Draguhn, 2004) (Porjesz *et al*, 2002). The description of the neurogenetic model (Benuskova and Kasabov, 2007) explicitly considers the influence of genetic dynamics on the behaviour of a neuron.

To model genetic dynamic behaviour, several genes considered directly relevant to neuronal activity are selected and grouped to form a gene regulatory network

(Dimitrov *et al*, 2004). Individual gene expression over time within the gene regulatory network can be extracted using linear/non-linear dynamic systems. In its simplest form, a gene regulatory network can be represented linearly as

$$g_j(t+\Delta t) = \sum_{k=1}^{n} w_{jk} g_k(t) \qquad \textbf{(2.6)}$$

where $g_j(t + \Delta t)$ gives the expression level of gene j at time $t + \Delta t$ when under the influence of k genes in the gene regulatory network.

Several parameters that control neuronal activity receive direct influence from a gene or set of genes according to

$$P_j(t) = c_j g_j(t) \qquad \textbf{(2.7)}$$

where $P_j(t)$ is the neuronal parameter under the influence of a gene/set of genes $g_j(t)$ at time t and c_j is a constant of proportionality. Figure 2.7 shows an illustration of the neurogenetic model.

Kasabov *et al* (Kasabov *et al*, 2005a) presents some of the parameters that are directly linked to genetic influence using the spike response model (SRM) of a neuron, e.g., time constants τ and maximum kernel amplitudes $C_{\max}$ (see Equation 2.4).

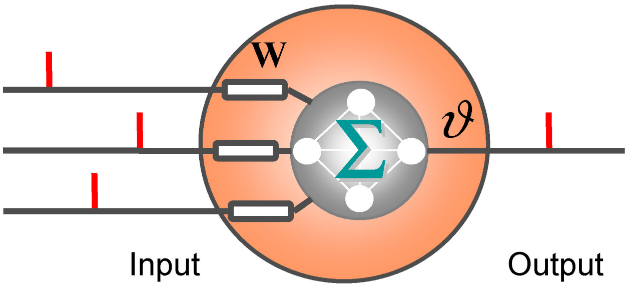


Fig. 2.7. Neurogenetic model of a neuron. Parameters that regulate neuronal activity are under constant influence of gene expression levels.

The neurogenetic model enables the simulation of: a) genetic interactions of neuronal related genes (genetic level); b) genetic influence on neuronal behaviour, i.e., the mechanisms that link gene expression level to neuronal behaviour; and c) an ensemble of neurons regulated by explicit genetic regulation (Kasabov and Benuskova, 2004) (Kasabov *et al*, 2005) (Kasabov *et al*, 2005a) (Kasabov *et al*, 2005b) (Wysoski *et al*, 2004).

7) Hodgkin-Huxley model

The Hodgkin-Huxley model (Hodgkin and Huxley, 1952) is often wrongly referred to as a model of a neuron. However, in reality, Hodgkin-Huxley tried to model the generation and propagation of action potentials in cells, primarily in the squid giant axon. Extended to neuronal and cardiac muscular cells, the model describes, with differential equations, the electrical activity of active membranes (axons and muscular cells) in terms of ion channels and chemical reactions. In the Hodgkin and Huxley model, there are three ionic channels: sodium (Na), potassium (K) and a general leakage channel. Voltage-dependent resistors represent the Na and K channels, whereas the general leakage channel is denoted by a passive resistor-capacitor parallel circuit (see Figure 2.8). The general equation of current that rules this circuit is

$$I_{total} = G(V - E) + G_{Na} m^3 h(V - E_{Na}) + G_K n^4 (V - E_K) \qquad \mathbf{(2.8)}$$

where G, G_{Na}, and G_K are the respective conductances, V is the difference of potential between inside and outside, E = -55.6 mV, E_{Na} = +50 mV, E_K = -77 mV. m, n and h variables that change with a set of differential equations. Whereas the complete mathematical formulation of the Hodgkin and Huxley model is out of the scope of this dissertation (Gerstner and Kistler, 2002) (Hille, 1992) (Koch, 1999) (Nelson and Rinzel, 1995), intuitively the combination of these differential equations, after being properly tuned, results in electrical activity that accurately represents biological cells. Particularly interesting to this research is the incoming stimulus (input spike), which results in a gradual increase of the inner potential. Further, if the inner potential reaches a certain level, a short-time duration current, now from inside-out (output spike), is generated, which is followed by an abrupt decrement of the inner potential.

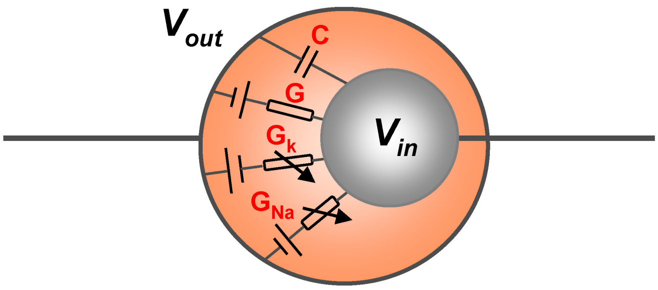


Fig. 2.8. Hodgkin-Huxley model representation.

8) FitzHugh-Nagumo and Izhikevich models

Over the years, the Hodgkin-Huxley equations demonstrated themselves to be computationally expensive, especially when simulating a large number of cells. For this reason, some models were proposed to simplify the Hodgkin-Huxley equations without losing precision. The FitzHugh-Nagumo model deserves special mention (FitzHugh, 1961) (Nagumo *et al*, 1962), a two-dimensional representation of the squid giant axons as does the Izhikevich model (Izhikevich, 2003). The latter is a simple and computationally inexpensive neuron model (suitable for large-scale simulation) that uses two coupled differential equations that prove to be able to reproduce several biologically realistic neuronal behaviours (brain-like activity, bursting, etc).

9) Compartmental models

Compartmental models are related to the description of a neuron into several functional compartments (see Figure 2.9). In particular, compartmental models attempt to explicitly describe the dendritic tree in its simplest form with the cable equation. The dendrites are considered passive compartments whose length and diameter are parameters. The spatial description of the dendritic tree can be very complex and can easily reach hundreds of compartments, e.g., the Blue Brain project (EPFL BlueBrain, 2007) considers around 400 compartments. In the compartmental models, the soma and its inherent non-linear processing capabilities, is considered as one or several compartments. See (Bower and Beeman, 1995) and (Gerstner and Kistler, 2002) for further analysis and references.

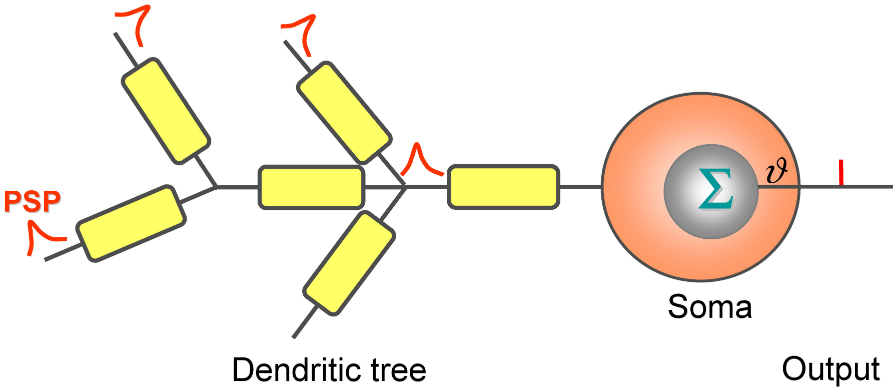


Fig. 2.9. Illustration of dendritic trees modelled into compartmental models. The soma and each dendrite are modelled as individual compartments.

Table 2.1 shows the different neuronal unit models according to their biological accuracy and their corresponding use.

Table 2.1. Classification of artificial neural network models according to biological relevance and corresponding use.

	Biologically motivated	**Moderate biological relevance**	**No biological relevance**
Model	Hodgkin-Huxley Compartmental models Neurogenetic model	Integrate-and-fire Spike response model Izhikevich model FitzHugh-Nagumo model	McCulloch-Pitts Perceptron
Usage	Simulate activity at the ionic channel level and dendritic trees. Commonly used by neuroscientists.	Simulate electrical activity of the neurons. Used to implement large networks, evaluate temporal properties and synchronicity of spiking neurons, and pattern recognition.	Basic processing unit for connectionist systems. Used in artificial intelligence to solve engineering problems.

2.1.2 Organization of neurons into specialized functional ensembles

Brain areas with different behavioural characteristics are formed according to how neurons are arranged and organized into ensembles. In a very simplified manner, the neurons connect to each other in two basic ways: through divergent and convergent connectivity. Divergent connectivity occurs when a given neuron has more output than input connections. Convergent connectivity, on the contrary, occurs when a neuron has more input than output connections, as illustrated in Figure 2.10.

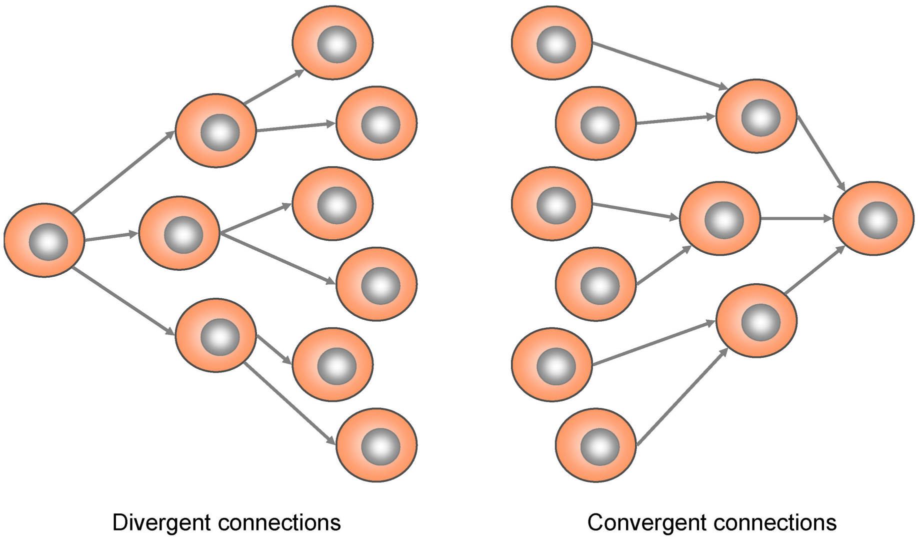


Fig. 2.10. Divergent and convergent neuronal connectivity. Modified from (Kandel, 2000).

The connectivity of neurons also follows a certain spatial order. Neurons are generally connected to neighbouring neurons and as the distance between neurons increases, connectivity decreases (there are few long distance connections). This principle is explored and modelled from the perspective of small world networks (Buchanan, 2003) (van den Berg and van Leeuwen, 2004) (Watts, 1999). The theory of small world networks enables the emulation of different network structures, from highly ordered networks to networks with a high level of randomness in their connections. This biologically inspired concept of connectivity has been also applied to other fields that require distributed processing (Dorogovtsev and Mendes, 2003).

Neuroscientists have also extensively studied the organization of neurons at higher instances (macroscopic level), which has enabled the understanding of the organization of neural units into functional areas (functional systems) (Bear *et al*, 2001) (Kandel, 2000). A functional system is responsible for performing specific tasks and processing specific types of information. Further experiments enabled the discovery of correct links between functional systems, i.e. building charts to describe the information pathways when different cognitive functions are executed by the brain.

To better illustrate the concept of functional areas, let’s take the example of sensory modalities. Cognitive functions take place mainly in the cerebral cortex. Anatomically, the cerebral cortex is a thin outer layer of the cerebral hemisphere with a thickness of around 2 to 4 mm and it is divided in four lobes: frontal/parietal/temporal/occipital (see Figure 2.11). Different functions are executed in different areas of the cerebral cortex as described in Table 2.2.

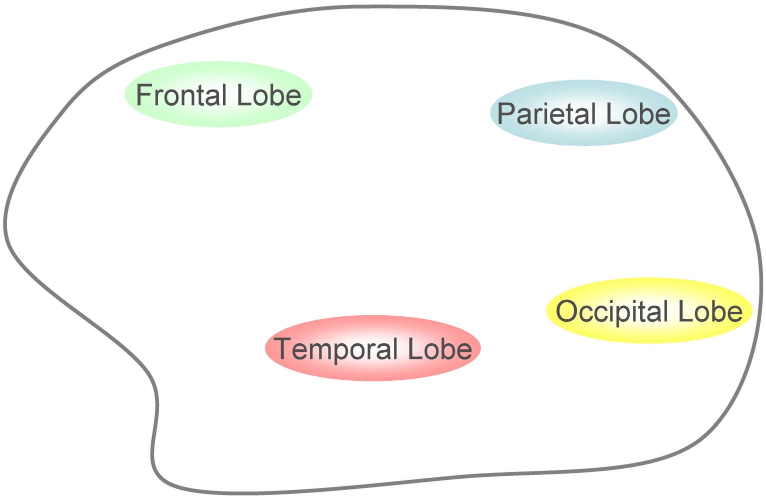


Fig. 2.11. Cerebral cortex according to (Kandel, 2000).

Table 2.2. Location of cognitive functions in the brain.

Cerebral cortex location	Function
Occipital Lobe	Visual sensory
Temporal Lobe	Auditory sensory
Occiptotemporal	Visual association
Temporal	Auditory association
Parietotemporal	Multimodal association (visuospatial location, language)
Temporal, frontal, parietal	Multimodal emotions, memory

Figure 2.12 shows a more detailed description of the visual information pathway. The retina acquires the image and performs the first processing. In the occipital lobe, the image is split into sub modalities that are processed in parallel (colour, motion, shape, and depth). The parietotemporal region is responsible for visuospatial location and the occiptotemporal region is where visual patterns are recognized.

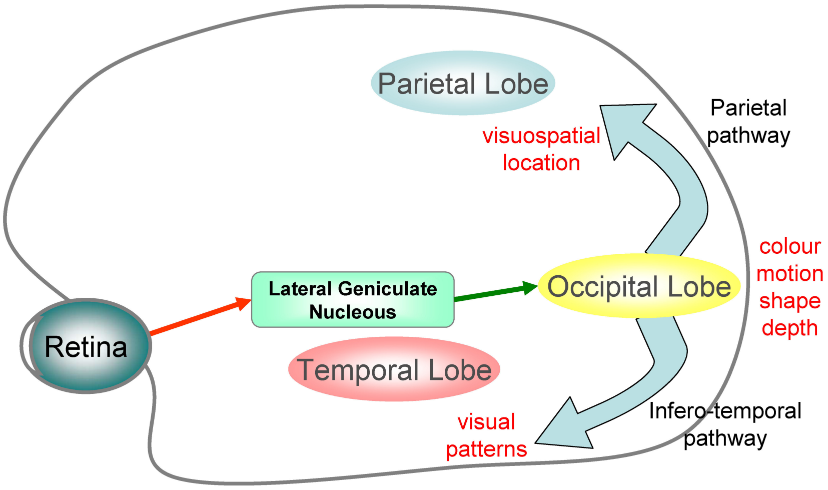


Fig. 2.12. Pathways of the visual system and organization of neurons into functional areas. Visual excitation propagates through the retina and follows parallel pathways in the occipital lobe, which process colour, motion, shape, and depth. Visuospatial location is then processed in the parietotemporal area whereas visual patterns are recognized in the occiptotemporal region. Modified from (Kandel, 2000).

In summary, when it comes to the organization of the neurons, the connectivity of single neuronal units is well understood and can be experimentally measured. At higher instances, there is much literature that divides the brain into functional areas and describes information pathways for most cognitive tasks (See (Kandel, 2000) for a extensive list of references).

2.2 Information processing in sensory pathways

The human brain mainly deals with five sensory modalities: vision, hearing, touch, taste and smell. Each modality has different sensory organs containing specialized receptors. After the receptors perform stimuli transduction, the information is propagated by the generation of neuronal action potentials. This mechanism follows a common pattern for all sensory modalities (despite it still not being clear how the information encoding process occurs). However, at the functional level, due to specialization caused by evolution or environmental adaptation (Stein and Meredith, 1986), ensembles of neurons present very peculiar processing differences.

Sensory systems are able to acquire four fundamental properties:

a) What. Indicates what is measured. For this property, there are specific cells for each modality (e.g. thermoreceptors, photoreceptors, chemoreceptors);

b) How much (intensity). The intensity of excitation, i.e., the ability to grade the incoming stimuli in levels;

c) Where (location). The location of the stimuli. Sensors responsible for this task are commonly topographically organized to represent spatial distances;

d) When (time). The time of an event or when stimuli occur. The temporal analysis in sensory systems is carried out by adaptation of receptors, which mostly respond at the onset/offset of stimulation.

In general terms, sensors actively capture variations of the fundamental properties (what, how much, where and when), e.g., ∂Intensity/∂t (variation of intensity in time), ∂Light/∂x (variation of illumination in a certain direction), etc. Additionally, sensors are not simply passive transducers. As the signals arrive at the higher levels of the sensory system, feedback is sent to the sensors, a phenomenon called centrifugal control of sensory relays (Stein and Stoodley, 2006), which effectively acts to filter the passage of sensory signals from the periphery to the higher levels to avoid redundant information being sent.

Another characteristic that is found in all sensory systems is that, for each sensory cell acquiring an input signal, another thousand neurons are behind the scenes analysing it. For instance, in the visual system, the signals of each ganglion cell (there are around one million in humans) are processed by approximately 1000 neurons in the primary visual cortex (Stein and Stoodley, 2006).

How information is processed in the auditory and visual system is summarized in the following subsections. Audiovisual integration is also described. Without covering all that is known about each modality, the most relevant differences in terms of information processing are stressed and references to artificial models are provided.

2.2.1 Processing visual information

On the subject of biological approaches to processing visual information, Hubel and Wiesel received many awards, including the Nobel Prize, for their description of the human visual system (Hubel and Wiesel, 1962). Through neurophysiological experiments, they were able to distinguish cell types that have different neurobiological responses according to the pattern of light stimulus. They identified the role that the retina has as a contrast filter as well as the existence of directionally selective cells in the primary visual cortex. Their findings have been widely implemented in biologically realistic image acquisition approaches. The idea of contrast filters and directionally selective cells can be considered a feature selection method that has a close correspondence with traditional ways of image processing, such as wavelets and Gabor filters (Sonka *et al*, 1998).

Visual Pathways

The visual information acquired in the ganglion cells of the retina are conducted through the optic chiasm to the lateral geniculate nucleous (LGN) in the Thalamus. The LGN propagates information to the primary visual cortex (V1) via optic radiations. From the primary visual cortex to the LGN (centrifugal control) feedback connections act to filter the input of LGN and provide selective attention (See Figure 2.13). From V1, the signal is sent to more than 30 different higher visual areas (Bear *et al*, 2001).

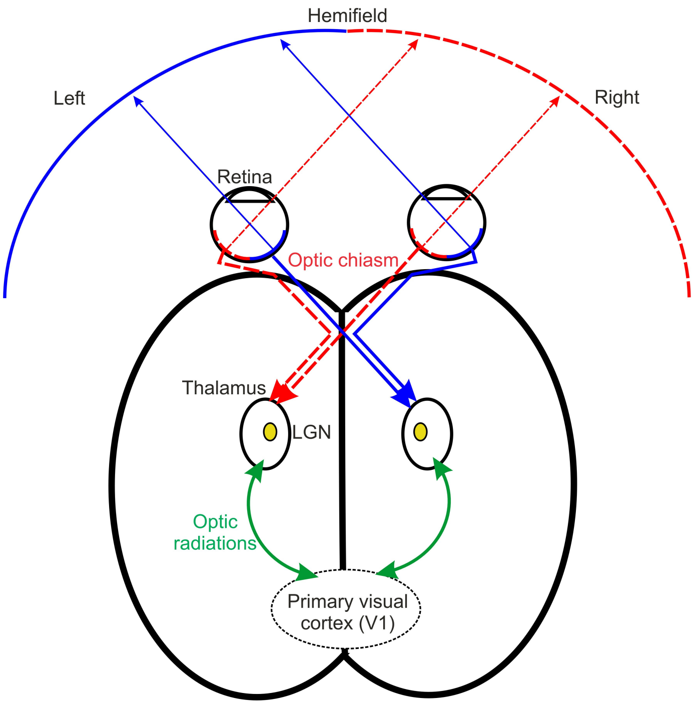


Fig. 2.13. Visual information is transduced in the ganglion cells of the retina and conducted through the optic chiasm to the lateral geniculate nucleous (LGN). LGN propagates information to the primary visual cortex (V1) via optic radiations. Feedback from the primary visual cortex to the LGN acts to filter and provide selective attention. Modified from (Kandel, 2000) and (Bear *et al*, 2001).

From here, and particularly interesting to this research, the information pathways responsible for processing shape/colour and location of a visual stimuli are described.

“What” pathway

The “what” pathway propagates the information from the LGN to the primary visual cortex V1, then to V2, V3 and V4 until it reaches the inferotemporal cortex (IFC). In these areas, the visual information is decomposed into different types (shape, colour, texture) (Figure 2.14).

In the LGN are found the On/Off cells that are sensitive to high contrast visual stimuli. In V1 the cells start to be more sensitive to elongated edges or parts of the image that have the same orientation. In V2 are found the complex cells, i.e., cells that are sensitive to complex shapes. In this area, cells start to become more selective of complex forms and are invariant to shifts in location. In the IFC, topographic maps can no longer be noticed and the big visual receptive fields are completely location invariant. Note that, as the information advances along the visual pathway the more specialized the neurons become. Using this theory, it is possible to imagine that at the end of the information pathway there is a cell to respond to a specific stimulus (commonly referred as a grandmother cell). However, this extreme scenario does not seem to happen in reality, as it could drastically limit the ability to memorize different patterns and could become too dependent on single processing units. Stein and Stoodley (Stein and Stoodley, 2006) provides some references that describe complex cells sensitivity to basic forms, e.g., circles, squares, parallel bars. When it comes to some very specific natural shapes, like hands, or faces, only specific areas are reported. It is further suggested that, in general, IFC neurons represent objects in terms of basic complex shapes, and through relatively small networks these shapes are associated to classify large amounts of objects.

A similar pathway describes the extraction of colour information. Cells in the retina, similarly to the On/Off contrast cells, are sensitive to the light's wavelength as a function of position (∂Wavelength/∂x). For instance, cells can be topographically organized with the centre sensitive to green colour and the surrounds sensitive to red. In the LGN there are still colour discriminative cells (green/red and red/green), but also yellow/blue organization appears (although not in a centre/surround discriminative organization). The blue/yellow mechanism does not follow spatial segregation. The yellow wavelength simply provides excitation to cells whereas blue colours inhibit them. Also in the LGN On/Off cells with higher spatial frequency are observed, which mainly act in detecting luminance contrast. In the primary visual cortex (V1), a slightly different type of cell can be observed. V1 cells still have the centre/surround spatial property, however, the centre cells are excited by a certain wavelength (e.g. green) whereas the surrounding cells are inhibited by another wavelength (e.g. red). These cells enable the reliable discrimination of colour (Stein and Stoodley, 2006). The areas after V1 have complex mechanisms that allow the perception of colour constancy, i.e., even with changes in the wavelength of the incoming light, the perception of colour remains the same.

Even though the idea of hierarchical pathways described by Hubel and Wiesel is still valid (Hubel and Wiesel, 1962), it is now considered too simple. Several studies followed, demonstrating the importance of adding numerous feedback and inhibitory processes, especially in the primary visual cortex (e.g., (Adini *et al*, 1997) (Cudeiro and Silito, 1996)).

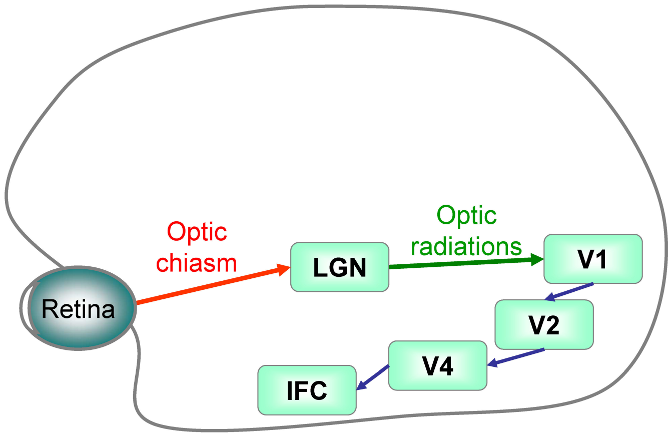


Fig. 2.14. Visual "what" pathway. Information propagates from the LGN to the visual cortex V1, V2, V3 and V4 until it reaches the inferotemporal cortex (IFC). Modified from (Bear *et al*, 2001).

"Where" pathway

The "where" visual pathway follows the LGN to V1, V2, MT (middle temporal) until it reaches the posterior parietal cortex (PPC) (Figure 2.15). In the primary visual cortex (V1) the information from both eyes comes together, and is processed by binocular cells, i.e., cells that are sensitive to the difference between the images generated by the left and right retina. These neurons are able to compute the binocular disparity (difference between left and right eye images).

However it is important to note that binocular disparity is not the only cue used by the brain to have a sense of depth and, consequently, 3D position. Some monocular cues that are processed at higher levels can also help in the position detection task, e.g., perspective, motion parallax (relative motion between objects), occlusion, etc. This mechanism takes place in the posterior parietal cortex (PPC).

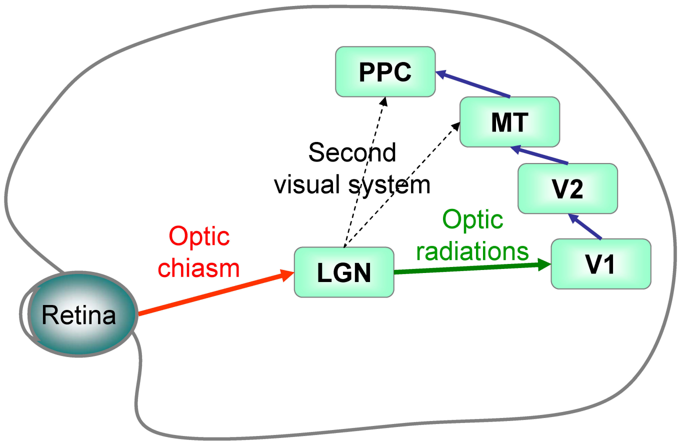


Fig. 2.15. Visual "Where" pathway. Modified from (Bear *et al*, 2001).

The detection of movement is done firstly by the magnocellular ganglion cells in the LGN, which are highly sensitive to the beginning and end of a light stimulation (100 Hz excitations can be detected). However, it is in the primary visual cortex where light direction and speed sensitive complex cells appear, here sensitive to slower rates (as few as 30 Hz). These cells are excited if certain objects move in a certain direction and inhibited if the object makes a contrary movement. The speed in which the object moves determines the level of neuronal excitation. The middle temporal (MT) area is where most neurons sensitive to movement are found, which include the processing of direction, speed, optic flow, change in size and disparity.

In addition to the normal visual pathway (LGN → V1 → V2 → MT → PPC), the processing of movement, has connections directly from LGN to MT and PPC areas, often referred as the "Second visual system" (Figure 2.15). The visual pathways are schematically illustrated in Figure 2.16.

There are several mathematical models that try to emulate the crucial features of the visual system (Fukushima, 1997) (Fukushima and Miyake, 1982) (Mel, 1998) (Thorpe, 1997). Some of these models are further explored in Chapter 3.

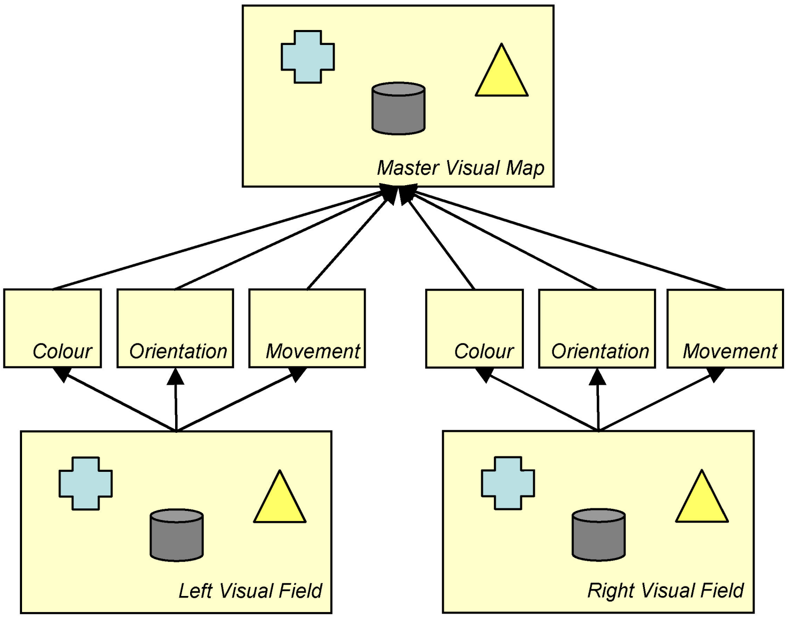


Fig. 2.16. Diagram of the visual pathway.

2.2.2 Processing auditory information

In the human ear, sounds waves are conducted to the external and middle ear, which effectively pre-process the sound signals before their arrival at the inner ear where transduction takes place in the cochlea (See Figure 2.17).

Auditory information is captured by biological systems with tonotopically-organized maps in the cochlea (cochlear hair cells transduce mechanical vibrations to action potentials). With the tonotopically-organised maps, frequency components of the sounds, tone, pitch and timbre can be extracted. Human ears are able to detect frequencies in the range of 20 to 20000 Hz. From spectral characteristics, it is possible to distinguish different sounds as well as to recognize the sound's source. In the cochlea, thousands of hair cells are sensitive to different frequencies. The auditory system, in addition to the analysis of spectral characteristics in time, also analyses amplitude changes (increase and decrease of sound amplitude) in time (Bear *et al*, 2001). Each human ear processes the incoming signals independently. Taking into consideration the signal's timing, amplitude and frequency, the integration occurs in later stages. In addition, the narrow difference in time between incoming signals from the left and right ear results in a reliable cue to perform sound source localization (Bear *et al*, 2001).

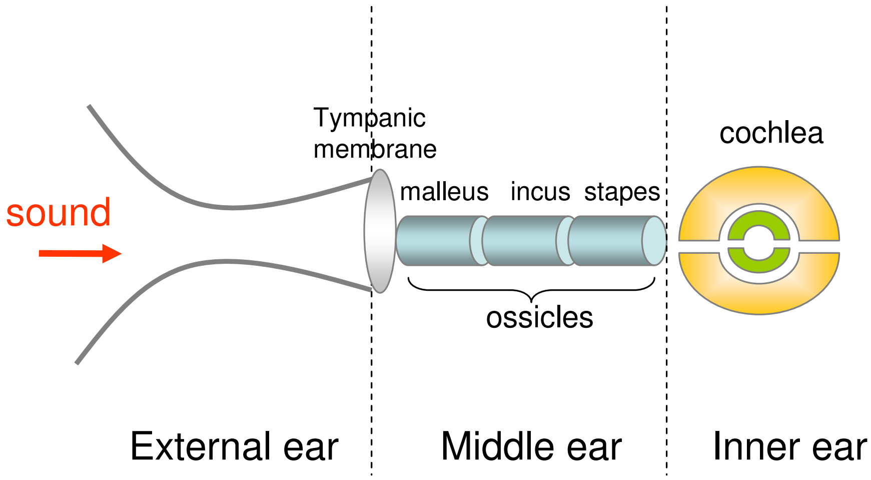


Fig. 2.17. Sound pathway before arrival in the cochlea for transduction. Modified from (Bear *et al*, 2001).

The spiral-like cochlear duct is composed of three tubes (scala vestibule, scala media and scala tympani) which contain the auditory transducers. Cochlear hair cells are tuned to respond at a single frequency. The output of the transduction of pressure changes into electrical discharges can reach 1000 pulses per second. In contrast to the retina cells (responsible for visual transduction), there is strong centrifugal control coming from higher levels of processing to the hair cells in the cochlea (adaptive feedback). This mechanism effectively acts to perform selective auditory attention, dynamically emphasizing the properties of interest in the sound signal (Bear *et al*, 2001).

After transduction is realized in the cochlea, where hair cells are placed tonotopically, the information advances to subsequent levels, the tonotopic organization loses selectivity and the cells become more specialized in detecting amplitude and frequency modulation. From the cochlea, the information flows to the cochlear nucleus, which is divided into ventral and dorsal, VCN and DCN, respectively. These two regions have different types of cells and effectively define two different pathways (See Figure 2.18):

- VCN signals flow to the contralateral superior olive, to the central nucleus of inferior colliculus and medial geniculate nucleous in a "what" pathway, mainly recognizing what the object of the auditory system is, through spectral analysis (tonotopic organization);
- The alternative route that starts in the DCN, passes to the nuclei of the lateral lemniscus, external nucleus of the inferior colliculus and the dorsomedial MGN, mainly responsible for detecting the dynamic variation of sound signals, defining the attention and movements resultant in the recognition of sound signals (mainly non-tonotopic organization).

These are the most important paths over which auditory signals flow, however, there are several other pathways, mainly providing feedback connections or integrating other sensory modalities. One example is the aforementioned auditory cortex feedback to the cochlear hair cells, MGN and inferior colliculus for selective attention. In terms of integration of modalities, the inferior colliculus is linked to the superior colliculus, where audiovisual integration occurs (Bear *et al*, 2001).

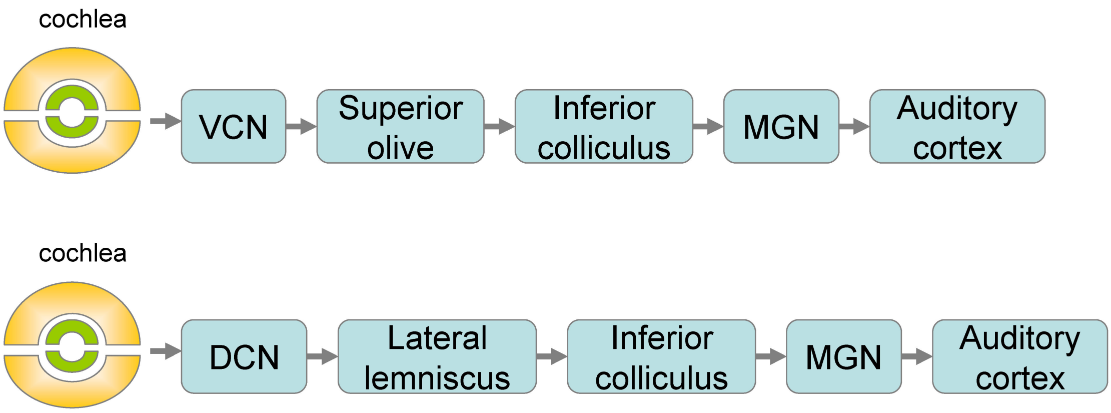


Fig. 2.18. Main auditory pathways. VCN has typically tonotopic organization and DCN is mainly non tonotopic.

In general terms, the intensity of sound stimulus is coded in firing rates (the higher the intensity, the higher the firing rate) and the number of active neurons (the higher the intensity, the more neurons are active). Frequency, on the other hand, is represented with tonotopic maps and phase locking, i.e., consistent firing of a cell at the same phase of a sound wave.

There are numerous artificial models seeking to reproduce the human ear (Kuroyanagi and Iwata, 1994) (Loiselle *et al*, 2005) (Shamma *et al*, 1986), and in most cases, modelling is divided into three different tasks: sound source location, sound recognition or source recognition. Artificial models for analysing sound commonly decompose the signal into various frequency bands that are processed in parallel pathways (Figure 2.19). In this respect, the most popular and biologically proven way to model the auditory system is to decompose the input sound in the frequency domain according to the MEL scale (Rabiner and Juang, 1993). The MEL scale models the non-linear way the human auditory system perceives sounds. From the MEL scale the MEL Filter bank can be derived, which consists of a bank of band-pass filters tuned according to the MEL Scale.

Based on MEL scale, MEL Frequency Cepstral Coefficients (MFCC) are extracted, which apply the discrete cosine transform (DCT) to the energy found in the MEL subbands. MFCCs are popularly used as a feature extraction method in a range of applications that involve the processing of sound and speech signals. Another feature extraction method widely applied in sound signals is Linear Predictor Coefficients

(LPC) and its cepstral counterpart (LPC cepstral coefficients) (Rabiner and Juang, 1993). Recently, wavelets have been successfully applied to compress and extract relevant sound characteristics (Ganchev, 2005). MFCC, LPC and wavelet approaches will be used as a baseline for the brain-like models for speaker authentication analysed and further developed in Chapter 4.

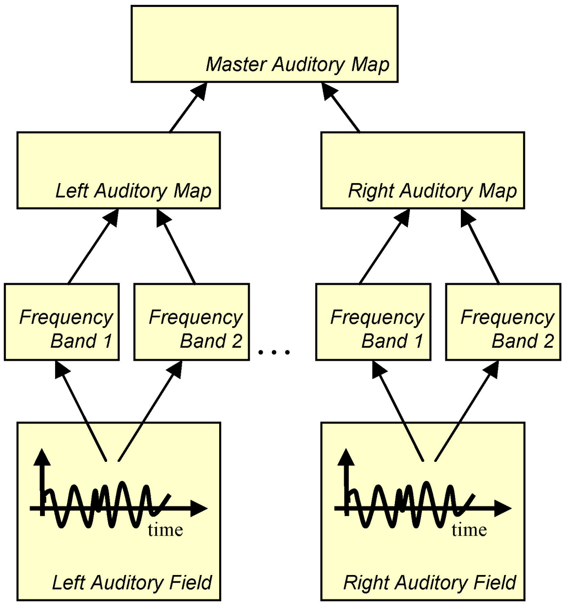


Fig. 2.19. Diagram of the auditory system pathway.

2.2.3 Integration of audiovisual information

As could be seen in the auditory and visual sensory systems, each sensory modality has mostly (not fully) distinct pathways where information is processed. Information is further split within a sensory modality. In the visual system, for instance, the information is divided into sub modalities (colour, shape and movement) that are independently processed in different pathways. In the auditory system, the ventral cochlear nucleous (VCN) and dorsal cochlear nucleous (DCN) also define different pathways. In different modalities and sub modalities, it is reasonable to think that the speed of transduction and the speed of information propagation in different pathways is not the same. If this is true, afferent stimuli from different sensory modalities arrive at the cerebral cortex at different times. The separation and integration of pathways within a modality as well as the integration of pathways from different modalities (and all the synchronizations implied in it) constitute a complex network that cannot be accurately described and reproduced. A simple illustration of the complexity of integrating the auditory and the visual senses is shown in Figure 2.20.

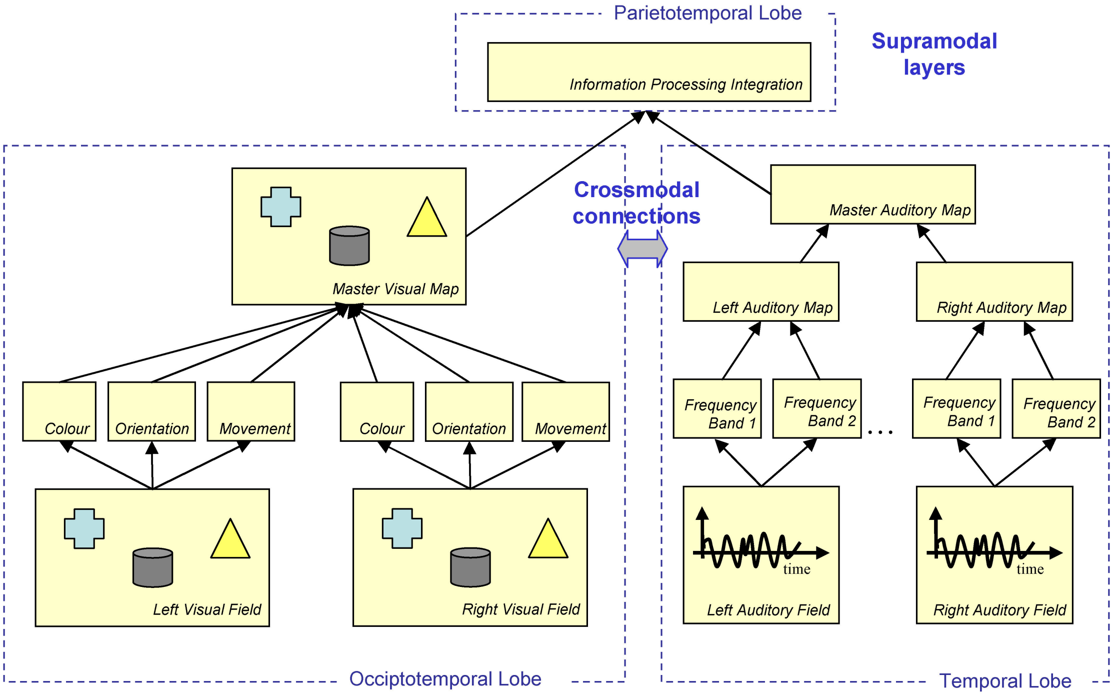


Fig. 2.20. Integration of sub modalities of the visual field (left side), the auditory pathways (right side) and the subsequent integration of modalities (above).

In general, the integration is done in brain areas composed of neurons sensitive to different types of sensory stimulation. Convergent connections convey different types of information to these areas (in the case of different modalities they are called supramodal areas). The convergence of information executed in parallel for accomplishing a specific task is called the binding problem. It is believed that the integration of the information is synchronized through the oscillation of thalamocortical circuits (at around 40-80 Hz, i.e., γ waves). These oscillations help to put together, for instance, the colour and the shape of the same object that is being processed in separate pathways (Stein and Stoodley, 2006).

However, the idea of having areas specifically responsible for integration, while not completely wrong, was shown to be very simplistic, mainly because the information pathways are not completely independent. There are numerous experiments which report influence from one information pathway to others, even cases where the information is acquired from different sensory modalities (crossmodal influences) and the nature of information seems to be completely uncorrelated.

Without aiming to completely explain the integration mechanism, the following describes three different viewpoints that can contribute to the understanding of the complexity and the inherent concatenation and overlapping between modalities.

The integration of modalities are analysed according to (Calvert, 2001) by:

a) the emergence of integration;

b) the evolutionary perspective, and;

c) the complexity of organisms.

The controversy about the integration of sensory modalities starts very early, with the *emergence of integration* or, as some would suggest, the segregation of modes. Some researchers believe that different senses originate from a primitive unity at birth and a person needs to learn to differentiate among them. Other groups believe that the senses are separated at birth and, through experience, can be interrelated.

Under the *evolutionary perspective*, it is believed that modalities have evolved from an undifferentiated (supramodal) system, not selective to specific sensory stimuli. The appearance of receptors sensitive to a specific sensory mode comes from the process of specialization and environmental adaptation.

In respect to the *complexity of different organisms*, the idea of multi-modes/sensory segregation appears even in the simplest multicellular invertebrates. Sponges, the lowest of the multicellular species, consist of a loose network of cells, where different sensory signals are transmitted in all directions simultaneously. However, it is with these species that nerve cells, synapses, and nerve nets first appear, with an aggregation of cells forming specialized sensory organs (detecting rotation, vibration, and optical stimuli). Within the platyhelminthes (flatworm), bimodal neurons have been identified, where neurons can be sensitive to illumination and vibration at the same time. In addition, in marine snails traces of crossmodal integration (association between modes) have been detected. Experiments in marine snails using conditioning paradigms successfully transferred the normal behavioural response to water turbulence to a light stimulus. After training, scientists were able to change the natural behaviour of the snails to move towards the light sources, to the attach-to-the-surface response, which is done naturally in the case of turbulent water. An in-depth analysis showed that the conditioning altered the ionic changes induced by light and rotation in a class of cells sensitive to both stimuli. This experiment demonstrated that, associations between modalities can exist in comparatively simple organisms that do not possess high levels of cognition. The highest level of segregation of sensors is found in the vertebrates. However, there are many works describing, for instance, nonauditory influences on neurons in the auditory cortex and nonvisual influences on thalamic and visual cortical neurons (Stein and Meredith, 1993).

In (Calvert, 2001) and (Stein and Meredith, 1993) an in-depth analysis is presented that considers the evolutionary aspects and the complexity of different organisms. From that discussion, the conclusion is that there is relative independence at the level of sensory signals in vertebrates (mainly on the primary projection pathways) which decreases with a high incidence of multisensory influence when higher levels of processing are reached. Such a statement accommodates the existence of specific sensory organs, supramodal areas and crossmodal interactions.

With advances in the understanding of the different pathways of information processing in the brain, several computational models to reproduce similar behaviours were promptly introduced. Models describing each stage of information processing are aimed at assisting the performance of functional analysis and suggest new theoretical frameworks for studying brain activity (Bruce and Young, 1986) (Burton *et al*, 1990) (Ellis *et al*, 1987), whereas other models are only aimed at using biological characteristics to increase performance of multimodal information processing systems (Kasabov *et al*, 2000).

2.3 Classification of neural-based information processing models

This section introduces a classification of artificial systems that use ensembles of neurons for modelling information processing. Artificial systems are divided into three classes. The first class, *connectionist systems* (COS), despite being based on processing units many times referred to as “neurons”, do not actually have any commitment to reproducing the biological behaviour of a neuron or biological system. The processing units in connectionist systems are simple mathematical equations and are optimized to perform mathematical operations. The second class encompasses the theory of *evolving connection systems* (ECOS). ECOS, when compared to COS, move towards a biological way of processing as they incorporate the ideas of continuous adaptation and evolution. In the final class, the *Brain-like Evolving Connectionist Systems* (*Brain-like* ECOS), as well as the ECOS properties, there is also a willingness to give a biological interpretation to the information processing mechanisms. In this classification, *Brain-like* ECOS is a subclass of ECOS, which is a subclass of COS, i.e., COS ⊃ ECOS ⊃ *Brain-like* ECOS (Figure 2.21).

The three classes of artificial systems are further described in the next sections. As this research concentrates on the *Brain-like* ECOS, the new systems to be described in next chapters are always evaluated considering the *Brain-like* ECOS aspects detailed in Section 2.3.3.

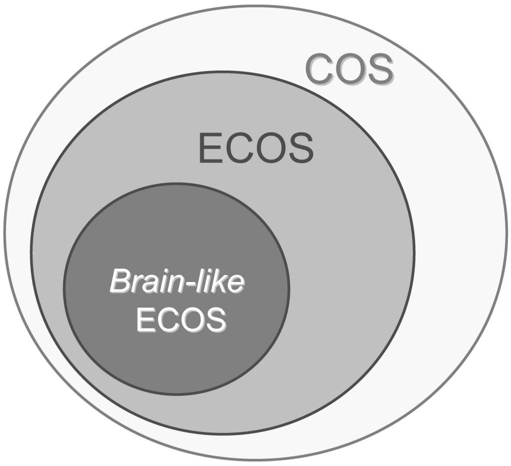


Fig. 2.21. *Brain-like* ECOS as a subclass of ECOS, which is a subclass of connectionist systems (COS).

2.3.1 Connectionist Systems (COS)

Connectionist systems have a broad definition that crosses different areas of knowledge. In this dissertation, the definition and the use of the term "connectionist systems" is restricted to the artificial intelligence field, where it is used to define systems that process information through the use of a network of simple processing units. In artificial intelligence, the term *connectionist systems* is commonly used only to refer to a network of neurons. However, the definition is broader and includes networks of any type of processing units, e.g., neuron models, units with step-wise, sigmoid, Gaussian-like functions, Fuzzy nodes, etc (Haykin, 1999) (Kasabov, 1996).

Two main properties that are inspired by the way brain processes information and make connectionist systems particularly appealing for solving a range of problems are: the distribution of the processing over simple processing units and the ability of the systems to learn to perform a specific task through examples.

As a consequence of these two properties, several characteristics can be derived (Kasabov, 1996):

1. connectionist systems are able to generalize and respond to unseen inputs, i.e., they are not limited exclusively to respond to previously learnt stimuli;
2. connectionist systems can be robust to local damage, i.e., when a few individual processing units fail to operate normally, the system can still offer satisfactory behaviour;
3. connectionist systems can be robust to noise and can still process reliably with missing information;
4. connectionist systems are intrinsically suitable for performing parallel processing.

Once the potential of the connectionist system paradigm was properly understood, a huge movement towards the design and implementation of new connectionist algorithms was initiated. Several processing units have been proposed as well as numerous architectures and learning rules. Haykin (Haykin, 1999) presents an extensive list of systems, deserving special mention:

- Multi-layer Perceptron Neural Networks;
- Recurrent Neural Networks;
- Radial Basis Function;
- Linear Vector Quantization;
- Fuzzy Neural Networks;
- Self-Organizing Maps.

In relation to learning rules, two main categories can be distinguished: supervised and unsupervised. The Hebbian rule (Hebb, 1949), in its unsupervised or supervised form, is the most used as it is more biologically plausible and because it can be interpreted to accommodate different processing unit mechanisms. Unsupervised rules are usually concerned with finding intrinsic properties from training data, which is obtained by searching for stable internal states within the system (Haykin, 1999). Supervised learning seeks to minimize output errors through gradient descent, general-purpose optimization algorithms, etc (Haykin, 1999).

2.3.2 Evolving Connectionist Systems (ECOS)

According to (Kasabov, 2007), "*ECOS is a framework that facilitates the modelling of evolving processes based on multimodal connectionist architectures*". ECOS mainly encompass the modelling of complex systems where boundaries cannot be easily defined and/or changed over time. ECOS gives freedom to the modelling process to expand and contract its coverage range as is required at any time. The ultimately goal in ECOS is the design of systems able to adapt and represent an open space and its changes over time.

ECOS use several different processing tools or units. Each tool in ECOS must allow functional and structural changes over a lifetime. Thus, several connectionist systems (COS) designed to model closed and static systems have been adapted to fulfil these requirements. Examples of tools adapted to ECOS requirements are the Evolving Self Organizing Maps (ESOM), Evolving Radial Basis Functions (ERBF), IPCA (Incremental Principal Component Analysis) to mention only a few (Kasabov, 2007). Other algorithms have been designed specifically with the ECOS paradigm in mind, e.g., Evolving Fuzzy Neural Networks (EFUNN) and Evolving Classifier Functions (ECF). The description of a wide collection of ECOS tools can be found in (Kasabov, 2007).

As described in (Kasabov, 2007), the ECOS paradigm is based on the way the human brain performs cognition: it is modular, it allows sequential, parallel, hierarchical processing and it allows lifelong adaptation. In addition, ECOS tools incorporate the use of brain-based learning, information acquisition, and knowledge storage, enabling even the forgetting of past experiences. Figure 2.22 is a simplified diagram that illustrates the ECOS framework.

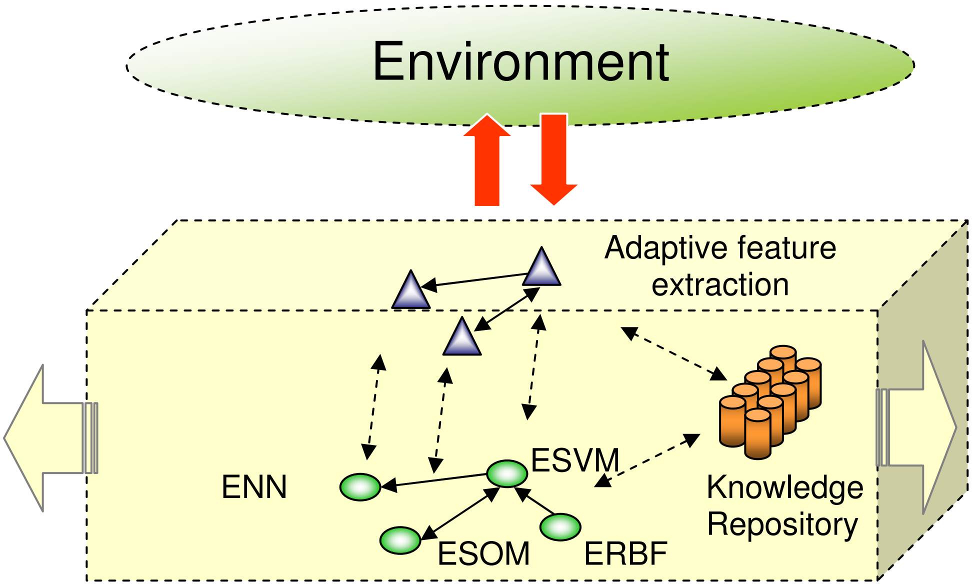


Fig. 2.22. An illustration of the ECOS framework. The figure represents a model in a hypothetical space, at a certain moment in time. The triangular nodes represent different adaptive ways of acquiring input stimuli (feature extraction). The circles represent some of the information processing tools (ENN = Evolving Neural Networks, ESVM= Evolving Support Vector Machines, ERBF = Evolving Radial Basis Function, ESOM = Evolving Self-Organizing Maps) (See (Kasabov, 2007) for details). The right-side nodes represent knowledge stored.

2.3.3 Brain-like Evolving Connectionist Systems (*Brain-like* ECOS)

Overall, the ECOS framework described in the previous section incorporates the biological experience into its components. However, the biological interpretation is not a *sine qua non* condition. A stricter version of ECOS (*Brain-like* ECOS) further requires each constituent tool to be evaluated in respect to its biological relevance. In addition to processing information with simple units integrated in a network trained by examples, which characterizes the connectionist systems (COS) (Section 2.3.1), and the extra requirement of adapting and evolving in an open space included in ECOS framework (Section 2.3.2)), *Brain-like* ECOS is further concerned with a consistent biological interpretation.

The biological interpretation required in *Brain-like* ECOS is mainly concerned with:

a) the use of biologically realistic processing units;
b) the biological way of connecting processing units in neuronal ensembles and functional areas (systemic level); and
c) the biological way of training and learning particular skills.

These three aspects seek to incorporate and evaluate the most recent discoveries from neuroscience for the design of new systems. In respect to the processing units, the spiking neurons theory is currently accepted as the most human-like manner of processing (Gerstner and Kistler, 2002). Therefore, they are the basis for all new designs presented in this thesis. At the systemic level, the behaviour of ensembles of neurons as well as the information processing pathways can be evaluated in biological terms. Of particular relevance to this dissertation are the auditory and visual systems that are discussed separately in Section 2.2. The auditory and visual pathways are considered in newly proposed audiovisual pattern recognition approaches. The learning theories and the corresponding algorithms to implement them are discussed from the perspective of computation with spiking neurons in Section 2.4.2. Table 2.3 shows some more specific aspects to be considered in *Brain-like* ECOS.

Table 2.3. Three main aspects to be considered in *Brain-like* ECOS.

Processing Units	Basic units for information processing follow biological rules. Example: biological neurons, communication using pulses, pulse stimuli generates action potentials, etc.
Structure	Information flow through hierarchical (sequential/parallel) pathways. Example: visual system: retina ⇒ directionally selective cells ⇒ complex cells.
Learning	Self adapt and evolve, lifelong learning. Example: functional and structural adaptation, Hebbian learning, Synaptic Time Dependent Plasticity (STDP).

2.4 Spiking Neural Networks (SNNs)

This section explores in detail spiking neural networks as they are composed of more realistic neuronal unit models. There are two main applications for a group of spiking neurons connected to each other in a network (see Section 2.1.1 for a description of various spiking neuron models). SNNs can be used for the modelling of brain function and as a tool in artificial intelligence. These two main streams are described in the following sections.

2.4.1 SNNs for modelling brain function

Traditionally SNNs have been used in computational neuroscience, usually in an attempt to evaluate the dynamics of neurons and how they interact in an ensemble (Benuskova *et al*, 2001). The Hodgkin-Huxley model of spiking generation (Hodgkin and Huxley, 1952) can be considered the pioneering work describing the action potentials in terms of ion channels and current flow (Nelson and Rinzel, 1995). Further studies expanded this work and revealed the existence of a wide number of ion channels and that the set of ion channels varies from one neuron to another (e.g., (Connor and Stevens, 1971) (Marom and Abbott, 1994)). Genesis and Neuron (Genesis, 2007) (Neuron, 2007) are examples of widely known simulation tools that use neurons described with ion channels.

A detailed description of ion channels demonstrated itself to be computationally expensive in experiments which simulated large numbers of units. Thus, new neuron models have been developed to produce similar internal and external behaviour (similar action potentials and output signals) at a lower computational cost. As examples of simplified models, the integrate-and-fire neuron (Gerstner and Kistler, 2002) for all intents and purposes has the properties of a single resistor-capacitor (RC) circuit and the Izhikevich model (Izhikevich, 2003) combines the Hodgkin-Huxley model with the integrate-and-fire model using a two-dimensional system of ordinary differential equations.

In addition to enabling the understanding of complex behaviours generated by ensembles of spiking neurons, spiking neural networks are also used as a tool to simulate the link at the neuronal and genetic level (Benuskova and Kasabov, 2007). In this arrangement, networks of genes (gene regulatory networks) have an explicit influence on the neuronal parameters, consequently affecting the entire dynamics of the spiking neural network (see Figure 2.23). Imprecise information at the genetic level and the difficulty of simulating large numbers of spiking neurons for long periods, as genetic expression levels have much slower dynamics than neurons, are currently the main challenges to be overcome in this approach (Kasabov and Benuskova, 2004) (Kasabov *et al*, 2005) (Kasabov *et al*, 2005a) (Kasabov *et al*, 2005b) (Wysoski *et al*, 2004).

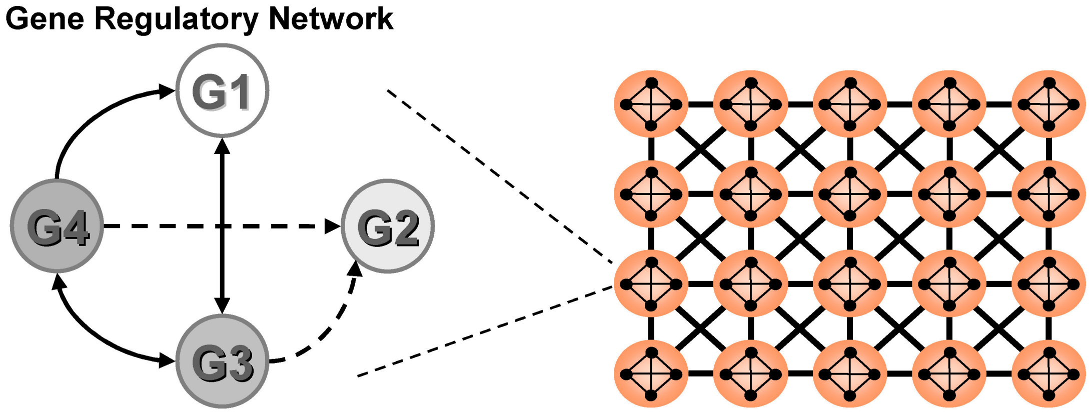


Fig. 2.23. A gene regulatory network (GRN) regulates the behaviour of spiking neurons and consequently the behaviour of the entire neural network. In the GRN illustration, the greyscale levels represent gene expression whereas the continuous/dashed lines represent respectively excitatory/inhibitory connections.

2.4.2 SNNs in artificial intelligence

Most neural networks which execute artificial information processing are described using processing units consisting of linear/non-linear processing elements (a sigmoid function is widely used) (Bishop, 2000) (Gallant, 1995) (Haykin, 1999) (Negnevitsky, 2002). Over the years, SNN has been considered too complex and difficult to analyze. Other reasons for leaving SNN aside in artificial intelligence tasks include:

1) biological cortical neurons have long time constants. Typically fast/slow inhibition can be in the order of dozens of milliseconds and fast/slow excitation can reach hundreds of milliseconds. This dynamic can considerably constrain applications that need fine temporal processing (Gewaltig, 2000).
2) unknown information encoding in time. Although it is known that neurons receive and emit spikes, whether neurons encode information using spike rate or precise spike time is still unclear (Thorpe and Gaustrais, 1998). For those supporting the theory of spike rate coding, it is reasonable to approximate the average number of spikes in a neuron with continuous values and consequently process them with traditional processing units (sigmoid, for instance). Therefore, it is not necessary to perform simulations with spikes, as the computation with continuous values is simpler to implement and evaluate.

However, new discoveries on the information processing capabilities of the brain and technical advances related to massive parallel processing, are bringing forward the idea of using biologically realistic networks in artificial intelligence systems. There are several works questioning rate coding, mainly under the assumption that rate coding can be very slow to provide reliable outputs (the average number of spikes needs to be computed over a certain period of time), which obviously disagrees with many perceptual experiments. For instance, a pioneering work has shown that the primate (including human) visual system can classify complex natural scenes in only around 100-150 ms (Thorpe *et al*, 1996). Similar numbers were obtained by (Potter, 1976) and (Subramaniam *et al*, 2000), which showed that unprimed views of common objects can be recognized at a rate of 10 Hz. This period of time for information processing is very impressive considering that billions of neurons are involved and the information is propagated through several areas of the brain before a decision is made. Such results culminated in a theory suggesting that a single neuron probably exchanges only one or a few spikes before the information processing task is concluded. As a result of Thorpe's work, a simple multi-layer feed-forward network of integrate-and-fire neurons that can successfully detect and recognize faces in real time was designed (Delorme and Thorpe, 2001) (Delorme *et al*, 1999). Other works (Bothe, 2003) (Gueorguieva *et al*, 2006) (Natschlager, 1998) also present systems using precise timing of spikes on pattern recognition (clustering, supervised and unsupervised training).

An important landmark in the use of SNNs in artificial information processing is the work of (Maass, 2001), which shows that, theoretically, SNN can be used as universal approximators of continuous functions. Lorenzo *et al* (Lorenzo *et al,* 2006) proposed an SNN of three-layers (input, generalization and selection layers) to perform unsupervised pattern analysis. Mishra (Mishra, 2006) gives examples of spiking neural networks applied to benchmark datasets (internet traffic data, EEG data, XOR problems, 3-bit parity problems, iris dataset) to perform function approximation and supervised pattern recognition. A comparison with a traditional Multi-Layer Perceptron Network (MLP) highlights the differences in performance between the systems in each specific dataset. Time series prediction using SNNs has been evaluated in (Sohn *et al*, 1999).

2.4.3 Properties of SNNs for pattern recognition

This section describes the particular properties of the spike paradigm for both static and dynamic pattern recognition.

1) Neuronal coding: How the information is encoded in spike trains is one of the most intriguing questions in neuroscience. Whereas information encoding based on spike rate (the number of spikes in a certain period of time) has been used for

some time, there is increasing evidence that information could be encoded in precise spiking time in a highly sparse network of neurons, or even through a combination of spike rate and spiking time. See (Abbott, 1994) (Dimitrov and Miller, 2001) (Reece, 1999) for further reference.

2) Learning: Learning takes place at a synaptic (the junction between the axon and the dendrites) level and is represented by the strength of inter neuronal connections. Learning can occur during development, at the time the network is generated and the connections are created, or as a process of adaptation, where the network responds to specific stimulus and reconfigures its connection strengths to memorize a certain task (Gerstner and Kistler, 2002).

3) Optimization: Computation with spikes opens up new optimization aspects that are not explicitly found in traditional neural networks:

   - Information encoding – Despite it still not being clear how information encoding effectively happens in the brain, there is strong evidence to suggest that spike encoding is optimal in terms of data transmission efficiency (**maximum data transmission**) (Bialek and Rieke, 1992) (Gerstner and Kistler, 2002) (Rieke *et al*, 1999);
   - Processing time – Experimental evidence shows that some types of cognitive processes are accomplished in the brain in a very short time (e.g. 150 ms for visual systems) and can be improved upon training (**minimum processing time**) (Thorpe *et al*, 1996);
   - Energy efficiency – Mammalian brains are known for having more than 10^{10} neurons operating at a very low spiking rate (1-3 Hz) (Gerstner and Kistler, 2002). These numbers suggest that the wiring and connectivity strength are set up in such a way that the processing is done with minimum energy consumption (**minimum neuronal activity**).

4) Processing spatio-temporal patterns: SNNs intrinsically spread information over time, processing it continuously. Due to this property, SNNs are particularly suitable for processing temporal patterns more naturally than recurrent networks. Examples of research in this direction are the spatio-temporal artificial neural network (STANN) proposed in (Mercier and Seguier, 2002) and the work of Holmberg *et al* (Holmberg *et al*, 2005) in detecting spatio-temporal characteristics of sound signals. When the task is to process static patterns, the static information is spread over time. Examples of this approach can be found in (Bothe, 2002) (Delorme and Thorpe, 2001) (Natschlager and Ruf, 1998).

2.4.4 Implementation of SNN models

SNNs have some properties that are particularly interesting in terms of implementation, being it in hardware or in simulations with general-purpose processing platforms.

There are two main approaches to implementing SNNs in general-purpose processing platforms: *event-driven* and *clock-driven* computation. In the *event-driven* approach, only the neurons that receive a spike are updated, whereas in the *clock-driven* approach all neurons are updated at a given clock step. Note that this property is purely concerned with computational implementation and does not affect the SNN's behaviour. The *event-driven* approach is particularly appropriate for computing very large numbers of neurons with low spiking rates.

The basic *clock-driven* computation of an SNN is illustrated with the following pseudo-code (Brette *et al*, 2007):

```
for all time steps
    {
    for all neurons
        {
        PropagateIncomingSpikes
        CalculatePostSynapticPotentialDecay
        }
    for all neurons
        {
        if PostSynapticPotential > Threshold
            {
            SetNeuronPotentialToRest
            for all neuron outputs
                EmitSpike
            }
        }
    }
```

Unlike the *clock-driven* approach, there is no universal recipe for *event-driven* computation. The implementation of *event-driven* systems will be highly dependent on the model of neuron used. The simplest scenario is that an output spike can only occur at the time an incoming spike arrives. Thus, the incoming spike triggers an event to initiate the computation of the state of a given neuron. In the simplest *event-driven* approach, a queue is used to store the issued spikes. A simple design to simulate an SNN using *event-driven* computation is exemplified as in (Brette *et al*, 2007):

```
Structure queue (has one item for each spike)
queue[i].SpikeTime
queue[i].TargetNeuron
queue[i].Weight

    while (queue.Size > 0 && time < SimulationPeriod)
    {
        GetLowestSpikeTime
        CalculatePostSynapticPotentialTargetNeuron
        PropagateIncomingSpikeToTargetNeuron
        if PostSynapticPotential > Threshold
            {
            for all neurons connected to TargetNeuron
                InsertItemInQueue
            SetTargetNeuronPotentialToRest
            }
    }
```

In some models of spiking neurons, output spikes can be generated several time steps after an incoming spike is received. This is the case for neurons that have a time constant for rising postsynaptic potential (Gerstner and Kistler, 2002). Brette *et al* (Brette *et al*, 2007) describes an efficient *event-driven* way to compute these cases. A more in-depth consideration of several variations of *event* and *clock-driven* computation of SNN models can be found in (Marian, 2002).

An important issue that needs to be carefully analysed during the implementation of SNNs, is the timing when each cell computes its new activation value and when the change of the cell's output take place (Gallant, 1995). In this respect, neurons can receive/emit spikes in a synchronous or an asynchronous way. In the asynchronous way, neurons are updated as the spikes arrive. The output of the neurons is changed immediately after the event. The neurons are processed sequentially one at a time. In the synchronous way, on the other hand, all the neurons' postsynaptic potential and output spikes are updated simultaneously at each time step. Synchronicity can be very important in some systems, and can have a great impact on the dynamic behaviour of the entire SNN. Synchronicity is especially relevant to the spiking neuronal models that are sensitive to the order of incoming spikes, e.g., integrate-and-fire neurons with a modulation factor (Delorme *et al*, 1999).

2.4.5 Mathematical facts about SNNs

It has been proven that spiking neural networks can, similar to traditional networks, act as Radial Basis Functions (RBF), performing weighted sum of temporal code, and be used for the universal approximation of continuous functions (Maass, 2001). In addition, a single neuron can act as a coincidence detector, which cannot easily be done with traditional neural networks. In (Gerstner and Kistler, 2002), after a detailed comparison, the authors concluded that traditional and biologically realistic networks are computationally equivalent. It is further suggested that any sigmoid neural net can be successfully reproduced using almost the same number of spiking neurons, whereas to simulate spiking neurons many more sigmoid units are necessary.

The equivalences with traditional networks have been translated to comparable performances in a series of artificial intelligence tasks, e.g. clustering in (Lorenzo *et al,* 2006), classification in (Mishra, 2006) and time-series prediction in (Sohn *et al*, 1999). At the end of this dissertation, the experimental results will also point to equivalent performances from SNN-based and traditional methods in the audiovisual scenario.

2.4.6 Limitations of SNN models

SNNs are networks that have a close association with what is known about the way the brain processes information. This statement is true if brain-like networks are compared to connectionist systems that use traditional processing units (e.g. McCulloch-Pitts, Perceptron, etc). However, in attempting to model the human brain, models of spiking neurons are still too simplistic. The simplification is mainly due to the lack of knowledge and/or lack of computational power to simulate complex models. Researchers are daily confronted with the trade-off between biologically accurate models and computational cost. In the recent Blue Brain Project (EPFL BlueBrain, 2007), a team of researchers are joining forces in an attempt to simulate the most up-to-date, biologically accurate, functional model of the brain in a supercomputer. It is too simplistic, though, to blame the lack of computational resources only. In reality, despite a lot being known about neurons and networks, there are still some key biological facts remaining to be understood. For instance, the use of spikes for transmitting information depends on having suitable ways of encoding the information in spike trains. Since how the encoding happens in human brains is still unknown, all the encoding heuristics employed in SNNs are hypothetical and it is not known which one, if any, represents reality.

2.4.7 Parameter optimization in neuronal models

There are several methods for finding values of neural network parameters to generate a desired behaviour (Achard *et al*, in press). They include: a) hand-tuning, used for instance, in (Nadim *et al,* 1995); parameter space exploration, applied in (Prinz *et al*, 2004); gradient descent, applied to a single neuron in (Bhalla and Bower, 1993) and at a network level in (Bothe *et al*, 2002) and (Tino and Mills, 2006); general-purpose evolutionary algorithms and particle swarm optimization (Achard and De Schutter, 2006) (Pavlidis *et al*, 2005); and bifurcation analysis (Guckenheimer *et al*, 1993).

Throughout the work presented in this dissertation, hand tuning is used to search for suitable parameters. In this approach, one or a number of model parameters are manually changed at a time, guided by trial-and-error, prior knowledge of the model, and experience, always keeping the parameters that provide the best behaviour in the explored parameter subspace.

Advantages of hand tuning: The obvious advantage of this approach is that it does not require the design of optimization algorithms nor is it computationally expensive. It also incorporates prior knowledge and insight from a specialist into the network behaviour.

Disadvantages of hand tuning: The main drawback is that there is no certainty that the set of parameters found is optimal at the end of the hand tuning procedure. However, this drawback can be also found, perhaps with less intensity, in for example, gradient descendent and evolutionary algorithms (Achard *et al*, in press).

Despite this drawback and the simplicity, it will be demonstrated in the experiments presented later in this dissertation that systems tuned by this approach can obtain accuracy levels comparable to results on the same tasks reported by other works.

More systematic parameter optimization procedures are out of the scope of this work and are left as a suggestion for future work.

2.5 Chapter conclusion

The chapter presented basic concepts of biological information processing. The neuron as a basic processing unit is explored as well as the ensemble of neuronal units that generates complex and dedicated behaviours in different areas of the brain. In addition, the auditory and visual sensory information pathways were reviewed.

A classification of artificial models into *Connectionist Systems* (COS), *Evolving Connectionist Systems* (ECOS) and *Brain-like* ECOS was proposed. As the remainder of this dissertation is concerned with *Brain-like* ECOS, three main aspects of Brain-like ECOS to be considered were suggested in terms of basic processing units, network structure and learning. These aspects are used throughout this dissertation to evaluate the new systems.

In the last part of the chapter, the spiking neuron theory was presented. Spiking neurons, as a more biologically plausible processing unit, are the basic components of the new systems presented in the following chapters.

Chapter 3 - Evolving SNNs for visual pattern recognition

This chapter explores different computational approaches to performing visual pattern recognition and presents a new learning procedure and a neural network structure to achieve visual pattern recognition with spiking neural networks.

First, the state-of-the-art algorithms used in visual pattern recognition are described, followed by a critical analysis of brain-inspired artificial systems used for visual information processing tasks (mainly covering the theory of spiking neurons, visual information pathways and lifelong learning). This work breaks new ground, exploring the use of a new SNN structure to deal specifically with two pattern recognition scenarios:

- Multiple view training: training samples are presented online to the network, which learns different views of an object through synaptic plasticity and structural adaptation;
- Multiple view recognition: several visual samples are presented to the network, which integrates opinions for the decision-making process.

The main innovations proposed can be summarized as:

a) an online training procedure for a hierarchical neural network of fast integrate-and-fire neurons. The training is done through synaptic plasticity and changes in the network structure. A mix of *clock-driven* and *event-driven* computation optimizes processing speed in order to simulate networks with large numbers of neurons. The training procedure is applied to a face recognition task. Preliminary experiments on a publicly available face image dataset show the same performance as the optimized offline method. A comparison with other classical methods of face recognition demonstrates the properties of the system.

b) an SNN architecture and its corresponding learning procedure to perform fast and adaptive multi-view visual pattern recognition. The network is composed of a simplified type of integrate-and-fire neuron arranged hierarchically in four layers of two-dimensional neuronal maps. During learning, the network adaptively changes its structure in order to respond optimally to different visual patterns. The network collects opinions over multiple frames to reach a final decision. The new network structure is tested with a benchmark dataset in order to recognize individuals using facial information from multiple frames.

Experimental results validate the two main novelties of the network: structural adaptation and integration of opinions over several frames.

3.1 Background and benchmarking

An image is a two-dimensional projection of the three-dimensional, real world. The human brain, and artificial visual systems as well, collects visual information about the environment through these two-dimensional projections. In humans, the integration of images acquired from both eyes, enables the retrieval of the third dimension. Stereo vision is the field that emulates this phenomenon artificially, with the use of dual cameras, where the sense of depth can be mathematically calculated through disparity maps. See (Dhond and Aggarwal, 1989) for a review. The computation of stereo disparity maps has also been proposed with biologically inspired networks (SNNs) (Henkel, 1997) (Henkel, 1997a).

Three-dimensional reconstruction methods began with the pioneering work of Horn (Horn, 1970), from which several directions were derived: stereo vision, shape from shading, shape from motion, and shape from texture (Zhang *et al*, 1999). The modelling of the third dimension of objects is still a very active field of research, mainly because reconstruction techniques normally add substantial noise to the representation while at the same requiring high computational effort. For these reasons, many artificial visual system applications prefer to use a single monocular two-dimensional representation.

Particularly interesting to this research, is the process of recognizing patterns in visual scenes. A pattern in an image can be described as having four properties (see Figure 3.1):

a) position, which can be expressed as the relative distance from a reference point or the distance from one pattern to another;

b) geometry (size, area, and shape);

c) colour and texture, and;

d) trajectory (the way an object moves can be used in its description).

With one or a combination of these properties, patterns can be artificially recognized and/or categorized. Common techniques for visual pattern categorization are:

a) filters, such as, colour filters, or filters in the frequency domain (Kakumanu *et al*, 2007) (Sigal *et al*, 2004);

b) template matching or connectionist systems for object classification (Fukushima and Miyake, 1982) (Kasabov *et al*, 2000) (Matsugu *et al*, 2002);

c) statistical methods (Fukunaga, 1990) for finding statistical properties of patterns, e.g., Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) (Belhumeur *et al*, 1997) (Martinez and Kak, 2001) (Zuo *et al*, 2007) and Independent Component Analysis (ICA) (Bartlett, 2007), etc.

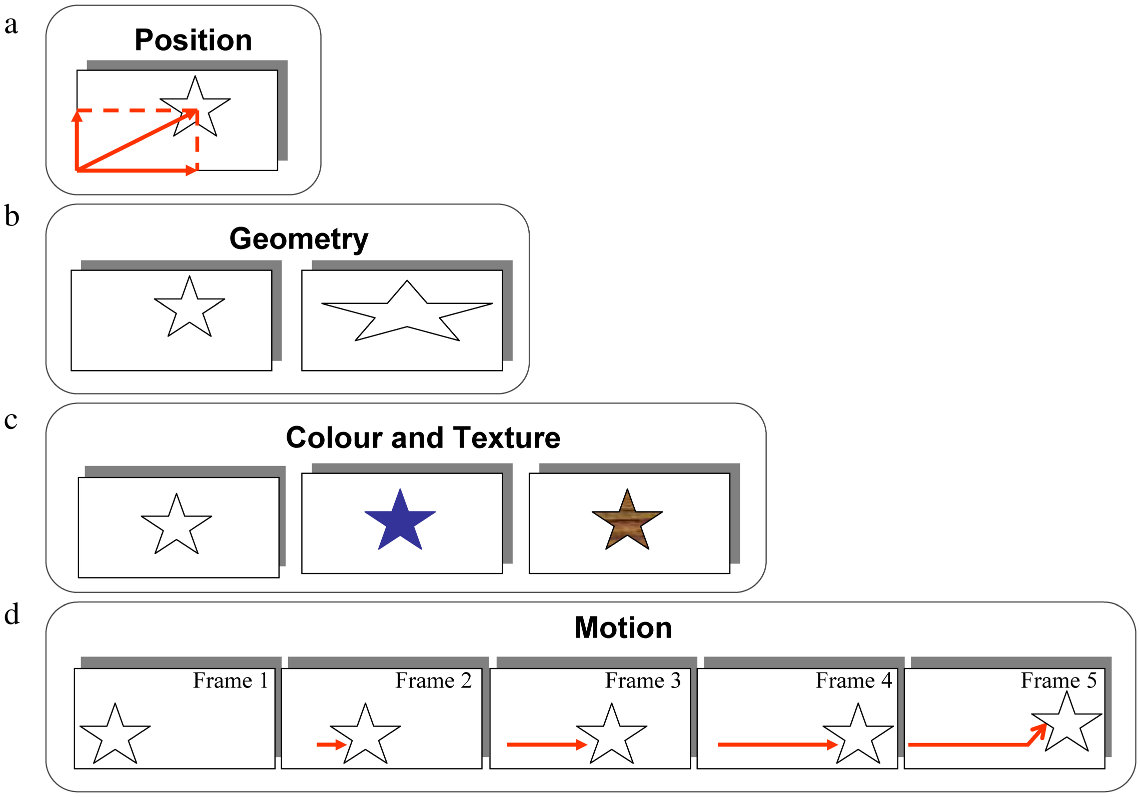


Fig. 3.1. Object description primitives: a) position; b) geometry (size, area, shape); c) colour and texture; d) trajectory (translation and rotation).

A combination of the above techniques is also common practice, e.g., colour filter followed by frequency domain analysis and morphological matching (Kruger *et al*, 2004) (Mel, 1998). This diversity of artificial methods for processing visual patterns try to, in one way or another, identify the most important characteristics which represent an object and increase the reliability of the identification of these characteristics. Thus, artificial visual pattern recognition systems use different features and classification tools tailored specifically to the application. For instance, in environments with constant illumination sources, it is usually sufficient to only use colour information to segment objects. This approach fails in unconstrained illumination conditions where the segmentation of an object is against complex backgrounds. A classical example is skin colour detection, where colour histograms of the skin and background colours are not easily separable (Kakumanu *et al*, 2007) (Sigal *et al*, 2004). Figure 3.2 shows an example where colour filters trying to retrieve only skin colours did not work properly. The existing overlap between skin/not skin colours can be seen in a colour histogram in the r (red) and g (green) space.

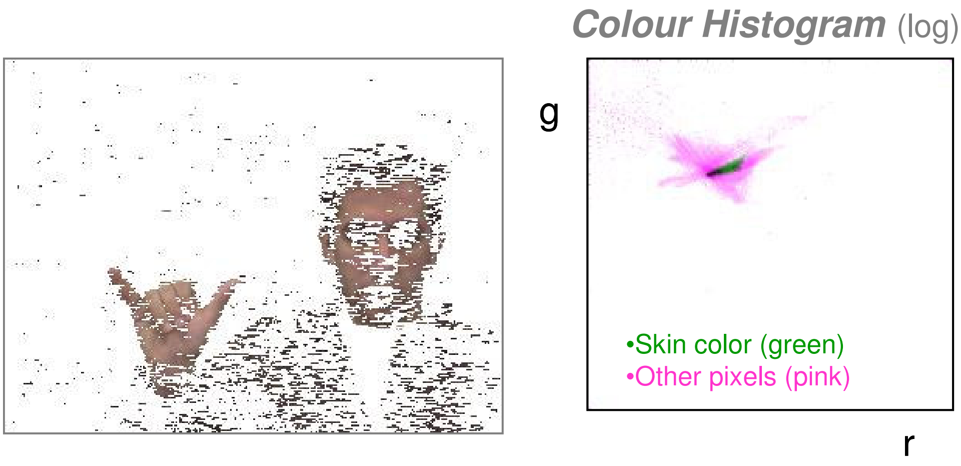


Fig. 3.2. Skin detection using colour information. For classification between skin/not skin colours, simple threshold values are used. Notice in the colour histogram that the skin and background colours are not easily separable.

When more complex visual shapes are to be detected/recognized, the most straightforward and standard way is to apply a linear transformation to the image and extract the object's principal components using PCA or LDA. Further description is provided in (Fukunaga, 1990) (Turk and Pentland, 1991) and a comprehensive comparative discussion is given in (Belhumeur *et al*, 1997) and (Martinez and Kak, 2001). Numerous researchers have been working on these directions (see (Chellapa *et al*, 1995) for a brief review). Zuo *et al* (Zuo *et al*, 2007) performed two LDA on distinctly different subspace methods with the results on FERET and ORL datasets superior to pFisherfaces and bi-directional PCA plus LDA in terms of recognition accuracy. A new method based on orthogonal discriminant locality preserving projections, comparable in results to Eigenface, Fisherface and Laplacianface methods, was suggested in (Zhu and Zhu, 2007). ICA has been shown to give better recognition performance than Eigenfaces at the same time maximizing information transfer for a more general set of input distributions (Bartlett, 2007).

In the nineties, the theory of wavelets started to receive a lot of attention, mainly because it enables a more general analysis when compared with the spectral analysis based on Fourier theory. The Gabor and Haar type wavelets were widely used on image processing tasks. Gabor function resembles the response of orientation selective cells found in the visual cortex. The convolution of an image with a set of Gabor functions is illustrated in Figure 3.3. Gabor wavelets are applied to the face recognition problem in (Escobar and Ruiz-del-Solar, 2002), (Garcia *et al*, 2000) and (Zhang *et al*, 2004) for example. In another approach, discriminant features can be extracted using a combination of Gabor-based image preprocessing and PCA (Yao *et al*, 2007). Haar wavelets, on the other hand, can be considered a coarse representation of Gabor functions, also computing directions (horizontal, vertical, diagonal) on different scales and positions, but are particularly interesting because they can be computed impressively fast using the Integral Image technique. Implemented in the OpenCV image processing library (Intel OpenCV, 2007), it has been demonstrated to be very useful for real time applications (Viola and Jones, 2001).

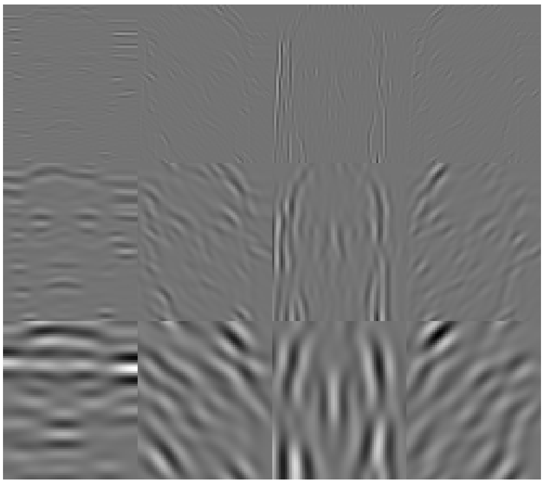


Fig. 3.3. Convolution of a frontal face with Gabor wavelets in different frequencies and orientations resembles orientation selective cells found in the visual cortex.

In another approach, Elastic Bunch Graph Matching uses Gabor jets after detecting the fiducial points of the face (eyes, mouth, etc). At each fiducial point the area is represented by a set of Gabor wavelet components (jets) (Figure 3.4) (Escobar and Ruiz-del-Solar, 2002) (Wiscott *et al*, 1999).

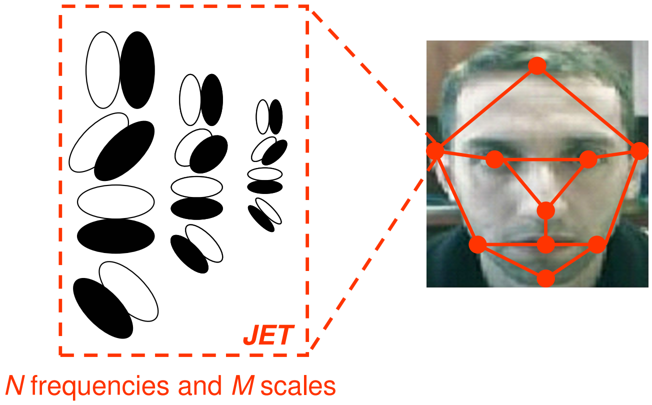


Fig. 3.4. Elastic Bunch Graph Matching, according to (Wiscott *et al*, 1999).

In respect to methods for performing visual pattern classification, Nearest Neighbour, *k-means*, Multi-layer Perceptron Neural Networks (MLP), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) (Haykin, 1999), Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2004), and weak classifiers (e.g., Adaboost (Freund and Schapire, 1997)) are examples that have been successfully applied to the problem. Potentially, most of the evolving connectionist methods (ECOS) described in (Kasabov, 2007) are also suitable, as demonstrated in (Ghobakhlou *et al*, 2004) (Kasabov *et al*, 2000) and (Wysoski *et al*, 2004a) on the face recognition problem.

3.2 Brain-like networks for visual information processing

Despite all the efforts to produce reliable visual pattern recognition systems, as described above, these models are far from being comparable to human vision. Each artificial model has its pros and cons, yet they are still not as general and accurate as the processing done by the brain. Thus, in another methodological approach the *emulation of brain-like processes* for the purpose of pattern recognition has emerged. In a pioneering attempt to create a network in which the information is processed through several areas resembling the visual system, Fukushima and Miyake proposed the Neocognitron, which processes information with rate-based neural units (Fukushima and Miyake, 1982). A new type of model for object recognition based on computational properties found in the brain cortex was described by Riesenhuber and Poggio (Riesenhuber and Poggio, 1999). This model uses hierarchical layers similar to the

Neocognitron and processing units based on MAX-like operation, to define the postsynaptic response, which results in relative position- and scale-invariant features. This biologically motivated hierarchical method is carefully analyzed by Serre *et al* (Serre *et al*, 2007) on several real-world datasets, extracting shape and texture properties. The analysis encompassed invariance on single object recognition and recognition of multiple objects in complex visual scenes (e.g. leaves, cars, faces, airplanes, motorcycles). The method presented comparable performance with benchmark algorithms.

In the same way, Mel (Mel, 1998) applies purely feed-forward hierarchical pathways to perform feature extraction, now integrating colour, shape, and texture. The hierarchical architecture enables the extraction of 102 features that are combined in a nearest-neighbour classifier. For a constrained visual world, the features were demonstrated to be relatively insensitive to changes in the image plane and object orientation, fairly sensitive to changes in object scale and non rigid deformation, and highly sensitive to the degradation (occlusion, noise, etc) of the visual stimuli. Kruger *et al* describes a rich set of primitive features that include frequency, orientation, contrast transition, colour and optical flow, which are integrated following semantic attributes (Kruger *et al*, 2004). Each attribute in practice, has a confidence level, which can be adapted according to visual context information.

Further in the attempt to explore the brain's way of processing, experimental results from neurobiology have led to the investigation of a third generation of neural network models which employ spiking neurons as computational units. Hopfield (Hopfield, 1995) proposed a model and learning algorithm for spiking neurons to realize Radial Basis Functions (RBFs) where spatial-temporal information is presented based on the timing of single spikes, i.e., not in a rate-based fashion. Natschlager and Ruf further extended the idea, by defining the pattern not only by the sequence of input spikes, but also by the exact firing time (Natschlager and Ruf, 1998) (Natschlager and Ruf, 1999). In these works, an input pattern representing a spatial feature is encoded in the temporal domain by one spike per neuron. It has also been shown how simple it is to modify the system to recognize sequences of spatial patterns by allowing the occurrence of more than one spike per neuron. Other conclusions of these works include: a) even under the presence of noise (in terms of spatial deformation or time warping) the recognition can be undertaken; and b) an RBF neuron can be used to perform a kind of feature extraction, i.e., a neuron can be designed to receive excitation/inhibition from a subset of features and be insensitive to others.

For simulating biological processes like vision and sensorimotor integration, biologically plausible Hebbian learning rules are used by Northmore (Northmore, 2004). In this work, a simulated vehicle with artificial whiskers and an eye was trained to navigate through an obstacle course. The SNN has learnt to visually guide the "cybermouse" based on detection of luminance and motion. The Hebbian rule used was spike-timing-dependent synaptic plasticity (STDP) discovered experimentally by Markram *et al* (Markram *et al*, 1997). Similar learning was also introduced in a two-layer SNN which performed character recognition tasks (Gupta and Long, 2007).

Maciokas (Maciokas, 2003) goes down to the level of ionic channels to describe a model of an audiovisual system that reproduces the responses of the GABAergic cells. Audio features were extracted using Short Term Fourier Transform and represented in tonotopic maps. The visual information of lip movement was extracted using Gabor filters. The two main results described in his work are: a) the accurate model of diverse firing behaviours of GABAergic cells; and b) proof that a large-scale network of the cortical processing preserves information in audiovisual modalities using an entropy measure. No attempts to test the classification abilities of the network have been made.

Thorpe (Thorpe, 1996) suggests that in order to be coherent with the time measured in certain classes of behavioural experiments on perceptual activities, the information processing mechanisms can afford to have a single neuron exchanging only one or a few spikes. The time between information acquisition and the cognitive response is too short to have rate-based neuronal encoding, since the information needs to travel sequentially over several different compartments located in distinct brain areas. Thus, the information needs to be sparsely encoded and, highly complex cognitive activities are reached through a complex wiring system that connects neuronal units. As an output of this work, the authors proposed a multi-layer feed-forward network (SpikeNet) using fast integrate-and-fire neurons that can successfully track and recognize faces in real time (Delorme and Thorpe, 2001) (Delorme *et al*, 1999). Coding of information in this model is based on the so-called rank order coding, where the first spike is the most important. It has been shown that using rank order coding and tuning the scale sensitivity according to the statistics of the natural images can lead to a very efficient retina coding strategy, which compares to image processing standards like JPEG (Perrinet and Samuelides, 2002).

Matsugu *et al* utilized a different coding strategy in a hybrid of a convolutional and SNN architecture for face detection tasks (Matsugu *et al*, 2002). In this hierarchical network, local patterns defined by a set of primitive features are represented in the timing structure of pulse signals. The training mentioned in the work is for the bottom feature-detecting layer to use the standard error back-propagation algorithm. The model implements hierarchical pattern matching by temporal integration of structured pulse packets. The packet signal represents intermediate or complex visual features (like an eye, nose, corners, a pair of line segments) that constitute a face model. As a result of the spatio-temporal dynamics the authors achieved size and rotation invariant internal representation of objects. Endowed with a rule-based algorithm for facial expression classification, this hybrid architecture achieved robust facial expression recognition together with robust face detection (Matsugu *et al*, 2003).

The remaining sections of this chapter follow the conceptual approach described in (Delorme and Thorpe, 2001) (Delorme *et al*, 1999), from which the basic building blocks of the model are borrowed, e.g., the fast integrate-and-fire neuron model and its respective learning rule, and the network structure, which is formed from hierarchical layers composed of neurons grouped in neuronal maps. In the following sections, the SNN model is presented and the evolving structure originating from a new online learning procedure is described. A layer responsible for integrating multi-view information is added in the face recognition task.

3.3 SNN model for face recognition

3.3.1 Model description

This section describes the biologically realistic model used in this work to perform online visual pattern recognition. The system is based on SpikeNet introduced in (Delorme and Thorpe, 2001) (Delorme *et al*, 1999) (Delorme *et al*, 2001) (Thorpe and Gaustrais, 1998). The neural network is composed of three layers of fast integrate-and-fire neurons.

Spiking neuron model

The system uses a simplified version of integrate-and-fire neurons, the dynamics of which were first described and analyzed in (Delorme and Thorpe, 2001) (Delorme *et al*, 2001). Compared to a standard integrate-and-fire neuron (Gerstner and Kistler, 2002), the main differences are the lack of postsynaptic potential (PSP) leakage, excitation dependent upon the order of spike arrival, and the inactivation of the neuron after the output spike (the PSP is permanently set to the resting potential level). Figure 3.5 illustrates the fast integrate-and-fire neuron. The main advantages of these neurons are that they are computationally very inexpensive, and they boost the importance of the first spikes to arrive.

As a brief formal description, the postsynaptic potential for neuron *i* at a time *t* is calculated as:

$$PSP\,(i,t) = \sum \mathrm{mod}^{\,order\,(j)}\, w_{j,i} \qquad \textbf{(3.1)}$$

where mod ∈ (0,1) is the modulation factor, *j* is the index for the incoming connection and $w_{j,i}$ is the corresponding synaptic weight. For instance, setting mod = 0.9 and considering $w_{i,j} = 1$, the first spike to arrive (*order* (*j*) = 0) changes the *PSP* by $0.9^{(0)} = 1$. The second spike (*order* (*j*) = 1) further influences the PSP by $0.9^{(1)} = 0.9$, the third spike by $0.9^{(2)} = 0.81$, and so on. An output spike is generated if

$$PSP(i,t) \geq PSP_{Th}(i) \qquad \textbf{(3.2)}$$

where PSP_{Th} is the postsynaptic threshold.

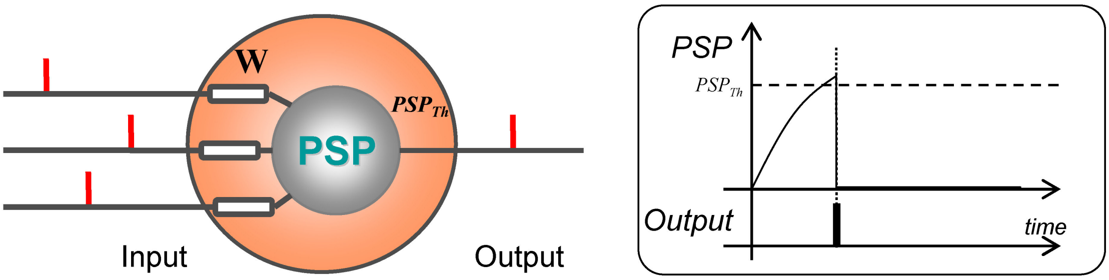


Fig. 3.5. Fast integrate-and-fire neuron with modulation factor.

Two-dimensional neuron grid (neuronal maps) and network structure

Each layer is composed of neurons grouped in two-dimensional grids forming neuronal maps. Connections between layers are purely feed-forward and each neuron can spike at most once when the PSP threshold is reached. The first layer (L1) neurons represent the On and Off cells of the retina, enhancing the high contrast parts of a given image (high-pass filter). To each pixel of an image (receptive fields), one neuron is allocated in each L1 neuronal map. L1 can have several pairs of neuronal maps of On and Off cells, each pair tuned to a different frequency scale. On and Off cells are implemented through weighted connections between the receptive fields and the L1 neurons. Weights are computed with a two-dimensional Difference of Gaussians, where different scales are chosen varying the standard deviation σ of the Gaussian curve (Vernon, 1991). Equation 3.3 describes the On/Off filters, where g normalizes the sum of weight elements to zero and the maximum and minimum convolution values to [+1, -1].

$$\nabla^2 G(x,y) = g\left(\frac{x^2+y^2-\sigma^2}{\sigma^4}\right)e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)} \qquad \textbf{(3.3)}$$

The output values of the first layer are encoded to pulses in the time domain. High output values of the first layer are encoded with short time delay pulses whereas pulses with long delays are generated in the case of low output values, according to the Rank Order Coding technique (Thorpe and Gaustrais, 1998) (Figure 3.6). Thus, the higher the amplitude, the shorter is the delay and vice-versa. L1 basically prioritizes the pixels with high contrast, which are consequently processed first and have a higher impact on neurons' PSP.

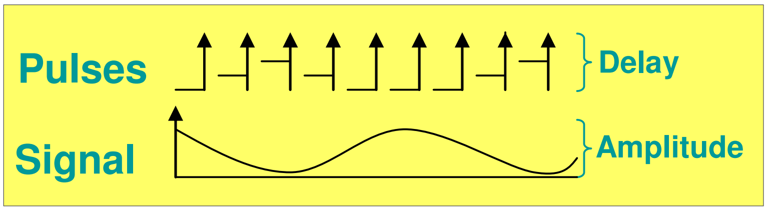


Fig. 3.6. Rank Order Coding. The amplitude of a signal is encoded over time. The higher the amplitude, the shorter the delay (and vice-versa). Modified from (Thorpe and Gaustrais, 1998).

The second layer (L2) is composed of eight orientation maps for each frequency scale, each one being selective of different directions (0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°). To compute the directionally selective filters the Gabor function is used

$$G(x,y) = e^{(\frac{x'^2+\gamma^2 y'^2}{2\sigma^2})} \cos(2\pi \frac{x'}{\lambda} + \varphi) \qquad \textbf{(3.4)}$$
$$x' = x\cos(\theta) + y\sin(\theta)$$
$$y' = -x\sin(\theta) + y\cos(\theta)$$

where φ is the phase offset, θ is the orientation [0,360], λ is the wavelength, σ is the standard deviation of the Gaussian factor of the Gabor function and γ is the aspect ratio which specifies the ellipticity of the support of the Gabor function. The Gabor filters are normalized globally for each frequency scale, in such a way that neurons having directionally selective cells as inputs can have PSPs that vary within that range [0, PSP_{max}], regardless of the scale of the filters.

The third layer is where the learning takes place. Maps in the third layer are trained to be sensitive to excitation of complex patterns (faces, as illustrated in the following experiments). See Figure 3.7 for the complete network architecture.

In (Delorme and Thorpe, 2001) the network has a fixed structure and the learning is done offline using the rule:

$$\Delta w_{j,i} = \frac{\text{mod}^{order(a_j)}}{N} \qquad \textbf{(3.5)}$$

where $w_{j,i}$ is the weight between neuron j of L2 and neuron i of L3, mod $\in$ (0,1) is the modulation factor, *order(a_j)* is the order of spike arrival from neuron *j* to neuron *i,* and *N* is the number of samples used for training a given class. Within this rule, there are two points to be highlighted:

a) the number of samples to be trained needs to be known *a priori*; and

b) after training, a map of a class becomes selective to the average pattern.

One of the properties of the system is the low activity of the neurons. It means that the system has a large number of neurons, but only few take active part during the retrieval process. Due to this property, the computational performance can be optimized through an *event-driven* approach (Delorme *et al*, 2001) (Mattia and del Giudice, 2000). Additionally, in most cases, the processing can be interrupted before the entire simulation is completed. Once a single neuron on the output layer reaches the PSP threshold and emits a spike, the simulation can be finished. The *event-driven* approach and early simulation interruption make this method suitable for real-time implementations.

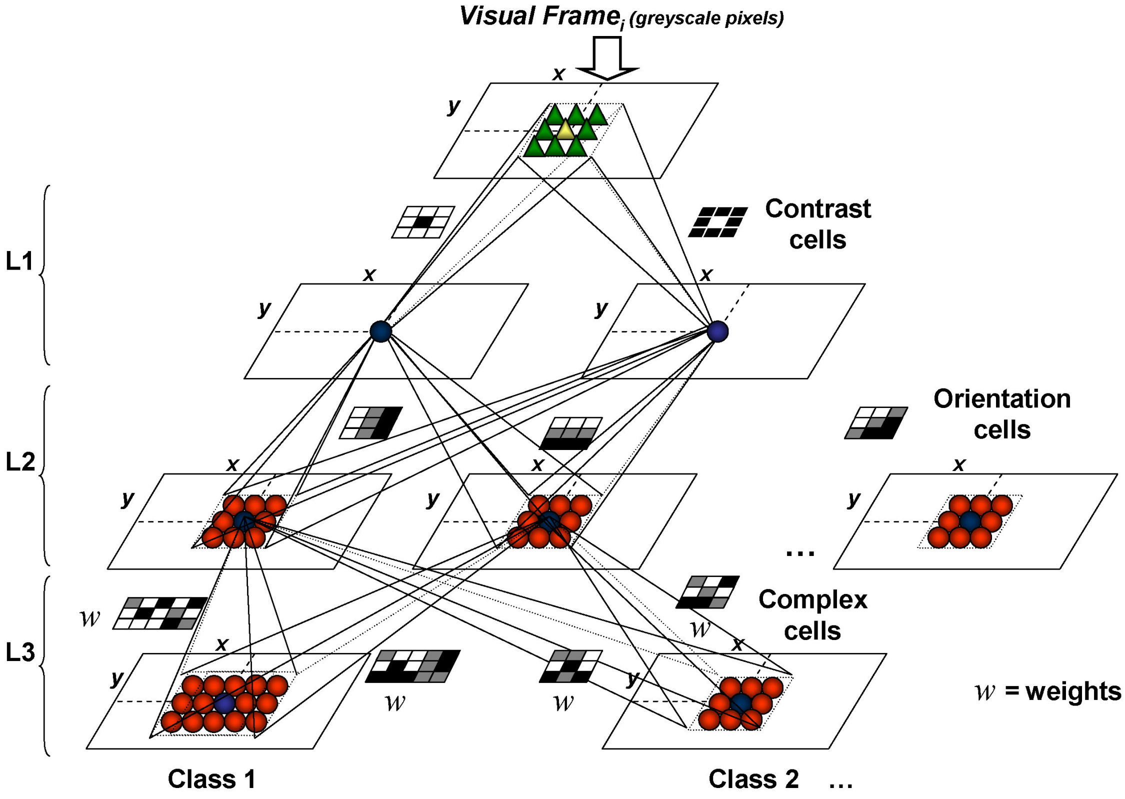


Fig. 3.7. Evolving spiking neural network (eSNN) architecture for visual pattern recognition.

3.4 A new online learning procedure with structural adaptation

3.4.1 General description

The new approach to learning with structural adaptation aims to give more flexibility to the system in scenarios where the number of classes and/or class instances is not known at the time the training starts. In addition, it intends to serve applications where new training samples can eventually be obtained and further or fine-tune training can be pursued without the need of completely retraining the system.

For this purpose, in the case of the spiking neural network described in Section 3.3, the output neuronal maps need to be created, updated or even deleted as the learning occurs. In (Kasabov, 2007) the ECOS framework dealing with adaptive and evolving problems is proposed and several methods and procedures describing adaptive systems are presented (see Section 2.3.2). To implement such a system, the learning rule needs to be independent of the total number of samples where the number of samples is unknown when the learning starts. Thus, the next section proposes the use of a modified equation to update the weights based on the dynamic average of the incoming patterns.

It is important to notice that, similar to the batch learning implementation of Equation 3.5, the outcome is the average pattern. However, the new equation calculates the average dynamically as the input patterns arrive.

There is a drawback to learning methods when, after training, the system responds optimally to the average pattern of the training samples. The average does not provide a good representation of a class in cases where patterns have high variance (Figure 3.8). A traditional way to attenuate the problem is the *divide-and-conquer* procedure. This procedure is implemented through structural modification of the network during the training stage. Specifically, a simple clustering procedure is integrated into the training algorithm, i.e., patterns within a class that comply with a similarity criterion are merged into the same neuronal map. If the similarity criterion is not fulfilled, a new map is created. The entire training procedure follows four steps described in the next section and summarized in the flowchart of Figure 3.9.

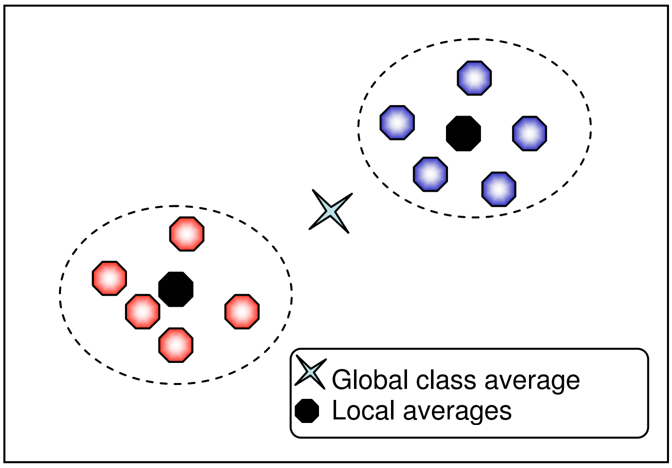


Fig. 3.8. *Divide-and-conquer* procedure to deal with high intra class variability of patterns in the hypothetical space of class *K*. The use of multiple maps that respond optimally to the average of a subset of patterns provides a better representation of the classes.

3.4.2 Learning procedure

The new learning procedure can be described in four sequential steps:

1. Propagate a sample *k* of class *K* for training within L1 (retina) and L2 (directionally selective cells);

2. Create a new map $Map_{C(k)}$ in L3 for sample k and train the weights using the equation:

$$\Delta w_{j,i} = \mathrm{mod}^{order(a_j)} \qquad \textbf{(3.6)}$$

where $w_{j,i}$ is the weight between neuron j of L2 and neuron i of L3, mod $\in$ (0,1) is the modulation factor, $order(a_j)$ is the order of spike arrival from neuron j to neuron i.

The postsynaptic threshold (PSP_{Th}) of the neurons in the map is calculated as a proportion $c \in [0,1]$ of the maximum postsynaptic potential (PSP) created in a neuron in $Map_{C(k)}$ with the propagation of the training sample into the updated weights, such that:

$$PSP_{threshold} = c\max(PSP) \qquad \textbf{(3.7)}$$

The constant of proportionality c is a measure of similarity between a trained pattern and a sample to be recognized. If c = 1, for instance, only an identical sample of the training pattern evokes the output spike. Thus, c is a parameter to be optimized in order to satisfy the requirements in terms of false acceptance rate (FAR) and false rejection rate (FRR). As a general rule, when the threshold increases so does the FRR while the FAR decreases.

3. Calculate the similarity between the newly created map $Map_{C(k)}$ and other maps belonging to the same class $Map_{C(K)}$. The similarity is computed as the inverse of the Euclidean distance between weight matrices.

4. If one of the existing maps for class K has similarity greater than a chosen threshold $Th_{simC(K)} > 0$, merge the maps $Map_{C(\mathrm{k})}$ and $Map_{C(K\mathrm{similar})}$ using arithmetic average as expressed in equation:

$$W = \frac{W_{Map_{C(k)}} + N_{samples}W_{Map_{C(Ksimilar)}}}{1+N_{samples}} \qquad \textbf{(3.8)}$$

where matrix W represents the weights of the merged map and $N_{samples}$ denotes the number of samples that have already being used to train the respective map. The PSP_{Th} is updated in a similar fashion as:

$$PSP_{Th} = \frac{PSP_{Map_{C(k)}} + N_{samples}PSP_{Map_{C(Ksimilar)}}}{1+N_{samples}} \qquad \textbf{(3.9)}$$

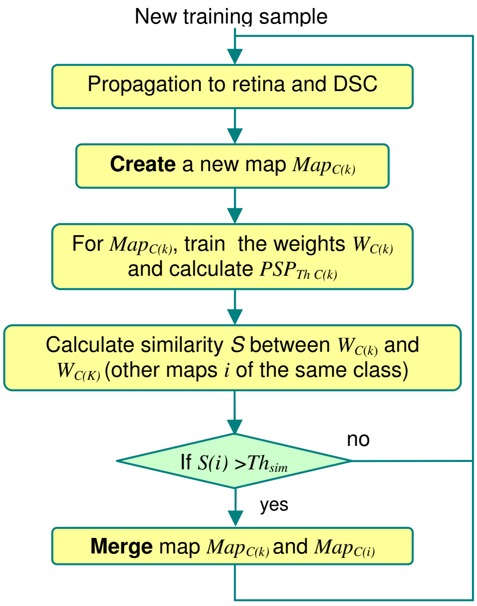

New training sample
Propagation to retina and DSC
Create a new map $Map_{C(k)}$
For $Map_{C(k)}$, train the weights $W_{C(k)}$ and calculate $PSP_{Th\ C(k)}$
Calculate similarity S between $W_{C(k)}$ and $W_{C(K)}$ (other maps i of the same class)
If $S(i) > Th_{sim}$
no
yes
Merge map $Map_{C(k)}$ and $Map_{C(i)}$

Fig. 3.9. Online learning procedure flowchart.

Notice that the learning procedure updates W and PSP_{Th} as well as enabling map merging for each incoming sample during training. For this reason, presenting the samples to the network in a different order can potentially lead to different network structures as well as different resultant W and PSP_{Th}. In other words, samples presented in a different order could potentially form slightly different clusters (different numbers of output maps for a given class), which can in turn affect the performance of the network. In this work, this property has not been explicitly explored as it was demonstrated as being negligible in experimental results.

3.4.3 Experimental setup

The performance of the adaptive learning procedure proposed in the previous section is evaluated and compared with benchmark work using the AT&T dataset available from (AT&T Face Dataset, 2007). The dataset is composed of 400 face views from 40 different individuals (10 views/individual). The frontal views of faces are taken with rotation angles varying in the range of [-30°, 30°] (rotation angles were not strictly controlled). The images were taken at different sessions for some individuals, without systematic control of light conditions and facial expression. Facial views in the dataset are in greyscale with 92 x 112 pixels.

Image Preparation

The position of eyes and mouth were manually annotated and the faces rotated to align the right and left eyes horizontally. The boundaries of the region of interest (ROI) were then defined as a function of the inter-ocular distance and the distance between the eyes and mouth. The ROI was then normalized to a size of 20 x 30 pixels in order to reduce the insertion of redundant information into the system, which is an inherent property of the representation of visual information using pixels. From the pattern recognition view point, the reduction of pixels speeds up the computation by avoiding the processing of similar information that can not further contribute to the recognition task and diminishes the effects of the well-known "curse of dimensionality" (Bishop, 2000).

The two-dimensional array (20 x 30) obtained from the size normalization was used as input to the SNN. No contrast or illumination manipulation was performed as previous work demonstrated the network's good response in the presence of noise and illumination changes (Delorme and Thorpe, 2001).

SNN Parameters for face recognition

The neuronal maps of retina (L1), directionally selective cells (L2) and output maps (L3) have a size of 20 x 30 pixels. The number of time steps used to encode the output of retina cells to the time domain is set to 100. The threshold for the directionally selective cells is set to 600, chosen in such a way that on average only 20% of neurons emit output spikes. The modulation factor, mod $\in$ (0, 1) is set to 0.98. In this way the efficiency of the input of a given neuron is reduced to 50% when 50% of the inputs receive a spike. The retina filters are implemented using a 5x5 Gaussian grid (calculated with Equation 3.3) and directionally selective filters are implemented using Gabor functions in a 7x7 grid according to Equation 3.4. The output of retina and directionally selective filters are shown in Figure 3.10.

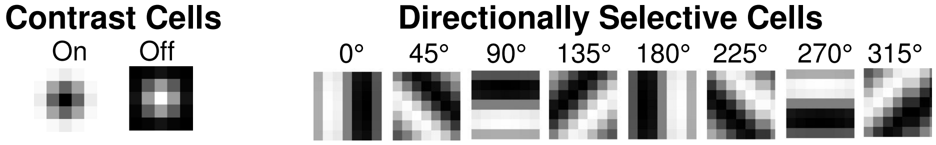


Fig. 3.10. Retina and directionally selective filters.

3.4.4 Comparative results

Comparison with previous work

Previous work demonstrated the high accuracy of the network and offline learning in coping with noise, contrast and luminance changes, reaching 100% in a training set composed of 10 samples (views) for each class (individuals) and 97.5% when testing generalization properties (Delorme and Thorpe, 2001). For the generalization experiment, the dataset was divided into 8 samples for training with the remaining two samples for testing. In this setup, the new adaptive learning procedure reached similar levels of accuracy as reported in (Delorme and Thorpe, 2001) with the training and test sets approaching 100%.

Testing the adaptive properties on unseen data

In another experimental setup, the ability of the new proposed system to add online output maps to achieve better generalization is demonstrated. Only three samples from each individual were used for training. The remaining seven samples of each person were used for testing. Three samples that appeared to be the most dissimilar were selected manually for training. The dissimilarity was mostly related to the facial views acquired from different angles. Thus, the training set was composed mostly of one view taken from the left side (30°), one frontal view and one view taken from the right side (-30°), as depicted in Figure 3.11.

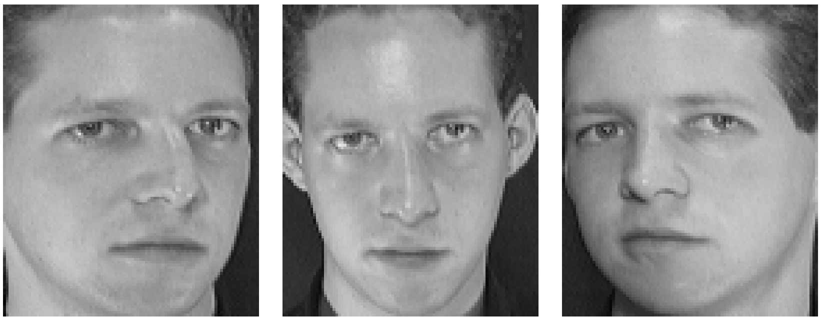


Fig. 3.11. Example of image samples used for training (30°, frontal and -30°) (AT&T Face Dataset, 2007).

Results are presented in terms of FAR (false acceptance rate) and FRR (false rejection rate) which are respectively calculated as:

$$FAR = \frac{NumFalseAcceptances}{TotalNumImpostorAttempts} \quad \textbf{(3.10)}$$

and

$$FRR = \frac{NumFalseRejected}{TotalNumClaimantAttempts} \quad \textbf{(3.11)}$$

On test data, we have used for each true claimant, the remaining 39 individuals to simulate impostor attempts. Thus, the total number of impostor attempts in the dataset is 40 claimants x 7 samples x 39 individuals = 10920 impostor attempts. The total number of true claimants is 40 x 7 samples = 280. The results are shown in Table 3.1. In column 2 of Table 3.1, Th_{sim} = 0.5 is set in such a way that only one output map for each class is created. With such a condition, the online learning procedure becomes equivalent to the original offline learning procedure described in Equation 3.5 (the same number of output maps are generated with the same weights). Tuning of Th_{sim} for performance in a scenario where the FAR is more important than the FRR (where FAR is around 10 times lower than FRR), the advantage of using more maps to represent classes that contain highly variant samples can be clearly seen, as the FRR decreases by 6% with a reduction of the FAR (last column of Table 3.1). In Table 3.1 only the best results achieved for different values of Th_{sim} are shown after scanning for different values of PSP threshold in L3 (according to Equation 3.7).

Table 3.2 and Table 3.3 present the network performance on the test set for different values of PSP threshold that are calculated as a function of the proportionality constant *c*. In all experiments, the constant *c* is the same for all maps. In a batch-mode operation the value of *c* can be optimized independently for each map after the training is completed, using Genetic Algorithms (GA) for example.

Table 3.1. Best result achieved on the test set (unseen samples) according to different similarity thresholds (Th_{sim}). Three samples of each claimant are used for training and the remaining seven for test. The systems are compared in conditions where FAR are around 10 times lower than FRR.

Similarity threshold Th_{sim} **(x10^{-3})**	**0.5**	**0.833**	**1.0**	**1.25**	**2**
Number of output maps	40	47	80	109	120
FRR (%)	25.7	22.5	21.4	20.0	**20.0**
FAR (%)	2.3	2.2	2.18	2.26	**1.77**

Table 3.2. Accuracy for different values of *c* keeping Th_{sim} = 0.5 x 10^{-3}. Number of output maps = 40. The systems are compared in conditions where FAR are around 10 times lower than FRR.

PSP threshold (PSP_{Th})	***c* = 0.30**	***c* = 0.35**	***c* = 0.40**	***c* = 0.45**
FRR (%)	27.9	**25.7**	26.8	28.6
FAR (%)	3.1	2.3	1.6	1.2

Table 3.3. Accuracy for different values of *c* keeping Th_{sim} = 2.0 x 10^{-3}. Number of output maps = 120. The systems are compared in conditions where FAR are around 10 times lower than FRR.

PSP threshold (PSP_{Th})	***c* = 0.30**	***c* = 0.35**	***c* = 0.40**	***c* = 0.45**
FRR (%)	25.00	21.4	20.0	**20.0**
FAR (%)	2.95	2.49	1.8	1.0

In a further comparison, to assess how difficult the dataset is and to have a better idea of the performance of the new learning algorithm, the face recognition system using adaptive SNN is compared with three other traditional methods of face recognition (Table 3.4).

Table 3.4. Comparison of different face recognition methods (experiments using NeuCom (Kedri NeuCom, 2007)).

Method	**Accuracy (%)**	**Properties**
PCA + SVM	90.7	Batch-mode
PCA + MLP	89.6	Batch-mode
PCA + ECF	74.0 (120 nodes)	One-epoch
Adaptive SNN	80.0 (109 maps)	One-pass online method

PCA (principal component analysis) is used to extract facial features. Classification is done using SVM (support vector machine) (Cristianini and Shawe-Taylor, 2004), MLP (multi-layer perceptron) neural network (Haykin, 1999) and ECF (evolving classifier function) (Kasabov, 2007). MLP and SVM are batch-mode methods while ECF presents adaptive learning characteristics similar to those proposed in this work. ECF can be trained in both one-pass and multi-pass mode (several epochs) (Kasabov, 2007). As expected, the batch-mode algorithms out performed the one-pass online methods. The reason for this is that for the batch-mode, the training samples are repeatedly presented to the classification method to minimize output errors. In one-pass online learning the adjustment of weights occurs only once, at the time the training samples are presented to the network. Therefore, the performance of the batch-mode methods can be considered roughly the target or the maximum accuracy that can be attained. When comparing the one-pass methods, the adaptive SNN presented better performance than ECF. Please note, this comparison cannot detect if the better performance is due to the learning method or to different representation of the features.

3.5 A new multi-view face recognition system using evolving SNNs

The previous section presented an SNN model that learns from multiple examples. Several training samples were presented to the network online, which learns different views of a class through synaptic plasticity and structural adaptation. Here, the SNN model for face recognition is further analysed, with the decision-making process now based on multiple views. A quantitative analysis on the use of different number of training samples is also presented.

In this scenario, the network architecture is extended, with the implementation of an additional layer. The main amendment to the architecture is the inclusion of a layer that collects opinions over several frames to recognize patterns in streams of video data.

3.5.1 Model description

The system uses the same neuronal model described in Section 3.3. The network structure, where neurons are placed in two-dimensional grids forming neuronal maps and consequent layers of maps, also follows the same pattern. The neural network is composed of four layers of integrate-and-fire neurons (See Figure 3.12). Layer 1 and Layer 2 are similar to those described in Section 3.3. In the first two layers (L1 and L2) there is no learning, they simply act as passive filters and time domain encoders. In the third layer (L3), where the learning takes place, maps are trained to be sensitive to incoming excitation of more complex patterns. Neuronal maps are created or merged during learning, according to the online learning procedure described in Section 3.4.2. There are lateral inhibitory connections between neuronal maps in the third layer, so that when a neuron fires in a certain map, other maps receive inhibitory pulses in an area centred in the same spatial position. An input pattern belongs to a certain class if a neuron in the corresponding neuronal map spikes first.

Layer 4 (L4), as the main addition to this structure, has one neuronal map containing a single neuron for each pattern class. The L4 neuron of a given class is connected to the corresponding L3 neuronal maps. There are excitatory connections (typically $w = +1$) between the L4 neuron and the neurons located close to the centre of L3 maps. Thus, L4 combines the results of a sequence of visual patterns, i.e. accumulates opinions from several frames.

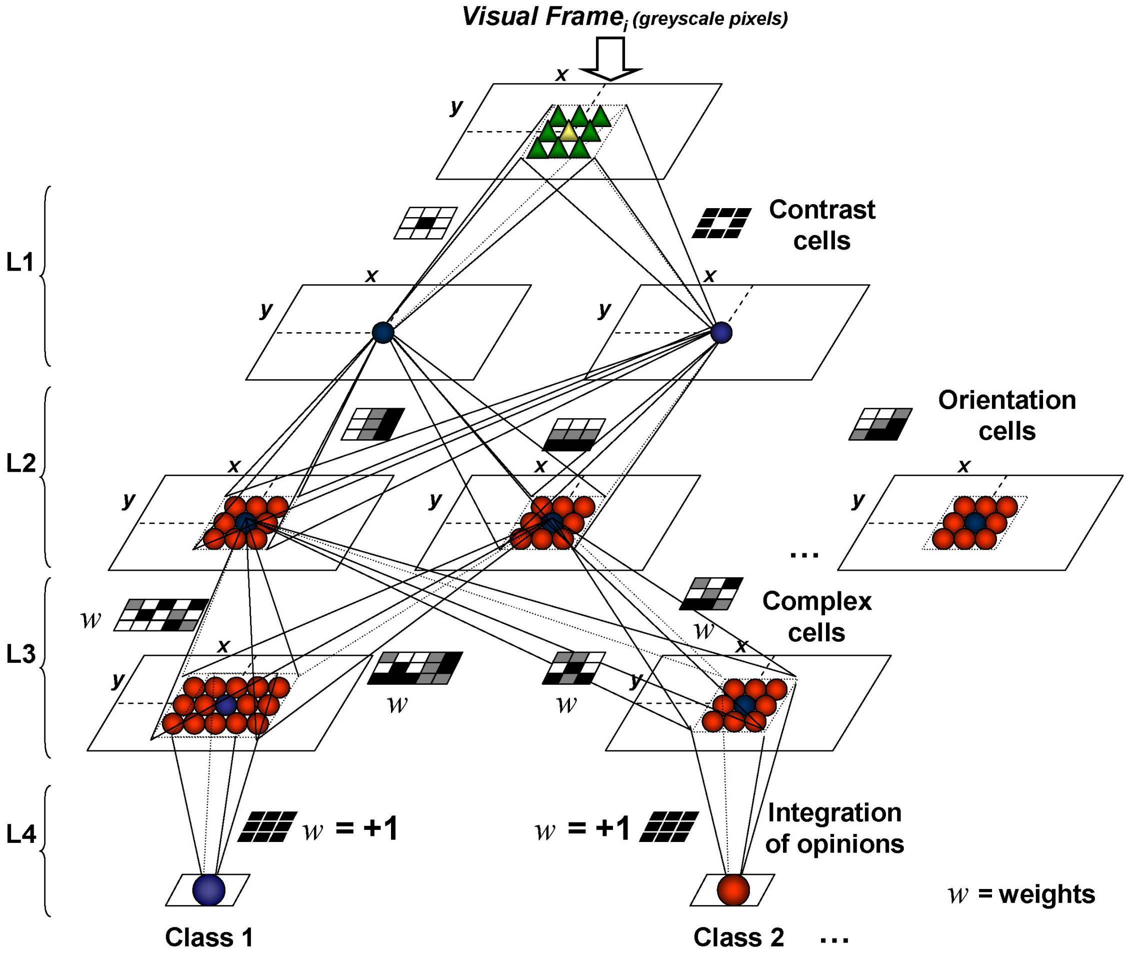


Fig. 3.12. SNN architecture composed of four layers. Neurons in L1 and L2 are sensitive to image contrast and orientation, respectively. L3 has the complex cells, trained to respond to specific patterns. L4 accumulates opinions over different input excitations in time.

In respect of the connection weights between L3 and L4, in the simplest case, they are not subject to learning. Excitatory connections with fixed amplitude can be used instead. In a more elaborate setup, connection weights with amplitude varying according to a Gaussian curve centred in the middle of each L3 map gives a sense of confidence regarding the L3 output spikes. This is because only the middle neuron in each L3 neuronal map is trained to respond optimally to a certain excitation pattern, decreasing in reliability as the neuron's location approach the map's extremities.

However, independent of the choice of weights, the PSP thresholds for L4 neurons need to be assigned. L4 PSP thresholds can be trained using a global optimization algorithm, or alternatively, as was done in the following experiments, a simple heuristic that defines L4 PSP thresholds as a proportion *p* of the number of frames used for testing can be used. With the inclusion of this simple procedure, it is possible to assess how many positive opinions from different frames are required to recognize a pattern successfully.

3.5.2 Network dynamics during test

Figure 3.13 illustrates the behaviour of the network in time, which can be described as follows: each frame of visual excitation is propagated to L1, which enhances the areas with high contrast and encodes it into spikes with Rank Order Coding (Delorme *et al*, 2001). Spikes of a given frame are propagated to L2 and L3 until a neuron belonging to a L3 map emits the first output spike, which is consequently propagated to L4. If a neuron in L4 generates an output spike, the simulation is truncated and the frames are labelled to the corresponding class. Otherwise, if there is no output spike in any L4 neuron and an L3 neuron has emitted a spike or there are no more spikes, the next frame is propagated. The following frame starts to be propagated after resetting the PSP in L2 and L3 neurons. L4 neurons retain their PSP levels which accumulate over consecutive frames, until a class is recognized with an L4 neuron output spike or until there are no more frames to be processed.

It is important to note that when resetting the PSP in L2 and L3 neurons, information about dynamic changes in the patterns is lost. Thus, this model does not keep track of the variations of a visual pattern nor the pattern's changes over time. Each visual pattern is considered independently, and L4 neurons effectively accumulate opinions of each frame being/not being similar to a trained pattern.

Computation with spikes, where neurons require a certain stimuli level to release output spikes, provides a distinctive difference when compared to traditional methods. Traditional networks usually return numeric values in each of the output nodes. The propagation of an excitatory input to the SNN described, on the other hand, can result in three different output conditions:

a) one or more output spikes occur in the same neuronal map: it occurs when the output neurons have enough excitation to issue output spikes. It suggests confidence of it being a single class is high;

b) silent output neurons: at the end of the simulation, none of the output neurons in any class were excited enough to produce output spikes. In this case, *no class* is assigned to the incoming patterns;

c) simultaneous spikes: when output spikes occur at the same time in neuronal maps with different class labels. The network outputs that the incoming pattern belongs to more than one class.

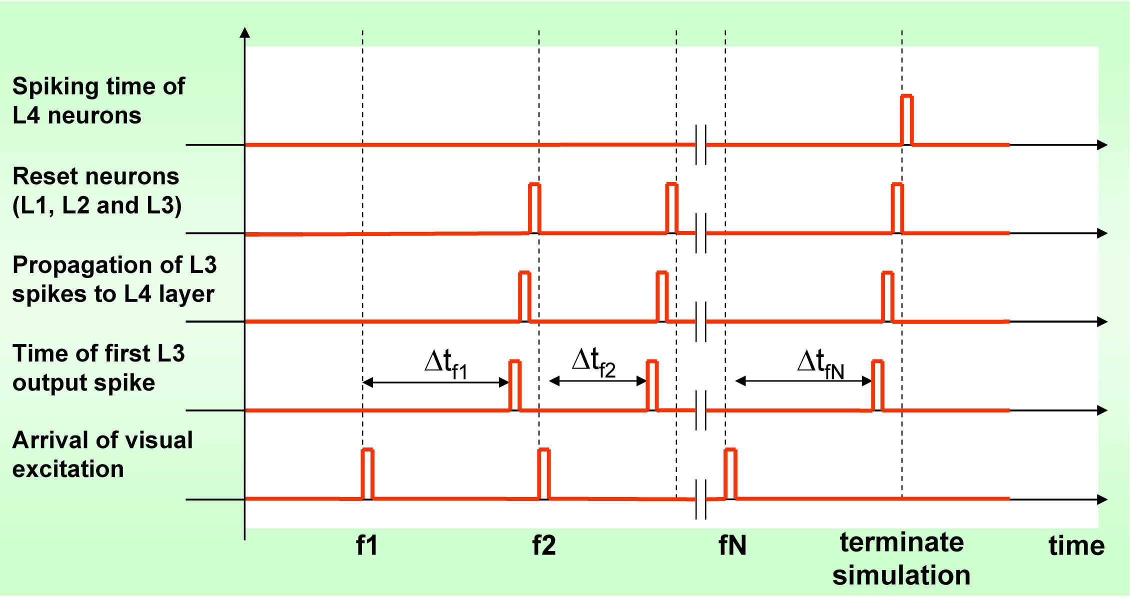

Spiking time of L4 neurons
Reset neurons (L1, L2 and L3)
Propagation of L3 spikes to L4 layer
Time of first L3 output spike
Arrival of visual excitation
Δt_{f1}
Δt_{f2}
Δt_{fN}
f1
f2
fN
terminate simulation
time

Fig. 3.13. Behaviour of the four layers of the SNN architecture over time. The visual excitation (frames f1, f2,…, fN) is propagated through L1, L2 to L3 until an L3 neuron generates an output spike (Δt_{fN}). L3 spikes are propagated to L4 and, if there is no output spike in any L4 neuron, L1, L2 and L3 neurons are reset to the rest potential. A new frame is then processed. The simulation is terminated when an L4 neuron spikes or there are no more frames to be processed.

These conditions represent the normal operation of the system and are used in all the experiments unless clearly stated otherwise.

One of the properties of this system is the sparse activity of the neurons, as described in Section 3.3.1, which enables the optimization of the computational performance. In general, the processing is interrupted before the entire simulation is completed. The simulation is terminated once a single neuron of the output layer (L4) reaches the threshold and emits an output spike. Thus, frequently, even though there are more frames that can provide further information, the process is interrupted because a certain level of confidence has been reached.

3.5.3 Experiments and results

Previous work demonstrated the high performance of the SNN when dealing with noise, contrast and illumination changes (Delorme and Thorpe, 2001) with a single frame. With the same publicly available dataset, the new online learning procedure was tested with similar results, reaching an accuracy of nearly 100% in both training and test sets (Wysoski *et al*, 2006) (Section 3.4.4). In a more challenging setup (only three samples for training in each class), the online addition of neuronal maps proved to be beneficial (Section 3.4.4).

Here, the extended spiking network model proposed is evaluated on a face recognition task. This time, the recognition with multiple views is tested on video streams of the VidTimit dataset created by Sanderson and Paliwal (Sanderson and Paliwal, 2004) (see Figure 3.14). The choice of this dataset was motivated by the aim to design and implement an integrated system to perform biologically-motivated audiovisual integration.

In the VidTimit dataset, video streams capture frontal views of individuals' faces while uttering predefined sentences. The dataset is composed of 10 streams of video (106 frames on average) from 43 different speakers, captured at 25 frames per second, recorded in 3 sessions. Individuals utter six sentences in the first session and two sentences each in the second and third sessions.

The algorithm developed by Viola and Jones (Viola and Jones, 2001 implemented in the OpenCV (Intel OpenCV, 2007) image processing library was used to perform automatic face detection. Detected faces were converted to greyscale, normalized in size (60 x 40 pixels) and convolved with an elliptical mask to decrease the amplitude of pixels at the image borders. Figure 3.15 shows examples of detected faces normalized in size. The dataset did not require face rotation and it was not necessary to normalize the faces in respect to illumination as this type of SNN has shown its robustness to illumination changes (Delorme and Thorpe, 2001).

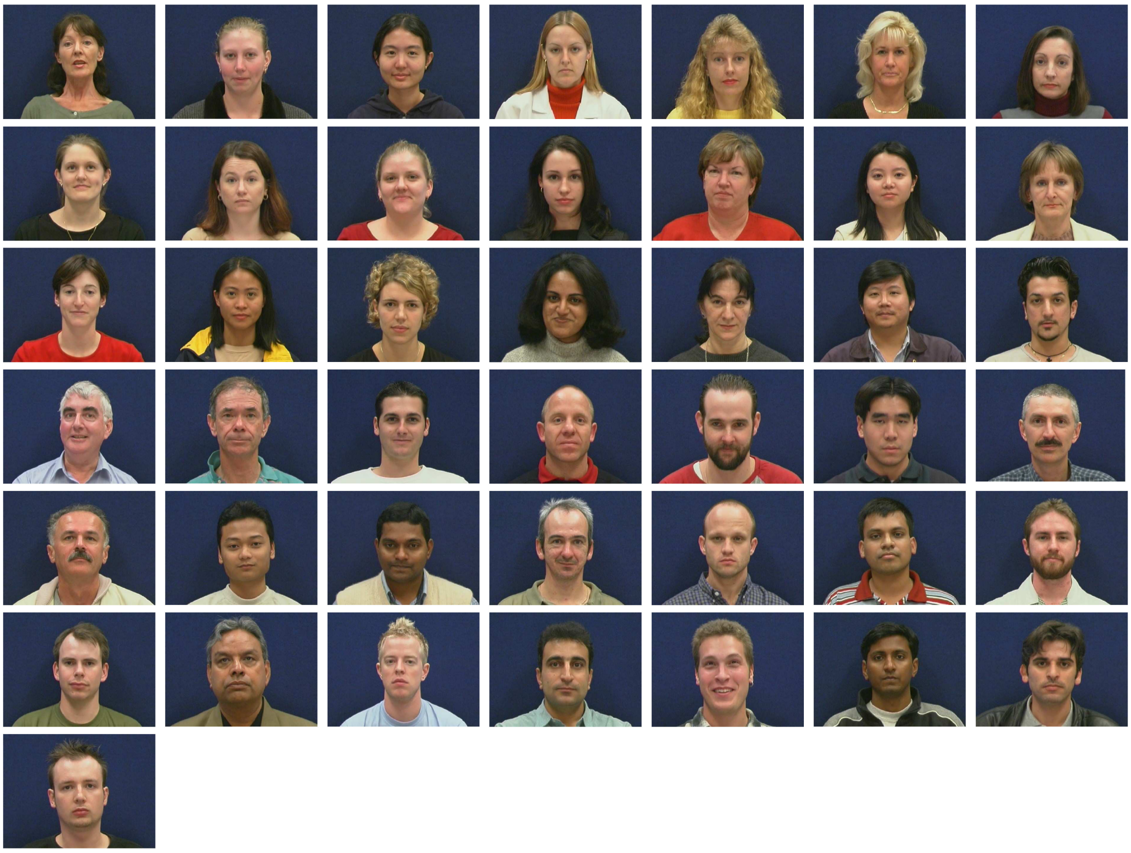


Fig. 3.14. VidTimit dataset composed of 43 individuals.

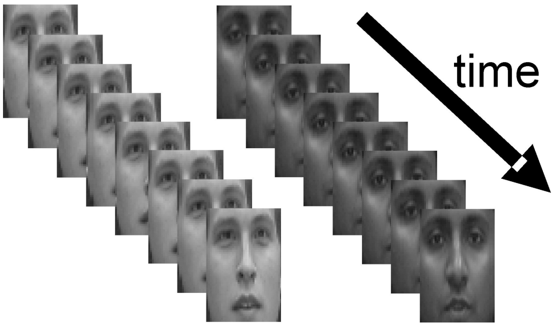


Fig. 3.15. Examples of VidTimit dataset after faces are detected, converted to greyscale and normalized in size.

On/Off cells with two frequency scales are used, so that the number of neuronal maps in L1 is set to 4 (2 pairs). In scale 1 the retina filters are implemented using a 3 x 3 Gaussian grid with σ= 0.9 and scale 2 uses a 5 x 5 grid with σ = 1.5. In L2, there are 8 different directions in each frequency scale with a total of 16 L2 neuronal maps. The directionally selective filters are implemented using Gabor functions with aspect ratio γ= 0.5 and phase offset φ = π/2. In scale 1 there is a 5 x 5 grid and wavelength λ = 5 and $\sigma = 2.5$ and in scale 2 a 7 x 7 grid with λ and σ set to 7 and 3.5, respectively. The Gabor functions and Gaussian filters were chosen based on experimentation with different combinations of values. The On/Off and directionally selective filters are shown in Figure 3.16. Figure 3.17 shows an example of PSP generated in the network maps by the propagation of an input face.

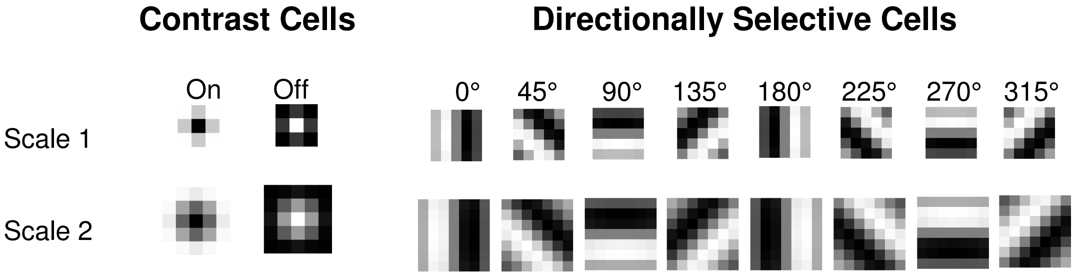


Fig. 3.16. Weights of On/Off and directionally selective neurons applied in L1 and L2 neuronal maps in two different frequency scales.

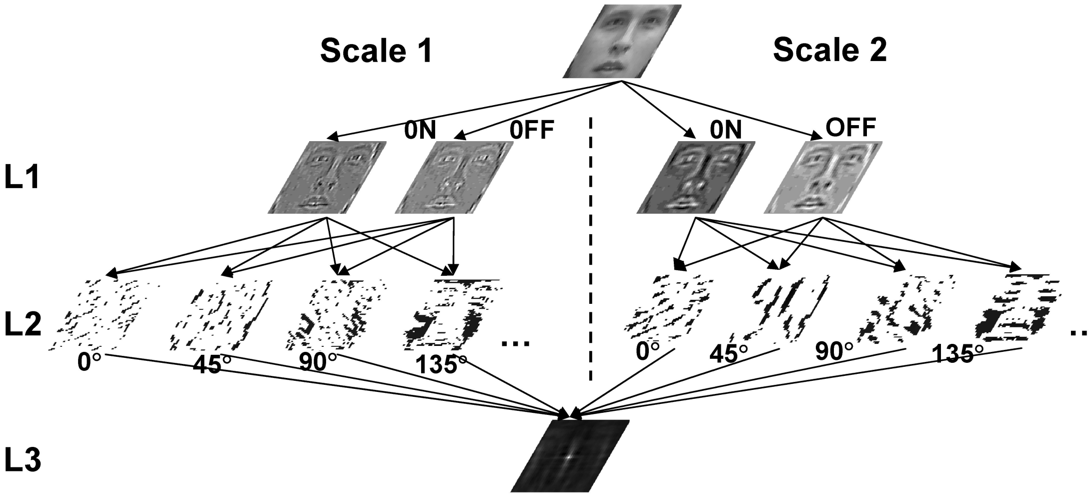


Fig. 3.17. Example of neuronal activity when the network is submitted to an external excitation (face extracted from VidTimit dataset created by (Sanderson and Paliwal, 2004)).

Experiment 1 - Evolvability

In the first experiment, in order to reproduce the same experimental setup described in (Sanderson and Paliwal, 2004), the system is trained to recognize 35 individuals. For testing, all 43 individuals are used. Thus, the testing set is composed of different frames of 35 individuals that have already participated in the training process and 8 completely unknown individuals. The modulation factor mod $\in$ (0, 1) was set to 0.995. The thresholds of the L2 cells were set to 0.3.

In the first experiment, the online learning procedure is evaluated on the VidTimit dataset, with particular focus on the adaptive addition of neuronal maps within a class to accommodate several training samples (views). For this, different numbers of samples (1, 3 and 5) were used to train on the 35 users. The training samples were chosen from different video streams (using the first frame from each video stream). The similarity threshold for merging neuronal maps was kept at a high level in order to inhibit map merging. Thus, each frame effectively originated a new neuronal map. For testing, one frame of the 43 individuals in the dataset was used, acquired in two different sessions (86 frames). The network was setup to give a decision for each test frame.

Figure 3.18 shows a comparison of the results on test frames for different numbers of training samples, varying the proportionally constant *c* in Equation 3.7, which effectively increases the firing threshold PSP_{Th} in L3 neurons.

Note that, varying the PSP_{Th} in L3 neurons, the system can have different operating points. As a general rule, when PSP_{Th} increases so does the FRR while the FAR decreases. Total error (TE) = FAR + FRR. In Figure 3.19 the performance of the network is plotted using different numbers of training samples (1, 3 and 5) for FRR with respect to FAR. It can be clearly seen that in the EER (equal error rate) region where FAR is equal to FRR the use of additional training samples enhances the performance. However, no further improvement was obtained with the inclusion of more than five training frames.

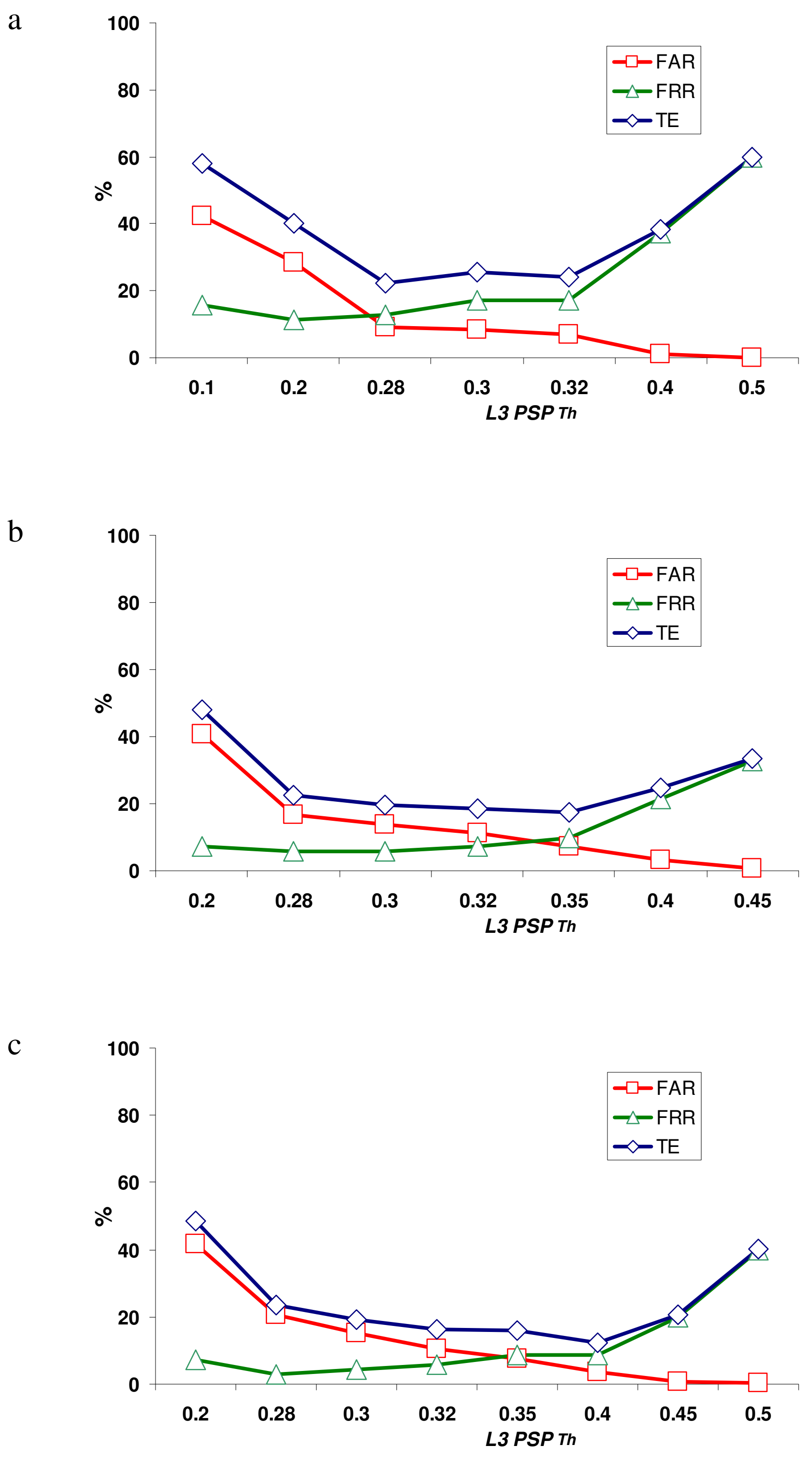


Fig. 3.18. Performance of the SNN network for various L3 firing thresholds PSP_{Th} using: a) 1; b) 3; c) 5 training samples per individual. As a general rule, when the threshold increases so does the FRR while the FAR decreases.

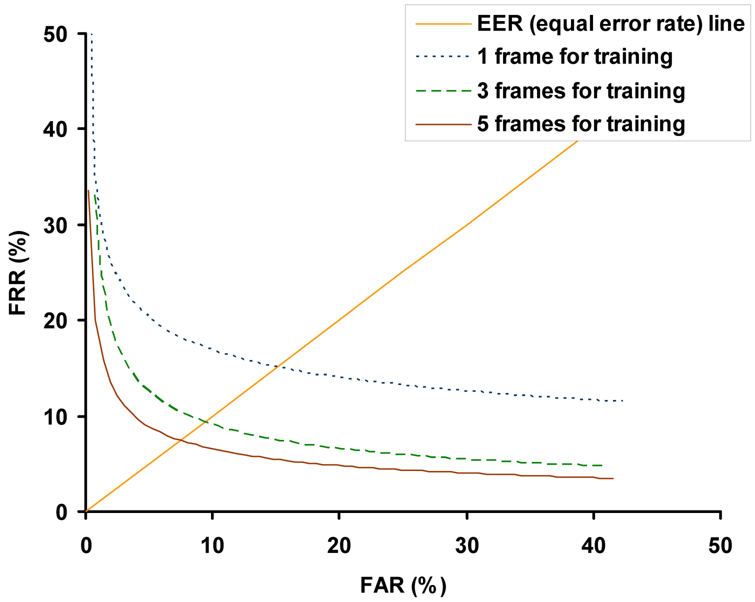


Fig. 3.19. Performance of the network using different numbers of training frames (1, 3 and 5) re-plotted with respect to FRR and FAR. Trend lines corresponding to the original points in Figure 3.18 are plotted.

Experiment 2 - Multi-view recognition

The following experiment evaluated a combination of multiple frames for decision-making. Here the number of training frames was kept constant to two in each class, only varying the number of frames used for recognition. Similar to the previous experiment, 43 individuals' frames acquired in two different sessions (86 frames) were used during the test. 1, 3 and 5 frames were evaluated here, with PSP_{Th} of L4 neurons set to 1, 2, and 3 respectively, which means that for recognition based on 3 frames, 2 frames need to be positively recognized and for recognition based on 5 frames, at least 3 positive opinions are required. In the experiments, these scenarios demonstrated themselves to be a good trade-off between accuracy and resources required (processing speed and memory). Frames were spaced 400 ms from each other to allow substantial changes in the acquired face. Figure 3.20 shows a comparison of the results for the different number of test frames, varying the proportionality constant *c* in Equation 3.7, which increases the firing threshold PSP_{Th} of L3 neurons. Plotting the results on a FAR x FRR plane shows that the results are very favourable for multi-view recognition use (Figure 3.21), which demonstrates the ability of the network to accumulate opinions over several frames.

These results are coherent with the results presented by Kittler *et al* (Kittler *et al*, 1997), where the fusion of multiple measurements of a single biometric modality in the framework of Bayesian estimation theory is formulated. In this framework, different fusion strategies were evaluated (average, maximum, minimum, median rules) with error rates decreasing by up to 40%. It has been reported that, in a dataset of 37 persons, the EER was reduced from 6.9% on considering only one frame, to 4.0% when considering the opinions of six frames. Similar to (Kittler *et al*, 1997), the experiments presented here also show that the gain in performance tended to saturate after the integration of a few opinions. More precisely, after five frames, no further enhancement could be noticed.

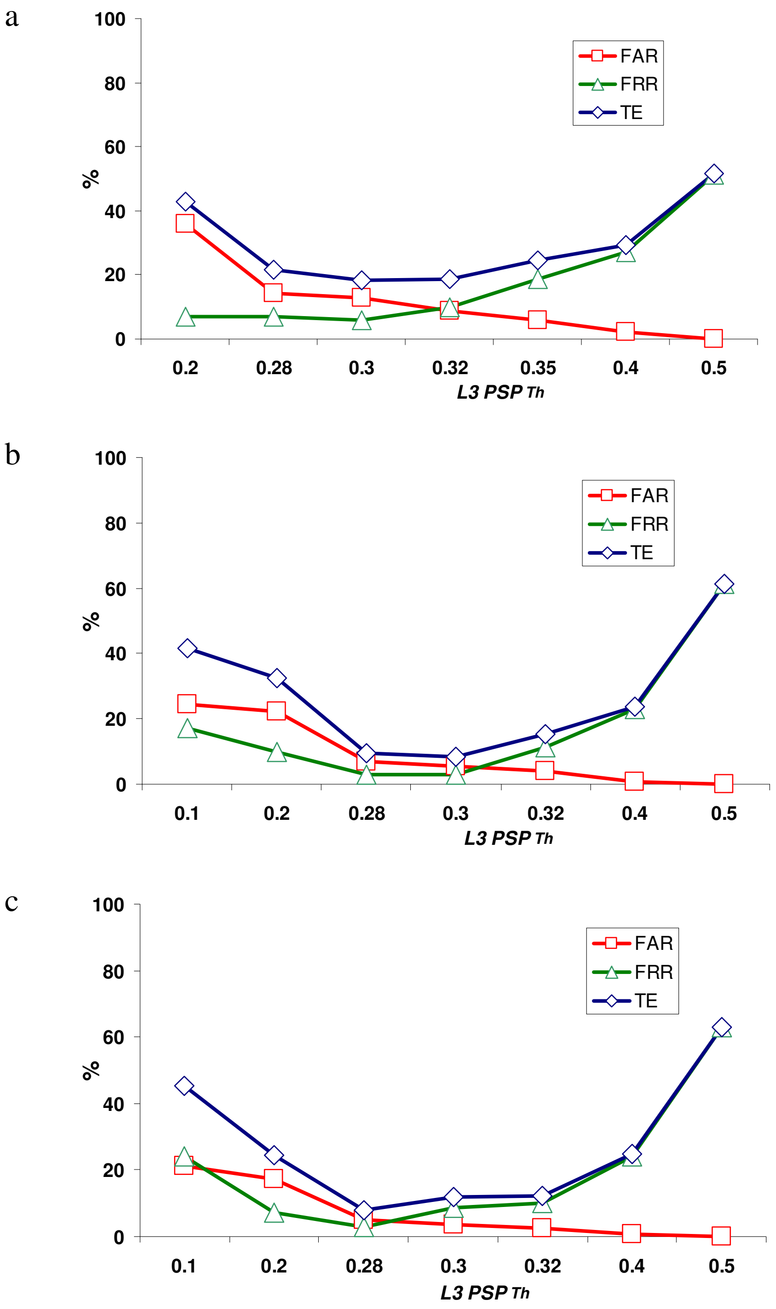


Fig. 3.20. Performance of the SNN network for different L3 PSP_{Th}. a) Test samples are composed of 1 frame and L4 $PSP_{Th} = 1$; b) Test samples are composed of 3 frames and L4 $PSP_{Th} = 2$; c) Test samples are composed of 5 frame and L4 $PSP_{Th} = 3$.

In the audiovisual recognition system described by Sanderson and Paliwal (Sanderson and Paliwal, 2004) is reported that the face recognition system alone reached a total error TE ≈ 8% based on PCA features and support vector machine (SVM) with the same VidTimit dataset. Figure 3.20.c shows an example of the same levels of performance being reached. For *c* of L3 $PSP_{Th} = 0.28$ and L4 $PSP_{Th} = 3$, a TE = 5.0 (FAR) + 2.9 (FRR) = 7.9 % was obtained.

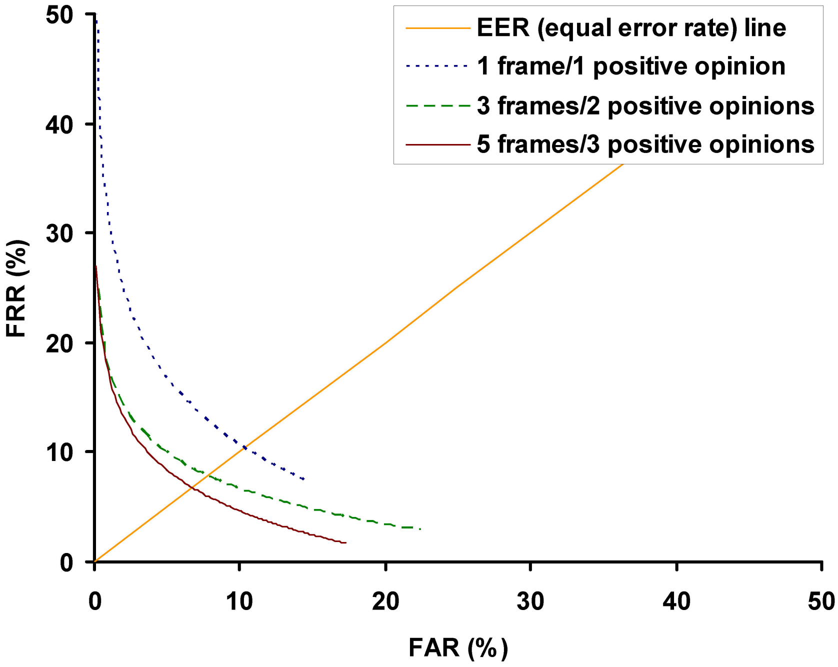


Fig. 3.21. Performance of the network using different number of testing frames (1, 3 and 5 frames) setting the L4 decision threshold to 1, 2, and 3, respectively. The number of training frames was kept constant at 3. Trend lines corresponding to the original points in Figure 3.20 are plotted.

Table 3.5 presents the results of multiple view recognition with different sizes of training and testing sets. All the results were obtained keeping the number of training samples for each user constant, i.e., two samples, and using 3 frames for recognition (with L4 PSP_{Th} set to 2). In the first scenario, the system is trained to authenticate 35 users and 8 users are used as impostors. In the second scenario, 22 users are trained and 21 impostors are used. Finally, the system is trained to authenticate 8 users and submitted to the test of 35 impostors.

Table 3.5. Performance of multi-view recognition under different training and testing conditions (lowest TE for each scenario is in bold).

		Proportionally constant c (PSP$_{Th}$ of L3 neurons) (See Equation 3.7)						
Scenario		**0.1**	**0.2**	**0.31**	**0.33**	**0.36**	**0.39**	**0.4**
35 users **8 impostors**	FAR	19.6	18.0	3.5	2.8	**1.9**	0.9	0.7
	FRR	17.1	10.0	8.6	14.3	**10**	18.6	22.8
	TE	36.7	28.0	12.1	17.1	**11.9**	19.5	23.5
22 users **21 impostors**	FAR	24.7	23.3	5.9	5.2	**3.1**	1.8	1.3
	FRR	9.1	9.1	11.4	11.4	**11.4**	22.7	27.2
	TE	33.8	32.4	17.3	16.6	**14.5**	24.5	28.5
8 users **35 impostors**	FAR	44.8	40.0	19.3	14.9	7.6	**4.2**	3.5
	FRR	0	6.2	12.5	12.5	12.5	**12.5**	18.7
	TE	44.8	46.2	31.8	27.4	20.1	**16.7**	22.2

The results in Table 3.5 and their corresponding trend lines depicted in Figure 3.22 suggest that the ratio of number of users/number of impostors does not influence performance.

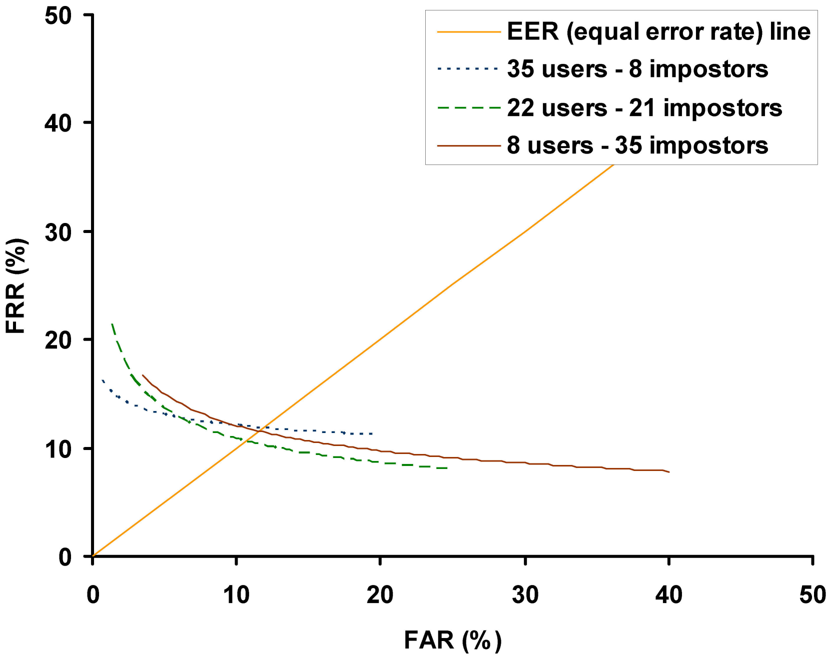


Fig. 3.22. Comparative performance for different training and testing conditions (varying number of users/number of impostors ratio). Trend lines corresponding to original points in Table 3.5 are plotted.

In another comparison, the network is trained with different numbers of users (10, 20 and 35) keeping the number of impostors in the test set to a constant 8, in order to investigate the effect of dataset size on performance of the network. For these experiments, 2 samples were used for training and 5 frames were used for recognition (with L4 PSP_{Th} set to 3). The results are presented in Table 3.6 and the trend lines in

Figure 3.23. It can be seen that changing the size of low/medium scale datasets does not have any significant effect on performance.

Table 3.6. Performance of multi-view recognition with different number of users (number of impostors are kept constant) (lowest TE for each scenario is in bold).

		Proportionally constant *c* (*PSP*$_{Th}$ of L3 neurons) (See Equation 3.7)					
Scenario		**0.1**	**0.2**	**0.3**	**0.33**	**0.36**	**0.4**
10 users	FAR	36.3	29.2	14.3	7.4	**4.6**	0.9
	FRR	15	10.0	10	10	**10**	10
	TE	51.3	39.2	24.3	17.4	**14.6**	10.9
20 users	FAR	20.9	20.3	**3.8**	3.5	2.8	0.8
	FRR	17.5	5	**10**	12.5	17.5	25
	TE	38.4	25.3	**13.8**	16.0	20.3	25.8
35 users	FAR	16.5	13.6	2.72	**2.2**	1.3	0.4
	FRR	24.3	7.1	8.6	**11.4**	15.7	24.3
	TE	40.8	20.7	31.8	**13.6**	17	24.7

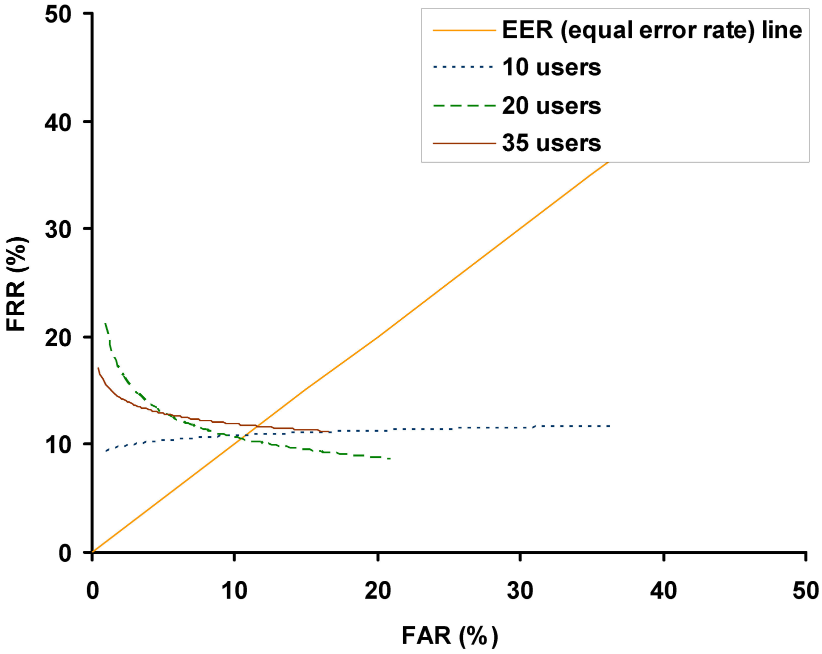


Fig. 3.23. Comparative performance for different dataset sizes. Trend lines corresponding to original points in Table 3.6 are plotted.

Experiment 3 - Comparison with baseline methods

Finally, the publicly available data analysis environment NeuCom (Kedri NeuCom, 2007) was used to compare the new SNN method with well-established basic feature extraction and pattern recognition techniques. On the VidTimit dataset, facial features were extracted using PCA (principal component analysis) and tested using three classification methods: Support Vector Machine (SVM), Multi-layer Perceptron Neural

Network (MLP), and standard Nearest Neighbour (Euclidean distance on the normalized feature space). For this comparison, three frames from 35 users (105 samples) were used for training, selected from different sentences in session 1. The same 35 users were used for testing, but with samples acquired from different sessions (session 2 and session 3 were used). One frame from each session, 70 samples in total. Differing from the previous two experiments, a closed setup was used, i.e., in testing, the classifiers had to always label an output class from the trained labels, i.e., there is no '*no class*' nor '*more-than-one class*' option. The best results obtained after hand-tuning parameters in NeuCom are shown in Table 3.7 for different numbers of principal components, where MLP and SVM were able to correctly label 95.7% of the samples. In the experiments with MLP, the best results were obtained with 50 hidden neurons. In SVM, best results used linear kernel.

Table 3.7. Accuracy of VidTimit dataset using traditional classifiers and PCA for feature extraction.

Number of Principal Components	**30**	**50**	**100**
SVM (%) (batch-mode)	94.3% (66/70)	94.3% (66/70)	**95.7% (67/70)**
Nearest Neighbour (%)	**90.0% (63/70)**	90.0% (63/70)	81.4% (57/70)
MLP (%) (batch-mode)	92.9% (65/70)	94.3% (66/70)	**95.7% (67/70)**

Accuracy % (correctly labelled/number of samples)

To implement a closed setup in the SNN model, the simulation was truncated at a certain time step (half of the total simulation period was used, i.e., 50 steps) and assigned the testing sample to the label of the L3 map that contained the highest activated neuron (neuron with highest *PSP*). Note that this is an artificial setup for comparison purposes and does not represent the normal way the network operates. Modulation factor = 0.997 was used for the experiments with a closed setup. The best results are shown in Table 3.8.

Table 3.8. SNN accuracy on a closed setup (35 users for training and testing) for different PSP_{Th} of L2 and L4 neurons. The simulation was truncated after 50 time steps and the class label was assigned to the L3 map that contained the highest activated neuron.

PSP_{Th} of L4 neurons	PSP_{Th} of L2 neurons (directionally selective cells)			
(number of positive opinions)	**0.4**	**0.5**	**0.6**	**0.7**
1	88.5% (62/70)	87.1% (61/70)	91.4% (64/70)	87.1% (61/70)
2	95.7% (67/70)	95.7% (67/70)	94.3% (66/70)	92.8% (65/70)
3	94.3% (66/70)	**97.1% (68/70)**	**97.1% (68/70)**	90.0% (63/70)

Accuracy % (correctly labelled/number of samples)

The network accurately labelled 91.4% of the samples using only one frame/opinion for recognition. However, it can be seen in Table 3.8 that performance increased to 97.1% when PSP_{Th} of L4 neurons were set to 3, i.e., when applying the integration of opinions from several frames. Thus, it can be concluded that SNN reaches similar levels of accuracy as traditional methods, despite several differences between the methods. In respect to the classifiers, SVM and MLP are batch-mode algorithms, where the training samples are repeatedly presented to the classification method to minimize output errors. In the SNN learning procedure, the adjustment of weights occurs only once at the time the training samples are presented to the network.

3.6 Implementation considerations

The SNN models described, and the corresponding learning procedure, have been implemented using C++ language in a mix of *event-driven* and *clock-driven* techniques. The mix of *event-driven* and *clock-driven* approaches was aimed at meeting the requirements of the current stage of the development process. As the main goal was conceptual modelling and the corresponding implementation of the model as a proof-of-concept, the focus was to have fast and intuitive implementation with reasonable computational performance. The *clock-driven* approach is very straightforward to implement, yet computationally expensive for networks with large numbers of neurons spiking sparsely, as each neuron needs to be visited each time step by the algorithm even if the neuron is not the subject of any activity. The *event-driven* approach, on the other hand, uses queues of spikes (received/emitted) to describe and organize the dynamics of the network. This is very efficient for networks with low activity neurons. However, how the queues are implemented is highly dependent on the dynamics of the neurons (Brette *et al*, 2007). Thus, a general design of the queue structure is practically impossible. In practice, in an *event-driven* approach, changing the neuronal dynamic, requires the structure of the network to be re-designed (Brette *et al*, 2007) (Marian, 2002).

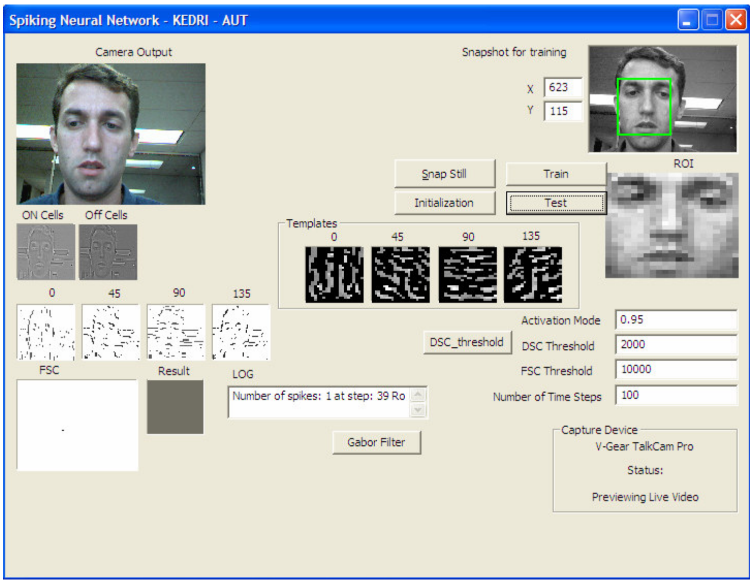
In this particular implementation, a good compromise was to implement the time steps in a *clock-driven* approach and the neuronal activity within a time step with the *event-driven* approach. Since the time steps in these simulations are in the order of hundreds only, there is no substantial waste of processing power visiting each time step, regardless of whether there is or is not neural activity. Within each time step, instead of visiting each neuron to check neuronal activity as suggested in the purely *clock-driven* approach (an order of millions), there is a queue that controls the occurrence of spikes in an event-based manner. The implementation can be summarized with the following pseudo-code:

```
Structure queue (has one item for each spike)
      queue[i].TargetNeuron
      queue[i].Weight



for all time steps                  (CLOCK-DRIVEN)
      {
      while (queue.Size > 0)        (EVENT-DRIVEN)
            {
            PropagateIncomingSpikeToTargetNeuron
            if PostSynapticPotential > Threshold
                  {
                  for all neurons connected to TargetNeuron
                        InsertItemInQueue
                  SetTargetNeuronPotentialToRest
                  }
            }
      }
```

As a result of the mix of *event-driven* and *clock-driven* techniques, computational performance able to take on the experiments described in this chapter has been possible on a single personal computer. However, it is important to note that for real time applications processing 30 frames per second with high resolution, a careful optimization of code is required, particularly when the number of trained neuronal maps starts to increase. Figure 3.24 illustrates the simulation of the evolving SNN for visual pattern recognition.



Spiking Neural Network - KEDRI - AUT

X 623

Y 115

Activation Mode 0.95

DSC Threshold 2000

FSC Threshold 10000

Number of Time Steps 100

Number of spikes: 1 at step: 39 Ro

Capture Device
V-Gear TalkCam Pro
Status:
Previewing Live Video

Fig. 3.24. Simulation of the evolving spiking neural network for visual pattern recognition.

3.7 Chapter conclusion

The chapter started with an overview of visual pattern recognition methods, with a particular emphasis given to brain-inspired algorithms. It was followed by a description and evaluation of a new procedure that performs online learning in a network of spiking neurons and a new extended SNN architecture that classifies visual patterns using multiple views from streams of video data.

In respect to the new learning procedure, new output maps are created and merged based on the clustering of intra-class samples during the learning stage. Experiments have shown that the learning procedure reaches similar levels of performance to the previously presented work (Delorme and Thorpe, 2001), and better performance can be attained in classes where samples have high variability with the *divide-and-conquer* approach. For this to occur, one more parameter needs to be tuned, i.e. Th_{sim} (similarity threshold for merging maps). In addition, more output neuronal maps require more memory to store trained weights.

With a benchmark dataset, VidTimit (Sanderson and Paliwal, 2004), further experiments indicate that the integration of several opinions (multi-view recognition) increases accuracy. With the VidTimit dataset, a comparison with traditional techniques shows that the method can be used for real visual pattern recognition problems.

In terms of normalization, the rank order codes are intrinsically invariant to changes in contrast and input intensities, basically because the neuronal units compute the order of the incoming spikes and not the latencies themselves (Delorme and Thorpe, 2001). It is suggested that, for this reason, adaptive SNN presents better result than PCA + ECF as feature extraction using PCA can degrade performance where there are illumination changes.

The adaptive SNN does not cope well with rotation of the patterns, requiring an additional step of performing rotation alignment prior to the propagation of visual excitation to the network. Alternatively, a certain degree of rotation invariance can be attained with the use of additional neuronal maps, in which each map needs to be trained to cover different angles. In this case, the adaptive learning procedure described, can automatically generate the new maps when it is required.

Table 3.9 contextualizes the results of the evolving SNNs in the three specific *Brain-like* ECOS aspects presented in Section 2.3.3, i.e., in terms of information processing units, information processing pathways and learning.

Table 3.9. Summary of results according to three specific aspects of *Brain-like* ECOS proposed in Section 2.3.3.

Processing Units	A fast and computationally inexpensive version of spiking neuron is used as processing unit in all stages of visual information processing.
Structure	Visual information propagates with feed-forward connections to four layers of two-dimensional grid of spiking neurons that represent the behaviour of various brain areas (retina cells, direction selective cells, complex cells).
Learning	The online evolving procedure enables the learning of external stimuli through synaptic plasticity and structural adaptation. The addition of new classes is done in a supervised way whereas the system adapts in an unsupervised fashion when new samples of a class are presented.

Overall, computation with pulses, contrast filters and orientationally selective cells has a close correspondence with traditional methods of image processing, such as wavelets, that have already proven to be very robust for feature extraction in visual pattern recognition problems. From the biological perspective, despite being a very simplified representation of what effectively happens in the brain, the use of pulses represents a reasonable starting point as it is stated in (Gerstner and Kistler, 2002).

As a future direction, aiming to improve the efficiency of biologically realistic neural networks for pattern recognition, it would be important to add adaptations to L1 and L2. It has been experimentally shown (Sharpee *et al*, 2006) (Tsukada and Pan, 2005) that neural filters change adaptively to increase the information carried out by the neural response. As a result, the contrast and directionally selective cells constitute optimized filters to describe natural scenes. A good exercise would be to explore how to find adaptively optimal filters in different types of data in a fashion similar to the human brain.

In addition, while the dynamics of the network have been demonstrated to be able to process streams of visual patterns, the system still needs to be evaluated further, particularly with respect to parameter optimization in consideration of three different criteria: data transmission performance (best encoding) (Bothe *et al*, 2002), speed (minimum processing time) and energy efficiency (lowest number of spikes and number of frames).

A large-scale problem also needs to be analysed in order to have a more in-depth idea of the behaviour of the system, not only in terms of accuracy, but also as concerns several engineering requirements, e.g., implementation speed, resources, scalability, etc. Despite some results showing the performance of the system with varying dataset sizes, this dissertation mainly concentrates on providing detailed behavioural analysis with a moderately low to medium numbers of samples.

Another direction worth exploring is the extension of the network to capture the dynamic changes in face expression and to evaluate whether the dynamic properties can contribute to the recognition task. Finally, other coding schemes could be explored and compared on the same visual pattern recognition task.

Chapter 4 - Evolving SNNs for auditory pattern recognition

This chapter begins with a review of several models of auditory information processing, in particular models that are inspired by the brain. The review is followed by a description of new systems that perform speaker authentication using spiking neural network architectures. The new systems are brain-inspired in terms of the information pathways, the information processing units (neuronal level), and the learning procedures.

The review focuses on auditory system models capable of achieving accurate pattern recognition when processing speech signals in both speech and speaker recognition problems. It can be seen from the review that among the most accurate models, only a few currently accomplish auditory information processing using techniques that resemble the brain (models that use brain-inspired processing units are particularly rare). Thus, the remainder of the chapter addresses the modelling of the auditory system on the speaker authentication problem. Here, the motivation is the same as for the research on visual pattern recognition described in Chapter 3, which is "*to closely replicate a sensory modality to the brain's way of processing to ultimately achieve a gain in performance*". As properly described in (Ghitza, 1994), models of the auditory system can assist in building computer models with similar performance to the human brain. Ghitza further suggests that the advantages these models can bring are only limited by how reliably the auditory system can be simulated.

The information processing of auditory signals is modelled using a new multi-layered spiking neural network architecture. Each speaker is represented by a set of prototypes that are trained with standard Hebbian rule and *winner-takes-all* approach. Incoming speech signals arrive at the network in the form of Mel Frequency Cepstrum Coefficients (MFCC) (Rabiner and Juang, 1993) in a frame-based representation where each frame contains a short time period of the speech signal. For every speaker a separate spiking network computes normalized similarity scores based on MFCC considering speaker and background models.

Experiments carried out with the VidTimit dataset (Sanderson and Paliwal, 2004) show similar performance of the system when compared with a benchmark method based on vector quantization. A procedure to create/merge neurons similar to the training procedure proposed for the visual model is also presented, which enables adaptive and online training in an evolving way.

A new speech signal feature extraction design is proposed using a multi-layer feed-forward network of spiking neurons, which uses wavelet-based filter banks. With this design, all the information processing stages, from pre-processing of the speech signal to higher levels of cognition (where the recognition is done), are accomplished with spiking neurons.

Finally, the way to represent low-dimensional data is also discussed in respect to the sparseness needed when processing information with biologically inspired system (Baddeley, 1996) (Baum *et al*, 1988) (Foldiak, 1995) (Olshausen and Field, 1997) (Perrinet and Samuelides, 2002). Population encoding of features is tested as an alternative on a two-dimensional dataset.

4.1 Background and benchmarking

Modelling of the auditory system can be developed for several stages of processing, from ear to the cortex. For instance, models can help decipher how signals are processed in the cochlea (in terms of encoding and information flow), can be applied to speech analysis, or even attempt to simulate higher levels of cognition (Holmes and Holmes, 2001).

Sound signals are pre-processed in the external and middle ear before reaching the inner ear where transduction takes place. More precisely, the transduction occurs in the cochlea. Several models describe this mechanism. In the work of Robert and Eriksson (Robert and Eriksson, 1999) a phenomenological model of the cochlea is presented, which is composed of a bank of filters (non-linear, time-varying and with active feedback). The main goal was to have a model with responsiveness to sound stimuli as similar as possible to that which has been measured in the cochlea. The model does not particularly attempt to provide an anatomical or physiological explanation. The output of the model represents the activity (in terms of spikes) of the inner hair cells.

Previous work in the same area include: a) a linear method (Jenison, 1991); and b) a nonlinear method (Patterson *et al*, 1995).

deCharms *et al* (deCharms *et al*, 1998), through an analogy to the primary visual system that decomposes visual excitation into basic features (edges, orientation, colour, movement), suggests that the primary auditory system can be modelled with the decomposition of auditory scenes into basic components. The list of these basic auditory components include stimulus edges, stimulus transition (both in frequency and time), and conjunction of features.

Lewicki (Lewicki, 2002) uses information theory to maximize the information given by auditory stimuli, i.e., to search for the most efficient coding strategy. This uses independent component analysis to extract efficient coding for three classes of sound: natural and environmental sounds and speech. To model the auditory system, 128 filters are used, where the parameters of the filters need to be set in order to maximize the information transmission. From this analysis, the form of coding depends on the sound class. When optimized for natural sounds, a Fourier-type of transformation is achieved. Environmental sounds resulted in wavelet-like filters whereas with a combined set, where speech has been included, the optimal coding resembles the characteristics extracted in biological measurements.

As summarized in (Holmes and Holmes, 2001), the modelling of the outer and middle ear is easy to simulate with electrical filters. The response of the cochlea (basilar membrane), despite being more complicated, is also well understood and models can reproduce measurements with high reliability. However, it is in neural transduction, i.e., in the transcription of the movements of the basilar membrane and the firing patterns, where the process remains largely unknown. Overall, the transduction process undergoes three steps (see (Holmes and Holmes, 2001) for more detail):

a) Rectification of the signal. The signal that arrives in the cochlea is filtered by the outer and middle ear. The inner hair cells in the basilar membrane are only stimulated by a filtered signal in one direction. This mechanism is modelled through half-wave rectification of the signal.

b) Compression. After half-wave rectification, the signal is compressed to reduce the dynamic range of the input signal.

c) Firing of the cells. The compressed signal is then directly related to the probability of the cells firing. Firing probability is also shown to be correlated to the timing of the previous spike (i.e., the closer the previous spike the lower the probability of a new spike occurring).

Tuning several variables presented in these three steps, reproducing firing properties after the transduction process under different types of auditory excitation was found to be possible. However, the main coding schemes that encode sound signal information into spikes are still unknown.

Further, at higher levels of the auditory information pathway, there are three traditional ways of assessing the spike trains generated by auditory stimuli:

a) place/rate: the information is retrieved using firing rates. In this case, the spectral characteristics of the signals are detected through the spiking rates of the cells sensitive to different frequencies, which are placed in different locations in tonotopic organization (e.g. (Holmberg *et al*, 2005)).

b) place/temporal: The tonotopic organization of the neurons defines different frequencies, which have their amplitudes measured by temporal representation of spiking signals (spiking time). Seneff (Seneff, 1988) describes a generalized synchrony detector, which compares the timing between spikes of a neuron under excitation with its corresponding preferred firing time.

c) non-place/temporal: The tonotopic organization of the cells is not considered and the information is retrieved based on the overall spiking responses. An example of the non-place/temporal model is the ensemble interval histogram (EIH) proposed in (Ghitza, 1988) (Ghitza, 1992).

When specifically considering speech signals, the auditory system and the intrinsic characteristic of the inner hair cells are traditionally modelled with filter banks (channels) in the frequency domain. To extract spectral characteristics in time, Short Term Fourier Transform is commonly applied to a short segment of the signal. The filter banks are set to represent the response to the spectrum range of the human ear. Experiments have detected that the human ear can perceive differences more accurately at low than at higher frequencies. This property was the origin of the MEL scale (Gold and Morgan, 2000), which is now commonly used in the form described in (Davis and Mermelstein, 1980) (See Figure 4.5). Equally common is to perform a further process called “cepstrum”, which mainly aims to separate the vocal-tract filtering from excitation (Holmes and Holmes, 2001). Effectively this process is done using Discrete Cosine Transform (DCT).

Another method used on short segments of speech signals after extracting spectral characteristics, is Linear Prediction Coding (LPC) (Atal and Hanauer, 1971), which comes together with its “cepstrum” counterpart (LPCC) (Atal, 1974). The rationale behind LPC is first, to perform a linear prediction of future samples based on previous data and then to compare the actual and predicted curves. The coefficients of the linear prediction model can be obtained by minimizing the difference between predicted and real samples.

More recent works use wavelet analysis for speech feature extraction in several different ways. (Tufekci and Gowdy, 2000) use wavelets instead of DCT to calculate the “cepstrum”, (Long and Datta, 1996) use high energy wavelet outputs as features, and (Sarikaya and Hansen, 2000) (Tufekci *et al*, 2006) calculate the energy over a spectrum resembling the Mel filter bank. Wavelets are discussed further in Section 4.5.

In respect of models which simulate higher levels of auditory signal cognition, Allen (Allen, 1994) recognizes the lack of a conclusive answer to the question “how do humans process and recognize speech”. However, according to Allen, the primary conceptual idea proposed by Fletcher (Fletcher, 1922), which describes the auditory system through a sequence of layers that recognizes in order: features, phones and phonemes, syllables, words, sentences, and meaning, still holds.

4.2 SNNs for speech processing

Robert and Eriksson (Robert and Eriksson, 1999) proposed a model of the auditory periphery to simulate the response to complex sounds. The model basically reproduces the filtering executed by the outer/middle ear, basilar membrane, inner hair cells, and auditory nerve fibers. The purpose of the model is to facilitate the understanding of signal coding within the cochlea and in the auditory nerve as well as analyse sound signals. The output of the inner hair cells and auditory nerve fibers are properly represented with trains of spikes. This model has been used in (Eriksson and Villa, 2006) to simulate the learning of synthetic vowels by rats reported in (Eriksson and

Villa, 2006a). In this latter work, based on experimental measurements, besides proving that rats are able to discriminate and generalize instances of the same vowel, it is further suggested that, similar to humans, rats use spectral and temporal cues for sound recognition.

An SNN model has been applied in sound localization (Kuroyanagi and Iwata, 1994) and in sound source separation and source recognition in (Iwasa *et al*, 2007). In (McLennan and Hockema, 2001) a simple SNN structure is proposed to extract the fundamental frequency of a speech signal online. The highlight of the latter system is that a Hebbian learning rule dynamically adjusts the behaviour of the network based on the input signal.

In (Holmberg *et al*, 2005) the importance of temporal and spectral characteristics of sound signals is described. The spectral properties are inherently represented with "rate-place code" during the transduction of the inner hair cells. Temporal information, on the other hand, provides additional cues, such as amplitude modulation and onset time. In the same work a multi-layer auditory model is presented, which emulates inner ear filtering, compression and transduction. The work mainly concentrates on using spiking neurons to model octopus neurons, which are neurons located at the cochlear nucleus. Octopus neurons enhance the amplitude modulations of speech signals and are sensitive to signal onsets. Preliminary experiments showed that the system performs in much the same way as Mel Frequency Cepstral Coefficients (MFCC) (Rabiner and Juang, 1993).

Rouat *et al* (Rouat *et al*, 2005) envisage the advantages of merging perceptual speech characteristics and biologically realistic neural networks. After a description of the perceptual properties of the auditory system and non-linear processing realized by spiking neural networks, a biologically inspired system to perform source separation on auditory signals is proposed. In the same work and in (Loiselle *et al*, 2005), a preliminary evaluation used SNN for recognition of spoken numbers.

Mercier and Seguier (Mercier and Seguier, 2002) proposed the use of the Spatio-Temporal Artificial Neural Network model (STANN) based on spiking neurons on the speech recognition problem (recognition of digits on the Tulips1 dataset (Movellan, 1995). STANNs were initially proposed to process visual information (Seguier and Mercier, 2001).

In summary, there are several models which describe the auditory information pathway, from the external, through to the inner ear. In addition, many methods describe signal transduction from waves to spikes. For transduction, most methods opt to encode the signal using "rate-place" representation and to process it with well established traditional models thereafter. There are fewer methods that further process the signal at the action potential level for speech recognition (e.g., (Mercier and Seguier, 2002) (Rouat *et al*, 2005)). In particular, this researcher is not aware of systems that use SNN to deal specifically with the speaker authentication problem. The next section explores in detail the use of SNNs to perform cognition at a phonetic level to extract properties that enable speaker authentication.

4.3 A new SNN-based method for text-independent speaker authentication

Computer-based speaker authentication presents a number of possible scenarios. Text-dependent, text-independent, long sentences, single words, speaker willing to be recognized, speaker trying to hide their identity are some examples. For each of these scenarios, different and specifically tuned processing techniques seem to be the most effective. Here, the focus is on the short-sentence text-independent problem, which is typically comprised of input utterances ranging from 3 seconds to 1 minute. In this scenario, a speaker being authenticated does not necessarily need to present the same word or sentence used during training. Moreover, due to the short length of the signal, it is not possible to acquire long-term dependencies of features that could eventually supply additional information that would enhance performance. Thus, state machines to detect phonemes, words, and bigrams cannot be setup at full strength.

Based on these properties, in recent years a convergence in the use of, first, Vector Quantization (VQ) (Burileanu *et al*, 2002) (Gray, 1984) and, later Gaussian Mixture Models (GMM) (Bimbot *et al*, 2004) (Reynolds *et al*, 2000) to tackle the text-independent speaker authentication problem can be seen. These methods are used as inspiration in the design of a new spike-based system. VQ is used as a benchmark for comparison purposes as well.

Two distinct network architectures that perform classification tasks using spiking neurons are presented. The highlight of the new architectures is the inclusion of two techniques that have already demonstrated themselves to be efficient in traditional methods (Gray, 1984) (Reynolds *et al*, 2000). They are:

- creation of prototype vectors through unsupervised clustering, and
- adaptive similarity score (*similarity normalization*).

The next section gives a general overview of the new speaker authentication system and the speech signal pre-processing stages. Section 4.3.2 presents the SNN models and Section 4.3.3 is devoted to experimental results.

4.3.1 Speech pre-processing

Voice activity detection

A simple algorithm based on the energy of a signal after being submitted to a low-pass Butterworth Filter as described in (Coleman, 2005) is used for voice activity detection (VAD). The filter has a cut-off frequency of 400 Hz, below which the RMS (root mean square) energy is calculated. A sample of the input signal is considered voice if the RMS energy of the signal at frequencies below 400 Hz is higher than a certain threshold. Otherwise, the sample is considered noise.

Obviously, this simple approach does not work in the presence of low-frequency environmental noise. In such conditions, low-frequency environmental noises are also classified as voice. However, for the dataset used in the following experiments, this approach gave a reliable performance. Figure 4.1 shows an example of the output of the VAD filter applied to a speech signal from the VidTimit dataset (Sanderson and Paliwal, 2004).

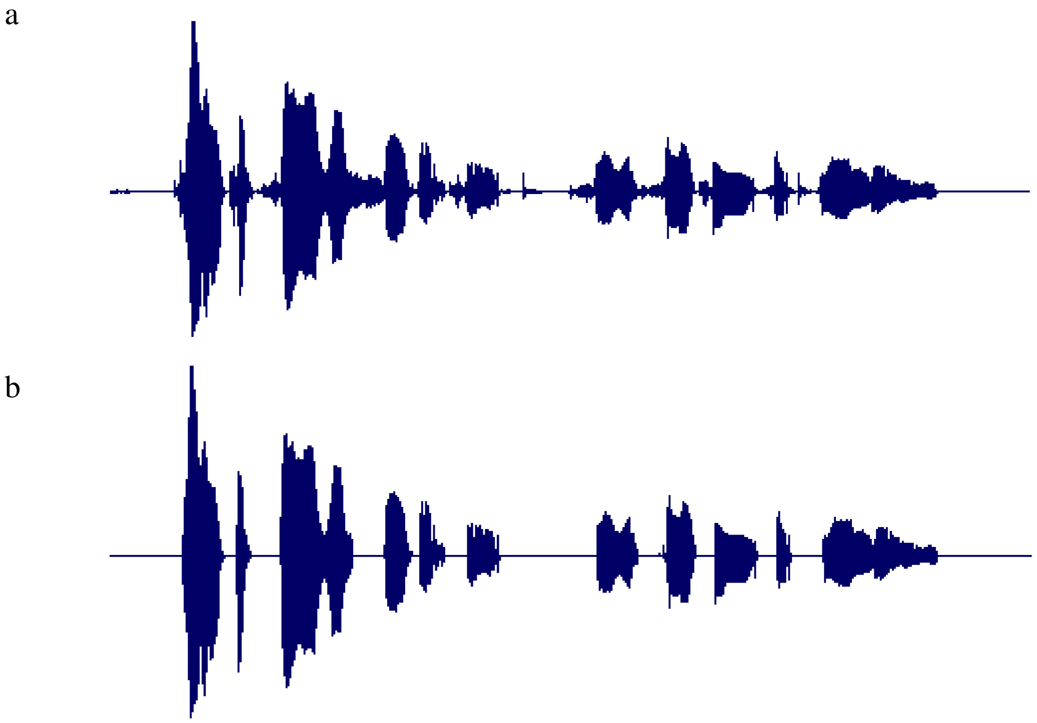


Fig. 4.1. Voice activity detection. a) Raw speech signal; b) Speech signal after VAD.

Feature extraction

Some of the most popular and efficient feature extraction methods for speaker authentication include Mel Scale Coefficients (MSC), Linear Predictive Coding, and Mel Frequency Cepstral Coefficients (MFCC) (Rabiner and Juang, 1993). MFCC is used in these experiments, mainly because it has a biological interpretation, detecting spectral characteristics in a manner similar to the human ear, and at the same time proving to be very efficient in speech recognition tasks. The computation of the Mel Frequency Cepstral Coefficients is described in detail below.

The entire feature extraction process is composed of five steps (see Figure 4.2).

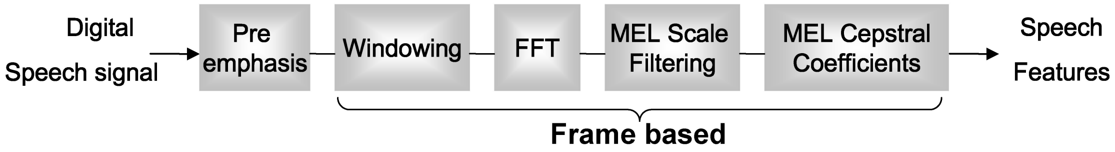


Fig. 4.2. Feature extraction process. Pre-emphasis filter is applied to the speech signal. The signal is sliced into frames and convolved with a hamming window (windowing). Spectral characteristics are calculated in a frame-based manner using Fast Fourier Transform (FFT) and filtered with Mel Scale filter banks. Cepstral coefficients are then calculated using Discrete Cosine Transform (DCT).

First, a pre-emphasis filter is applied to the speech signal

$$y(t) = x(t) - \alpha x(t-1) \qquad \textbf{(4.1)}$$

where α is a number between zero and one, typically 0.97 to emphasize the high frequency components of the signal spectrum, *t* is time, *x*(*t*) is the input signal and *y*(*t*) is the filtered output signal.

In the next step, windowing, the signal is divided into small segments, called frames with 50% overlap (See Figure 4.3). In each window, the signal is assumed stationary for the purpose of spectral analysis. The frame signal is then multiplied by a hamming window function (Figure 4.4), which has the purpose of increasing the accuracy of the computation of spectral characteristics. The equation for the hamming window function is given as

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \qquad \textbf{(4.2)}$$

where *N* is the length of a frame and *n* is an index varying from 0 to *N*.

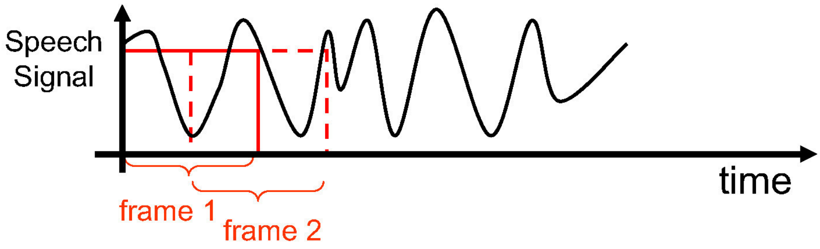


Fig. 4.3. Windowing. Frames with 50% overlap.

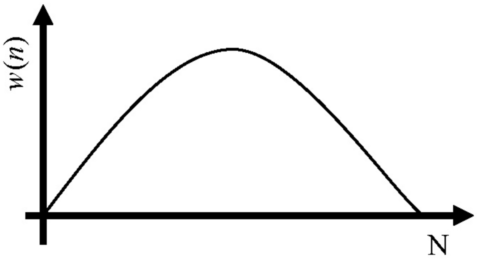


Fig. 4.4. Hamming window (according to Equation 4.2).

Fast Fourier Transform (FFT) converts the signals of each frame from the time to the frequency domain (*X*). The power spectrum is then computed from the complex result of the FFT as

$$P(k) = \mathrm{Re}(X(k))^2 + \mathrm{Im}(X(k))^2 \qquad \textbf{(4.3)}$$

where Re(X) and Im(X) are respectively the real and imaginary parts of the signal X in the frequency domain and k is the index of the signal (harmonic).

The frequency f_k, corresponding to each harmonic in the frequency domain can be calculated as

$$f_k = \frac{kf_s}{N} \qquad \textbf{(4.4)}$$

where k is the harmonic, f_s is the sampling frequency and N is the total number of point signals used on the FFT conversion (length of a frame).

The next step consists of filtering the spectrum with Mel Filter Banks, which consists of applying different band-pass filters to the signal to isolate different sub-bands. Table 4.1 presents each sub-band with its respective centre and cut-off frequencies. For simplicity, a filter is used, which follows a triangular waveform (See Figure 4.5). The Mel filter bank presented in Table 4.1 has frequency components up to 6800 Hz. This is due to the intrinsic characteristics of a speech signal, where the dominant frequencies are confined to this upper frequency limit.

Table 4.1. MEL Filters with overlapping bands. Frequency up to 6800 Hz.

Band	Lower Edge	Centre	Upper Edge
1	0	50	100
2	50	150	200
3	150	250	350
4	250	350	450
5	350	450	550
6	450	570	690
7	570	700	830
8	700	840	980
9	840	1000	1160
10	1000	1170	1340
11	1170	1370	1570
12	1370	1600	1830
13	1600	1850	2100
14	1850	2150	2450
15	2150	2500	2850
16	2500	2900	3300
17	2900	3400	3900
18	3400	4000	4600
19	4000	4800	5600
20	4800	5800	6800

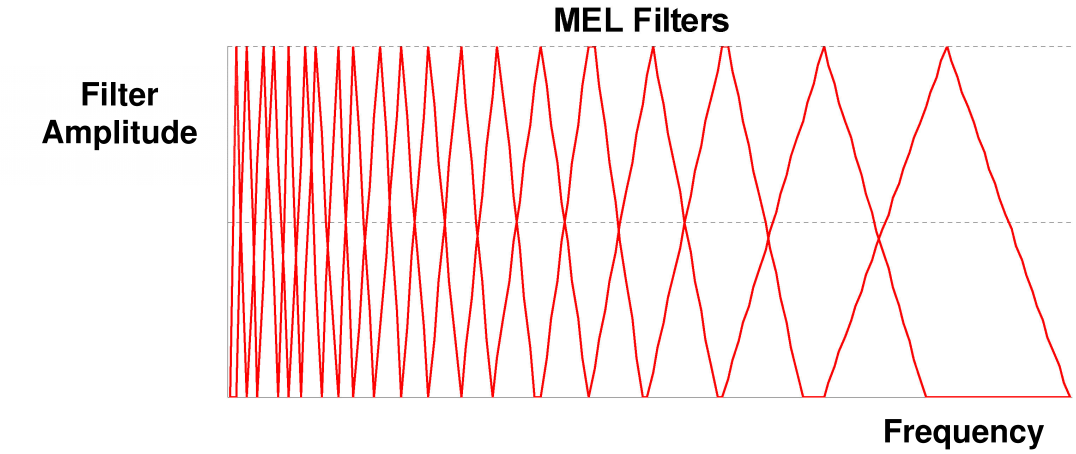


Fig. 4.5. MEL Filter Bank. The abscissa represents frequency and the ordinate filters' amplitude.

After applying Mel filter banks, Mel Scale Coefficients (MSC) are calculated. MSC corresponds to the energy of the signal in each of the Mel frequency bands as

$$C(m) = \sum_{k=0}^{k=N/2} P(k)Mel_m(k) \tag{4.5}$$

where m is the index of the filter bank, taking values between 0 and (number of filter banks – 1), $P(k)$ is the power spectrum at the harmonic k, and $Mel(k)$ is the amplitude of the corresponding filter m at the harmonic k.

At this point, there is one value for each MEL filter in each frame of the speech signal. Discrete Cosine Transform (DCT) is then applied to the natural logarithm of the Mel Scale Coefficients, producing the Mel Frequency Cepstrum Coefficients. This process is calculated with

$$S(n) = \sum_{m=0}^{m=n\_banks-1} \ln(C(m))\cos(\frac{\pi n}{2n\_banks}(2m+1)) \tag{4.6}$$

where n represents the DCT coefficient, $C(m)$ is the Mel Scale Coefficients and m is the index of the filter bank. The first coefficient is excluded because it denotes the energy of the signal. Thus, there are 19 coefficients in each frame. Published experimental results show that the relevant discriminatory information of the speech signal is preserved with MFCC (Ghobakhlou *et al*, 2003).

The first and second derivatives (∇ and ∇^2) of MFCC are also features commonly used on speaker identification, aiming mainly to provide information about the variation of the signal in time (Becchetti and Ricotti, 1999) (Coleman, 2005). They are calculated as:

$$\begin{aligned} &\nabla^{i}\{MFCC_{t}\} = \nabla^{i-1}\{MFCC_{t+1}\} - \nabla^{i-1}\{MFCC_{t-1}\} \\ &\nabla^{0}\{MFCC_{t}\} = MFCC_{t} \end{aligned} \qquad \textbf{(4.7)}$$

where *i* is the order of the derivative and *t* is the time index of an MFCC vector (frame).

Table 4.2 shows the list of parameters used on speech signals at the pre-processing stage and Figure 4.6 illustrates the frame-by-frame extraction of MFCC on a speech signal.

Table 4.2. Overall properties of the speech signal at the pre-processing stage

Property	Value
Speech sampling	16000 Hz
Pre-emphasis constant α	0.97
Hamming window size	512
Speech window	32 msec
Number of MEL filters	20
DCT coefficients	19 (except fundamental)
Frequency range	up to 6800 Hz
Window shift	16 msec (1/2 window overlap)
Total Number of Features	19 DCT coefficients + 19 ∇DCT = 38

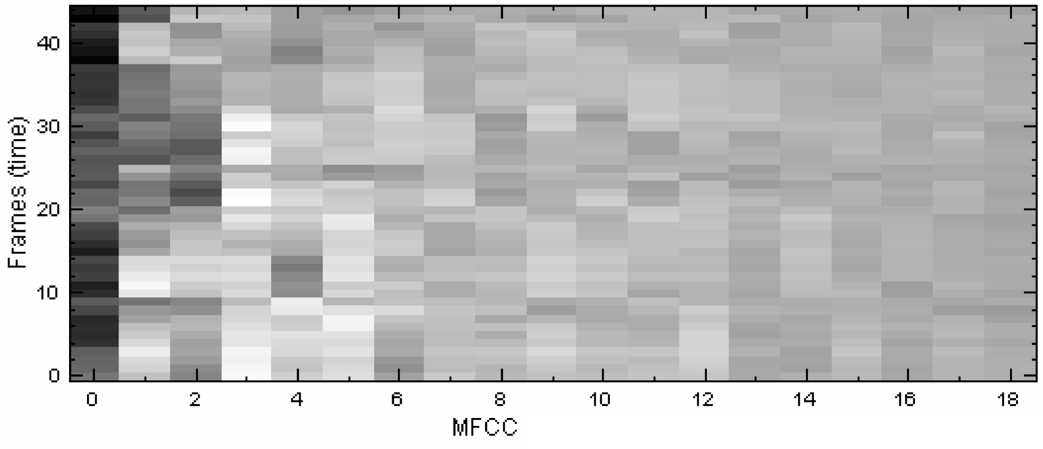


Fig. 4.6. Frame-by-frame MFCC of a given speech signal. The grey levels denote MFCC levels ($S(n)$ according to Equation 4.6).

Spike encoding

Each frame of the signal containing speech fragments generates an MFCC vector (see Figure 4.6) that is translated into spikes using Rank Order Coding (Delorme *et al*, 1999) (Figure 4.7). In the experiments presented in this chapter, one input neuron represents one MFCC vector. The encoding of the features onto a population of neurons is further discussed in a later section, which seems to provide a more biologically plausible, sparse representation (see Section 4.4).

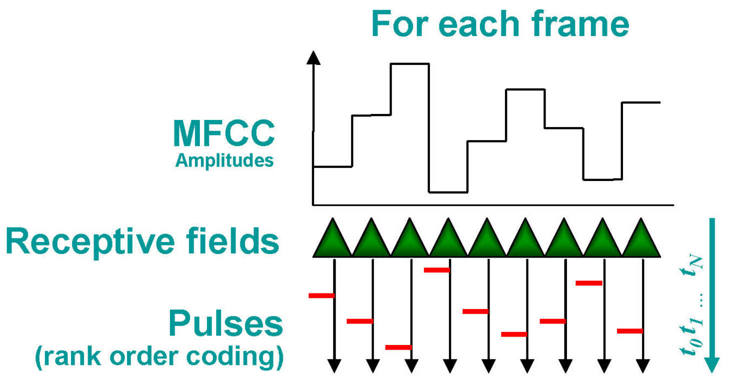


Fig. 4.7. MFCC encoded as spiking time with Rank Order Coding (Delorme *et al*, 1999). The higher the amplitude the shorter is the spike delay.

Normalizations

Speaker authentication is well-known as having a high variation between training and test conditions. In order to attenuate this problem, several techniques have been used. The majority of recent attempts usually normalize the features and/or the method of computing similarity (Burileanu *et al*, 2002) (Reynolds *et al*, 2000). For the normalization of features (*parameter domain*), cepstral mean subtraction of the MFCC levels is used in this implementation (Burileanu *et al*, 2002).

In the *similarity domain*, the spiking neural network model has a similarity normalization technique embedded, in which the authentication score is calculated not only based on the similarity between a test sample and the speaker model, but on the relative similarity between the test sample and the speaker model and between the test sample and a background model. Figure 4.8 illustrates the normalization in similarity domain. With this procedure, the variations between train and test conditions are taken into account when computing similarity. Normalization in the similarity domain has already being extensively implemented in traditional methods of speaker verification and is currently found in most of state-of-the-art speaker authentication methods (Bimbot *et al*, 2004). In the new SNN-based implementation described in this dissertation, normalized similarity is computed allocating excitatory connections to neurons representing the claimant model and inhibitory connections to neurons representing the background model.

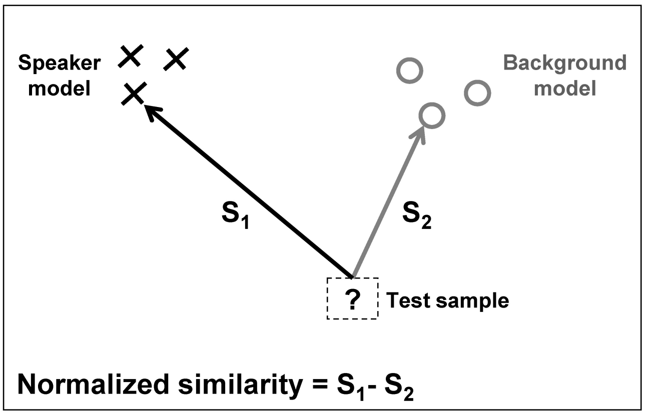


Fig. 4.8. Normalization in the similarity domain in a hypothetical two-dimensional space.

4.3.2 Evolving spiking neural network models

The design of the speaker authentication uses two/three layers feed-forward networks of integrate-and-fire neurons where each speaker has their own network. Each layer is composed of integrate-and-fire neurons with a modulation factor, as described in Section 3.3.1, that lends greater importance to the earliest spikes.

4.3.3 Architecture 1 - Integration of binary opinions

Figure 4.9 illustrates Architecture 1. Receptive field neurons encode each feature of a frame, typically MFCC, to the time domain using Rank Order Coding (Delorme *et al*, 1999) (See Figure 4.7). One neuron encodes each coefficient. The output of the receptive field neurons is a spike time pattern for every frame. Layer 1 (L1) is composed of two neuronal maps. One neuronal map has an ensemble of neurons representing a speaker model (speaker prototypes). Each neuron in the neuronal map is to be trained to respond optimally to different segments of the training utterances, i.e., different speech phones (minimal unit of speech segmentation). The second neuronal map in L1 is trained to represent the background model. Several ways to represent background models, that can be universal or unique for each speaker, are described and analysed in (Bimbot *et al*, 2004).

Similar to L1, L2 has two neuronal maps representing the speaker and the background model. Each L2 neuronal map is composed of a single neuron. L1 and L2 are connected to each other as follows:

a) excitatory connections between neurons corresponding to neuronal maps with the same label, i.e., L1 speaker to L2 speaker and L1 background to L2 background, and;

b) inhibitory connections between neurons with differing neuronal map labels, i.e., L1 speaker to L2 background and L1 background to L2 speaker. Effectively, L2 neurons accumulate opinions of each frame of being/not being a speaker and being/not being the background.

The dynamic behaviour of the network is described as:

a) For each frame of a speech signal, features are generated (MFCC) and encoded into spiking times using receptive field neurons.

b) The spikes are then propagated to L1 until an L1 neuron emits the first output spike, which is propagated to L2. If a neuron in L2 generates an output spike, the simulation is terminated. If not, the next frame is propagated.

c) Before processing the next frame, L1 PSPs are reset to the rest potential whereas L2 neurons retain their PSPs, which are accumulated over consecutive frames, until an L2 output spike is generated.

The classification is completed when a neuron in L2 generates an output spike or all frames and all spikes in the network have been propagated. If the L2 neuron representing the speaker releases an output spike, the speaker is authenticated. The authentication fails in a case where no spikes occur in L2 after all frames have been processed or an L2 neuron representing background releases an output spike.

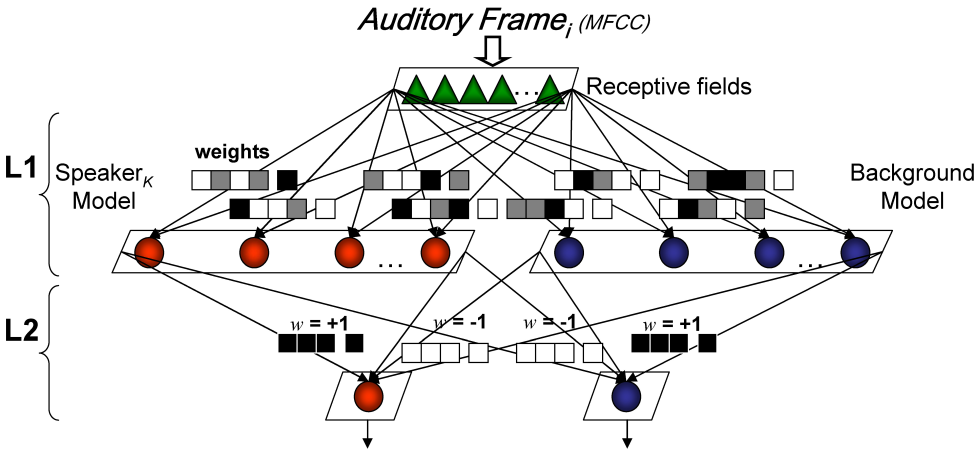


Fig. 4.9. Evolving SNN Architecture 1. Frame-by-frame integration/accumulation of binary opinions (Wysoski *et al*, 2007a).

Note that, in the architecture described, L2 neurons accumulate opinions of being/not being a given speaker over several frames. The propagation of each frame into the network leads to a binary opinion (yes/no), based on two criteria:

a) high similarity of the input frame to a certain prototype represented by an L1 neuron, in such a way that the similarity causes an L1 neuron to fire;

b) competition between a speaker and a background model, i.e., the frame needs to be more similar to a speaker prototype than to a prototype representing the background in order to fire earlier.

The latter effectively implements *similarity domain* normalization and enables the network to adapt to variations inherently present in the speaker authentication problem (Bimbot *et al*, 2004).

However, the output after each frame is propagated does not give a notion of how similar the input frame is to a previously trained prototype. In general, traditional methods that apply *similarity domain* normalization, compute the relative distance between the closest prototype of a speaker model and the closest prototype of the background model. To overcome this constraint, i.e., to extract the normalized similarity scores of each frame, a second network architecture is proposed in next section (Architecture 2).

4.3.4 Architecture 2 - Integration of similarity scores

In this more complex configuration, the network is composed of three layers as illustrated in Figure 4.10. Similar to the previous architecture, the encoding of features from each frame into precise spike times is carried out by receptive field neurons. Layer 1 (L1) has two neuronal maps (speaker and background model) where each neuron is trained to respond optimally to a certain input excitation. The neurons in L1 are set to detect the closest prototype in both speaker and background model (the learning procedure to adjust L1 weights is described in the following section). Only one neuron in each L1 map is allowed to spike.

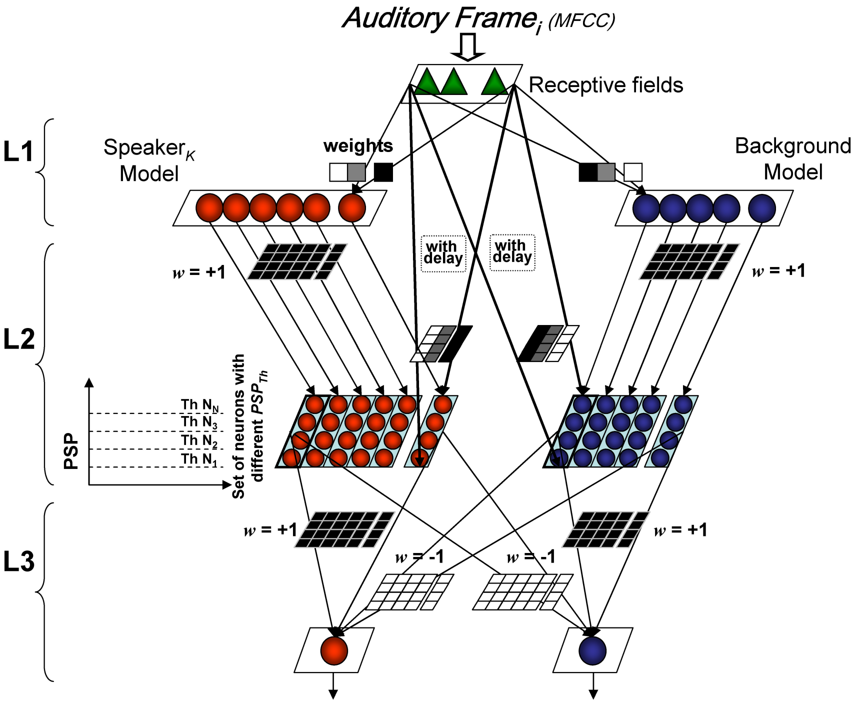


Fig. 4.10. Evolving SNN Architecture 2. Frame-by-frame integration/accumulation of similarity scores (Wysoski *et al*, 2007a).

Each L1 neuron is connected to a set of layer 2 (L2) neurons. The set of L2 neurons are connected to the receptive field neurons with the same connection weights as the corresponding L1 neuron, however, they receive the spike train after a certain delay. The delay is set in such a way that L1 output spikes arrive at L2 before the arrival of incoming spikes from the receptive field neurons. L1 output spikes are effectively used to raise the PSP of all neurons in the set to a level where spikes can occur. Thus, in L2, only the neurons belonging to the winner set (closest prototype) become active and can generate output spikes with the arrival of spikes from the receptive fields. The main characteristic of each set of L2 neurons related to an L1 neuron is that each neuron has the same incoming weight connection from the receptive field neurons, but different PSP thresholds. Therefore, the neurons in a L2 set generate output spikes at different levels of PSP. Upon the arrival of input trains of spikes on L2 neurons, several neurons from the winner set are expected to fire. The neurons with the lowest PSP_{Th} fire first, followed by neurons with higher PSP_{Th} levels, and so on.

Layer 3 (L3) integrates L2 spikes with excitatory connections between neuronal maps with the same labels and inhibitory connections between neuronal maps with differing labels. With this mechanism, PSP levels in L3 denote normalized similarity between the most similar speaker prototype and the most similar background prototype.

Similar to the behaviour of the previous network, PSPs on L1 and L2 are reset after every frame whereas on L3 the PSPs are accumulated over several frames. The simulation is terminated when L3 emits an output spike or there are no more frames to be processed. Each frame is processed until all the spikes are propagated or until all L2 neurons representing the speaker or background emit output spikes.

4.3.5 Learning procedure

Training is done in the synapses connecting the receptive field and L1 neurons in a similar fashion for both network architectures. To update weights during training, the simple rule used in the visual system model (See section 3.4.2) is applied:

$$\Delta w_{j,i} = \text{mod}^{order(j)} \tag{4.8}$$

where $w_{j,i}$ is the weight between the receptive field neuron *j* and neuron *i* of the L1, mod $\in$ (0,1) is the modulation factor, *order*(*j*) is the order of arrival to neuron *i* of a spike produced by neuron *j*. For each training sample, the *winner-takes-all* approach is used, in such a way that only the neuron with the highest *PSP* value in L1 has its weights updated.

The postsynaptic threshold (PSP_{Th}) of a neuron is calculated as a proportion $c \in$ [0, 1] of the maximum postsynaptic potential (*PSP*) generated with the propagation of the training sample into the updated weights, such that:

$$PSP_{Th} = c \max(PSP) \tag{4.9}$$

The adaptive online procedure for training the network and creating new neurons is adapted from the visual pattern recognition model presented in Chapter 3 and is summarised with the following pseudo-code:

```
For all phrase samples in the training set
  For each frame
    Create a new neuron
    Propagate the frame into the network
    Train the newly created neuron using Equation 4.8 and 4.9
    Calculate the similarity between weight vectors of newly
    created neuron and existent neurons within the neuronal map
    If similarity > Threshold
       Merge newly created neuron with the most similar neuron
       using Equation 4.10 and 4.11
```

To merge a newly created neuron with an existing neuron, the weights *W* of the existing neuron *n* are updated calculating the average as

$$W = \frac{W_{new} + N_{Frames} W}{1 + N_{Frames}} \tag{4.10}$$

where N_{Frames} is the number of frames previously used to update the neuron in question. Similarly, the average is also computed to update the corresponding PSP_{Th}:

$$PSP_{Th} = \frac{PSP_{Thnew} + N_{Frames} PSP_{Th}}{1 + N_{Frames}} \tag{4.11}$$

Alternatively, the network structure and the number of desired prototypes (neurons) can be defined *a priori*, using a *k-means*-like clustering algorithm to update the weights of the winner neuron (for more information about a basic version of *k-means* algorithm see, for instance, (Deller Jr. *et al*, 2000)). In this case, a simple iterative heuristic can be described in two steps:

1. Initialization of the neurons' weights

```
 For each neuron
   Propagate a random frame of the training set into the network
   Update the neuron's weights using Equation 4.8 and 4.9
```

2. Recursive training

```
Until weights converge
  For all phrase samples in the training set
    For each frame
      Propagate each frame into the network
      Find the maximally activated neuron (the neuron with maximum
      PSP)
      Create a new neuron and train it using Equation 4.8 and 4.9
      Update weights of the maximally activated neuron merging it to
      the new neuron (using Equation 4.10 and 4.11)
```

The latter method is used in the experiments to be described in the following sections, attempting to reproduce as closely as possible the scenario of the benchmarking algorithm (VQ with *k-means* clustering).

In SNN Architecture 1 (See Figure 4.9), L1 neurons are fully connected to neurons in L2. The weights are set in order to accumulate positive or negative opinions of each input frame for each speaker (W = 1 for the links between each L1 neuronal map and its corresponding L2 neuron. W = -1 when the label of the L1 neuronal map differs from the label of the L2 neuron.

In SNN Architecture 2 (See Figure 4.10), the connections between an L1 neuron and the corresponding set of neurons in L2 are excitatory (W = 1). Neurons in L2 are fully connected to L3 neurons. There are excitatory connections (W = 1) between neurons belonging to the neuronal maps with the same label, otherwise the connections are inhibitory (W = -1).

4.3.6 Experiments and results

The spiking network models proposed in the previous sections were implemented and the speech part of the VidTimit dataset (Sanderson and Paliwal, 2004) was used for performance evaluation. VidTimit contains 10 utterances from 43 different speakers. In order to make a comparison with the experiments described in (Sanderson and Paliwal, 2004), the system was set to authenticate 35 individuals, each individual trained with 6 utterances. The remaining 4 utterances of each individual was used as a test. In addition, 4 utterances of the 8 remaining individuals were used to simulate impostor access. Thus, the number of true claims for each individual model is 4 (each utterance is taken individually), and the number of impostors that try to break into each model is (35 - 1 remaining user x 4 utterances) + (8 impostors x 4 utterances), which gives a total of 168 impostors. For all individual models of the entire dataset, there are (35 users x 4 utterances), totalling 140 true claimants and (35 users x 168 utterances) = 5880 impostors.

The speech signals are sampled at 16 kHz, and features are extracted using standard MFCC with 19 MEL filter sub-bands ranging from 200 Hz to 7 kHz. MFCC is then encoded into spikes spread across 19 receptive field neurons. See Section 4.3.1 for details of the speech signal pre-processing steps.

A specific background model for each speaker is trained. For the sake of simplicity, the background model of a speaker *i* is trained using the same number of utterances used to train its corresponding speaker model (6 utterances), with the utterances randomly chosen from the remaining individuals in the dataset.

For comparison purposes, a standard vector quantization (VQ) algorithm (Burileanu *et al*, 2002) with *k-means* clustering was used. Training was done with the same features (19 MFCCs) and the same strategy for selecting background models was applied. The performance for different numbers of prototypes was tested. Figure 4.11 reports the best performance of the VQ algorithm obtained with 32 prototypes for speakers and 32 prototypes for the background model. These results are comparable with the work presented by (Sanderson and Paliwal, 2004), where, with the same dataset, the authors reported total error TE = false acceptance rate (FAR) + false rejection rate (FRR) = 22 % in slightly different setup conditions using Gaussian Mixture Model. The VQ implementation presented here obtained TE = 25 %.

In respect to the SNN implementation, the number of neurons in the L1 neuronal maps for the speaker and background models (80 neurons each) was defined *a priori*.

The modulation factor (mod) was set to 0.9 for L1 neurons in Architecture 1 and L1 and L2 neurons in Architecture 2. The other layers are composed of neurons with mod = 1.

In the experiments with SNN Architecture 1 (Figure 4.9), PSP_{Th} of L2 neurons were defined as a proportion p of the number of frames used for identification. For instance, if an utterance used for authentication is composed of 40 frames and p is 0.2, the PSP_{Th} used for authentication is 40 x 0.2 = 8. The PSP_{Th} of L1 neurons were calculated as a proportion c of the maximum PSP obtained during training according to Equation 4.9. The performance for p = 0.2 and the different values of c are shown in Figure 4.12 (top). The minimum TE reached was 31.1%.

In SNN Architecture 2 (Figure 4.10), the PSP_{Th} of L3 neurons were defined as a proportion p (0.2 was used) of the number of frames used for identification. PSP_{Th} of L1 neurons were calculated as a proportion c of the maximum PSP obtained during training according to Equation 4.9. A set of L2 neurons had PSP_{Th} levels ranging from 0 to the maximum PSP_{Th} of their corresponding L1 neuron (equally spaced). Figure 4.12 (bottom) shows a typical performance using 20 PSP_{Th} levels for different c. The minimum TE reached was 36.0 %. Note that, in this scenario, for values of c below 0.4, the FRR starts to rise again. This trend occurs when the system reaches an operating point where the set of PSP_{Th} levels in L2 are not acting properly to compute normalized similarities.

Figure 4.11 and Figure 4.12 clearly show that VQ and SNN manifest a similar error trend, with a slightly better performance from VQ when the FAR and FRR curves intersect each other (equal error rate point). From the experiments, as a proof-of-concept, it can be concluded that both network architectures proposed are able to process frames of speech data using spiking times, they can accumulate opinions over many frames and can discern whether they are similar to previously trained patterns. Despite comparable results, it is important to clarify that more extensive experiments are required to assert which system presents the best performance.

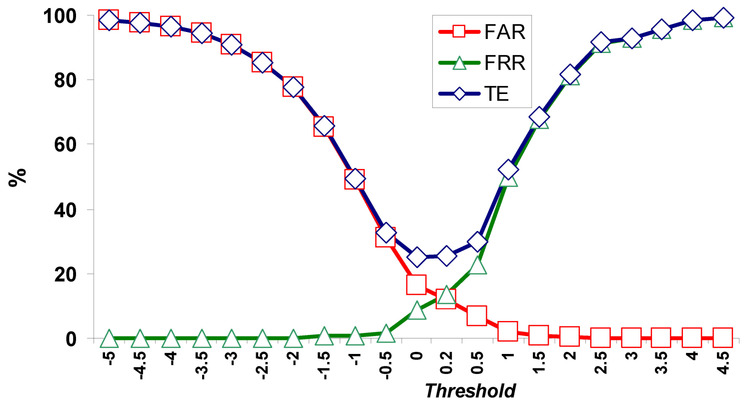


Fig. 4.11. Vector Quantization (VQ) performance on VidTimit dataset. FAR is the false acceptance rate, FRR is the false rejection rate, and TE is total error (FAR+FRR).

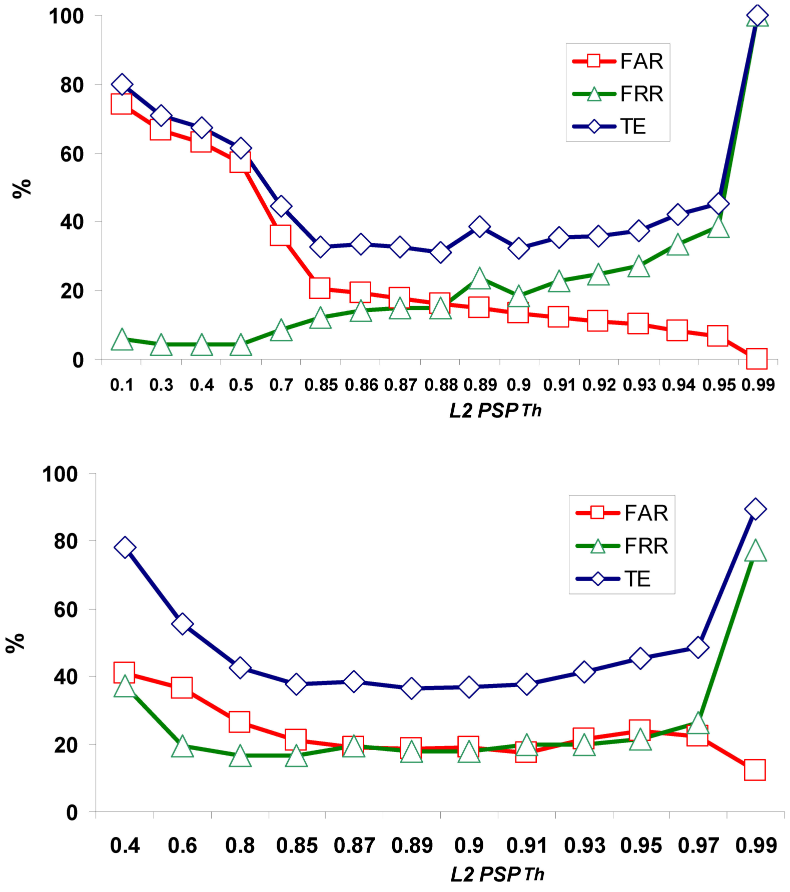


Fig. 4.12. Typical SNN performance for different values of *c* (proportion of the maximum PSP generated by a training sample). Top: SNN Architecture 1. Bottom: SNN Architecture 2. The same trained weights are used in both architectures.

4.3.7 Implementation considerations

The SNN architectures and the learning procedure were implemented in C# language with a mix of *event-driven* and *clock-driven* techniques (see Section 3.6). The number of neurons needed to process auditory information is relatively low when compared with the visual system. To process the information from one individual, the number of neurons is in the order of hundreds in Architecture 1 (Figure 4.9) and increases substantially in Architecture 2 (Figure 4.10) depending on the number of neurons used in each L2 set. The number of spikes to be propagated to the network can also vary significantly, for instance, for the implementation that compresses the auditory information with Mel Frequency Cepstral Coefficients (MFCC), the number of spikes is relatively low, whereas, if the population encoding technique is applied the number of spikes increases according to the number of receptive fields. The population encoding technique is discussed in the next section.

With a mix of *event-driven* and *clock-driven* techniques and the fast integrate-and-fire neuron model, a single personal computer has been shown to be sufficient to undertake all the experiments presented in this chapter.

4.4 A new wavelet-based speech feature extraction design with evolving SNNs

This section presents the design of a new network architecture based on fast spiking neurons performing feature extraction on speech signals. The network simulates the task of the inner hair cells of the cochlea, which perform the transduction of waves into spikes with tonotopically-organized ensembles. The systemic behaviour of the ensemble of inner hair cells is simulated with biologically inspired basic processing units (spiking neurons) to be used in artificial speech processing systems. Note that, this design does not aim to accurately reproduce the activity of the inner hair cells, despite the dynamics of spiking neurons being more biologically reasonable than other computing methods.

Sound signals are described with spectral characteristics. Cochlear fibers are sharply tuned to specific frequencies (Kiang *et al*, 1965), which are commonly modelled with the Short Term Fourier Transform (STFT) or wavelets. STFT as a discrete mathematical method has the intrinsic characteristic of being able to provide high spectral resolution of low frequency signals and low spectral resolution at high frequencies. This property does not affect the extraction of speech features for speech recognition. The Mel scale that forms the Mel filter banks also has sharply tuned filters at low frequencies and broadly tuned filters at higher frequencies.

Nonetheless, as properly described in (Rabiner and Juang, 1993) and the main object of research for (Ganchev, 2005), Mel filter banks and consequently MFCC, extract features particularly suitable for speech recognition. MFCC is also used successfully for speaker authentication, but it may occlude other features that can facilitate a unique description of a speaker. Ganchev (Ganchev, 2005) further argues that capturing the uniqueness of the speaker may need higher spectral resolution at high frequency bands, at the same time requiring flexibility to precisely capture sharp variations in time. The same work explores in detail more general properties of wavelets when compared with STFT on the speaker recognition problem, and gives a comprehensive evaluation of wavelet-based approaches through a comparison with several variations of MFCC based systems and probabilistic neural networks.

In this new design, for being more general than STFT, wavelets are used in a conceptual description of a speech signal pre-processing method using SNNs. This pre-processing of speech signals with spiking units uses the integrate-and-fire neurons with the modulation factor described in Section 3.3.1 and is composed of the following steps:

1) A pre-emphasis filter according to Equation 4.1 is applied to the speech signal;

2) The filtered signal is divided into small segments (frames);

3) Receptive fields convert each frame to the time domain using Rank Order Coding (Delorme *et al*, 1999). One neuron represents each frame position. From hereafter the processing is done through spikes;

4) Layer 1 (L1) neurons of the pre-processing network have weights calculated according to the wavelet mother function $\psi(t)$, for different scales s (expansion and compression of the wavelets) and different spatial shifts τ. The mother wavelet function is described as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-\tau}{s}) \qquad \textbf{(4.12)}$$

On L1, the shape of the mother function, the number of scales, and the number of shifts are parameters to be chosen or optimized.

5) Layer 2 (L2) neurons integrate the energy of different L1 filters representing spectral and spatial properties. This step resembles filter banks, where the number of banks and filter shapes are also subject to optimization. Figure 4.13 shows the general pre-processing network architecture. The output of L2 is a train of spikes that extracts spectral and spatial characteristics of an input frame that mimics wavelet computation.

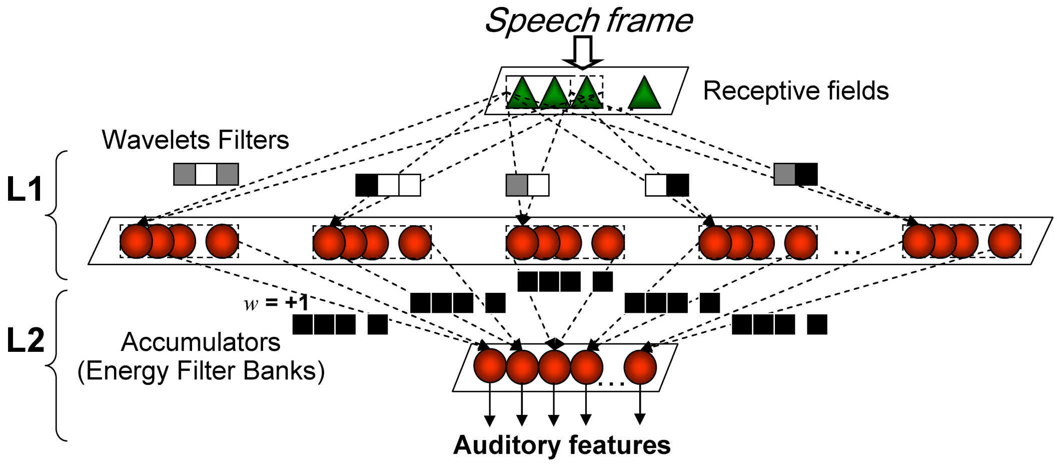


Fig. 4.13. Integrated design of a two-layer SNN that performs speech signal pre-processing.

The pre-processing can be integrated effortlessly into the classification procedures described in Section 4.3.3 and Section 4.3.4. Thus, the entire process, from the extraction of characteristics to the cognitive decision to identify a speaker, is done using processing units that communicate with spikes. Figure 4.14 shows a diagram of the integrated system.

Note that, despite of the filters in L1 being built using wavelet functions, due to the dynamics of the spiking neurons, more precisely, due to the non-linearity inserted during the computation of the post synaptic potentials, the resultant features provide only a coarse representation of wavelet output. The advantage of this design is that the entire process (pre-processing stage and recognition) is done with the same basic processing unit (spiking neurons).

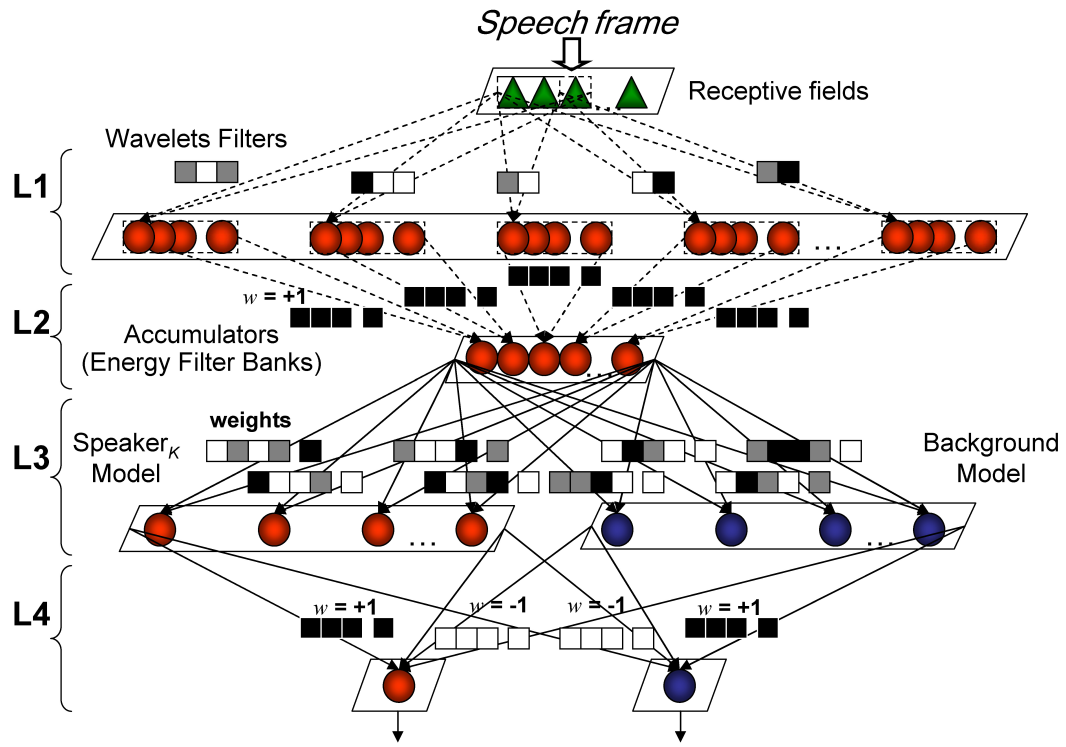


Fig. 4.14. Integrated design of an evolving SNN that performs speech signal pre-processing and speaker authentication.

4.5 Sparse representation of low-dimensional data

This section discusses the use of biologically inspired systems on processing low-dimensional artificial data. Rank Order Coding, using the relative time of spikes across a set of N neurons, has a nominal capacity for encoding patterns of N! (possible spiking orders) (Thorpe and Gaustrais, 1998) if one neuron can spike only once. Fast integrate-and-fire neurons with a modulation factor described in Section 3.3.1 are used as decoders and the weights are set to make the neurons selective to a specific train of spikes. Adjusting the threshold level (PSP_{Th}), the selectivity of a neuron to a specific pattern can be changed. At one extreme, the PSP_{Th} can be set in such a way that only an incoming spike order identical to the level of weights can make a neuron fire. In this unique setup, Rank Order Coding has the capacity to decode N! different patterns. However, in reality, for pattern recognition, PSP_{Th} operates at lower levels to take into account noisy patterns, deformed patterns, generalization ability, etc. This reduces greatly the capacity for information encoding using rank order.

Bothe (Bothe, 2003), following Natschlager and Ruf (Natschlager and Ruf, 1998) and Zhang and Sejnowski (Zhang and Sejnowski, 1999), proposed the encoding of input variables in spike times with a population of neurons, with the main aims of increasing the temporal distance between patterns, increasing information transmission capacity, enhancing the separability of clusters, and enabling scale sensitivity. Other reasons for sparse representation of data in biological systems are described in the work of Baum *et al* (Baum *et al*, 1988), which suggests that it increases the ability of performing associative memory, and Baddeley (Baddeley, 1996), which recommends sparse representation to achieve a gain in efficiency. In Foldiak (Foldiak, 1995), the main purpose of sparse representation is to facilitate associations and decrease the length of connections. Also in favour of sparse representation are (Perrinet and Samuelides, 2002) and (Olshausen and Field, 1997).

With rank order coding, population encoding can have the same effect as increasing information transmission capacity as presented in Bothe (Bothe, 2003), since the nominal capacity of information is directly proportional to the number of neurons (nominal information capacity is N!). In addition, with more neurons, separability of temporal patterns increases and the PSP_{Th} can operate at lower levels to account for noisy information and generalization ability.

The need for population encoding is particularly strong in low-dimensional data, which is exemplified by the two-spiral dataset (Lang and Witbrock, 1988). The two-spiral dataset is composed of two sets of two-dimensional data forming interlaced spirals (see Figure 4.15). First, lets consider a scenario where each dimension is represented by one receptive field neuron that encodes the information to the time domain using Rank Order Coding. Layer 1 (L1) is composed of fast integrate-and-fire neurons (described in Section 3.3.1) that are sensitive to the order of incoming spikes (Figure 4.16). Having two input neurons, the capacity of information transmission is 2!, i.e., there are only two possibilities:

a) Neuron 1 spikes earlier than neuron 2. This always happens when the amplitude of dimension 1 is higher than dimension 2;

b) Neuron 2 spikes earlier than neuron 1 because the amplitude of dimension 2 is higher than dimension 1.

Consequently, the SNN network described in Figure 4.16 is able to distinguish only two states: dimension 1 > dimension 2 and dimension 1 < dimension 2. Such a system cannot separate the interlaced spirals accurately. To overcome this constraint, the following section presents a modified version of population encoding proposed in (Bothe, 2003) to better suit the dynamics of fast integrate-and-fire neurons with modulation factor, called *Rank Order Population Encoding*. Soltic *et al* (Soltic *et al*, 2008) further analyzes *Rank Order Population Encoding* and applies this technique to the recognition of gustatory patterns.

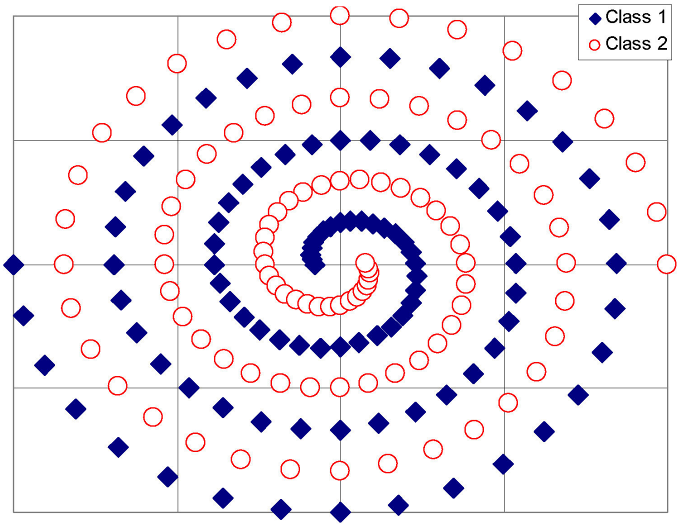


Fig. 4.15. Two-spiral dataset.

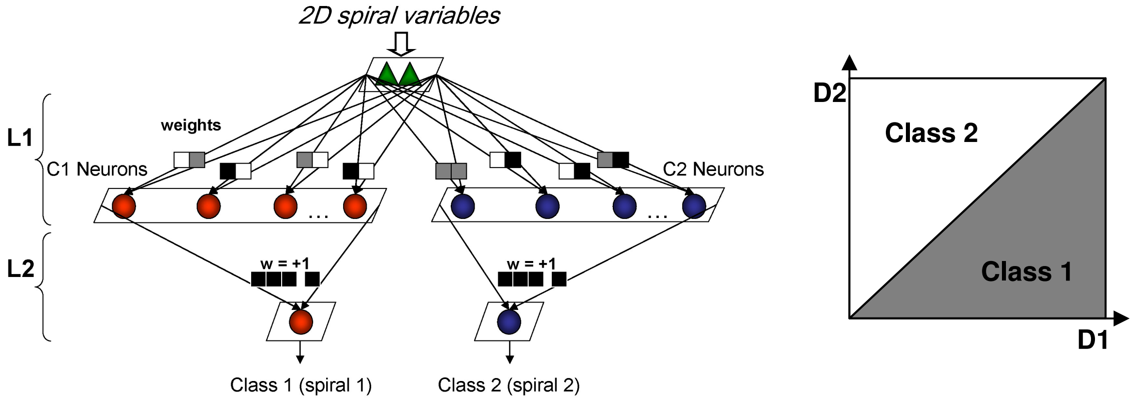
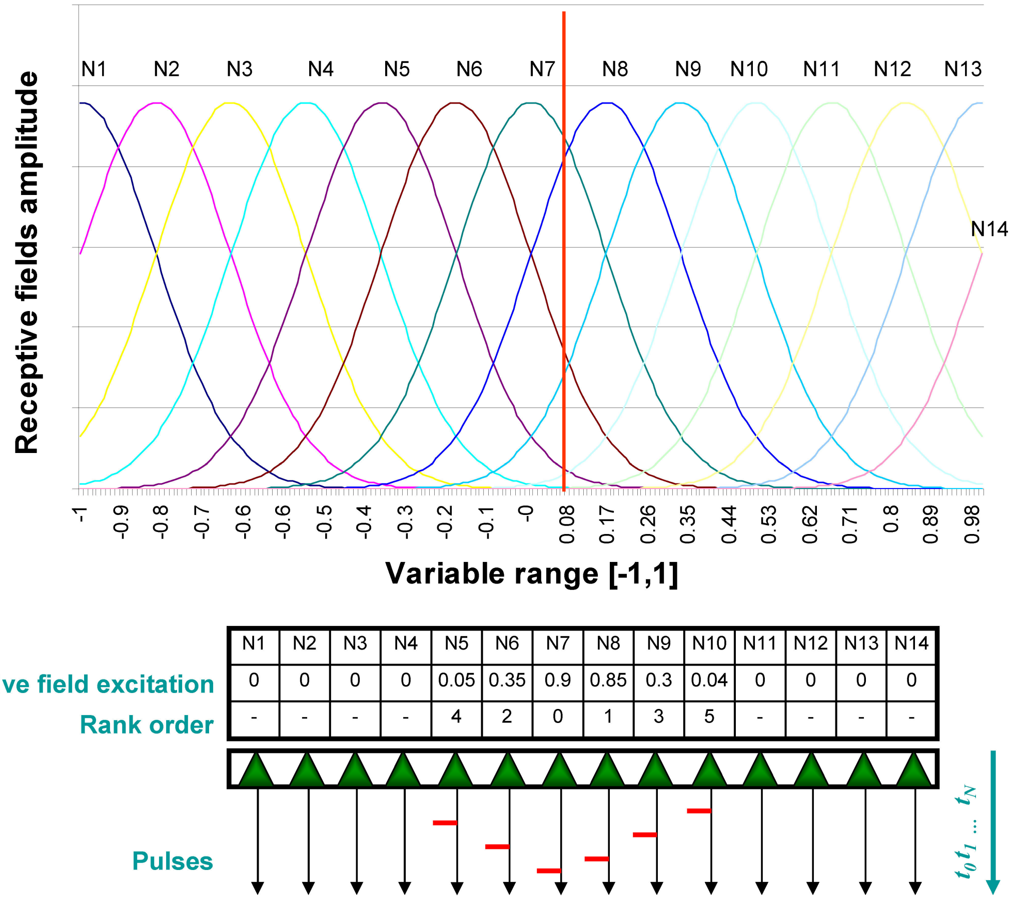


Fig. 4.16. Two-layer SNN classification of two-dimensional data using fast integrate-and-fire neurons with modulation factor and Rank Order Coding. The network is only able to separate two states: dimension 1 > dimension 2 and dimension 1 < dimension 2.

4.5.1 Population encoding of features

Population encoding splits the information of a single variable in a population of neurons to obtain sparse representation. For this purpose, overlapping Gaussian-like receptive fields placed in different locations in the range of a variable are commonly used (Baldi and Heiligenberg, 1988) (Pouget *et al*, 1999). Figure 4.17 shows the encoding process of the value 0.08 of a given variable defined in the range [-1, 1]. The variable range is covered by 14 Gaussian receptive fields equally spaced. For the value 0.08 (red vertical line) the highest excitation occurs in the receptive field neuron N7 with the value 0.9, which is followed by N8 = 0.85, N6 = 0.35, N9 = 0.3, N5 = 0.05 and N10 = 0.04. As opposed to other techniques previously described, which calculate spike times proportional to the levels of receptive field excitation, the coding is done according to rank order. Thus, in the example of Figure 4.17, N7 > N8 > N6 > N9 > N5 > N10 (the other receptive fields have very low excitation and can be disregarded) becomes the corresponding order of spikes for the value 0.08.



	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14
Receptive field excitation	0	0	0	0	0.05	0.35	0.9	0.85	0.3	0.04	0	0	0	0
Rank order	-	-	-	-	4	2	0	1	3	5	-	-	-	-

Fig. 4.17. Gaussian receptive fields and their corresponding encoding in spiking time (Rank Order Population Encoding).

Population encoding has been applied to a two-spiral dataset, where each dimension has been independently encoded with 15 sharply tuned receptive fields and 4 broadly tuned receptive fields (Bothe, 2003). Figure 4.18 shows the resultant encoding of 97 samples that form one spiral. A two-layer neural network (the architecture shown in Figure 4.19) having been trained with the adaptive online learning procedure described in Section 4.3.5, has a threshold setting for merging neurons set in such a way merging never occurs. Therefore, for each training sample a neuron in L1 is generated. The weights of L1 neurons after training are depicted in Figure 4.20.

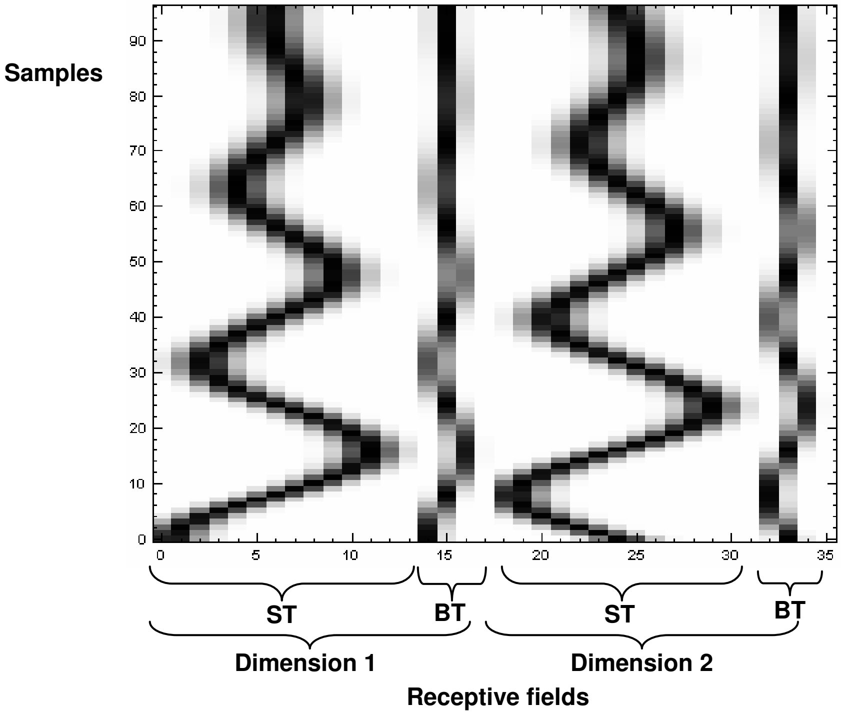


Fig. 4.18. 97 samples of one spiral data encoded with 15 sharply tuned (ST) and 4 broadly tuned (BT) receptive fields. Each row represents the encoding of one sample with rank order population encoding. The greyscale level represents the spiking time (The darker, the earlier a spike occurs).

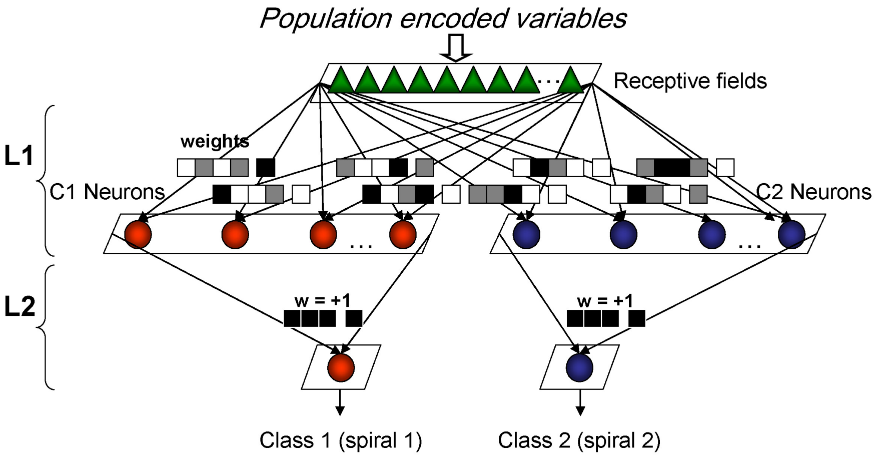


Fig. 4.19. Two-layer SNN for classification of a two-spiral dataset with population encoding.

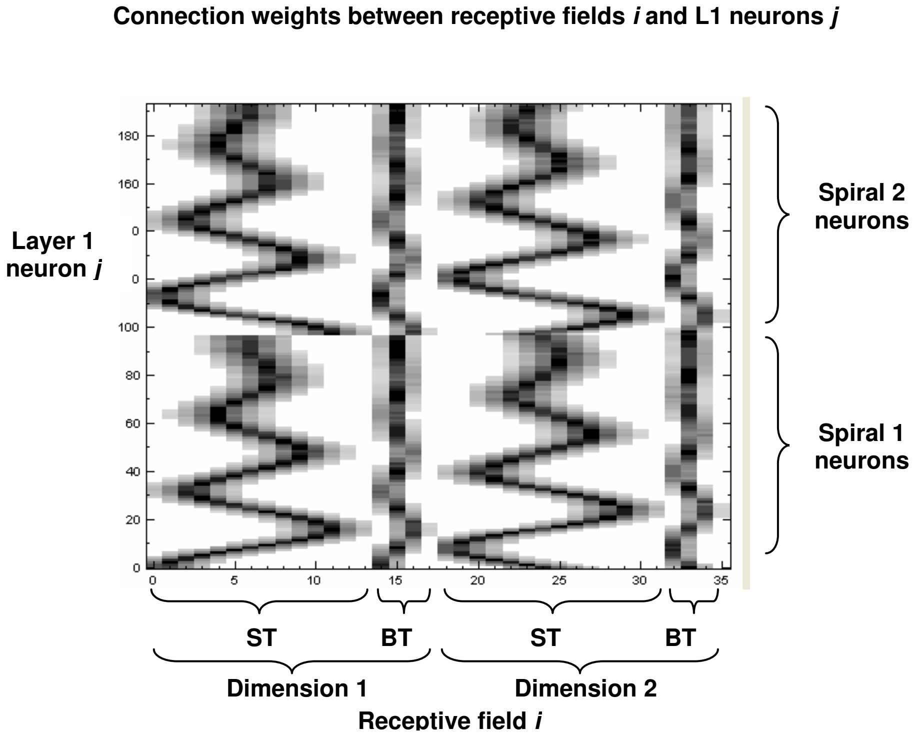


Fig. 4.20. Trained SNN network weights to classify the two-spiral dataset. The greyscale level represents the weights' strength (The darker, the stronger the weight connection). ST and BT are sharply and broadly tuned receptive fields respectively.

After training, the network was able to correctly classify all the training samples. To visualize the spatial division attained by the network, a two-dimensional space has been covered with a grid of equally spaced points (used 21 x 21). The points were population encoded and tested on the trained network. The resultant classification generated by the network is shown in Figure 4.21. In Figure 4.21 there are areas where dots are not plotted. In these areas, the network does not provide any output despite input excitation, i.e., the input excitation is too dissimilar from any pattern previously trained. There are also a few occasions where output spikes are generated in both classes of neurons (depicted with triangular shapes). Overall, the surrounding areas that were described by the trained samples are correctly classified with the help of population encoding.

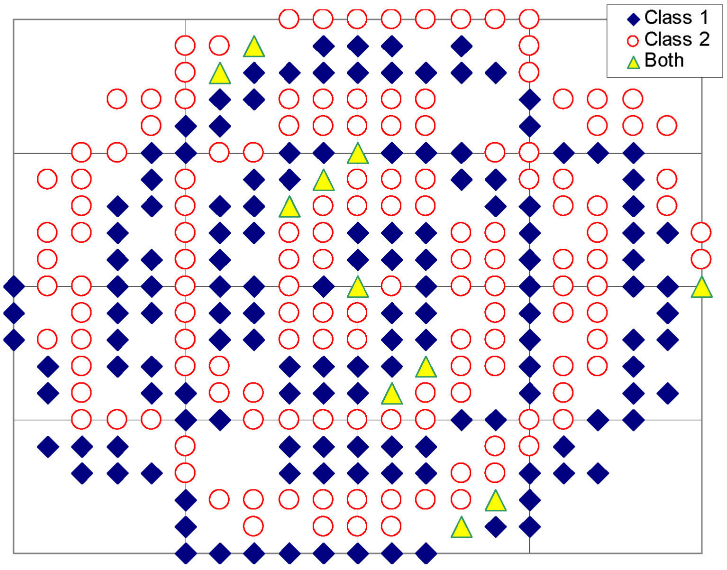


Fig. 4.21. Space division of two classes using 97 samples where each dimension is encoded with a population of neurons (15 sharply and 4 broadly tuned). Yellow triangles illustrate areas where both classes release an output spike whereas empty spaces are where none of the output neurons released spikes.

In another experimental setup, MFCC encoding was used with a population of neurons to work on the speaker authentication problem. However, no further enhancement in performance was noticed when compared with the results described in Figure 4.12.

4.6 Chapter conclusion

In this chapter, the models of the auditory system are reviewed, with a special focus on models that emulate information processing in recognizing auditory patterns. To the speaker authentication problem, a new method that enables a high level of cognition of auditory information using spiking neural networks is presented. Two spiking neuron architectures for speaker authentication are described for the first time (see also (Wysoski *et al*, 2007a)). These networks process streams of speech signals in a frame-based manner. The output layers accumulate positive and negative opinions on whether it is a certain speaker or a background. The main difference between the architectures is that, for each frame, Architecture 1 outputs a binary opinion while Architecture 2 gives a notion of similarity between the incoming frame and the closest prototypes.

Connection weights between receptive fields and L1 can be trained to respond to different parts of an utterance, closely corresponding to the use of *k-means* algorithm to create codebooks (Burileanu *et al*, 2002), or a set of Gaussians in GMM (Reynolds *et al*, 2000). The new models also incorporate the idea of *normalized similarity*, which demonstrated itself to be effective in several classical models (Burileanu *et al*, 2002) (Reynolds *et al*, 2000).

The procedures suggested in Section 4.3.5 based on *k-means* and network structural adaptation enable continuous and adaptive training. The main properties of these procedures are:

a) *k-means*: needs to define the number of neurons in advance, can present initialization and local minima problem;

b) network structural adaptation: an additional parameter for merging neurons needs to be tuned (merging threshold), and a different division of the feature space can be obtained according to the order of the training samples (Gallant, 1995) (Wysoski *et al*, 2006).

Experiments carried out with the VidTimit dataset indicate that both network architectures proposed are able to process frames of speech data using spiking times. Further, SNNs manifest a similar error trend when compared with Vector Quantization (VQ).

The final part of the chapter describes a novel multi-layer network design based on wavelets to perform pre-processing and extraction of features using spiking neurons. With this model, a complete system can be integrated and processed using spiking units, from the processing of raw signals to the mechanism that achieves higher levels of cognition.

The architectures proposed have their capacity for information processing substantially reduced in low-dimensional data. The encoding of variables in population of neurons is presented as an alternative and exemplified in a two-dimensional dataset.

Table 4.3 describes the results in respect to the three specific aspects of *Brain-like* ECOS proposed in Section 2.3.3 (information processing units, information processing pathways and learning ability).

Table 4.3. Summary of results according to three specific aspects of *Brain-like* ECOS proposed in Section 2.3.3.

Processing Units	Spiking neurons are used as information processing unit in the decision-making stage. A new design of feature extraction using spiking neurons is also described.
Structure	Auditory information propagates with feed-forward connections into four-layers neuronal maps of spiking neurons that represent the behaviour of various auditory areas (see Figure 4.14) (tonotopically organized cells, spectral filter banks, phonetic association).
Learning	The online evolving procedure enables the learning of external stimuli through synaptic plasticity and structural adaptation. The addition of new classes is done in a supervised way. The adaptive learning creates/merges neurons that respond optimally to different speech phones in a supervised or unsupervised fashion when new utterances of a class are presented.

While the dynamics of the network architectures implemented with spiking neurons have proven suitable for performing speaker authentication, further development is needed in the direction of representing the auditory pathways in a manner closer to biology. In this direction, further effort is needed to integrate features processed with spiking units (Section 4.5) into the decision-making levels (Section 4.3). The implementation of an integrated system is the first step towards the simulation of the complete auditory information pathway. Simulation of a complete system can provide the means of optimizing both features and recognition parameters. Further, it may be possible to use multi-criteria parameter optimization procedures to reach better data encoding (Bothe *et al*, 2002) (Smit and Barnard, 2004), minimize processing time and reduce the overall number of spikes.

The experiments conducted in this dissertation used a dataset with a few short sentences. Under the assumption that learning of the temporal dynamics can not be reliably done with small training sets of data, the dynamic properties of speech signals were not considered. However, speech signals are well known for carrying a great deal of information in their temporal dynamics, being it at a phonetic or at a semantic level

(Rabiner and Juang, 1993). Networks of spiking neurons have already proven themselves to be very suitable for the computation of temporal dynamics (Lysetskiy *et al*, 2002) (Natschlager and Ruf, 1998). Therefore, extending and testing the architectures proposed to learn temporal variations with longer speech utterances is a promising task.

Chapter 5 - Evolving SNNs for audiovisual pattern recognition

This chapter covers the most relevant models for performing integration of sensory information and describes a new brain-inspired system that integrates audiovisual information on a person authentication task.

First, a concise review presents the models that emulate the human-way of information processing with respect to the combination of specialized sensory pathways to attain a particular ability. In the second part of the chapter, a new brain-inspired system integrates the individual auditory and visual information processing modules described in Chapter 3 and Chapter 4. The integrated system is built so that it can accommodate the individual characteristics of each modality, i.e., specific processing areas are allocated for each modality at lower processing levels. However, communication pathways further enable the exchange of signals between sensory modalities at higher levels of information processing. Such a characteristic is not considered in already proposed artificial models, e.g., (Kasabov *et al*, 2000) (Ross and Jain, 2003) (Sanderson and Paliwal, 2004).

Individual sensory pathways as well as the integrative modules are implemented using a fast version of spiking neurons grouped in spiking neural network architectures capable of lifelong adaptation. A new crossmodal integration mechanism enables individual modalities to influence others before individual decisions are made, a function that resembles some characteristics of biological brains.

The system is applied to the person authentication problem. Preliminary results, with the VidTimit dataset (Sanderson and Paliwal, 2002), show that the integrated system can improve the accuracy in many operating points as well as enable a range of multi-criteria optimization of parameters. A discussion on the main properties of the integrated system and future directions conclude the chapter.

5.1 Background and benchmarking

There is strong experimental evidence showing that integration of sensory information occurs in the brain (Calvert, 2001) (Kriegstein and Giraud, 2006) (Kriegstein *et al*, 2005) (Stein and Meredith, 1993) (Ghazanfar *et al*, 2005) and a lot is known about the location in the brain where different modalities converge. A more conservative theory asserts that the integration occurs in *supramodal* areas that contain neurons sensitive to more than one modality, i.e., neurons that process different types of information (Ellis *et al*, 1997). Nonetheless, behavioural observations and electrophysiological experiments have demonstrated the occurrence of another integrative phenomenon: *crossmodal* coupling, which is related to the direct influence of one modality to areas that intrinsically belong to other modalities (Calvert, 2001) (Ghazanfar *et al*, 2005).

As studies of the neuronal mechanisms that underlie interaction among modalities at the level of single or an ensemble of neuronal cells are still inconclusive (Kriegstein *et al*, 2005), computational models of neuronal interactions inspired by perceptual and neurophysiologic studies are used to test theories of modular interdependencies. These computational models can also be applied to multimodal pattern recognition in an attempt to enhance the performance of traditional pattern recognition algorithms. The latter is the immediate objective in this part of the research.

First, this review describes some models that emulate the biological integration. It is followed by some insights into traditional ways of integrating multimodal information.

5.1.1 Insights from psychology

The work of Bruce and Young (Bruce and Young, 1986), which is further analysed in (Burton *et al*, 1990), describes a functional model based on the compilation of a series of psychological observations of the process involved in identifying an individual. The model is composed of face recognition units (FRU) that store visual structural descriptions. Each individual has a separate FRU. The output of an FRU activates the appropriate person identity node (PIN). The link to semantic and associative information about the identity of each known individual occurs in the person identity node occurs, i.e., the facial information is associated with other types of information, e.g., gender (male, female), profession (engineer, soccer player), etc. Different from the

FRU, which is completely dedicated to facial processing, PIN can be accessed by other routes, e.g., voice, written or heard name, etc. The model further incorporates the "identity priming" concept. With identity priming, the recognition of a familiar face is faster if it is preceded by the 'prime' face of a closely associated person, e.g., Prince Charles followed by Princess Diana. Another interesting point related to the processing speed covered by the model is that familiar faces are recognized more rapidly if the faces have been rated as "distinctive" in appearance compared with those rated as more "typical" in appearance (Valentine and Bruce, 1986). This effect seems to be a genuine consequence of the relationship between a face and the majority of known faces. Thus, in summary, the model based on perceptual evidences covers: a) familiarity decisions; b) identity priming; c) cross and within domain semantic priming; and d) speed of processing.

After acknowledging that it is still unclear how cortical modules interact, in (Kriegstein *et al*, 2003) (Giraud and Truy, 2002) (Giraud *et al*, 2001) some results of neuroimaging experiments are presented, replicating earlier findings that human voices are specifically processed along the Superior Temporal Sulcus (STS) and Fusiform Face Area (FFA) is inherently the area for processing faces. These areas are anatomically segregated. However, it is further identified that voices of familiar individuals generate responses in the FFA, which suggests that, besides the existence of a supra-modal layer responsible for the integration of modes, there might be additional and earlier mechanisms of coupling between modalities. Other previously proposed models (Burton *et al*, 1990) (Ellis *et al*, 1997) (Schweinberger and Burto, 2003) describe the reciprocal influence among modalities with a top-down approach, where the individual modes are linked in the supramodal layer which is responsible for integrating information for each modality and redistributing the information back to individual modes through feedback connections (see Figure 5.1).

Kriegstein *et al* (Kriegstein *et al*, 2005) specifically report the voice-to-face crossmodal effects on the task of speaker recognition. The top-down approach described in Figure 5.1 could not be observed on a task that emphasizes speaker over speech recognition. The top-down approach would imply a higher correlation of activity between supramodal areas and the STS (voice area) than between STS areas and FFA (face area). However, these experimental results effectively demonstrated a higher correlation between STS and FFA, which indicates a direct link between individual modalities where the signals from one modality do not pass through the supramodal areas to reach the other. A diagram of supramodal area and crossmodal coupling of information is presented in Figure 5.2.

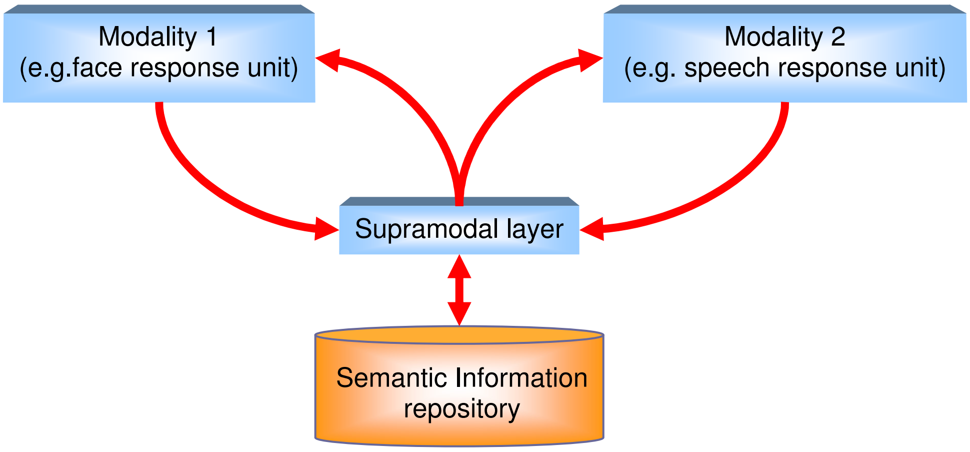


Fig. 5.1. Sensory integration routes according to (Burton *et al*, 1990) (Ellis *et al*, 1997) models. The integration of modalities follows a top-down approach where the individual modes are linked by the supramodal layer with feedback connections. (Diagram modified from (Kriegstein *et al*, 2005)).

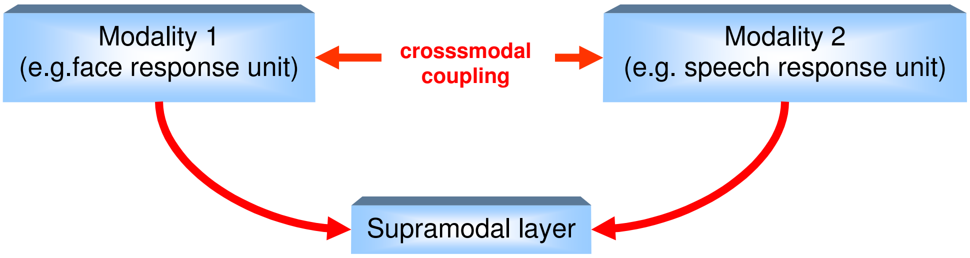


Fig. 5.2. Audiovisual sensory integration according to (Kriegstein *et al*, 2005) on the speaker recognition task. The supramodal region and a direct link between individual modalities (crossmodal coupling).

Further research with neuroimaging techniques by the same research group (Kriegstein and Giraud, 2006) suggest that even unimodal retrieval is facilitated if a pattern has been trained with more than one modality. In the process of recognizing a person using speech information that has been trained with image and speech information, the FFA (face area) is also activated. These multisensory associations are effective in increasing the recognition abilities.

5.1.2 Insights from physiology

Physiological studies are mostly in line with the psychological models, agreeing that the crossmodal coupling occurs before the information processing reaches higher levels. Ghazanfar *et al* (Ghazanfar *et al*, 2005) reports the crossmodal effects through audiovisual experiments in monkeys. In (Calvert and Campbell, 2003) (Calvert *et al*, 1997) the observations in human brains are described.

At the neuronal level, several experiments in different animals (cat, rat, monkey) demonstrated areas (in the superior colliculus and in the cerebral cortex) where the cells can respond to more than one type of excitation (Benevento *et al*, 1977) (Bruce *et al*, 1981) (see (Calvert, 2001) for more references). Investigating the underlying mechanisms of neurons sensitive to different sensorial stimuli, it has been found that some neurons are able to combine sensory inputs as integrated product, i.e., when two sensory stimuli from different sources are presented in temporal proximity, the activity of the neuron increases by more than a simple summation (almost in a multiplicative fashion). Another property detected is the so-called *inverse effectiveness*, a mechanism which substantially increases the crossmodal influence when a stimulus of a single modality is not as effective (Stein and Meredith, 1993). Lastly, multisensory neurons can present *response depression* activity, a phenomenon described in (Kadunce *et al*, 1997) that decreases incoming stimuli for different modalities if they are not spatially correlated (see (Stein and Meredith, 1993) for a detailed explanation).

5.1.3 Insights from pattern analysis

The integration of modalities for the purpose of pattern recognition often targets tasks that cannot be solved by a single system or can be facilitated by using more than one source (generally where there is unimodal ambiguity, unimodal lack of data and/or correlation among modes). Many studies report considerable performance improvement (Kasabov *et al*, 2000) (Ross and Jain, 2003) (Sanderson and Paliwal, 2004) (Sharkey, 1999) as well as state that the use of modularity results in systems that are easy to understand and modify. In addition, modular approaches are well known for preventing modular damage, facilitating training and the inclusion of prior knowledge (Ross and Jain, 2003).

There are two classic issues when dealing with multimodal systems: how to perform the decomposition and recombination of modes:

- **Decomposition:** Decomposition can occur with modules and sub-modules, e.g. a visual can be decomposed into colour and shapes, which can be further decomposed into edges and borders, and so on. For the decomposition, the problems are not always well known and explicit as is the case with the visual and auditory modalities. In some cases, the decomposition can be done by automatically breaking down the problem based on intrinsic properties of the information provided (Sharkey, 1999).
- **Recombination:** The recombination of the modules can be cooperative (all modules contribute to the result), competitive (only the most reliable module is responsible for the decision), sequential (the computation of one module depends on the output of the other), and supervised (one module is used to supervise the performance of others) (Ross and Jain, 2003) (Sanderson and Paliwal, 2004).

Sometimes in order to avoid the recombination process, systems perform the combination of information from different modalities before the recognition process is undertaken. One unique module is then used for recognition. While this approach is easier to design, often the unique module encounters difficulties during the learning process. Also in this configuration, the designer cannot include or extract explicitly any knowledge related to individual modalities during the recognition process.

As an example of multimodal models that use traditional methods of computation, Haller *et al* (Haller *et al*, 2006) present an audiovisual system to detect a TV anchor-person. The system processes modalities separately. The speech part uses MFCC (Mel Frequency Cepstral Coefficients) as features and Gaussian Mixture Model for classification. The face detection presented better accuracy with principal component analysis (PCA) and Gaussian Mixture Model. The decision fusion is made with AND and OR gates. Park *et al* (Park *et al*, 2006) perform audiovisual human authentication using PCA on visual information and cepstral LPC (linear prediction coefficients) to describe auditory features. Hidden Markov Model (HMM) is used for classification. The integration of modalities is done through a non-linear approach using Fuzzy logic. In another system for audiovisual person authentication, Ben-Yacoub *et al* (Ben-Yacoub *et al*, 1999) presents a multimodal fusion of experts to add robustness and to increase the performance of the authentication process. The individual modalities are processed separately. The face is represented with Elastic Graph Matching (EGM) and features extracted using Gabor functions. Features between two face models are matched with a given distance equation. The speech information is represented with cepstral LPC and a statistical method to calculate the similarity between speakers called Arithmetic-Harmonic Sphericity (Bimbot *et al*, 1995) is used in the text-independent setup. In the same work, a text-dependent system based on HMM is also presented. Nonetheless, Ben-Yacoub *et al* concentrate on the fusion component, with five different post integration methods being proposed and compared. They are: SVM-based (support vector machine) fusion, minimum cost Bayesian classifier, Fisher linear discriminant, C4.5 classifier, and neural network-based (multi-layer perceptron) classifier. Experiments were carried out with the XM2VTSDB database (Messer *et al*, 1999), which contains synchronized recordings of facial properties as the individual utter digit from 0 to 9.

Brunelli and Falavigna (Brunelli and Falavigna, 1995) present a system where two classifiers are used to process speech signals and three others to recognize visual inputs. The results of these individual classifiers are connected to the input of a new integrative module based on HyperBF networks (Poggio and Girosi, 1990). MFCC and the corresponding derivatives are used as features, and each speaker is represented by a set of vectors based on Vector Quantization (VQ) (Rosenberg *et al*, 1987). A local template matching approach at the pixel level, where particular areas of the face (eyes, nose, mouth) are compared with a previously stored data, is used for face authentication.

Attempting to further improve the performance of the multimodal systems, several methods propose adaptation of the fusion mechanisms (Chibelushi *et al*, 1999) (Sanderson and Paliwal, 2002) (see (Chibelushi *et al*, 2002) for an extensive and comprehensive list).

Maciokas and Goodman (Maciokas and Goodman, 2002), based on brain-like approaches, tackle the problem of integrating the visual information of lip movements with the corresponding speech generated by it. It uses a biologically realistic spiking neural network with 25,000 neurons placed in 10 columns and several layers. Tonotopic maps fed from Short Term Fourier Transform (STFT) with a neural architecture that resembles MEL scale filters are used for converting audio signals to spikes. Gabor

Filters extract the lip movements. The encoding of three distinct sentences in three distinct spiking patterns was demonstrated. In addition, after using the Hebbian rule for training the output spiking patterns were also distinguishable from each other.

Seguier and Mercier (Seguier and Mercier, 2002) also describe a system for integrating lip movements and speech signals to present a one-pass learning with spiking neurons. The performance achieved is favourable to the integrated system, mainly when audio signals are deteriorated with noise. The system was intended to produce real-time results, therefore simple visual features are used and auditory signals are represented by 12 cepstral coefficients. Vector quantization is applied individually to extract vector codes, which are then encoded into pulses to be processed by the Spatio-Temporal Artificial Neural Network (STANN) (Mozayani *et al*, 1988) (Vaucher, G., 1998).

Chevallier *et al* (Chevallier *et al*, 2005) present a system based on SNN to be use in a robot capable of processing audiovisual sensory information in a prey-predator environment. In reality, the system is composed of several neural networks (prototype-based incremental classifier), one for each sensorial modality. A centralized compartment for data integration is implemented as a bidirectional associative memory. A network (also incremental) is used to perform the final classification (this architecture is described in detail in (Crepet *et al*, 2000)). Particularly interesting in the prey-predator implementation is the spike-based bidirectional associative memory used. As properly suggested by the authors, the implementation using spikes enables the flow of information over time. The integration of these streams of incoming data is also processed on the fly as soon as the data from different modalities are made available. Furthermore, the bidirectional associative memory implemented with the spiking mechanism enables the simulation of crossmodal interaction.

Kittler *et al* (Kittler *et al*, 1998), after providing a review, tries to find a common basis for the problem of combining classifiers through a theoretical framework. It is argued that most of the methods proposed so far can be roughly classified in one of the following types: product rule, sum rule, min rule, max rule, median rule and majority voting. After performing error sensitivity analysis on several combined systems, it is further suggested that the sum rule outperforms the other combination procedures.

A more specific review of the speech-based audiovisual integration problem (speech and speaker recognition) is provided in (Chibelushi and Deravi, 2002).

Among all the systems mentioned before, whether using traditional techniques or brain-like networks, none of them demonstrated a degradation of performance of multimodal systems. The integration, in a synergistic way, achieves higher accuracy levels when compared with single modalities alone.

The next section presents a simple attempt to process multimodal sensory information with a new architecture of fast spiking neurons. Besides the inherent ability of the neurons to process information in a simple and fast way (Delorme *et al*, 1999), the main property of the system is the ability to receive and integrate information from several different modules on the fly, as the information becomes available. Because the entire system is based on the same principle of computation (spiking units) and the processing time of the information is also meaningful, back and forth connections as well as connections that emulate crossmodal influences are able to be simulated in a more biologically realistic manner. The crossmodal connections enrich the architecture of the current multimodal systems (depicted in Figure 5.3) that are based traditionally on the decomposition and consequent recombination of modalities. The illustration of multimodal systems with crossmodal connections is shown in Figure 5.4. In particular, the system tackles the person authentication problem with the integration of audiovisual cues.

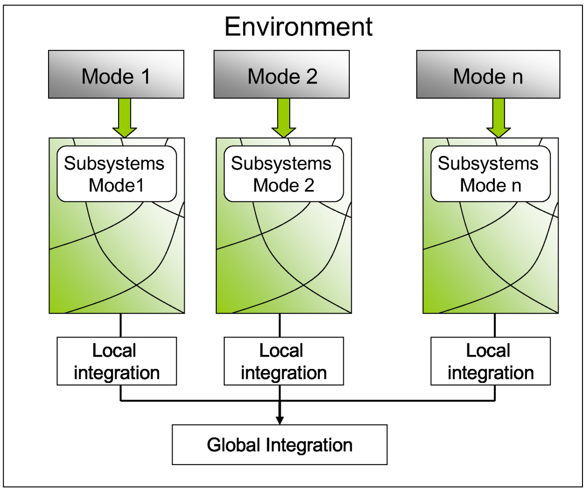


Fig. 5.3. Traditional architecture for *a posteriori* integration of multimodal systems.

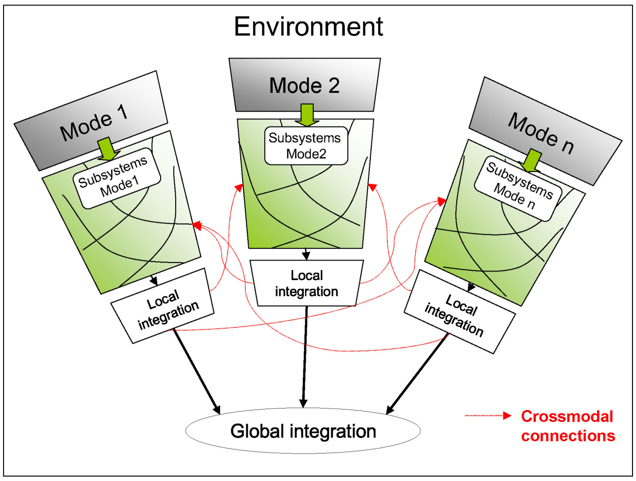


Fig. 5.4. Multimodal integration with crossmodal influences, which is employed in the new model described in this dissertation.

5.2 A new evolving SNN model for audiovisual integration

The biologically inspired integration of modalities for pattern recognition uses the theory of spiking neural networks, where the individual modes and the integration procedure are implemented with spiking neurons. The same fast integrate-and-fire neuron described in Chapter 3 and Chapter 4 is used (see Section 3.3 for a complete description).

Each individual modality has its own network of spiking neurons. In general, the output layer of each modality is composed of neurons that authenticate/not authenticate a class they represent when output spikes are released.

The approach for integrating modalities consists of attaching a new layer onto the output of the individual modes. This layer (supramodal layer) represents the supramodal region and contains neurons that are sensitive to more than one modality (Stein and Meredith, 1993). In the implementation proposed here, the supramodal layer contains two spiking neurons for each class label. Each neuron representing a given class C in the supramodal layer has incoming excitatory connections from the output of class C neurons of each individual modality. The two neurons have the same dynamics, yet different thresholds for spike generation (PSP_{Th}). For one neuron, the PSP_{Th} is set in such a way that an output spike is generated after receiving incoming spikes from any single modality (effectively it is a spike-based implementation of an OR gate). The other neuron has PSP_{Th} set so that incoming spikes from all individual modalities are necessary to trigger an output spike (AND gate). AND neuron maximizes the accuracy and OR neuron maximizes the recall (See Figure 5.5).

In addition to the supramodal layer, a simple way to perform crossmodal coupling of modalities is designed. The crossmodal coupling is set as follows: when output neurons of an individual modality emit spikes, the spikes not only excite the neurons in the supramodal layer, but also excite/inhibit other modalities that still have ongoing processes. Effectively the excitation/inhibition influences the decision on other modalities, biasing (making it easier/more difficult) the other modality to authenticate/not authenticate a pattern.

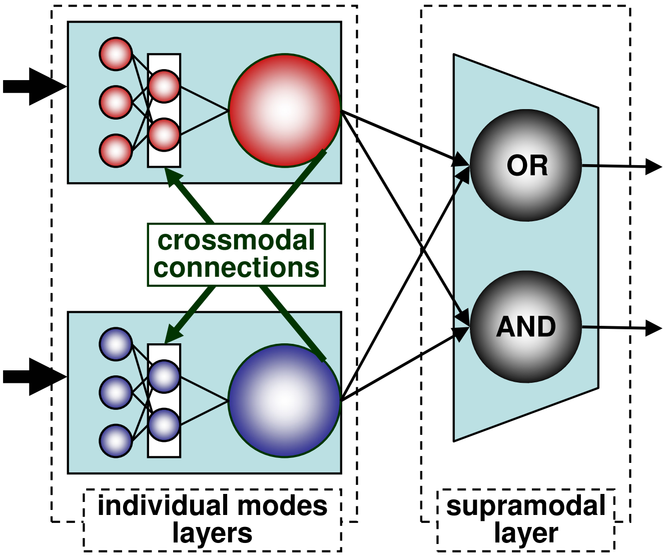


Fig. 5.5. Integration of individual layers with a supramodal layer and crossmodal connections. The individual and supramodal layers are implemented using spiking neurons.

For the crossmodal coupling, different from the supramodal layer connections that are only excitatory, both excitatory and inhibitory connections are implemented. With this configuration, the output of a given class C in one modality excites the class C neuronal maps in other modalities. In contrast, the output class $\hat{C}$ (not class C) in one modality has an inhibitory effect on class C neuronal maps in other modalities.

In the following section, the supra/cross modal concepts are applied to the case of audiovisual integration in a person authentication problem based on face and speech information. The implementation of the visual model follows the description given in

Section 3.5 and the auditory model uses the architecture described in Section 4.3.3. A more detailed explanation of the implementation is also given.

5.2.1 Visual system model

The visual system is modelled with a four-layer feed-forward network of spiking neurons, with the same configuration as described in Section 3.5 and in (Wysoski *et al*, 2008). Figure 5.6 shows the network architecture, which combines opinions of being/not being a desired face over several frames (multi-view face recognition). Basically, the network receives in its input several frames that are processed in a frame-by-frame manner. Neurons in the first layer (L1) represent the On and Off cells of the retina, enhancing the high contrast parts of a given image (high-pass filter). The second layer (L2) is composed of orientation maps for each frequency scale, each one being selective to different directions. They are implemented using Gabor filters in eight directions (0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°) and two frequency scales. Maps in the third layer are trained to be sensitive to complex visual patterns (faces in the case study evaluated here). In L3, neuronal maps are created or merged during learning in an adaptive online way. See Section 3.4.2 for a description of the learning procedure.

Different from the architecture described in Chapter 3, besides having incoming connection from the L2 neurons, L3 neurons receive crossmodal influences from other modalities. In other words, instead of L3 being composed of exclusively unimodal neurons sensitive to visual excitation, L3 has multisensory capabilities. L3 neurons are still mainly visual, but are also sensitive to stimuli from other modalities.

Neurons in layer 4 (L4) accumulate opinions about being a certain class over several frames. If the opinions are able to trigger an L4 neuron to spike, the authentication is completed.

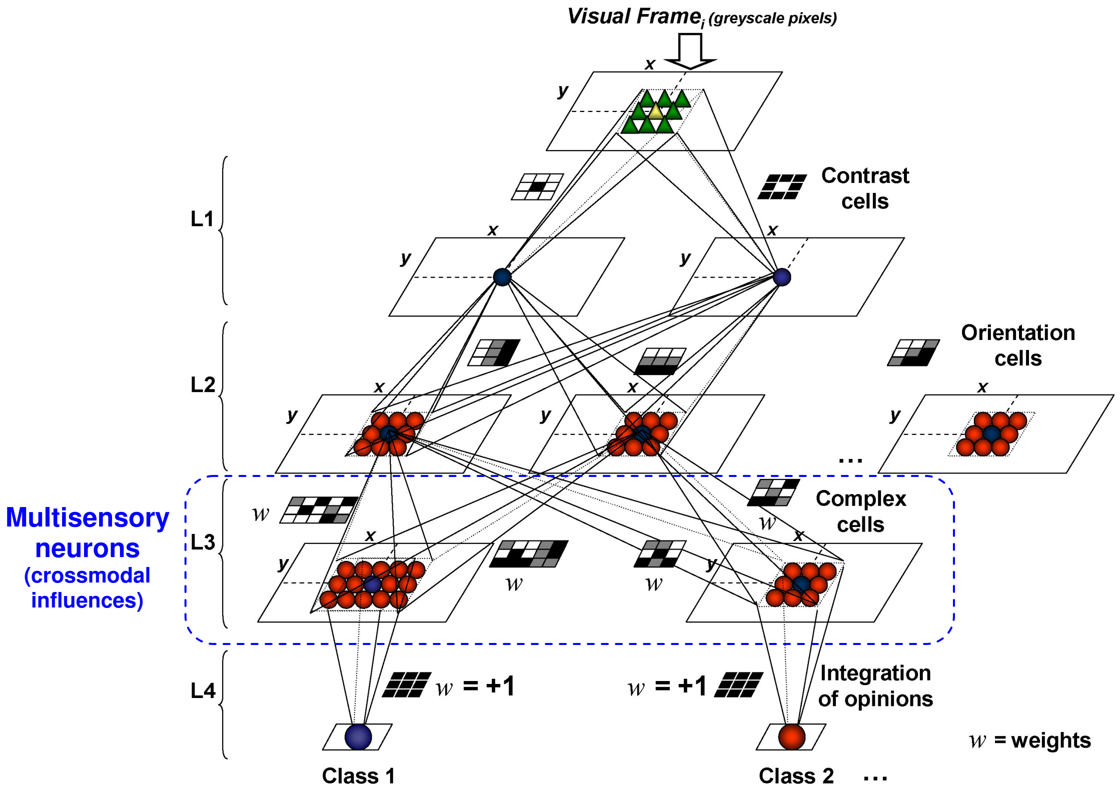


Fig. 5.6. Evolving spiking neural network (eSNN) architecture for visual pattern recognition. Neurons in L1 and L2 are sensitive to image contrast and orientations, respectively. L3 has the complex cells, trained to respond to specific patterns. It is in L3 that crossmodal coupling occurs. L4 accumulate opinions from different input excitations over time.

5.2.2 Auditory system model

The auditory system is modelled with a two-layer feed-forward network of spiking neurons with the same architecture and behaviour as described in Section 4.3.3 (integration of binary opinions) (Wysoski *et al*, 2007a). Each speaker is represented by a set of prototype vectors that compute normalized similarity scores of MFCC (Mel Frequency Cepstrum Coefficients) considering speaker and background models. Prototypes of a given class are memorized in the connection weights of L1 neurons. For the integrative approach described here, L1 neurons are also the recipients of crossmodal influences, in the form of excitation or inhibition. Thus, L1 neurons, besides being primarily responsible for processing auditory information, can be affected by other modalities (therefore multisensory units) to a lower degree. The network architecture is illustrated in Figure 5.7.

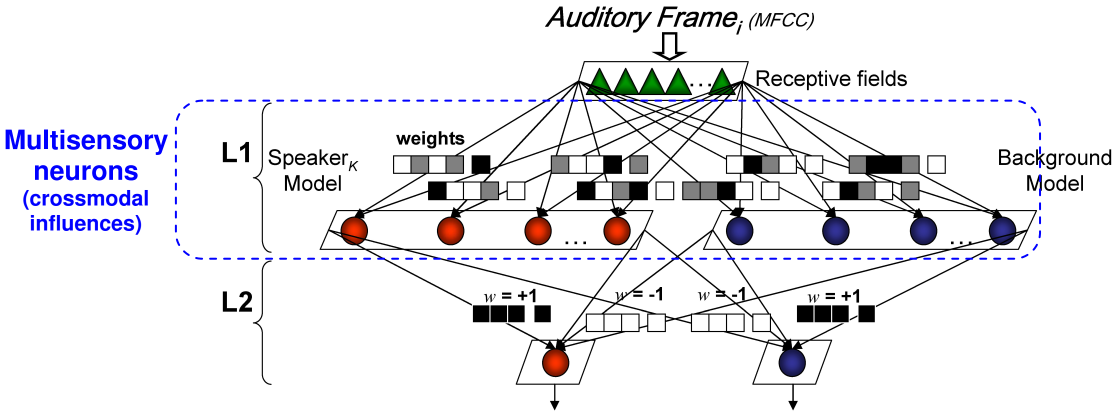


Fig. 5.7. Speaker authentication with spiking neural networks. L1 neurons, with their respective connection weights, implement the prototypes of a given class. L1 neurons also are receivers of crossmodal excitation/inhibition. L2 neurons accumulate binary opinions about being a claimant over several frames of speech signals.

There are two neurons in L2 for each speaker accumulating opinions over several frames of speech signals. One neuron is triggered if the speaker is authenticated and the other if the input excitation is more likely to be the background model.

5.2.3 Audiovisual integration

The detailed audiovisual crossmodal integration architecture is shown in Figure 5.8. The bottom part of Figure 5.8 shows two neurons (OR and AND) representing the supramodal layer. Each spiking neuron in the supramodal layer operates in the same way as the neurons that compose the SNNs of individual modalities (fast integrate-and-fire neurons with modulation factor described in Section 3.3).

Even this simple configuration of the supramodal layer can have quite a complex behaviour that cannot be easily described in an analytical way. However, to facilitate the illustration of the integrative system, a particular case is described. The supramodal neurons are set with a modulation factor of mod = 1, and all the incoming excitatory connection weights (W) are set to 1. Thus, the PSP_{Th} that implements the OR integration for two modalities is equal to 1. The neuron implementing the AND integration receives PSP_{Th} = 2. Notice that it is only possible to set these parameters deterministically because the neurons can spike only once during the entire simulation period.

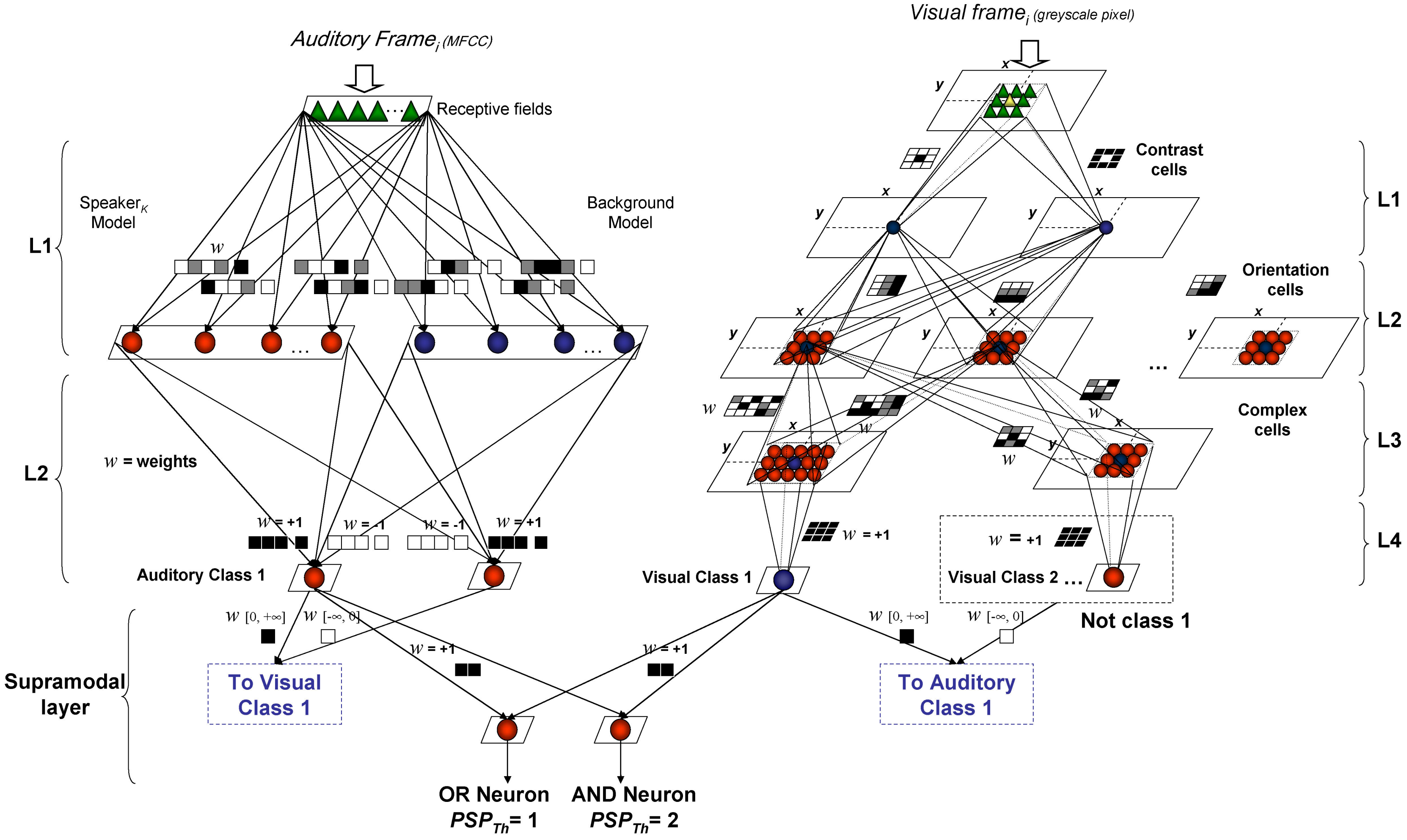


Fig. 5.8. Integration of modalities using evolving SNNs. The supramodal layer integrates incoming sensory information from individual modalities and crossmodal connections enable the influence of one modality upon the other.

Once again, to facilitate the analysis, crossmodal influences between modalities are effectively modelled through the modification in the PSP_{Th} of the crossmodal neurons, namely L3 neurons in the visual system and L1 neurons in the auditory system. Thus, instead of simulating crossmodal influences with spikes that will consequently excite/inhibit a neuron (increase/decrease neuron's PSP), which corresponds to the biological method, the crossmodal influence is implemented by increasing/decreasing the PSP threshold of the neurons. The effect in terms of network behaviour is the same, however it is found to be easier to parameterize the amount of crossmodal influence through the variation on the PSP thresholds. Thus, the strength of the crossmodal influences can be denoted with the following crossmodal parameters: CM_{AVexc} (audio to video excitation), CM_{AVinh} (audio to video inhibition), CM_{VAexc} (video to audio excitation), CM_{VAinh} (video to audio inhibition), which are implemented as a proportional change in the usual PSP_{Th} values as:

$$PSP_{ThNew} = PSP_{ThOld}\,(1 + CM_{exc\,/\,inh}) \qquad \textbf{(5.1)}$$

where $CM_{exc/inh}$ is negative for crossmodal excitatory influence and positive for inhibitory influence.

In the simplest case, setting crossmodal coupling parameters to zero effectively means that each modality is processed separately, with a simple OR/AND fusion of opinions. Increasing the absolute value of crossmodal coupling parameters effectively increases the crossmodal influences.

Note that the definition of supramodal layer here is related only to the layer that effectively combines sensory information to make the final decision. It does not include all the areas where multisensory neurons are located. L3 neurons of the visual system and L1 neurons of the auditory system, despite being multisensory neurons, are considered a part of the individual pathways outside the supramodal layer. Thus, individual pathways could be more appropriately named as "mainly" visual and "mainly" auditory pathways.

5.2.4 Network dynamics during test

Figure 5.9 illustrates the behaviour of the network over time. The dynamic behaviour of the integrated network is described as follows: each frame of the visual and auditory excitation (frames f1, f2,…, fN) are propagated through their corresponding individual architectures until the supramodal layer. Spikes of a given visual frame are propagated to L2 and L3 until a neuron belonging to a L3 map emits the first output spike, which is propagated to L4. L4 neurons accumulate opinions over several frames, whereas L1, L2 and L3 neurons are reset to their resting potential on a frame basis. The same occurs with auditory frames. Spikes are propagated to L1 neurons until a L1 neuron emits the first output spike, which is propagated to L2. L2 neurons accumulate opinions over several frames whereas auditory L1 neurons are reset to their resting potential before each frame is processed.

When auditory L2 neurons and/or visual L4 neurons release an output spike, the spikes are propagated to the supramodal layer. If there is no output spike in any visual L4 neuron and a visual L3 neuron has emitted a spike or there are no more spikes to be processed, the next visual frame can be propagated. In a similar fashion, if there is no output spike in any auditory L2 neuron and an auditory L1 neuron has emitted a spike or there are no more spikes to be processed, the next auditory frame can be propagated.

Visual L4 neurons and auditory L2 neurons retain their PSP levels that are accumulated over consecutive frames, until a class is recognized with an L4 neuron output spike or until there are no more frames to be processed. Crossmodal influences, if existent, are propagated synchronously before a new frame is processed. The crossmodal influence starts when one individual modality produces a result (output spike in a auditory L2 neuron or in a visual L4 neuron) and lasts until the processing is completed in all modalities.

In this model, the processing time for auditory and visual frames are considered the same, i.e., the supramodal layer receives synchronous information in a frame basis, although it is well known that auditory stimuli are processed faster than visual (Stein and Meredith, 1993).

Note that when resetting the PSP in the visual L2 and L3 neurons and auditory L1 neurons in each frame, information about dynamic changes of the patterns are lost, i.e., the model does not keep track of the variations of a visual pattern nor how the pattern changes over time. Each visual frame is considered independently and the last layer of each individual modality effectively accumulates opinions about whether it is a trained pattern.

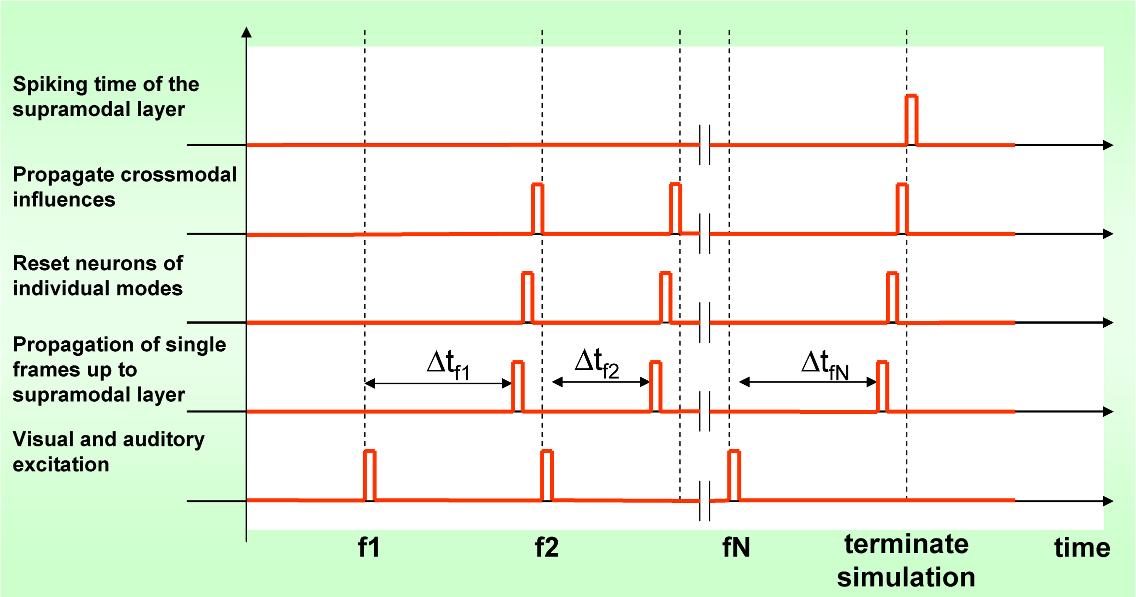

Spiking time of the supramodal layer
Propagate crossmodal influences
Reset neurons of individual modes
Propagation of single frames up to supramodal layer
Visual and auditory excitation
Δt_{f1}
Δt_{f2}
Δt_{fN}
f1
f2
fN
terminate simulation
time

Fig. 5.9. Typical behaviour of the integrated SNN architecture over time. The visual and auditory excitation (frames f1, f2,…, fN) are propagated through their corresponding individual architectures until the supramodal layer. Neurons of individual modalities are reset to their resting potential, namely L1, L2 and L3 neurons of the visual and L1 neurons of the auditory architecture. Crossmodal influences are propagated and a new frame is processed. The simulation is terminated when the supramodal layer spikes, both individual modes have released their opinions or there are no more frames to be processed.

With respect to the processing speed, in principle, the crossmodal connections decrease the time required to authenticate true claimants and increase the time needed to reject false claims when compared with a purely AND integration. In other words, it speeds up the processing of correlated information from different modes because once an individual modality finishes its analysis and labels a pattern, it exerts excitatory influence on the neurons of other modalities with the same label. The bias effect towards the second modality facilitates its decision in case true information about the claimant is also provided, which causes a resultant decision to be achieved quickly. On the other hand, the time needed to reject false claimants increases. Should the first modality results in a negative opinion about the claimant, the crossmodal connections send inhibitory signals to the claimant's neurons on other modalities, making its authentication harder. If the claimant provides true information on the second modality, due to the negative opinion given by the first modality, the second modality will be more rigorous on the authentication process, which consequently affects the time required to release the overall result.

5.2.5 Experiments and results

The integration of audiovisual modalities with a network of spiking neurons is evaluated with the VidTimit dataset (Sanderson and Paliwal, 2002), which contains video and audio recordings of 43 individuals. The same dataset is used in the experiments described in Chapter 3 and Chapter 4. The test setup deals specifically with the audiovisual person authentication problem. A person is authenticated based on spoken phrases and the corresponding facial information as the utterances are recorded (faces are captured in frontal view).

The following items present the configuration details of each individual system as well as the parameters used on the integration mechanism:

- **Visual**: Face detection is accomplished with the Viola and Jones algorithm (Viola and Jones, 2001) implemented in the OpenCV library (Intel OpenCV, 2007). Faces are converted into greyscale, normalized in size (height = 60 x width = 40), convolved with an elliptical mask, and encoded into spikes using rank order coding (Delorme *et al*, 2001). SNN does not require illumination normalization (Delorme and Thorpe, 2001). There are two scales of On/Off cells (4 L1 neuronal maps). In scale 1, the retina filters are implemented using a 3 x 3 Gaussian grid with $\sigma = 0.9$ and scale 2 uses a 5 x 5 grid with $\sigma = 1.5$. In L2, there are eight different directions in each frequency scale with a total of 16 neuronal maps. The directionally selective filters are implemented using Gabor functions with aspect ratio $\gamma = 0.5$ and phase offset $\varphi = \pi/2$. In scale 1 a 5 x 5 grid with a wavelength of $\lambda = 5$ and $\sigma = 2.5$ is used and in scale 2 a 7 x 7 grid with λ and σ set to 7 and 3.5, respectively. The modulation factor for the visual neurons was set to 0.995.
- **Auditory**: Speech signals are sampled at 16 kHz, and features extracted using standard MFCC with 19 MEL filter sub-bands ranging from 200 Hz to 7 kHz. Each MFCC is then encoded into spikes using rank order coding (Delorme *et al*, 2001). One receptive field neuron is used to represent each MFCC (19 input receptive fields). A specific background model is trained for each speaker model. For the sake of simplicity, the following procedure is applied: the background model of a speaker *i* is

trained using the same amount of utterances used to train the speaker model. The utterances are randomly chosen from the remaining training speakers. For the experiments, the number of neurons in the auditory L1 neuronal maps for the speaker and background model are defined *a priori* (50 neurons each). The modulation factor for auditory neurons is set to 0.9.

- **Integration**: The crossmodal parameters according to Equation 5.1 are set as: $CM_{AVexc} = CM_{VAexc} = 0.1$ and $CM_{AVinh} = CM_{VAinh} = 0$. Results that do not take into account the crossmodal coupling are also presented, i.e., $CM_{AVexc} = CM_{VAexc} = CM_{AVinh} = CM_{VAinh} = 0$, which effectively correspond to AND or OR integration.

The system is trained to authenticate 35 persons using six utterances from each individual. To train the visual part, only two frames from each individual are used, collected when uttering two distinct phrases from the same recording session were uttered.

The test uses two phrases (each phrase corresponding to one sample) recorded in two different sessions, therefore 35 users x 2 samples = 70 positive claims. Acting as impostors, the eight remaining users attempt to deceive each of the 35 users' models with two utterances, which give a total of 560 false claims.

The test is carried out frame-by-frame keeping the time correspondence between speech and visual frames. However, to speed up the computational simulations, the visual frames are downsampled. Five visual frames per second are used whereas the speech samples have a rate of 50 frames per second (Figure 5.10). The downsampling of the visual frames does not affect the performance, as for a period lower than 200 ms no substantial differences between one facial posture and another can be noticed in the VidTimit dataset.

Figure 5.11 shows typical input streams to the SNN-based audiovisual person authentication system, where frames of detected faces are sampled at 200 ms (5 frames/second) and 19 MFCC extracted from the detected speech parts are processed every 20 ms (50 frames/second).

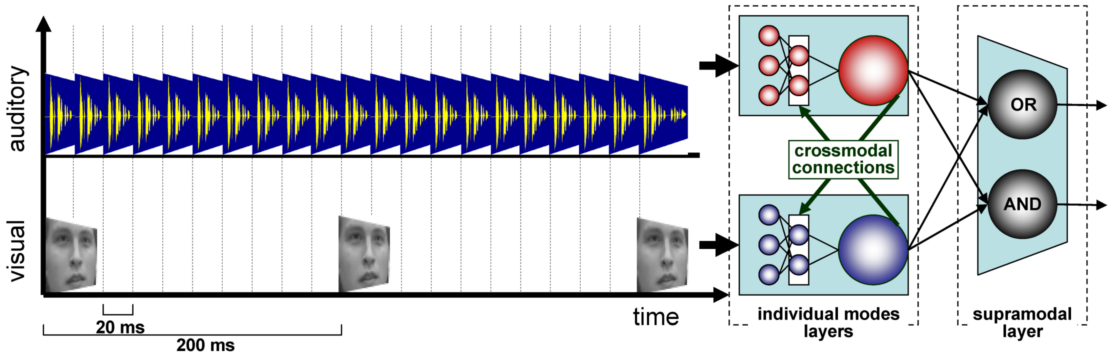


Fig. 5.10. Frame-based integration of modalities.

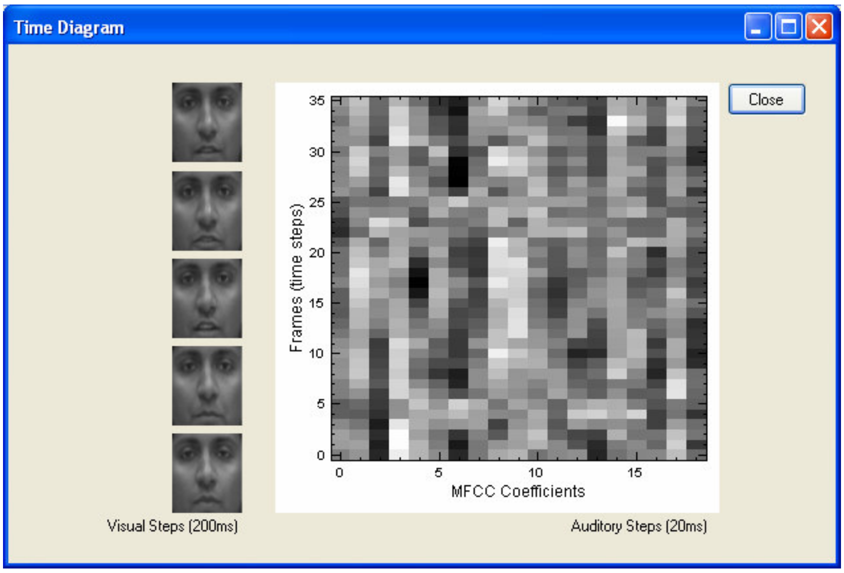


Fig. 5.11. Input data streams used on audiovisual person authentication. On the left, frames of faces sampled at 200 ms. On the right, 19 coefficients of MFCC features are extracted every 20 ms of speech.

The supramodal layer and the crossmodal coupling are updated when an individual modality outputs a spike, which may occur once in every frame (see Section 5.2.4 which describes the dynamic behaviour of the network). Here, it is assumed that the processing time for one frame is the same, regardless of the modality, although it is well known that auditory stimuli are processed faster than visual (difference of approximately 40 to 60 ms (Stein and Meredith, 1993)).

For the speech mode, the number of opinions to validate a person is set proportionally to the size of a given utterance (20% of the total number of frames in an utterance is used). For the visual mode, the number of opinions to authenticate a person is set to two (two frames). Figure 5.12 shows the best performance obtained on each individual modality. While the best total error (TE) for the face authentication is 21%, the auditory authentication is TE ≈ 38% (varying values of L1 PSP_{Th} in the auditory system and L3 PSP_{Th} in the visual system).

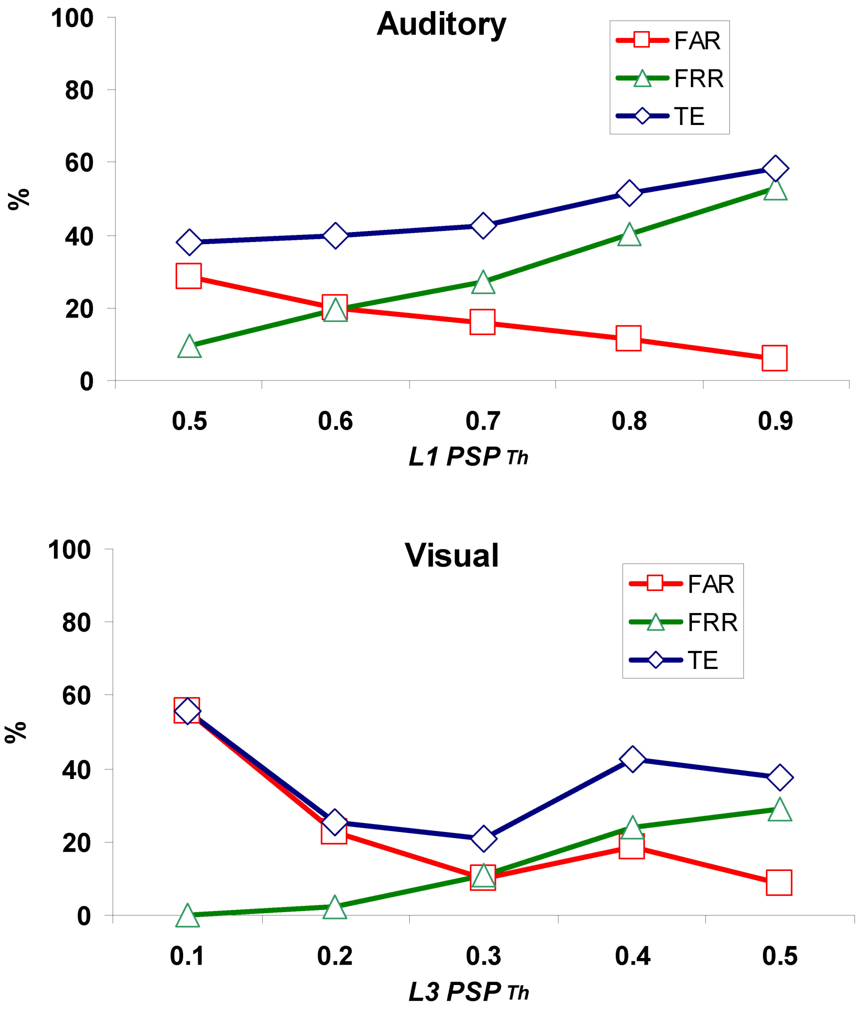


Fig. 5.12. Performance of individual modalities for different values of auditory (L1 PSP_{Th}) and visual parameters (L3 PSP_{Th}). Top: auditory system. Bottom: visual system. FAR is the false acceptance rate, FRR is the false rejection rate and TE is the total error (FAR + FRR).

Figure 5.13 shows the best performance of the system considering the type of integration held in the supramodal layer. First, the crossmodal coupling parameters are set to zero, simulating only the OR and AND integration of individual modalities done by the supramodal layer. Then, the crossmodal coupling is made active ("Crossmodal AND"), setting CM_{AVexc} = CM_{VAexc} = 0.1 and CM_{AVinh} = CM_{VAinh} = 0. The same parameters are used for individual modalities in this experiment, i.e., auditory parameters (L1 PSP_{Th}) and visual parameters (L3 PSP_{Th}) ranging from [0.5, 0.9] and

[0.1, 0.5], respectively. The x-axis represents different combinations of L1 and L3 PSP_{Th} ordered according to the performance.

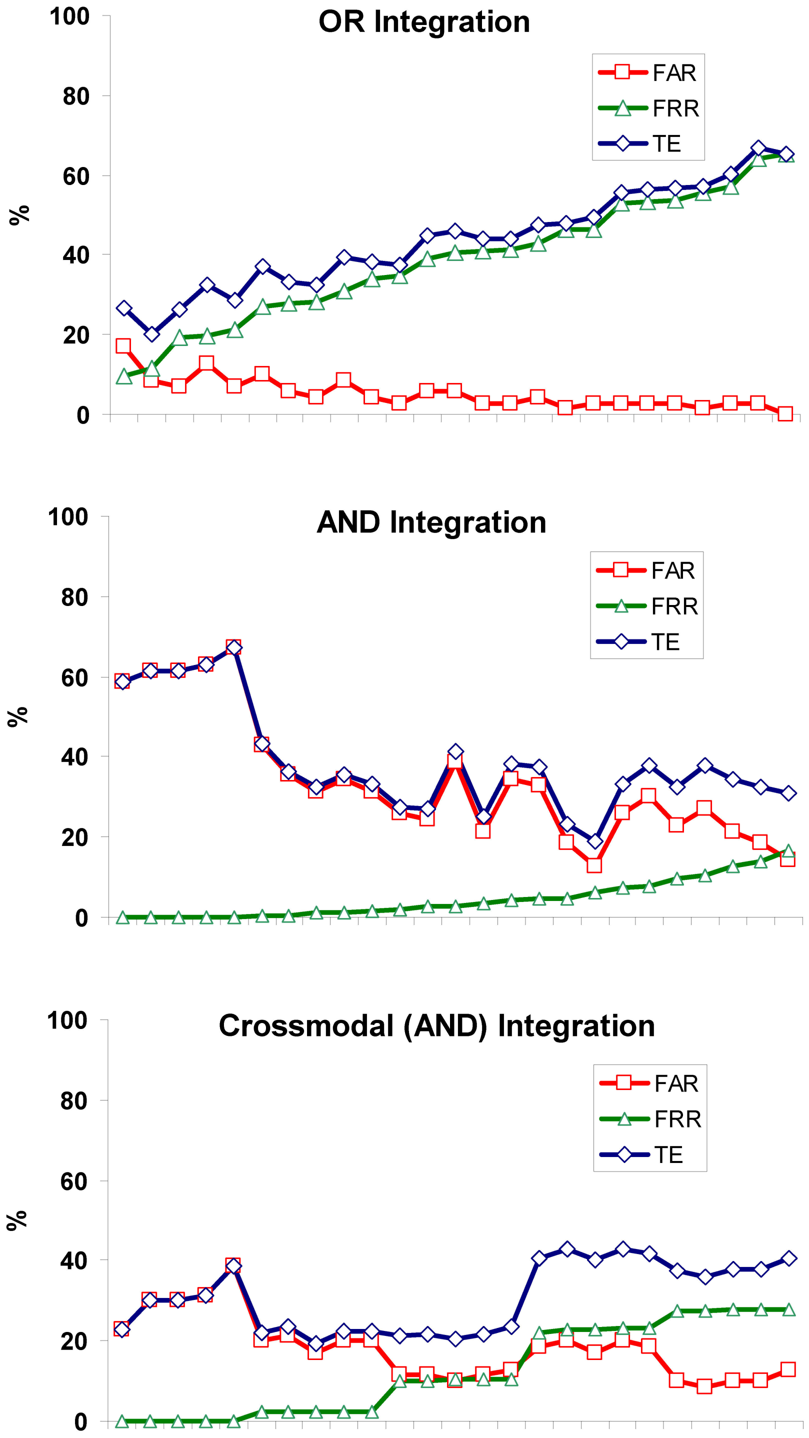


Fig. 5.13. Performance of the OR and AND integration of modalities with a supramodal layer of spiking neurons (upper and middle graphs, respectively). The bottom graph, when excitatory crossmodal influences are activated "Crossmodal AND" (for auditory L1 PSP_{Th} and L3 PSP_{Th} ranging from [0.5, 0.9] and [0.1, 0.5], respectively).

Figure 5.14 shows the potential advantages of the integration module. When the system needs to operate with low FAR levels (below 10%), AND and "Crossmodal AND" provide lower FRR than any singular modality. When the system is required to operate with low FRR (below 10%), OR integration can be used instead, providing lower FAR for the same FRR levels.

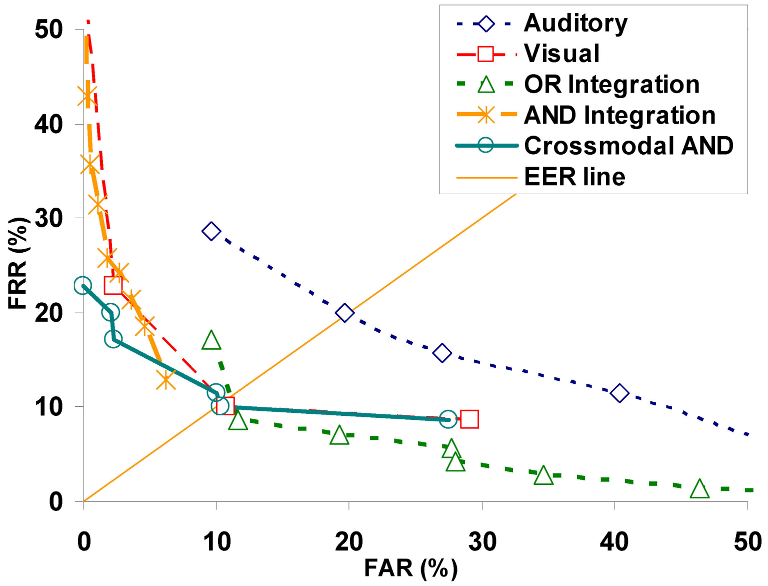


Fig. 5.14. Comparison between individual modes (auditory and visual) and the corresponding integration. Overall, the integration presents better performance than individual modes. OR, AND, "Crossmodal AND" alternate in the best position for different operating points. EER is the equal error rate (where FAR = FRR).

In another scenario, the influence of crossmodal connections on the integrated system is evaluated. A subset of the VidTimit dataset is used for this purpose. The setup for training is composed of six utterances from 10 individuals, whereas 12 individuals (10 that participated in the training stage and two completely unknown individuals) are used for testing. Each of the 10 individuals has 4 attempts at the test in a total of 40 positive claims. Acting as impostors, two individuals attempt to authenticate each of the 10 trained models four times, with a total of 2 x 4 x 10 = 80 impostor attempts (false claims). Similar to the previous experiments, the authentication threshold is set proportionally to the size of an utterance (20% of the total number of frames needs to provide positive opinions) and only two visual frames are necesary to authenticate a person based on the face. Figure 5.15, shows the performance of the integrated network for different values of crossmodal excitation. From the graph it is not possible to detect the best crossmodal parameter values, which means that a range of parameter values can be used with the same result. However, once again is clear that OR integration works better for high FAR than any single modality, and AND integration works better for low FRR than any single modality.

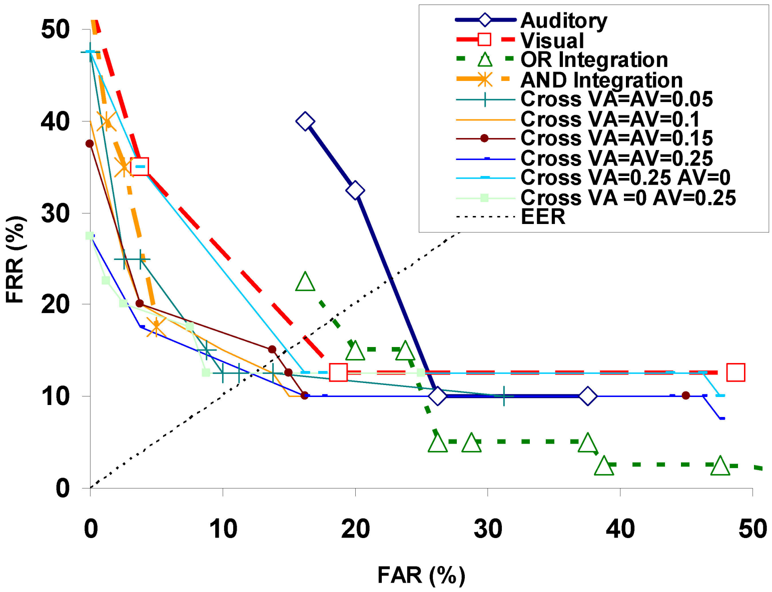
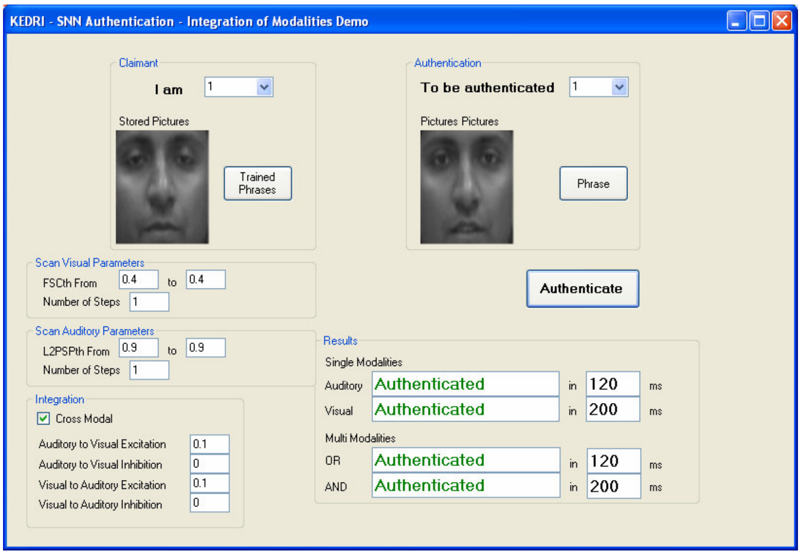


Fig. 5.15. Performance of the network for different values of crossmodal excitation. There is a range of values of crossmodal influence for which the model gives similar performance, however, for all values, the integration presents better performance than individual modes and ANDs and OR configuration alternate as the best choices for different operating points.

5.2.6 Implementation considerations

The integrative system, i.e., the supramodal layer and crossmodal connections, is implemented in C# language with a mix of *event* and *clock-driven* technique (see Section 3.6). The SNNs architecture for visual modality is implemented in C++ as explained in Section 3.6 and the auditory system in C# (Section 4.3.7).

Figure 5.16 illustrates a common usage of the audiovisual system for pattern recognition. A claimant presents its identification details (e.g., name). On the top left side of the user interface the claimed identification details as well as the data recorded during the training phase related to the claimed identity is shown. Multiple frames of the claimant's face are then recorded when the claimant utters a certain sentence (top right side of the user interface). After running the authentication procedure, the result is displayed according to different criteria (single or multimodalities). In Figure 5.16 the claimant is correctly authenticated whereas in Figure 5.17 one example where an individual tries to authenticate with a false identity (false claimant) can be seen. This is promptly detected by the system.


KEDRI - SNN Authentication - Integration of Modalities Demo

Claimant

I am 1

Stored Pictures

Authentication

To be authenticated 1

Pictures Pictures

Scan Visual Parameters

FSCth From 0.4 to 0.4

Number of Steps 1

Scan Auditory Parameters

L2PSPth From 0.9 to 0.9

Number of Steps 1

Integration

☑ Cross Modal

Auditory to Visual Excitation 0.1

Auditory to Visual Inhibition 0

Visual to Auditory Excitation 0.1

Visual to Auditory Inhibition 0

Results

Single Modalities

Auditory Authenticated in 120 ms

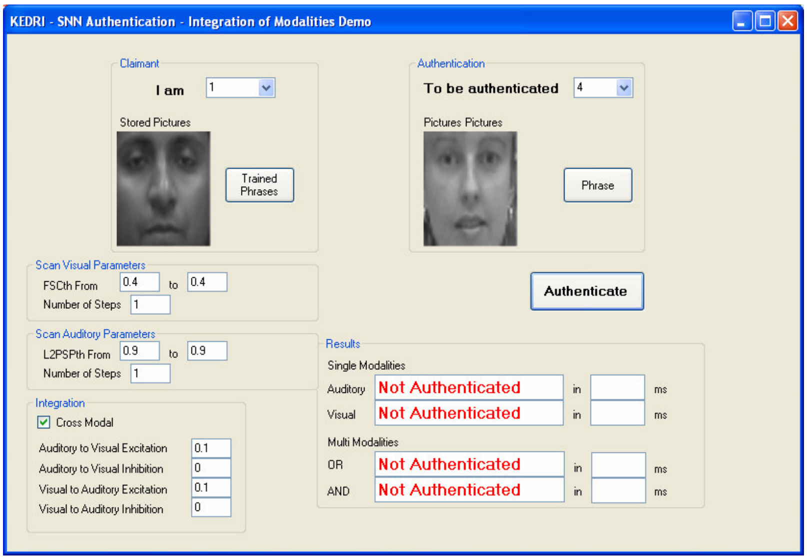
Visual Authenticated in 200 ms

Multi Modalities

OR Authenticated in 120 ms

AND Authenticated in 200 ms

Fig. 5.16. Example of correctly authenticated individual.


KEDRI - SNN Authentication - Integration of Modalities Demo

Claimant

I am 1

Stored Pictures

Authentication

To be authenticated 4

Pictures Pictures

Scan Visual Parameters

FSCth From 0.4 to 0.4

Number of Steps 1

Scan Auditory Parameters

L2PSPth From 0.9 to 0.9

Number of Steps 1

Integration

☑ Cross Modal

Auditory to Visual Excitation 0.1

Auditory to Visual Inhibition 0

Visual to Auditory Excitation 0.1

Visual to Auditory Inhibition 0

Results

Single Modalities

Auditory Not Authenticated in ms

Visual Not Authenticated in ms

Multi Modalities

OR Not Authenticated in ms

AND Not Authenticated in ms

Fig. 5.17. Example of successful detection of an impostor.

5.3 Chapter conclusion

This chapter covered the integration of modalities for the purpose of audiovisual pattern recognition. Of particular interest was the compilation of biological findings that inspire the proposal of models to explain the way brains effectively process and integrate different sensory information. Through an evaluation of several models and theories describing brain activity, the focus is given to the understanding of two properties that can be useful in enhancing artificial pattern recognition tasks, in particular:

- the supramodal area, and;
- crossmodal connections between modalities.

The second part of the chapter describes a new simple way to integrate modalities using fast spiking neurons (See also (Wysoski *et al*, 2007)). In the new system, each individual modality utilizes specialized adaptive SNNs, the same as presented in detail in Section 3.5 (visual) and Section 4.3.3 (auditory). The integration is done in a supramodal layer composed of multisensory neurons. In addition, one modality can influence another using a crossmodal mechanism.

The model also enables the strength of crossmodal connections to be set individually for each pair of single modes. In biology, audiovisual crossmodal learning has been experimentally observed in (McIntosh *et al*, 1998). In their experiments, after a training session with visual and auditory stimuli, when auditory stimuli alone were presented, areas of the visual cortex were also activated. In (Gonzalo *et al*, 2000) the areas of neuronal changes (time-dependent plasticity) that may be related to the crossmodal operations are further investigated. However, there was no attempt to quantify or to define the rules for neuronal changes. In this respect, new neuronal models for exploring the mechanisms that govern such activities can underpin new discoveries. In the model proposed in this chapter, a proper training procedure for crossmodal connections can be explored and evaluated.

The new model has several aspects that require further development, namely:

a) the model cannot take into account some biological behaviours detected by psychological experiments, e.g., cannot cover familiarity decisions, semantic information, identity priming, and within and across domain semantic priming (Burton *et al*, 1990) (Ellis *et al*, 1987) (Ellis *et al*, 1997).

b) with respect to the implementation, the use of frames and the respective frame-by-frame synchronization seems to be very artificial, truncating the natural flow of information. In addition, at this stage, the difference in processing time in each modality (Stein and Meredith, 1993) is ignored.

c) the model can not emulate the mechanism that facilitates unimodal recognition when the training is done with more than one modality, behaviour which has been described in (Kriegstein and Giraud, 2006).

Under the pattern recognition perspective, the network was tested on the person authentication problem. Experiments clearly showed that the integration of modes enhances the performance in several operating points of the system when the learning is done with the same training examples. For a comparative analysis, in (Sanderson and Paliwal, 2002), the integration of modalities is explored with the VidTimit dataset using a combination of mathematical and statistical methods. The auditory system alone, using MFCC features and GMM in a noise-free setup, reached TE (total error) = FAR (false acceptance rate) + FRR (false rejection rate) ≈ 22%. The visual system is reported to have TE ≈ 8% with features extracted using PCA (principal component analysis) and SVM (support vector machine) for classification. After testing several adaptive and non adaptive systems to perform integration, the best performance is obtained with a new approach that builds the decision boundaries for integration with consideration of how the distribution of opinions are likely to change under noisy conditions. The accuracy with the integration reached TE ≈ 6% involving 35 users for training and 8 users acting as impostors. Despite some differences between the experimental setup when compared to (Sanderson and Paliwal, 2002), the results shown in Figure 5.14 are clearly not as good. Nonetheless, to extract the best performance from the system and evaluate the crossmodal influence specifically on the pattern discrimination ability, an optimization mechanism needs to be incorporated. Similarly important is to explore different information coding schemes.

As pointed out in (Wysoski *et al*, 2008) and in Chapter 3 and Chapter 4, one of the promising properties of the computation with spiking neurons is that it enables the multi-criteria optimization of parameters according to accuracy, speed and energy efficiency. Since the integration is also based on spiking neurons, the optimization can be extended to cover the parameters used on integration as well (supramodal layer and crossmodal connection parameters).

Table 5.1 describes the integrative system in respect to the three specific aspects of *Brain-like* ECOS proposed in Section 2.3.3, namely information processing units, information processing pathways and learning ability.

Table 5.1. Analysis of integrative system according to three specific aspects of *Brain-like* ECOS.

Processing Units	Spiking neurons are used as processing units in the individual and integrative information processing areas.
Structure	The information of individual sensory modalities propagates with feed-forward connections into multiple layers composed of spiking neurons, representing the behaviour of various auditory and visual areas. Crossmodal connections and a supramodal layer integrate the systems (See Figure 5.8).
Learning	Online evolving procedures enable the learning of external stimuli through synaptic plasticity and structural adaptation separately for each modality. Algorithms to train the strength of crossmodal connections and weights of the supramodal layer still need to be designed.

Since the integrated system is composed of the unimodal systems described in Section 3.5 and Section 4.3.3, the improvements suggested in the end of the respective chapters are also valid here and will not be mentioned again. The following paragraphs mainly focus on the integrative properties that can be further evaluated.

The supramodal layer, as a first step, is implemented in this work with only two neurons. Two neurons were demonstrated to be sufficient to integrate incoming information from different modalities and to provide the system with complex dynamics that are difficult to evaluate analytically. In the simplest scenario, OR and AND integration has been simulated. Although a single neuronal unit can be interpreted as representing an entire ensemble of neurons, a more realistic implementation could be considered.

The underlying mechanisms that rule crossmodal activities remain the subject of further inquiry. The optimization of such connections and/or how to perform crossmodal learning is still an open field (a good introduction can be found in (Gonzalo *et al*, 2000) (McIntosh *et al*, 1998)). The experiments presented earlier in this chapter only illustrate as a proof-of-concept how the crossmodal connections can be set up in a network of spiking neurons. Further evaluation, such as sensitivity analysis with respect to different performance criteria, and exploration of the best values of crossmodal influence (excitatory and inhibitory) still deserve special attention.

Chapter 6 - Conclusion and future directions

In this final chapter, the main achievements of the research are summarised and the individual topics covered throughout the dissertation are related to the main objectives of the research. Several directions for future work are also discussed that would improve the performance and biological relevance of the research. Scalability and expansion to other sensory modalities as well as prospects for the design and its portability to lower level hardware platforms are also briefly covered.

The chapter is organized as follows: first, the main achievements of this dissertation are summarized from three perspectives: in terms of basic information processing units, information processing areas (inter and intra modality pathways), and system adaptability (learning rules). The section closes stressing the original contributions of this work and giving a summary of the experimental results. In a separate section other coding schemes that can be analysed in future extensions of this model are discussed. Several scalability aspects of the system are followed by a brief description of the olfactory and gustatory modalities, which are two natural choices for further testing the unimodal and integrative biological methods of processing. The dissertation closes with hardware implementation considerations and possible commercial applications of computational systems based on spiking neurons.

6.1 Introduction

A good number of systems have used the terms *biologically realistic* or *brain-like* to define a new generation of neural networks that try to process information in a way similar to the human brain. The main motivation for using brain-like systems is that artificial information processing systems, despite enormous effort, are still struggling to deliver general and reliable systems. Each application requires a uniquely tailored artificial system whereas the human brain effortlessly processes information, integrates sensory modalities, controls motor activities while taking care of vital involuntary functions using only a few kJoules of energy per hour.

Brain-like artificial information processing systems started to appear in attempts to reproduce the information pathways executed by the brain. The first attempts were mainly for the visual and auditory systems, perhaps because of the strong appeal for neuroscience and industrial applications. Historically, the visual and auditory systems have been the most studied of the sensory systems. This has resulted in a huge repository of information about visual and auditory sensory receptors and the pathways undertaken by the corresponding information. On the other hand, there is a strong industrial interest in more intelligent visual and acoustic computer systems in a wide variety of sectors dominated by giant conglomerates with huge budgets, e.g., car manufacturing, aerospace, medicine.

In (Fukushima and Miyake, 1982) (Mel, 1998) (Riesenhuber and Poggio, 1999) is described some models of visual system for the purpose of pattern recognition, which reused the Hubel and Wiesel model (Hubel and Wiesel, 1952) of the primary visual cortex with contrast, directionally selective and complex cells placed in an hierarchical pathway. Examples of auditory models can be found in (Ghitza, 1988) (Shamma *et al*, 1986).

Also under the *biologically realistic* label, many approaches showed how artificial systems could adapt and evolve in a intelligent and autonomous way. In this direction, networks of processing units learn what is the best structural configuration based on a few soft constraints and self-growing/shrinking procedures (see (Gallant, 1995) and (Kasabov, 2007) for extensive reviews on adaptive methods and procedures).

Thus, up to this point, there are *brain-like* models of network structures and *brain-like* ways to perform network connectivity and reconfiguration. Recently another factor has added to the momentum. The fact that neurons exchange information using spikes and processing mechanisms use action potentials is another addition to the *biologically realistic* realm (Hopfield, 1995). In (Gerstner and Kistler, 2002) this concept is properly clarified, stating that, in order to avoid any prior assumptions on neural computation, neurons need to process at the level of action potentials. Thus, spiking neurons and spiking neural networks (SNNs), historically used as a tool for neuroscientists to study the dynamics of single or ensembles of neuronal units, emerged as a new generation of neural network models for pattern recognition.

This dissertation unifies in a single pattern recognition system these three brain-like aspects. As a result, it presents an integrated system with **biologically inspired**

processing units arranged in **biologically inspired pathways** under **biologically inspired adaptive rules**. From these viewpoints, the main achievements of this dissertation are highlighted.

6.2 Summary of achievements

In a nutshell, in this work an integrated biologically inspired audiovisual pattern recognition system was designed and implemented. The system was applied to the person authentication problem. The main achievements of this work, classified according to information processing units, pathways and adaptive rules, are described in the next sections and summarized in Table 6.1.

Table 6.1 Summary of achievements.

Biologically inspired processing units	• Spiking neurons have been used throughout the system. SNN enables a more natural integration of feature extraction, decision-making and multiple modalities, where the processing time has meaning.
Biologically inspired pathways	• Major processing areas of the visual and auditory system have been implemented in a simplified manner with feed-forward connections. • New spike-based integration of modalities. Supramodal layer and crossmodal connections.
Biologically inspired adaptive rules	• Online learning through structural adaptation and synaptic plasticity.
Performance	• On the person authentication problem, individual systems (audio and visual) as well as the integrated approach (audiovisual) achieved similar results when compared to traditional methods.

6.2.1 Biologically inspired processing units

Striving to be closer to the biological way of processing, this work integrated several stages of information processing with a single type of processing unit. From the lower levels of sensory processing to the higher levels of cognition, a simple model of spiking neurons was used.

Additional to the biological appeal, SNNs enable a close integration of feature extraction and decision-making modules as well as the integration of multiple modalities. This close integration is mainly possible because the processing time has a meaning in spiking neuron systems. In other words, with spiking neurons, the time a spike takes to travel from one neuron to another can be explicitly set up. The generation of postsynaptic potential also occurs in time, set up through the excitatory/inhibitory time constants of a neuron (τ). These values can be set in accordance with biological measurements. Having the processing time of single units and the time spent in communication between units, the time taken by an area for processing can also be defined. This process can ultimately lead towards the simulation of an entire pathway where the information flows in a relevant time scale.

The implication of achieving information processing where the time matters for pattern recognition is that it breaks the existing hard separation between feature extraction and classification. Features are propagated as soon as they are processed and they can arrive at different times in areas where classification is undertaken. Similarly, processing time in different modalities vary. Thus, the individual modalities asynchronously feed a global decision-making process. Computation with real processing time also enables the implementation of crossmodal connections between modalities, where one modality can influence others according to its partial opinions in time. This phenomena can effectively increase overall accuracy (as proved to be the case in the human brain) or make the decision-making process faster (Stein and Meredith, 1993).

Note that, this work only attempts to point towards computing with meaningful processing time. However, in order to perform a realistic simulation of information processing where the processing time of different areas and pathways are biologically coherent, there are still some hurdles to overcome. There is a clear opportunity to use more elaborate spiking neuronal models, perhaps even to simulate neurons at the level of ionic channels. Further, in this dissertation only one information coding mechanism is evaluated (one spike per neuron where the highest importance is given to the first spike to arrive). An extension to this work is the reproduction of more natural patterns of spiking activity and other coding schemes.

6.2.2 Biologically inspired pathways

Neuroscientists have been drawing very accurate and detailed maps of the pathways taken by sensory information. In this research, a very simplified version of the major levels of processing is implemented.

For the visual system, the functional behaviour of retina cells, directionally selective cells and complex cells are implemented with a two-dimensional grid of spiking neurons. Only feed-forward connections are used and no adaptation at lower levels is applied.

In respect to the auditory speaker recognition process, features extracted from a functional model that resembles the characteristics of the human ear (MFCC) are used during the design and evaluation of the decision-making process for speech signals. A subsequent design using tonotopic organization of the spiking neurons (wavelet-based) is proposed that amounts to the entire processing of sound signals being undertaken with spiking neurons.

The integration of modalities is also accomplished with spiking neurons. Supramodal layers of spiking neurons as well as crossmodal connections were implemented.

As the biological pathways are more and more clearly understood, a more detailed description of the biological pathways can be incorporated into the model, e.g., the addition of redundant pathways, new layers, feedback connections, etc.

6.2.3 Biologically inspired adaptive rules

Biological systems are capable of life long functional and structural modifications, which enable learning of new tasks as well as memorization in an online fashion. Learning can occur in a supervised or an unsupervised fashion, such that changes can occur during sleep as well as with new external stimuli.

This work considers learning through structural adaptation and synaptic plasticity upon the event of external stimuli. The system automatically adds new classes, when in training mode, or further fine-tunes the training when new samples of a class are presented. The procedure is applied to two networks of spiking neurons that process visual and auditory information over multiple frames. In both cases, the learning procedure demonstrated its suitability, achieving results comparable with traditional methods (See Section 6.2.4).

In the future, the learning procedure can be further elaborated to reproduce memory consolidation and forgetting. On another front, it is necessary to define learning rules for integrative modules as well as a systematic procedures to train crossmodal connections.

6.2.4 Experimental results

In that which is concerned with quantitative analysis, the main highlights of this dissertation are:

1. **Visual system**. A SNN-based multi-view face authentication system demonstrated:
   a) the ability to adaptively learn from multiple frames. More frames for training of a class increased the accuracy. A peak in performance is reached after five frames.
   b) the ability of the system to accumulate opinions from several frames for decision-making. More test frames increased accuracy. The accuracy level flattens after five frames.
2. **Auditory system**. In the text-independent speaker authentication scenario using SNNs, the adaptive learning procedure was used to create speaker codebooks. Neuronal maps representing background models were also introduced to achieve similarity normalization. Two SNN architectures were proposed, which achieved similar levels of performance when compared with a traditional Linear Vector Quantization (LVQ) model to authenticate 43 users uttering short-sentences.
3. **Audiovisual system**. A supramodal area as well as crossmodal connections were used to process audiovisual features for person authentication. Different

configurations of the integrated system clearly outperformed individual modalities.

6.2.5 Original contributions

In this section, the original contributions of this work are highlighted and summarized. They are:

a) **Online adaptive learning procedure for SNNs of integrate-and-fire neurons**. The simple online learning procedure changes network structure and connection weights online, as new samples and/or classes are presented. The learning procedure is described in detail in Chapter 3 for visual pattern recognition (see also (Wysoski *et al*, 2006) (Wysoski *et al*, 2006a) (Wysoski *et al*, 2008), in Chapter 4 for auditory data analysis (see also Wysoski *et al*, 2007a), and in Chapter 5 for the integrated audiovisual system. Experiments demonstrated that the online procedure achieves comparable results with traditional methods.

b) **Extension of a visual system to perform multi-view pattern recognition**. An extended version of a visual system model described in (Delorme and Thorpe, 2001) has been designed and implemented. The main innovations are the incorporation of adaptive learning and integration of multiple views for pattern recognition (Chapter 3 and Wysoski *et al*, 2008). Experimental results demonstrate that decision-making based on multiple views increases performance when compared to single view.

c) **SNNs for processing auditory information**. A new network architecture to process auditory information using spiking neurons has been designed to recognize speakers. In addition to the adaptive learning procedure, the system has neuronal maps (representing background models) to normalize similarities. A detailed description is found in Chapter 4 and in (Wysoski *et al*, 2007a). Experimental results show that the new system is comparable to traditional methods.

d) **Integration of audiovisual information for pattern recognition using SNNs**. An integrative system combining the auditory and visual systems has been designed and implemented. The main novelties are the addition of a supramodal layer composed of spiking neurons as well as crossmodal connections linking modalities. A complete discussion of the integration can be found in Chapter 5 and in (Wysoski *et al*, 2007). Overall, the integrative system increases performance when compared to the single modalities.

6.3 Future directions

6.3.1 Scalability

There are several aspects that need to be considered in order to define the scalability of a multi-modal pattern recognition system. Particularly important is to evaluate how scalable the system is terms of number of training samples, number of classes to be authenticated, number of sensory modalities that can be integrated, or even in respect to the size or resolution of input patterns. This work concentrates on a detailed behavioural analysis of the newly proposed networks with a moderately low number of samples and classes. Despite some results clearly indicating that varying low-scale dataset sizes and class numbers does not have any significant effect on performance (see Figure 3.22 and Figure 3.23) and no theoretical limitation can be found in the learning procedures to adjust to new samples and new classes, the behaviour of the system must be analysed in large-scale problems, e.g., to authenticate thousands of individuals.

Two sensory modalities (auditory and visual) have been integrated in this work with the addition of supramodal layer and crossmodal connections. This approach can be extended to accommodate other modalities. The main challenge here is to develop procedures to find the optimal influence among modalities.

In order to reduce the insertion of redundant visual information, the visual region of interest was reduced in size. With the same purpose, auditory patterns are sampled at 16 kHz and 20 MEL filter banks are spread over the spectral range. These parameters were chosen to reach good accuracy and satisfy computational constraints. Being it required, the network can be further extended to accommodate larger visual region of interest and larger frequency range of auditory signals. Once again, there is no theoretical limitation for the network size, however processing speed and resources required in conjunction with the effects of the "curse of dimensionality" (Bishop, 2000) need to be considered.

6.3.2 Information coding methods

Spiking time theory (in opposition to the spiking rate theory) was used in this work for the conceptual design and implementation of algorithms. In particular, a spiking neuron model was used that privileges early spikes and a constraint that enabled the occurrence of only one spike per neuron (Thorpe, 1990). Based on these assumptions, concrete models were implemented and validated. However, as coding schemes utilized by the brain are still not clearly understood other spike-based coding mechanisms can be evaluated.

A good introduction to the issues related to the encoding of information in neuronal activity can be found in (Reece, 2001). A traditional theory, suggests that information is transmitted by firing rates (see (Gerstner and Kistler, 2002) (Mazurek and Shadlen, 2002)). This theory is gradually proving not to be sufficient, as several independent neurophysiological experiments demonstrate the existence of spike-timing patterns in both single and in ensembles of neurons. For instance, in (Villa *et al*, 1999), *in vivo* measurements enabled to the prediction of rat's behaviour responses through the analysis of spatio-temporal patterns of neuronal activity. Izhikevich (Izhikevich, 2006) created the term "polychronization" to define the spatio-temporal behaviour of a group of neurons that are "time-locked" to each other, a term to distinguish it from synchronous, asynchronous or polysynchronous spiking activity behaviour. Abeles, in 1982 (Abeles, 1982), first launched the term "synfire chains" to describe neuronal maps organized in a feed-forward manner with random connections between maps showing synchronous activity. This phenomenon has been experimentally verified in a series of independent works (See (Abeles and Gat, 2001)) and computational models explored the storage and learning capabilities of this theory (e.g., Bienenstock (1995) and (Gutig and Sompolinsky, 2006)). From all these theories, it is also reasonable to believe that different areas in the brain can utilize different coding schemes. If this is the case, combined approaches would be needed to better represent a given information pathway.

6.3.3 Olfactory and gustatory sensory modalities

Two natural choices for further extending the unimodal and integrative biological processing methods are the olfactory and gustatory sensory modalities.

Smell, similar to taste, is a chemical sense and does not maintain spatial relations with the input receptors. Olfactory discrimination capacity in humans varies greatly and can reach 5000 different odorants in trained people. The number of odorant receptors can reach 1000. Different to other neurons, olfactory sensory neurons have a relatively short life, between 30 to 60 days, then being replaced by newborn cells (Kandel, 2000).

In terms of information pathway, chemical stimuli are acquired by millions of olfactory sensory neurons. In the olfactory bulb, there is a convergence of sensory neurons to units called glomeruli (approximately 25000 to 1), organized in such a way that one glomerulus can receive only one type of receptor. Each odorant (smell) is recognized using several glomeruli. Glomeruli are not odour specific and a specific odour is described by a unique set of glomeruli. Glomeruli can be roughly considered to represent the neural image of the odour stimuli (Zigmond, 1999). The information is then sent to different parts of the olfactory cortex for odour recognition (Figure 6.1).

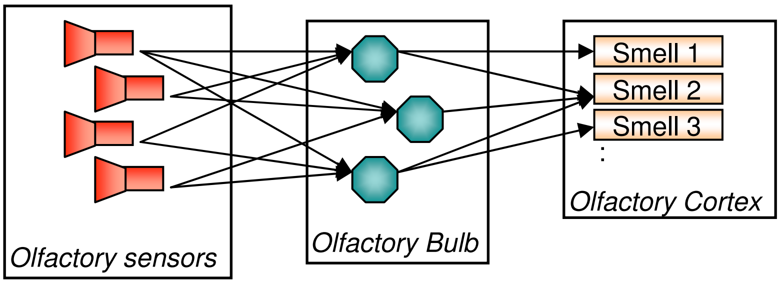


Fig. 6.1. Olfactory System Pathway. Olfactory chemical sensors collect information that is sent to the olfactory bulb and consequently to the olfactory cortex for recognition.

Taste, on the other hand, is much less sensitive than smell and humans can discriminate only five classes: sweet, sour, salt, bitter, and umami (taste receptor for glutamate). Gustatory cells are located in the taste buds (20 cells in each bud) with a lifespan of less than 7 days. Usually all buds present some response to all five tastes, to different degrees however. Sensitivity changes according to the position of the buds in the tongue (Figure 6.2). Thus, recognition of a taste pattern is based on a combination of the responses of receptors tuned to the five basic tastes. Despite the mechanism for information processing still not being completely known, it is believed that lateral inhibition occurs at the receptor level and the map, with a range of tastes, is transferred to another level of processing. As the information advances along the pathway, the neurons become more and more selective to particular tastes.

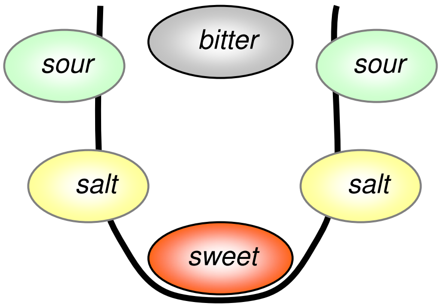


Fig. 6.2. Map of tastes on the tongue.

In terms of artificial implementation, artificial olfactory sensors (artificial noses) are widely available and include metal oxide sensors, polymer resonating sensors and optical bead sensors. There are many models aiming to describe the olfactory information pathways, from the oversimplified, in terms of biological plausibility, (e.g., White and Kauer, 1999) to the more complex and detailed (e.g. (Mombaerts *et al*, 1996)). Valova *et al* (Valova *et al*, 2004) provides an example of the use of spiking neural networks for the task.

Electronic tongues have been used for several practical applications, e.g., wine selection (Legin *et al*, 2003), water quality measurement (Lindquist and Wide, 2001), tea tasting (Iwarsson *et al*, 2001). An investigation of the use of spiking neural networks for taste recognition on two real-world datasets is provided in (Soltic *et al*, 2008).

As described in (Allen *et al*, 2002), an artificial system that processes olfactory and gustative information needs to have three main properties: the ability to process many sensor inputs, the ability to discern a large number of different odours, and handle noisy data. These are interesting properties to be tackled on a unimodal evolving spiking neural network system as well as with a combination of modalities where crossmodal interactions can be further explored.

6.3.4 Hardware implementation of evolving SNNs

The spiking model and network used in this dissertation has been carefully designed to consider computational constraints (Delorme and Thorpe, 2001). With a very simple processing unit (a simplified version of integrate-and-fire neurons) with a low level of activity (one spike per neuron) implemented with an event-based approach, networks of millions of neurons can be simulated in standard general purpose processing platforms (networks with up to 16 millions neurons in this work). However, as complexity is added to neural models, computational resources become bottlenecked, especially with the simulation of networks with large numbers of neurons. In this respect spiking neurons, in their various forms, are particularly attractive for hardware implementation mainly because communication using pulses can be associated with the binary properties of digital systems. With the maturation of FPGA (Field Programmable Gate Arrays) technology (Altera, 2007) (Xilinx, 2007), which at present can have several thousands of hardware logic units easily reconfigurable by software embedded, the prospect of implementing SNNs in a much faster way has emerged.

Since the SNN structures are composed mostly of processing units (neurons) placed in parallel, to each neuron a specific hardware area can be allocated, which results in parallelization at the hardware level. Obviously, hardware resources on FPGA are limited and the implementation of large SNNs purely in parallel is generally not possible, so finding a balance between speed and resources is required (Maguire *et al*, 2007). In (Glackin *et al*, 2005) an approach to implement large-scale SNNs in a limited hardware area is described. The trade-off between speed and resources required was reached by multiplexing the hardware model of a neuron in time. Other examples of successful implementation of SNN on FPGA include (Glackin *et al*, 2004) (Maya *et al*, 2000) (Ros *et al*, 2006).

6.3.5 Evolving SNNs for commercial applications

The new pattern recognition systems and algorithms based on spiking neurons evaluated in this dissertation require one additional iteration in the product development cycle. In order to reach a commercial scale, three important points need to be tackled. In order of design, implementation, and test stages, they are:

1) Development of efficient methods for automatic optimization of parameters. This will enable parameters to be easily adjusted for different applications and environmental conditions.

2) Optimization of code. After design, the experiments were implemented primarily considering code readability and reusability. Even though the implementation reached speed levels that enabled the experiments presented in this work to run on ordinary platforms, more efficient implementation is required. In addition, the algorithms could be wrapped into a comprehensive library with a Software Development Kit (SDK) to facilitate developer and end-user access.

3) Comparison of optimized SNN-based methods with traditional techniques on benchmarking datasets. For speaker authentication, for instance, NIST annually runs the NIST Speaker Recognition Evaluation (NIST-SRE, 2007) context. In these contexts, commercial application providers can test and compare their methods. Standard benchmark datasets for face recognition are FERET (Phillips *et al*, 2000) and XM2VTS (XM2VTS, 2007).

SpikeNet (SpikeNet, 2007) is a pioneering commercial product based on spiking neural networks for dealing with the visual pattern recognition problems on a range of applications. Using the fast type of integrate-and-fire neurons described in Section 3.3.1, it joins the biological concepts of computation with spikes developed from over 15 years of research with an efficient software implementation to reach real-time commercial requirements. SpikeNet is a good example of the maturation of the spiking neurons theory for solving engineering problems, proving that, indeed, spiking networks are already useful for real pattern recognition applications.

6.4 Overall conclusion

This dissertation presented new artificial systems to execute audiovisual pattern recognition that applies brain-like principles to the way information is processed. Following the research design, conceptual modelling and implementation was carried out after an evaluation of the state-of-the-art systems. The new models, as belonging to the class of *Brain-like* ECOS, are analyzed under three proposed perspectives: information processing units, information processing pathways, and adaptive learning. Experiments validated the applicability of the new models on a person authentication task.

References

Abbott, L. F., Decoding neuronal firing and modeling neural networks. Quart. Rev. Biophys. 27 (1994) 291-331

Abeles, M., Local Cortical Circuits: An Electrophysiological study. Springer, Berlin (1982)

Abeles, M., Gat, I., Detecting precise firing sequences in experimental data. Journal of Neuroscience. Methods. 107 (2001) 141-154

Achard, P., De Schutter, E., Complex parameter landscape for a complex neuron model. PLoS Computational Biology. 2 7 (2006) e94

Achard, P., van Geit, W., LeMasson, G., Parameter searching. In: Computational modeling methods for neuroscientists, De Schutter, E. (ed). (in press)

Adini, Y., Sagi, D., Tsodyks. M., Excitatory-inhibitory network in the visual cortex: Psychophysical evidence, Proceedings of the National Academy of Sciences, USA. 94 (1997) 10426-10431

Albiol, A., Torres, L., Delp, E. J., Optimum color spaces for skin detection, Proceedings of the IEEE International Conference on Image Processing. (2001) 122-124

Allen, J. B., How do humans process and recognize speech? IEEE Transactions on Speech and Audio Processing. 2 4 (1994) 567-577

Allen, J. N., Abdel-Aty-Zohdy, H. S., Ewing, R. L., Chang, T. S., Spiking networks for biochemical detection. 45th Midwest Symposium on circuits and systems. 3 4-7 (2002) 129-132

Altera. http://www.altera.com/. Last accessed on 03/12/2007

Ashford, J. W., Coburn, K. L., Fuster, J. M., Functional cognitive networks in primates, In: Parks, R. W., Levine, D. S., Long, D. L. (eds), Fundamentals of Neural Network Modeling. Neuropsycology and Cognitive Neuroscience, MIT Press, Cambridge. (1998)

AT&T Face Dataset. www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html. Last accessed on 03/12/2007

Atal, B.S., Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, Journal of the Acoustical Society of America. 55 (1974) 1304-1312

Atal, B.S., Hanauer, S. L., Speech analysis and synthesis by linear prediction of the speech wave, Journal of the Acoustical Society of America. 50 (1971) 637-655

Back, T., Evolutionary Algorithms in Theory and Practice: Evolutionary Strategies, Evolutionary Programming, Genetic Algorithms, Oxford University Press, New York. (1996)

Baldi, P., Heiligenberg, W., How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers, Biological Cybernetics. 59 (1988) 313-318

Bartlett, M. S., Information maximization in face processing, Neurocomputing. 70 (2007) 2204-2217

Bear, M. F., Connors, B. W., Paradiso, M. A., Neuroscience. Exploring the brain, 2 ed. Lippincott Williamns & Wilkins. (2001)

Becchetti, C., Ricotti, L. P., Speech recognition. Theory and C++ Implementation, John Wiley and Sons. (1999)

Beijsterveldt, C. E. M. van, Baal, G. C. M. van, Twin and family studies of the human electroencephalogram: a review and meta-analysis, Biological Psychology. 61 (2002) 111-138

Belhumeur, P., Hespanha, P., Kriegman, D., Eigenfaces vs fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence. 19 7 (1997) 711-720

Ben-Yacoub, S., Luetting, S., Jonsson, J., Matas, K., Kittler, J., Audio-visual person verification, Computer Vision and Pattern Recognition, Los Alamitos, California. (1999) 580-585

Benevento, L. A., Fallon, J., Davis, B. J., Rezak, M., Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey, Experimental Neurology. 57 (1977) 847-872

Benuskova, L., Kanich, M., Krakovska, A., Piriform cortex model of EEG has random underlying dynamics, In: Rattay, F. (ed), Proc. 1st World Congress on Neuroinformatics, Vienna. (2001) 287-292

Benuskova, L., Kasabov, N., Computational Neurogenetic Modeling, Springer-Verlag, New York. (2007)

Bhalla, U. S., Bower, J. M., Exploring parameter space in detailed single neuron models: Simulations of the mitral and granule cells of the olfactory bulb, Journal of Neurophysiology. 69 6 (1993) 1948-1965

Bialek, W., Rieke, F., Reliability and information transmission in spiking neurons, Trends in Neuroscience. 15 (1992) 428-434

Bienenstock E., A model of neocortex, Network: Computation in Neural systems. 6 (1995) 179-224

Bimbot, F. *et al.*, A tutorial on text-independent speaker verification, EURASIP Journal on Applied Signal Processing. 4 (2004) 430-451

Bimbot, F., Magrin-Chagnolleau, I., Mathan, L., Second-order statistical measure for text independent speaker identification, Speech Communication. 17 1-2 (1995) 177-192

Bishop, C., Neural networks for pattern recognition, Oxford, University Press, New York. (2000)

Bothe, S. M., Spiking Neural Networks, PhD Thesis, University of Leiden. (2003)

Bothe, S. M., La Poutre, H. A., Kok, J. N., Unsupervised clustering with spiking neurons by sparse temporal coding and multi-layer RBF networks, IEEE Transactions on Neural Networks. 10 2 (2002) 426-435

Bower, J. M., Beeman, D., The book of GENESIS: exploring realistic neural models with the General Neural Simulation System, Springer, New York. (1995)

Brette, R. *et al.*, Simulation of networks of spiking neurons: A review of tools and strategies, Journal of Computational Neuroscience. (2007 in press)

Bruce, C., Desimone, R., Gross, C. G., Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque, Journal of Neurophysiology. 46 (1981) 369-384

Bruce, V., Young, A., Understanding face recognition, British Journal of Psychology. 77 (1986) 305-327

Brunelli, R., Falavigna, D., Person identification using multiple cues, IEEE Transactions on Pattern Analysis and Machine Intelligence. 17 10 (1995) 955-966

Buchanan, M., Nexus: Small worlds and the groundbreaking theory of networks, Norton, W. W. & Company. (2003)

Burileanu, C., Moraru, D., Bojan, L., Puchiu, M., Stan, A., On performance improvement of a speaker verification system using vector quantization, cohorts and hybrid cohort-world models, International Journal of Speech Technology. 5 (2002) 247-257

Burton, A. M., Bruce, V., Johnston, R. A., Understanding face recognition with an interactive activation model, British Journal of Psychology. 81 (1990) 361-380

Buzsaki, G., Draguhn, A., Neuronal oscillations in cortical networks, Science. 304 (2004) 1926-1930

Calvert, G. A., Crossmodal processing in the human brain: insights from functional neuroimaging studies, Cerebral Cortex. 11 (2001) 1110-1123

Calvert, G. A., Campbell, R., Reading speech from still and moving faces. The neural substrates of visible speech, Journal of Cognitive Neuroscience. 15 (2003) 57-70

Calvert, G. A., Hansen, P. C., Iversen, S. D., Brammer, M. J., Detection of multisensory integration sites by application of electrophysiological criteria to the BOLD response, NeuroImage. 14 (2001) 427-438

Calvert, G. A. *et al.*, Activation of auditory cortex during silent lipreading, Science. 276 (1997) 593-596

Chellapa, R., Wilson, C., Sirohey, S., Human and machine recognition of faces: a survey, Proceedings of IEEE. 83 5 (1995) 705.741

Chevallier, S., Paugam-Moisy, H., Lemaitre, F., Distributed processing for modelling real-time multimodal perception in a virtual robot, International Multi-Conference Parallel and Distributed Computing and Networks, Innsbruck. (2005) 393-398

Chibelushi, C. C., Deravi, F., Mason, J. S. D., A review of speech-based bimodal recognition, IEEE Transactions on Multimedia. 4 1 (2002) 23-37

Chibelushi, C. C., Deravi, F., Mason, J. S. D., Adaptive classifier integration for robust pattern recognition, IEEE Transactions on Systems, Man, and Cybernetics - Part B. 29 6 (1999) 902-907

Coleman, J., Introducing speech and language processing, Cambridge University Press. (2005)

Connor, J. A., Stevens, C. F., Prediction of repetitive firing behaviours from voltage clamp data on an isolated neurone soma, Journal of Physiology. 213 (1) (1971) 31-53

Crepet, A., Paugam-Moisy, H., Reynaud, E., Puzenat, D., A modular neural model for binding several modalities, International Conference on Artificial Intelligence, ICAI. (2000) 921-928

Cristianini, N., Shawe-Taylor, J., An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press. (2004)

Cudeiro, J., Silito, A. M., Spatial frequency tuning of orientation-discontinuity-sensitive corticofugal feedback to the cat lateral geniculate nucleus, Journal of Physiology. 490 (Pt 2) (1996) 481-492

Darwin, C., The Origin of Species by Means of Natural Selection, John Murray, London. (1859)

Davis, S. Mermelstein, P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing. 28 (1980) 357-366

Deller Jr., J. R., Hansen, J. H. L., Proakis, J. G., Discrete-time processing of speech Signals, IEEE Press, Piscataway. (2000)

Delorme, A., Gautrais, J., van Rullen, R., Thorpe, S., SpikeNet: a simulator for modeling large networks of integrate and fire neurons, Neurocomputing. 26-27 (1999) 989-996

Delorme, A., Perrinet, L., Thorpe, S., Networks of integrate-and-fire neurons using Rank Order Coding, Neurocomputing. (2001) 38-48

Delorme, A., Thorpe, S., Face identification using one spike per neuron: resistance to image degradation, Neural Networks. 14 (2001) 795-803

Dhond, U. R., Aggarwal, J. K., Structure from stereo a review, IEEE Transaction on Systems, Man, and Cybernetics. 19 6 (1989) 1489-1510

Dimitrov, A. G., Miller, J. P., Neural coding and decoding: communication channels and quantization, Network: Computation in nervous systems. 12 4 (2001) 441-472

Dimitrov, D., Sidorov, I., Kasabov, N., Computational Biology, In: Rieth, M., Sommers, W. (eds), Handbook of theoretical and computational nanotechnology, American Scientific, Los Angeles. 1 (2004)

Dorogovtsev, S. N., Mendes, J. F. F., Evolution of networks: from biological networks to the Internet and WWW, Oxford University Press. (2003)

Ellis, A. W., Young, A. W., Hay, D. C., Modelling the recognition of faces and words, In: Morris, P. E. (ed), Modelling Cognition, Wiley, London. (1987)

Ellis, H. D., Jones, D. M., Mosdell, N., Intra- and inter-modal repetition priming of familiar faces and voices, British Journal of Psycology. 88 (1997) 143-156

EPFL BlueBrain. http://bluebrain.epfl.ch/. Last accessed on 03/12/2007

Eriksson, J. L., Villa, A. E. P., Artificial neural networks simulation of learning of auditory equivalence classes for vowels, International Joint Conference on Neural Networks, IJCNN, Vancouver. (2006) 1253-1260

Eriksson, J. L., Villa, A. E. P., Learning of auditory equivalence classes for vowels by rats, Behavioural processes. 73 (2006a) 358-359

Escobar, M. J., Ruiz-del-Solar, J., Biologically-based face recognition using Gabor filters and Log-Polar images. International Joint Conference on Neural Networks, IJCNN. 2 (2002) 1143-1147

FitzHugh R., Impulses and physiological states in theoretical models of nerve membrane. Biophysical Journal. 1 (1961) 445-466

Fletcher, H., The nature of speech and its interpretation, Journal of the Franklin Institute. 193 6 (1922) 729-747

Freund, Y., Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences. 55 1 (1997) 119-139

Fukunaga, K., Introduction to Statistical Pattern Recognition, Elsevier, Amsterdam. (1990)

Fukushima, K., Active vision: neural network models, In: Amari, S., Kasabov, N (eds), Brain like Computing and Intelligent Information Systems, Springer-Verlag. (1997)

Fukushima K., Miyake, S., Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition, In: Amari, S. Arbib, M. A. (eds), Competition and Cooperation in Neural Nets, Lecture Notes in Biomathematics. Springer-Verlag, Berlin, Heidelberg, New York. (1982) 267-285

Gallant, S. I., Neural network learning and expert systems, MIT Press, Cambridge, MA. (1995)

Ganchev, T., Speaker Recognition, PhD Thesis, Dept. of Electrical and Computer Engineering, University of Patras, Greece. (2005)

Garcia, C., Zikos, G., Tziritas, G., Wavelet packet analysis for face recognition, IVC. 18 4 (2000) 289-297

Genesis. http://www.genesis-sim.org/GENESIS/. Last accessed on 03/12/2007

Gerstner, W., Kistler, W. M., Spiking Neuron Models, Cambridge University Press, Cambridge, MA. (2002)

Gewaltig, M.-O., Evolution of Synchronous Spike Volleys in Cortical Networks. Network Simulations and Continuous Probabilistic Models, Springer-Verlag, Aachen, Germany. (2000)

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., Logothetis, N. K., Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex, Journal of Neuroscience. 25 (2005) 5004-5012

Ghitza, O., Auditory models and human performance in tasks related to speech coding and speech recognition, IEEE Transactions on Speech and Audio Processing. 2 1 Part II (1994) 115-132

Ghitza, O., Auditory nerve representation as a basis for speech processing, In: Furui, S., Sondhi, M. M., Advances in Speech Signal Processing, New York. (1992) 453-485

Ghitza, O., Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment, Journal of Phonetics. 16 (1988) 109-124

Ghobakhlou, A., Watts, M. J., Kasabov, N., Adaptive speech recognition with evolving connectionist systems, Information Science. 156 1-2 (2003) 71-83

Ghobakhlou, A., Zhang, D., Kasabov, N., An evolving neural network model for person verification combining speech and image. ICONIP, Lecture Notes in Computer Science. 3316 (2004) 381-386

Giraud, A. L., Price, C. J., Graham, J. M., Truy, E., Frackowiak, R. S., Cross-modal plasticity underpins language recovery after cochlear implantation, Neuron. 30 (2001) 657-663

Giraud, A. L., Truy, E., The contribution of visual areas to speech comprehension: A PET study in cochlear implants patients and normal-hearing subjects, Neuropsychologia. 40 (2002) 1562-1569

Glackin, B., Maguire, L. P., McGinnity, T. M., Intrinsic and extrinsic implementation of a bioinspired hardware system, Information Sciences. 161 1-2 (2004) 1-19

Glackin, B., McGinnity T. M., Maguire, L. P., Wu, Q. X., Belatreche, A., A novel approach for the implementation of large scale spiking neural networks on FPGA hardware, Lecture Notes in Computer Science. 3512 (2005) 552-563

Gold, B., Morgan, N., Speech and Audio Signal Processing, John Wiley & Sons. (2000)

Gonzalo, D., Shallice, T., Dolan, R., Time-dependent changes in learning audiovisual associations: a single-trial fMRI study, NeuroImage. 11 (2000) 243-255

Gray, R. M., Vector quantization, IEEE Acoustics, Speech, and Signal Magazine. (1984) 4-28

Guckenheimer, J., Gueron, S., Harris-Warrick, R. M., Mapping the dynamics of a bursting neuron, Philosophical Transactions of the Royal Society B. 341 (1993) 345-359

Gueorguieva, N., Valova, I., Georgiev, G., Learning and data clustering with an RBF-based spiking neuron, Journal of Experimental & Theoretical Artificial Intelligence. 18 1 (2006) 73-86

Gupta, A., Long, L. N., Character recognition using spiking neural networks. International Joint Conference on Neural Networks, Orlando, Florida. (2007) 53-58

Gutig, R., Sompolinsky, H., The tempotron: a neuron that learns spike timing-based decisions, Nature Neuroscience. 9 (2006) 420-429

Haller, M., Hyoung-Gook, K., Sikora, T., Audiovisual anchorperson detection for topic-oriented navigation in broadcast news, IEEE International Conference on Multimedia and Expo, Toronto. (2006) 1817-1820

Hebb, D. O., The Organization of Behavior, Wiley, New York. (1949)

Henkel, R. D., A simple and fast neural network approach to stereovision, Advances in Neural Information Processing Systems. (1997) 808-814

Henkel, R. D., Fast stereovision by coherence detection, International Conference on Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, New York. 1296 (1997a) 297-304

Hille, B., Ionic channels of Excitable Membranes, 2 ed., Sinauer Associates, Sunderland. (1992)

Hodgkin, A. L., Huxley, A. F., A quantitative description of membrane current and its application to conduction and excitation in nerve, Journal of Physiology. 117 (1952) 500-544

Holmberg, M., Gelbart, D., Ramacher, U., Hemmert, W., Automatic speech recognition with neural spike trains, Interspeech. (2005) 1253-1256

Holmes, J., Holmes, W., Speech synthesis and recognition, Taylor and Francis, New York. (2001)

Hopfield, J. J., Pattern recognition computation using action potential timing for stimulus representation, Nature. 376 6535 (1995) 33-36

Horn, B. K. P., Shape from shading: a method for obtaining the shape of smooth opaque object from one view, PhD Thesis, MIT. (1970)

Huang, X. D., Ariki, Y., Jack, M. A., Hidden Markov models for speech recognition, Edinburgh University Press, Edinburgh. (1990)

Hubel, D.H., Wiesel, T.N., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, Journal of Physiology. 160 (1962) 106-154

Izhikevich, E. M., Simple model of spiking neurons, IEEE Transactions on Neural Networks. 14 (2003) 1569-1572

Izhikevich, E. M., Polychronization: computation with spikes, Neural Computation. 18 2 (2006) 245-282

Intel OpenCV. http://www.intel.com/technology/computing/opencv/. Last accessed on 03/12/2007

Ivarsson, P., Holmin, S., Hojer, N.-E., Krantz-Rulcker, C., Winquist, F., Discrimination of tea by means of a voltametric electronic tongue and different applied waveforms, Sensors and Actuators B. 76 (2001) 449-454

Iwasa, K., Inoue, H., Kugler, M., Kuroyanagi, S., Iwata, A., Separation and recognition of multiple sound source using pulsed neuron model. ICANN, Lecture Notes in Computer Science. 4669 (2007) 748–757

Jenison, R. L., A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers, Journal of the Acoustical Society of America. 90 (1991) 773-786

Kakumanu, P., Makrogiannis, S., Bourbakis, N., A survey of skin-color modeling and detection methods, Pattern Recognition. 40 (2007) 1106-1122

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Principles of Neural Science, 4 ed., McGraw-Hill, New York. (2000)

Kanduce, D. C., Vaughan, J. W., Wallace, M. T., Benedek, G., Stein, B. E., Mechanisms of within- and cross-modality suppression in the superior colliculus, Journal of Neurophysiology. 78 (1997) 2834-2847

Kasabov, N., Evolving Connectionist Systems, Springer-Verlag. (2007)

Kasabov, N., Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering, MIT. (1996)

Kasabov, N., Benuskova, L., Computational neurogenetics, Journal of Computational and Theoretical Nanoscience. 1 (2004) 47-61

Kasabov, N., Benuskova, L., Wysoski, S. G., A computational neurogenetic model of a spiking neuron, IEEE International Joint Conference on Neural Networks, Montreal, Canada. 1 (2005) 446-451

Kasabov, N., Benuskova, L., Wysoski, S. G., Biologically plausible computational neurogenetic models: modeling the interaction between genes, neurons and neural networks, Journal of Computational and Theoretical Nanoscience. 2 (2005a) 569-573

Kasabov, N., Benuskova, L., Wysoski, S. G., Computational neurogenetic modeling: integration of spiking neural networks, gene networks, and signal processing techniques, ICANN, Lecture Notes on Computer Science, Springer-Verlag, Berlin Heidelberg. 3697 (2005b) 509-514

Kasabov, N., Postma, E., van den Herik, J., AVIS: A Connectionist-based framework for integrated auditory and visual information processing, Information Sciences. 123 (2000) 127-148

Kedri NeuCom. www.theneucom.com. Last accessed on 03/12/2007

Kiang, N. Y-S., Watanabe, T., Thomas, E. C., Clark, L. F., Discharge patterns of single fibers in the cat's auditory nerve, MIT Press, Cambridge (1965)

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 3 (1998) 226-239

Kittler, J., Matas, J., Johnsson, K., Ramos-Sanchez, M. U., Combining evidence in personal identity verification systems, Pattern Recognition Letters. 18 (1997) 845-852

Koch, C., Biophysics of Computation, Oxford University Press, New York. (1999)

Kriegstein, K. von, Eger, E., Kleinschmidt, A., Giraud, A. L., Modulation of neural responses to speech by directing attention to voices or verbal content, Cognitive Brain Research. 17 (2003) 48-55

Kriegstein, K. von, Giraud, A., Implicit multisensory associations influence voice recognition, PLoS Biology. 4 10 (2006) 1809-1820

Kriegstein, K. von, Kleinschmidt, A., Sterzer, P., Giraud, A., Interaction of face and voice areas during speaker recognition, Journal of Cognitive Neuroscience. 17 3 (2005) 367-376

Kruger, N., Lappe, M., Worgotter, F., Biologically motivated multi-modal processing of visual primitives, Interdisciplinary Journal of Artificial Intelligence the Simulation of Behaviors, AISB. 15 (2004) 417-428

Kuroyanagi, S., Iwata, A., Auditory pulse neural network model to extract the inter-aural time and level difference for sound localization, Transactions of IEICE. E77-D 4 (1994) 466-474

Lang, K. J., Witbrock, M. J., Learning to tell two spirals apart, Proceedings of the Connectionist Summer School Morgan Kauffman. (1988)

Legin, A., Rudnitska, A., Lvova, L., Vlasov, Y., Di Natale, C., D'Amico, A., Evaluation of italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception, Analytica Chimica Acta. 484 (2003) 33-44

Lewicki, M. S., Efficient coding of natural sounds, Nature Neuroscience. 5 4 (2002) 356-363

Lindquist, M., Wide, P., Virtual water quality tests with an electronic tongue, Proceedings of IEEE Instrumentation and Measurement Technology Conference, Budapest. (2001) 1320-1324

Loiselle, S., Rouat, J., Pressnitzer, D., Thorpe, S., Exploration of Rank Order Coding with spiking neural networks for speech recognition, IJCNN, Montreal. (2005) 2076-2080

Long, C. J., Datta, S., Wavelet based feature extraction for phoneme recognition, Proceedings of ICSLP. (1996) 264-267

Lysetskiy, M., Lozowski, A., Zurada, J. M., Invariant recognition of spatio-temporal patterns in the olfactory system model, Neural Processing Letters. 15 (2002) 225-234

Maass, W., Computing with spiking neurons, In: Maass, W., Bishop, C., Pulsed Neural Networks, The MIT Press, Cambridge, MA. 2 (2001) 55-85

Maass, W., Bishop, C. M. (eds), Pulsed Neural Networks, MIT Press. (2001)

Maciokas, J. B., Towards an understanding of the synergistic properties of cortical processing: a neuronal computational modeling approach, PhD Thesis, University of Nevada. (2003)

Maciokas, J., Goodman, P. H., Harris Jr., F. C., Large-scale spike-timing dependent-plasticity model of bimodal (audio/visual) processing, Technical Report of Brain Computation Laboratory, University of Nevada, Reno. (2002)

Maguire, L. P., McGinnity, T. M., Glackin, B., Ghani, A., Belatreche, A., Harkin, J. Challenges for large-scale implementations of spiking neural networks on FPGAs, Neurocomputing. 71 1-3 (2007) 13-29

Marian, I. D., A biologically inspired model of motor control of direction, PhD Thesis, University College, Dublin. (2002)

Marom, S., Abbott, L. F., Modeling state-dependent inactivation of membrane currents, Biophysical Journal. 67 2 (1994) 515-520

Markram, H., Lubke, J., Frotscher, M., Sakmann, B., Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs, Science. 275 (1997) 213-215

Martinez, A. M., Kak, A. C., PCA vs LDA, IEEE Transactions on Pattern Analysis and Machine Intelligence. 23 2 (2001) 228-233

Matsugu, M., Mori, K., Ishii, M., Mitarai, Y., Convolutional spiking neural network model for robust face detection. International Conference on Neural Information Processing, ICONIP. (2002) 660-664

Matsugu, M., Mori, K., Mitari, Y., Kaneda, Y., Subject independent facial expression recognition with robust face detection using a convolutional neural network, Neural Networks. 16 (2003) 555-559

Mattia, M., del Giudice, P., Efficient event-driven simulation of large networks of spiking neurons and dynamical synapses, Neural Computation. 12 10 (2000) 2305-2329

Maya, S, Reynoso, R., Torres, C., Estarda, M. A., Compact spiking neural network implementation in FPGA, Lecture Notes on Computer Science. 1896 (2000) 270-276

Mazurek, M. E., Shadlen, M. N., Limits to the temporal fidelity of cortical spike rate signals, Nature Neuroscience. 5 (2002) 463-471

McCulloch, W. S., Pitts, W. H., A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics. 5 (1943) 115-133

McIntosh, A. R., Cabeza, R. E., Lobaugh, N. J., Analysis of neural interactions explains the activation of occipital cortex by an auditory stimulus, Journal of Neurophysiology. 80 (1998) 2790-2796

McLennan, S., Hockema, S., Spike-V: an adaptive mechanism for speech-rate independent timing. IULC Working Papers Online 02-01

Mel, B. W., SEEMORE: Combining colour, shape, and texture histogramming in a neurally-inspired approach to visual object recognition, Neural Computation. 9 (1998) 777-804

Mercier, D., Seguier, R., Spiking neurons (STANNs) in speech recognition, 3rd WSES International Conference on Neural Networks and Applications, Interlaken. (2002)

Messer, K., Matas, J., Kittler, J., Luttin, J., Maitre, G., XM2VTSDB. The extended M2VTS database, 2nd International Conference on Audio-Video Based Biometric Person Authentication, Washington. (1999) 72-77

Mishra, D., A neural network using single multiplicative spiking neuron for function approximation and classification, Proceedings of International Joint Conference on Neural Networks, IJCNN, Vancouver, Canada. (2006) 1079-1086

Mombaerts, P., Wang, F., Dulac, C., Chao, S. K., Nemes, A., Mendelsohn, M., Edmondson, J., Axel, R., Visualizing an olfactory sensory map, Cell. 87 (1996) 675-686

Movellan J. R., Visual speech recognition with stochastic networks, In: Tesauro, G., Toruetzky, D., Leen, T. (eds), Advances in Neural Information Processing Systems. 7 (1995) 851-858

Mozayyani, N., Baig, A. R., Vaucher, G., A fully neural solution for on-line handwritten character recognition. (1998) 160-164

Nadim, F., Olsen, O. H., De Schutter, E., Calabrese, R. L., Modeling the leech heartbeat elemental oscillator. I. Interactions of intrinsic and synaptic currents. Journal of Computational Neuroscience. 2 (1995) 215-235

Nagumo J., Arimoto S., Yoshizawa S., An active pulse transmission line simulating nerve axon, Proceedings of IRE. 50 (1962) 2061–2070

Natschlager, T., Ruf, B., Pattern analysis with spiking neurons using delay coding, Neurocomputing. 26-27 (1999) 463-469

Natschlager, T., Ruf, B., Spatial and temporal pattern analysis via spiking neurons, Network: Computation in Neural Systems. 9 3 (1998) 319-338

Negnevitsky, M., Artificial intelligence. A guide to intelligent systems, Addison Wesley. (2002)

Nelson, M., Rinzel, J., The Hodgkin-Huxley model, In: Bower, J. M., Beeman, D. (eds), The book of GENESIS, Springer-Verlag, New York. 4 (1995) 27-51

Neuron. http://www.neuron.yale.edu/neuron/. Last accessed on 03/12/2007

NIST-SRE. http://www.nist.gov/speech/. Last accessed on 03/12/2007

Northmore, D. P. M., A network of spiking neurons develops sensorimotor mechanisms while guiding behavior, Neurocomputing. 58-60 (2004) 1057-1063

Olshausen, B. A., Field, D. J., Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research. 37 (1997) 3311-3325

Park, C., Choi, T. Kim, Y., Kim, S., Namkung, J., Paik, J., Multi-modal human verification using face and speech, IEEE International Conference on Computer Vision Systems, ICVS. (2006) 54-59

Patterson, R. D., Allerhand, M. H., Giguere, C., Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform, Journal of the Acoustical Society of America. 98 (1995) 1890-1894

Pavlidis, N. G., Tasoulis, O. K., Plagianakos, V. P., Nikiforidis, G., Vrahatis, M. N., Spiking neural network training using evolutionary algorithms, IEEE International Joint Conference on Neural Networks, IJCNN. 4 (2005) 2190- 2194

Perrinet, L., Samuelides, M., Sparse image coding using an asynchronous spiking neural network, European Symposium on Artificial Neural Networks, Bruges. (2002) 313-318

Phillips, P. J., Moon, H., Rauss, P. J., Rizvi, S., The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence. 22 10 (2000) 1090-1104

Poggio, T., Girosi, F., Regularization algorithms for learning that are equivalent to multilayer networks, Science. 247 (1990) 978-982

Porjesz, B., Almasy, L., Edenberg, H. J., Linkage disequilibrium between the beta frequency of the human EEG and a $GABA_A$ receptor gene locus, Proceedings of the National Academy of Sciences, USA. 99 (2002) 3729-3733

Potter, M., Short-term conceptual memory for pictures, Journal Exp. Psychology: Human Learning and Memory. 2 (1976) 509-522

Pouget, A., Deneve, S., Ducom, J-C., Latham, P.E., Narrow versus wide tuning curves. What is the best for a population code? Neural Computation. 11 1 (1999) 85-90

Prinz, A. A., Bucher, D., Marder, E., Similar network activity from disparate circuit parameters, Nature Neuroscience. 7 2 (2004) 1345-1352

Rabiner, L., Juang, B., Fundamentals of Speech Recognition, Prentice Hall, New Jersey. (1993)

Reece, M., Encoding information in neuronal activity, In: Maass, W., Bishop, C. (eds), Pulsed Neural Networks, MIT Press. (2001)

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., Speaker verification using adapted Gaussian Mixture Models, Digital Signal Processing. 10 (2000) 19-41

Riano, L., Rizzo, R., Chella, A., A new unsupervised neural network for pattern recognition with spiking neurons, Proceedings of International Joint Conference on Neural Networks, IJCNN, Vancouver, Canada. (2006) 7634-7641

Rieke, F., Warland, D., van Steveninck, R. R., Bialek, W., Spikes. Exploring the Neural Code, MIT Press. (1999)

Riesenhuber, M., Poggio, T., Hierarchical models of object recognition in cortex, Nature Neuroscience. 2 11 (1999) 1019-1025

Robert, A., Eriksson, J. L., A composite model of the auditory periphery for simulating responses to complex sounds, Journal of Acoustics Society of America. 106 4 (1999) 1852-1864

Ros, E., Ortigosa, E. M., Agis, R., Carillo, R., Arnold, M., Real-time computing platform for spiking neurons (RT-Spike), IEEE Transactions on Neural Networks. 17 4 (2006) 1050-1063

Rosenberg, A. E., Soong, F. K., Evaluation of a vector quantization talker recognition system in text independent and text dependent modes, Computer Speech and Language. 2 3-4 (1987) 143-157

Rosenblatt, F., Principles of neurodynamics, Spartan, New York. (1962)

Ross, A., Jain, A. K., Information fusion in biometrics, Pattern Recognition Letters. 24 13 (2003) 2115-2125

Rouat, J., Pichevar, R., Loiselle, S., Perceptive, non-linear speech processing and spiking neural networks, In: Chollet, G. *et al* (eds), Nonlinear speech modelling. Lecture Notes on Artificial Intelligence. 3445 (2005) 317-337

Rumelhart, D. E., Hinton, G. E., Williams, R. J., Learning internal representations by error propagation, In: Rumelhart, D. E., McClelland, J. L. (eds), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press/Bradford Books, Cambridge, MA. 1 (1986) 318-363

Sanderson, C., Paliwal, K. K., Identity verification using speech and face information, Digital Signal Processing. 14 (2004) 449-480

Sarikaya R., Hansen, H.L., High resolution speech feature parameterization for monophone-based stressed speech recognition, IEEE Signal Processing Letters. 7 (2000) 182-185

Schweinberger, S. R., Burton, A. M., Covert recognition and the neural system for face processing, Cortex. 39 (2003) 9-30

Seguier, R., Mercier, D., A generic pretreatment for spiking neuron. application on lipreading with STANN (Spatio-Temporal Artificial Neural Networks), 5th International Conference on Artificial Neural Networks and Genetic Algorithms. (2001)

Seguier, R., Mercier, D. Audio-visual speech recognition one pass learning with spiking neurons. Lecture Notes on Computer Science, Springer-Verlag, Berlin Heidelberg, New York. 2415 (2002) 1207-1212

Seneff, S., A joint synchrony/mean-rate model of auditory speech processing, Journal of Phonetics. 16 (1988) 55-76

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T., Robust object recognition with cortex-like mechanisms, IEEE Transactions on Pattern Analysis and Machine Intelligence. 29 3 (2007) 411-426

Shamma, S. A., Chadwick, R. S., Wilbur, W. J., Morrish, K. A., Rinzel, J., A biophysical model of cochlear processing: intensity dependence of pure tone responses, Journal of the Acoustical Society of America. 78 (1986) 1612-1621

Sharkey, A., Combining artificial neural nets: ensemble and modular multi-net systems, Springer-Verlag. (1999)

Sharpee, T. *et al*, Adaptive filtering enhances information transmission in visual cortex, Nature. 439 (2006) 936-942

Sigal, L., Sclaroff, S., Athitsos, V., Skin color-based video segmentation under time-varying illumination, IEEE Transactions on Pattern Analysis and Machine Intelligence. 26 7 (2004) 863–877

Smit, W., Barnard, E., Efficient coding leads to novel features for speech recognition, 15th Annual Symposium of the Pattern Recognition Association of South Africa, UCT Press. (2004) 99-102

Sohn, J-W., Zhang, B-T., Kaang, B-K., Temporal pattern recognition using a spiking neural network with delays, International Joint Conference on Neural Networks, IJCNN, Washington. (1999) 2590-2593

Soltic, S., Wysoski, S. G., Kasabov, N., Evolving spiking neural networks for taste recognition. IEEE World Congress on Computational Intelligence, Hong Kong. (2008) 2092-2098

Sonka, M., Hlavac, V., Boyle, R., Image Processing, Analysis, and Machine Vision, 2 ed., Brooks/Cole Publishing Company. (1999)

SpikeNet. http://www.spikenet-technology.com/. Last accessed on 03/12/2007

Stein, B. E., Meredith, M. A., The merging of the senses, MIT Press. (1993)

Stein, J. F., Stoodley, C. J., Neuroscience. An introduction, John Wiley and Sons. (2006)

Subramaniam, S., Biederman, I., Madigan, S., Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences, Visual Cognition. 7 4 (2000) 511-535

Thorpe, S., Spike arrival times: a highly efficient coding scheme for neural networks, In: Eckmiller, R., Hartman, G., Hauske, G. (eds), Parallel processing in neural systems, Elsevier. (1990) 91-94

Thorpe, S., Fize, D., Marlot, C., Speed of processing in the human visual system, Nature. 381 (1996) 520-522

Thorpe, S., Gaustrais, J., Rank Order Coding, In: Bower, J. (ed), Computational Neuroscience: Trends in Research. Plenum Press, New York. (1998)

Tino, P., Mills, A., Learning beyond finite memory in recurrent networks of spiking neurons, Neural Computation. 18 3 (2006) 561-613

Tsukada, M., Pan, X., The spatiotemporal learning rule and its efficiency in separating spatiotemporal patterns, Biological Cybernetics. 92 (2005) 139-146

Tufekci, Z., Gowdy, J., Feature extraction using discrete wavelet transform for speech recognition, Proceedings of South East Conference. (2000) 116–123

Tufekci, Z., Gowdy, J. N., Gurbuz, S., Patterson, E., Applied Mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition, Speech Communication. 48 (2006) 1294-1307

Turk, M., Pentland, A., Face recognition using eigenfaces, Proceeding of IEEE Conference on Computer Vision and Pattern Recognition. (1991) 586-591

Valentine, T., Bruce, V., Recognizing familiar faces: the role of distinctiveness and familiarity, Canadian Journal of Psycology. 40 (1986) 300-305

Valova, I., Gueorguieva, N., Kosugi, Y., An oscillation-driven neural network for the simulation of an olfactory system, Neural Computation and Applications, 13 (2004) 65-79

van den Berg, D., van Leeuwen, C., Adaptive rewiring in chaotic networks renders small-world connectivity with consistent clusters, Europhysics Letters. 65 4 (2004) 459-464

Vaucher, G., An algebraic interpretation of PSP composition, Biosystems. 48 (1998) 241-246

Vernon, D., Machine Vision, Prentice-Hall. (1991) 98-99, 214

Vezhnevets, V., Sazonov, V. Andreeva, A., A survey on pixel-based skin color detection techniques, Proceedings of the International Conference on Computer Graphics. (2003) 85–92

Villa, A. E., Tetko, I. V., Hyland, B., Najem, A., Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioned task, Proceedings of the National Academy of Sciences, USA. 96 (1999) 1106-1111

Viola, P., Jones, M. J., Rapid object detection using a boosted cascade of simple features, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1 (2001) 511-518

Watts, D. J., Small worlds: the dynamics of networks between order and randomness, Princeton University Press. (1999)

White, J., Kauer, J. S., Odour recognition in an artificial nose by spatio-temporal processing using an olfactory neuronal network, Neurocomputing. 26 27 (1999) 919-924

Wiskott, L., Fellous, J.-M., Kruger, N., von der Malsburg, C., Face recognition by Elastic Bunch Graph matching, In: Jain, L. C. *et al* (eds), Intelligent Biometric Techniques in Fingerprint and Face Recognition, CRC Press. (1999) 355-396

Wysoski, S. G., Benuskova, L., Kasabov, N., Simulation of neurogenetic models, Proceedings of Neuro-Computing and Evolving Intelligence (NCEI), Auckland. (2004) 30-31

Wysoski, S. G., Benuskova, L., Kasabov, N., On-line learning with structural adaptation in a network of spiking neurons for visual pattern recognition, ICANN, Lecture Notes in Computer Science, Springer-Verlag, Berlin. 4131 (2006) 61-70

Wysoski, S. G., Benuskova, L., Kasabov, N., Adaptive spiking neural networks for audiovisual pattern recognition, ICONIP, Lecture Notes in Computer Science, Springer-Verlag, New York. 4985 (2007) 406-415

Wysoski, S. G., Benuskova, L., Kasabov, N., Text-independent speaker authentication with spiking neural networks, ICANN, Lecture Notes in Computer Science, Springer-Verlag, New York. 4669 (2007a) 758-767

Wysoski, S. G., Benuskova, L., Kasabov, N., Fast and adaptive network of spiking neurons for multi-view visual pattern recognition, Neurocomputing. 71 13-15 (2008) 2563-2575

Wysoski, S. G., Zhang, D., Ghobakhlou, A., Shukla, V., Handheld adaptive speech and face based person verification, Proceedings of Neuro-Computing and Evolving Intelligence (NCEI), Auckland. (2004a) 97-98

Xilinx. http://www.xilinx.com/. Last accessed on 03/12/2007

XM2VTS. http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/. Last accessed on 03/12/2007

Yao, Y.-F., Jing,, X.-Y., Wong, H.-S., Face and palmprint feature level fusion for single sample biometrics recognition, Neurocomputing. 70 (2007) 1582-1586

Zhang, B.-L., Zhang, H., Ge, S. S., Face recognition by applying wavelet subband representation and kernel associative memory, IEEE Transactions on Neural Networks. 15 1 (2004) 166-177

Zhang, R., Tsay, P.-S., Cryer, J. E., Shah, M., Shape-from-shading: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 21 8 (1999) 690-706

Zhao, W., Chellappa, R., Phillips, P. J., Rosenfeld, A., Face recognition: a literature survey, ACM Computing Surveys. 35 4 (2003) 399-458

Zhu, L., Zhu, S., Face recognition based on orthogonal discriminant locality preserving projections, Neurocomputing. 70 (2007) 1543-1546

Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L., Squire, L. R., Fundamental neuroscience, Academic Press. 25 (1999)

Zuo, W., Wang, K., Zhang, D., Zhang, H., Combination of two novel LDA-based methods for face recognition, Neurocomputing. 70 (2007) 735-742