Neural Networks 23 (2010) 819–835

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Evolving spiking neural networks for audiovisual information processing

Simei Gomes Wysoski[a,*], Lubica Benuskova[a,b], Nikola Kasabov[a]

[a] Knowledge Engineering and Discovery Research Institute,[1] Auckland University of Technology, 1051 Auckland, New Zealand

[b] Department of Computer Science, University of Otago, Dunedin, New Zealand

ARTICLE INFO

Article history:
Received 7 July 2008
Received in revised form 26 January 2010
Accepted 27 April 2010

Keywords:
Spiking neural network
Audio and visual pattern recognition
Face recognition
Speaker authentication
Online classification

ABSTRACT

This paper presents a new modular and integrative sensory information system inspired by the way the brain performs information processing, in particular, pattern recognition. Spiking neural networks are used to model human-like visual and auditory pathways. This bimodal system is trained to perform the specific task of person authentication. The two unimodal systems are individually tuned and trained to recognize faces and speech signals from spoken utterances, respectively. New learning procedures are designed to operate in an online evolvable and adaptive way. Several ways of modelling sensory integration using spiking neural network architectures are suggested and evaluated in computer experiments.

1. Introduction

A number of systems have used the terms *biologically realistic* or *brain-like* to define a new generation of neural networks that attempt to process information in a way similar to the human brain. What mainly motivates researchers in this direction is that artificial information processing systems, despite enormous effort, still struggle to deliver general and reliable solutions. The majority of the attempts were on modelling the visual and auditory systems, perhaps because their inputs and outputs are better defined than for somatosensory or olfactory/gustatory systems and also because there is a strong interest in more intelligent visual and acoustic computer systems in a wide variety of industrial sectors (e.g., car manufacturing, aerospace, medicine, etc.).

Several models of visual systems used the Hubel and Wiesel model of the primary visual cortex with contrast, directionally selective and complex cells placed in a hierarchical pathway (Hubel & Wiesel, 1962) for the purpose of pattern recognition (Fukushima & Miyake, 1982; Mel, 1998; Riesenhuber & Poggio, 1999). Examples of brain-like auditory models can be found in Ghitza (1988) and Shamma, Chadwick, Wilbur, Morrish, and Rinzel (1986). Also under the *biologically realistic* label, many approaches showed how artificial systems could adapt and evolve in an 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; 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 principle that neurons can perform pattern recognition using spike timings is another addition to *biologically realistic* information processing (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 and exchange information at the level of spikes. 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.

Although SNNs are mathematically more complex than traditional artificial neural networks, they are potentially better suited for hardware implementation due to the "integrate-and-fire" nature of spiking neurons (Tikovic, Vöros, & Durackova, 2001). There are several advantages of the hardware implementations of spiking neurons, e.g. no multiplications as in traditional models, pulse processing can be implemented using shifts and adds, interconnections transmit only a single bit instead of real numbers. Sparse and asynchronous communication can also be easily implemented. However, it is important to note that this prospective advantage does not manifest itself yet when implementing SNNs in a general purpose computer platform.

As the theory of spiking neurons is currently most accepted to describe brain-like way of processing (Gerstner & Kistler, 2002), and is the most promising with respect to the future super-fast and reliable hardware implementations, it is the basis for all

[*] Corresponding author. Tel.: +64 9 526 4486.
E-mail address: wysoski@hotmail.com (S.G. Wysoski).

[1] http://www.kedri.info.

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.04.009

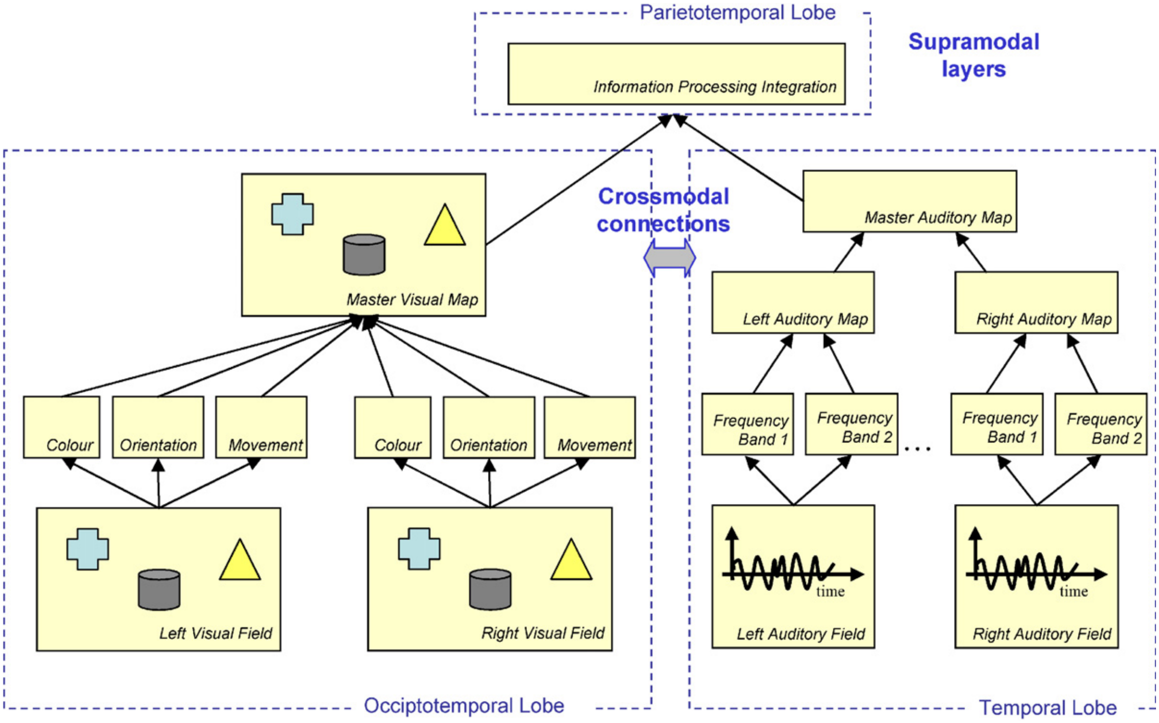

Fig. 1. Integration of sensory modalities of the visual system (left side), the auditory system (right side) and the subsequent integration of modalities (above).

new designs presented in this paper. At the systemic level, the behaviour of ensembles of neurons and the information processing pathways are also evaluated under the biological perspective. Of particular relevance to this research are the auditory and visual systems that are discussed separately in Sections 2 and 3. The auditory and visual pathways are considered in a new integrative audiovisual pattern recognition approach. The learning theories and the corresponding algorithms to implement them are discussed from the perspective of computation with spiking neurons.

The main original contributions of this paper are:

(a) Design of a new spiking neural network architecture to perform person authentication through the processing of signals from auditory and visual modalities. The integrative architecture combines opinions from individual modalities within a supramodal layer, which contains neurons sensitive to multiple sensory inputs. 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.
(b) Extension of adaptive online learning procedure to audiovisual pattern recognition. An online learning procedure that enables the system to change its structure by creating and/or merging neuronal maps of spiking neurons is presented and evaluated.
(c) Experimental evaluation of a new architecture that integrates sensory modalities on a person authentication problem, and comparison with traditional approaches.

A schematic illustration of the complexity of integrating the auditory and the visual senses is shown in Fig. 1. Each sensory modality has distinct pathways where information is processed. Within a sensory modality, information is decomposed, e.g., in the visual system, the information is divided into submodalities (colour, shape, motion, etc.) that are independently processed in different pathways. In the auditory system, the ventral cochlear nucleus with mainly tonotopical organization of cells and dorsal cochlear nucleus (mainly non-tonotopical) also define different pathways. In different modalities and submodalities, 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 in principle accurately described and reproduced in a computational model. Therefore, many simplifications and abstractions need to be introduced to the design of sensory integration models.

The visual and the auditory models, and the novel online learning procedure that were described in Wysoski, Benuskova, and Kasabov (2006, 2007, 2008a) are evaluated in Sections 2 and 3 of this paper. Section 4 explores the integration of modalities (with preliminary description presented in Wysoski, Benuskova, & Kasabov, 2008b), which is followed by experimental evaluation of the adaptive properties of the system. Section 5 concludes the paper and points to further directions to explore in order to achieve more biologically realistic and reliable pattern recognition systems.

2. The visual model

2.1. Short literature review

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 & 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 (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 analysed by Serre, Wolf, Bileschi, Riesenhuber, and Poggio (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 showed comparable performance with benchmark algorithms, like the constellation models, hierarchical SVM-based face detection, systems that use Ullman et al. 's fragments and GentleBoost.

In the same way, 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 demonstrated to be relatively insensitive to changes in the image plane and object orientation, fairly sensitive to changes in object scale and nonrigid deformation, and highly sensitive to the quality of the visual objects. 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, Lappe, & Worgotter, 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 (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 implemented this idea, by defining the pattern not only by the sequence of input spikes, but also by the exact firing time (Natschlager & Ruf, 1998, 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.

Maciokas, Goodman, and Harris (2002) goes down to the level of ion 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. Though, no attempts to rigorously test the classification abilities of the network have been made.

Thorpe, Fize, and Marlot (1996) suggest that in order to be coherent with the time measured in psychophysical experiments on fast perceptual classification, the information processing mechanisms can afford to have neurons 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 tens of 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, Gautrais, van Rullen, & Thorpe, 1999; Delorme & Thorpe, 2001). 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 is compared to image processing standards like JPEG (Perrinet & Samuelides, 2002).

Matsugu et al. utilized a different coding strategy in a hybrid of a convolutional and SNN architecture for face detection tasks (Matsugu, Mori, Ishii, & Mitarai, 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 method used standard error back-propagation algorithm for the bottom feature-detecting layer. 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, etc.) 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, Mori, Mitari, & Kaneda, 2003).

Next section follows the conceptual approach described in Delorme et al. (1999) and Delorme and Thorpe (2001), 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 comprised of hierarchical layers of neurons grouped in neuronal maps.

2.2. SNN architecture for visual information processing

Architecture, training and experimental evaluation of SNN for visual information processing was described in detail in Wysoski et al. (2008a) with some preliminary results reported in Wysoski et al. (2006). For the sake of completeness of this paper, we will repeat the main points here. The visual system uses the integrate-and-fire neuron model described in Delorme, Perrinet, and Thorpe (2001) and Delorme and Thorpe (2001). The neuron's excitation depends on the order of arrival of spikes and the postsynaptic potential (PSP) for neuron i at a time t is calculated as

$$\mathrm{PSP}(i,t) = \sum_j \mathrm{mod}^{\mathrm{order}(i,j)} w_{j,i} \qquad (1)$$

where mod $\in (0, 1)$ is the modulation factor, j is the index for the incoming connection and $w_{j,i}$ is the corresponding synaptic weight and order(i, j) is the order of spike arrival from neuron j to neuron i. For instance, setting mod $= 0.9$ and considering $w_{i,j} = 1$, the first spike to arrive (order$(i, j) = 0$) changes the PSP by $0.9^{(0)} = 1$. The second spike (order$(i, j) = 1$) 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

$$\mathrm{PSP}(i,t) \geq \mathrm{PSP}_{\mathrm{Th}}(i) \qquad (2)$$

where $\mathrm{PSP}_{\mathrm{Th}}$ is the postsynaptic firing threshold. The main advantages of these neurons are that they are computationally very inexpensive and they boost the importance of the first pre-synaptic spikes. The network structure, where neurons are placed in two-dimensional grids forming neuronal maps and consequently, layers of maps, also follows the same pattern as introduced in SpikeNet (Delorme & Thorpe, 2001). However, we have introduced few modifications related to online training on

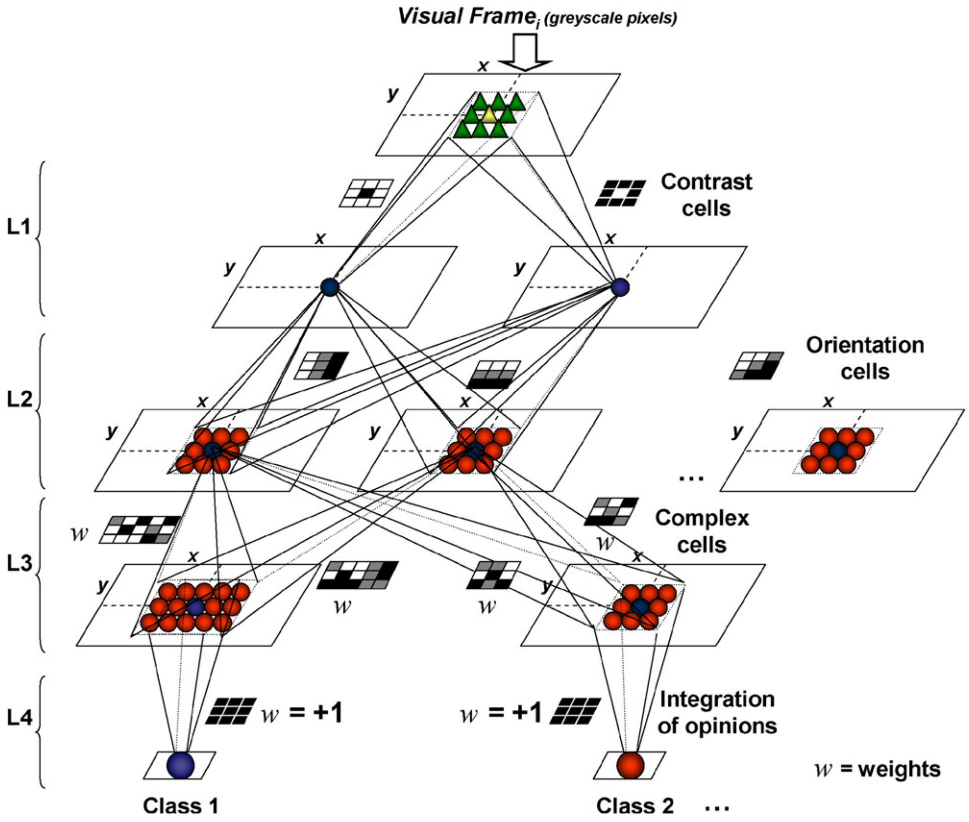

Fig. 2. Rank Order Coding. The amplitude of a signal is encoded over time. The higher the amplitude, the shorter the delay (and vice versa). Modified from Delorme et al. (2001).

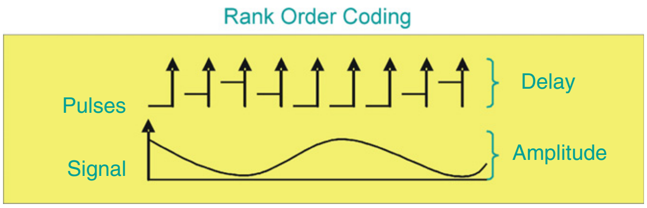

Fig. 3. Visual SNN architecture composed of four layers. Neurons in L1 and L2 are sensitive to image contrast and orientation, respectively. L3 has the complex cells, trained to respond to specific patterns. L4 accumulates opinions over different input excitations in time.

multi-view images, and later to multimodal integration. Thus, our SNN is composed of four layers of integrate-and-fire neurons (see Fig. 2). In the first two layers (L1 and L2) there is no learning, they simply act as passive filters and time domain encoders. Neurons in L1 represent the contrast 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, each pair tuned to a different frequency scale. Contrast 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. Eq. (3) describes the contrast filters, where g normalizes the sum of weight elements to zero and the maximum and minimum convolution values to $[+1, -1]$.

$$\nabla^2 G(x,y) = g\left(\frac{x^2+y^2-\sigma^2}{\sigma^4}\right) \mathrm{e}^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}. \tag{3}$$

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 (Delorme et al., 2001) (Fig. 3). L1 basically prioritizes the pixels with high contrast, which are consequently processed first and have a higher impact on neurons' PSP.

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

$$G(x,y) = \mathrm{e}^{\left(\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)} \cos\left(2\pi\frac{x'}{\lambda}+\varphi\right) \tag{4}$$

$$x' = x\cos(\theta) + y\sin(\theta)$$
$$y' = -x\sin(\theta) + y\cos(\theta)$$

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 $[0, \mathrm{PSP}_{\max}]$, regardless of the scale of the filters.

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 next section and in Wysoski et al. (2008a). An input pattern belongs to a certain class if a neuron in the corresponding neuronal map spikes first. There are lateral inhibitory connections between neuronal maps in the third layer L3, so that when a neuron fires in a certain map, other maps receive inhibitory pulses in an area centred in the same spatial position. Layer 4 (L4) has one neuronal map containing a single neuron for each pattern class. The L4 neuron of a given class is fed by the corresponding L3 neuronal maps. There are excitatory connections (typically $w = +1$) between neurons located close to the centre of L3 maps and the corresponding L4 neuron. In fact, L4 combines the results of a sequence of visual patterns, i.e. accumulates opinions from several frames. This layer and an online learning are new features in comparison with the original SpikeNet model (Delorme & Thorpe, 2001).

The connection weights between L3 and L4, in the simplest case, 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 employed. 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.

In terms of network dynamics, spikes of a given visual frame are propagated to L2 and L3 until any neuron belonging to any of L3 map emits the first output spike, which is consequently propagated to L4. If any neuron in L4 generates an output spike, the simulation is truncated and the frame is labelled to the corresponding class. The next frame follows. Otherwise, if there is no output spike in any L4 neuron or there are no spikes from L3, the next frame is propagated. The next frame starts to be propagated always only after resetting the PSPs and the order of arrival of spikes order(i, j) in L2 and L3 neurons. On the other hand, L4 neurons retain their PSP levels and the order of arrival of spikes order(i, j), to accumulate influences over consecutive frames, until a class is recognized with an L4 neuron output spike or until there are no more frames to be processed.

2.3. Visual pattern learning procedure

For the sake of a complete description of the model, we also summarise here the learning procedure presented in detail in Wysoski et al. (2008a). The learning procedure follows 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 $\mathrm{Map}_{C(k)}$ in L3 for sample k and train the weights using the equation:

$$\Delta w_{j,i} = \mathrm{mod}^{\mathrm{order}(i,j)} \tag{5}$$

3. 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(i, j) is the order of spike arrival from neuron j to neuron i.
4. The postsynaptic threshold ($\mathrm{PSP}_{\mathrm{Th}}$) 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 $\mathrm{Map}_{C(k)}$ with the propagation of the training sample into the updated weights, such that:

$$\mathrm{PSP}_{\mathrm{threshold}} = c \max(\mathrm{PSP}). \tag{6}$$

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).
5. Calculate the similarity between the newly created map $\mathrm{Map}_{C(k)}$ and other maps belonging to the same class $\mathrm{Map}_{C(K)}$. The similarity is computed as the inverse of the Euclidean distance between weight matrices.
If one of the existing maps for class K has similarity greater than a chosen threshold $\mathrm{Th}_{\mathrm{sim}C(K)} > 0$, merge the maps $\mathrm{Map}_{C(k)}$ and $\mathrm{Map}_{C(K\mathrm{similar})}$ using arithmetic average as

$$W = \frac{W_{\mathrm{Map}_{C(k)}} + N_{\mathrm{samples}} W_{\mathrm{Map}_{C(K\mathrm{similar})}}}{1 + N_{\mathrm{samples}}} \tag{7}$$

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 $\mathrm{PSP}_{\mathrm{Th}}$ is updated in a similar fashion as:

$$\mathrm{PSP}_{\mathrm{Th}} = \frac{\mathrm{PSP}_{\mathrm{Map}_{C(k)}} + N_{\mathrm{samples}} \mathrm{PSP}_{\mathrm{Map}_{C(K\mathrm{similar})}}}{1 + N_{\mathrm{samples}}}. \tag{8}$$

Note that the learning procedure updates W and $\mathrm{PSP}_{\mathrm{Th}}$ as well as enables 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 structure as well as different final W and $\mathrm{PSP}_{\mathrm{Th}}$. 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. The training can be summarised with the following pseudo-code:

```
For all samples in the training set
  For each sample
    Create a new map in L3
    Propagate the sample into the network
    through L1 and L2
    Train the newly created map using Equation 5
    and Equation 6
    Calculate the similarity between resulting
    weight vectors of
    newly created map and existent maps within
    L3
    If similarity > Threshold
       Merge newly created map with the most
       similar map using
       Equation 7 and Equation 8
```

Table 1 gives a summary of the main properties of the visual SNN.

2.4. Experimental evaluation of visual SNN

The system has been extensively evaluated on two datasets and compared with other methods (SVM, MLP and NN) in our previous experiments (Wysoski et al., 2006, 2008a). Here we present some relevant results on the visual part of the VidTimit dataset from (Wysoski et al., 2008a), in order to later appreciate a comparison with the integrated audiovisual setup. The visual system is trained on greyscale pictures of 35 individuals. For testing, 43 individuals are used, in such a way that 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. The online learning procedure was

Table 1
Properties of the visual information processing system.

Processing units	A fast and computationally inexpensive version of spiking neuron is used as processing unit in all stages of visual information processing (Delorme & Thorpe, 2001).
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, orientation selective cells, hierarchically higher complex cells).
Learning	The online evolving procedure allows ongoing learning of patterns 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 (see Wysoski et al., 2008b, for details).

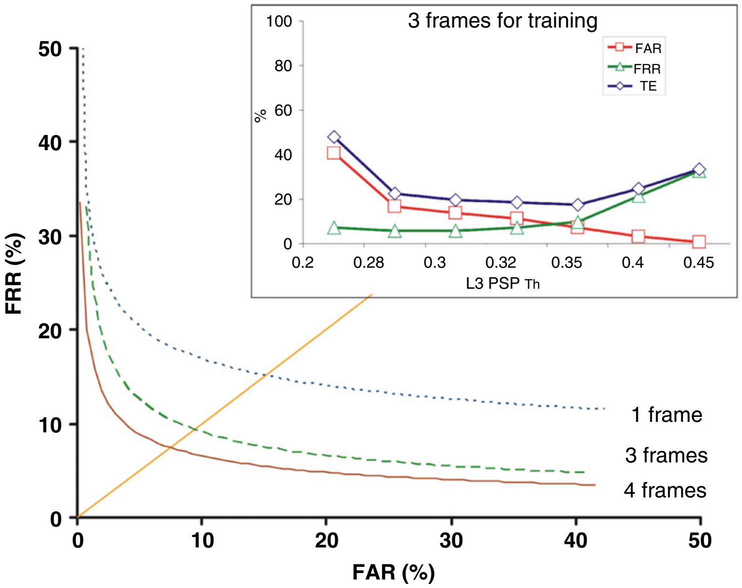

Fig. 4. Performance of the visual SNN for increasing number of training samples per individual user in terms of EER (equal error rate), which is the value where FAR is equal to FRR. The inset shows the performance of the SNN network for 3 training samples and different L3 firing thresholds PSP_{Th}. As expected, when the threshold increases so does the FRR while the FAR decreases.

evaluated, with adaptive addition of neuronal maps within a class to accommodate several training samples (views) as described above. For this, different numbers of view samples (1, 3 and 5) were used to train for each of 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 any merging activity. 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. Fig. 4 shows the results on test frames for different numbers of training samples. The inset shows results for 3 frames for different firing threshold PSP_{Th} in L3 neurons. Note that, varying the PSP_{Th} in L3 neurons, the system can have different operating points. As expected, when PSP_{Th} increases so does the FRR, while the FAR decreases, which was the case in every training setup. Total error (TE) = FAR + FRR. It can be 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. Previously we showed that classification performance of our visual SNN is comparable, or even slightly better than results obtained using traditional classifiers like SVM, Nearest Neighbour or MLP with PCA for feature extraction (Wysoski et al., 2008a).

3. The auditory model

3.1. Short literature review

Robert and Eriksson (1999) proposed a biologically plausible 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 Robert and Eriksson's model is to facilitate the understanding of signal coding within the cochlea and in the auditory nerve as well as analyse sound signals. The outputs 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 (2006b) 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 & Iwata, 1994) and in sound source separation and source recognition in Iwasa, Inoue, Kugler, Kuroyanagi, and Iwata (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, Gelbart, Ramacher, and Hemmert (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

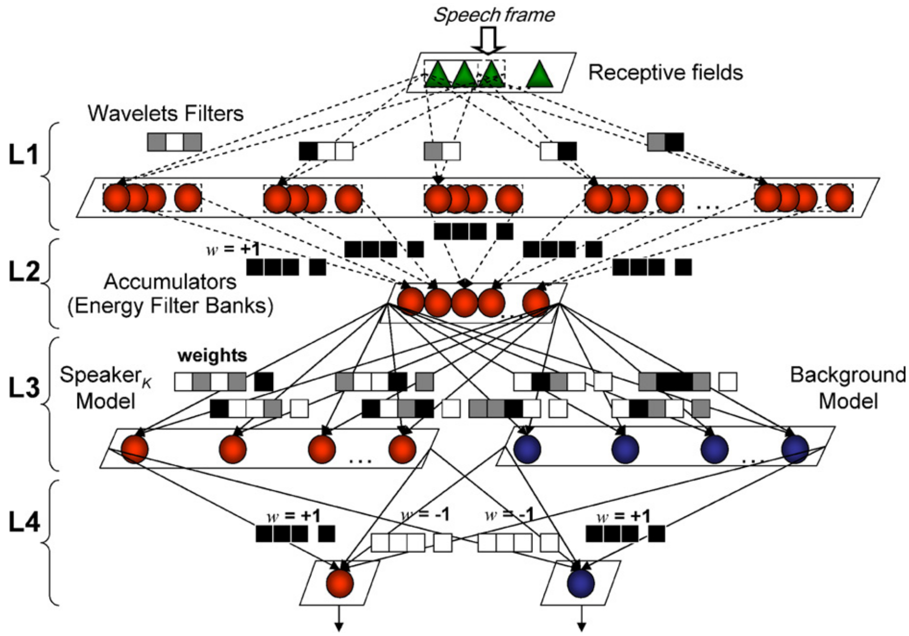

Fig. 5. Integrated design of an evolving SNN that performs speech signal pre-processing and speaker authentication.

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 & Juang, 1993).

Rouat, Pichevar, and Loiselle (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 performed by spiking neural networks, a biologically inspired system to perform source separation on auditory signals is proposed. In the same work and in Loiselle, Rouat, Pressnitzer, and Thorpe (2005), a preliminary evaluation used SNN for recognition of spoken numbers. These works can be considered to be an extension of principles used in the SpikeNet for auditory domain.

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 & Mercier, 2001).

The next section presents the design of a new network architecture based on fast spiking neurons (Delorme & Thorpe, 2001) 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.

3.2. *SNN architecture for auditory information processing*

This is a novel architecture (Fig. 5), which uses principles of SpikeNet and our visual SNN to build a new model for auditory domain. It is an extension of the auditory SNN for text-independent speaker authentication presented in Wysoski et al. (2007).

The systemic behaviour of the ensemble of inner hair cells is simulated with biologically inspired basic processing units (spiking neurons). However, note that this design does not aim to reproduce the activity of the inner hair cells in full detail. Sound signals are described with spectral characteristics. Cochlear fibers are sharply tuned to specific frequencies (Kiang, Watanabe, Thomas, & Clark, 1965), which are commonly modelled with the Short Term Fourier Transform (STFT) or wavelets. STFT as a discrete mathematical method has intrinsic characteristics 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 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 has been used successfully for speaker authentication, but it may occlude other features that can facilitate a unique description of a speaker. 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 our 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 Eq. (1) and is composed of the following steps:

(1) A pre-emphasis filter 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. One neuron represents each frame position. From hereafter the processing is done through spikes;
(4) Layer 1 (L1) neurons (see Fig. 5) 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:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right). \tag{9}$$

In 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. The output of L2 is a train of spikes that extracts spectral and spatial characteristics of an input frame that mimics wavelet computation.

The pre-processing layers are integrated into the classification procedure described in Wysoski et al. (2007) and summarised in the following paragraphs. 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 PSP, 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).

The network architecture to perform classification tasks of auditory patterns using spiking neurons includes two techniques that have already proven to be efficient in traditional methods (Gray, 1984; Reynolds, Quatieri, & Dunn, 2000). They are:

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

These techniques are implemented in Layer 3 (L3) and Layer 4 (L4). L3 is composed of two neuronal maps. One neuronal map has an ensemble of neurons trained by positive examples (prototypes). Each neuron in the neuronal map is created and/or merged and trained to respond optimally to different segments of the correct training utterances, i.e., different speech phones (minimal unit of speech segmentation). The second neuronal map in L3 is trained also adaptively with negative examples (background model). Several ways to represent background models that can be universal or unique for each class are described and analysed in Bimbot et al. (2004).

Similar to L3, layer L4 has two neuronal maps representing the correct class and the background model. Each L4 neuronal map is composed of a single neuron. L3 and L4 are connected to each other as follows:

(a) excitatory connections between neurons corresponding to neuronal maps with the same label, i.e., L3 correct class to L4 correct class and L3 background to L4 background, and;
(b) inhibitory connections between neurons with differing neuronal map labels, i.e., L3 correct class to L4 background and L3 background to L4 correct class. Effectively, L4 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 by L1 and L2 layers.
(b) The spikes are then propagated to L3 until an L3 neuron emits the first output spike, which is propagated to L4. If a neuron in L4 generates an output spike, the simulation is terminated. If not, the next frame is propagated.
(c) Before processing the next frame, L3 PSPs and order of arrival of spikes *order*(i, j) are reset whereas L4 neurons retain their PSPs, which are accumulated over consecutive frames, until an L4 output spike is generated.

The classification is completed when a neuron in L4 generates an output spike or all frames and all spikes in the network have been propagated. If the L4 neuron representing the correct class releases an output spike, the class is authenticated. The authentication fails in a case where no spikes occur in L4 after all frames have been processed or an L4 neuron representing background releases an output spike.

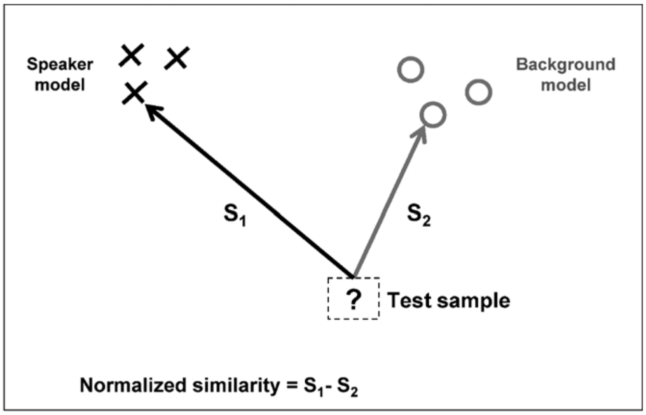

Fig. 6. Normalization in the similarity domain in a hypothetical two-dimensional space.

The authentication score of a class is calculated not only based on the similarity between a test sample and the correct class model, but on the relative similarity between the test sample and the class model and between the test sample and a background model. This normalization process is illustrated in Fig. 6. 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 our experiments a SNN-based implementation, normalized similarity is computed allocating excitatory connections to neurons representing the speaker model and inhibitory connections to neurons representing the background model.

3.3. Auditory pattern learning procedure

Training is done on the synapses connecting L2 and L3 neurons. To update weights during training, the simple rule described in Eqs. (5) and (6) used in the visual system model is applied. 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 adaptive online procedure for training the network and creating new neurons is similar to the visual pattern recognition model described in Section 2.3 and can be summarised with the following pseudo-code (see also Wysoski et al., 2007, for more details):

```
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 5 and Equation 6
    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 10 and Equation 11
```

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

Table 2
Main properties of the SNN-based auditory information processing system.

Processing units	Features are extracted using spiking neurons as basic information processing unit. Spiking neurons are also used in the decision-making stage.
Structure	Auditory information propagates with feed-forward connections into four layers of neuronal maps of spiking neurons that represent the behaviour of various auditory areas (tonotopically organized cells, spectral filter banks, phonetic association).
Learning	The online evolving procedure enables the ongoing learning of patterns 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.

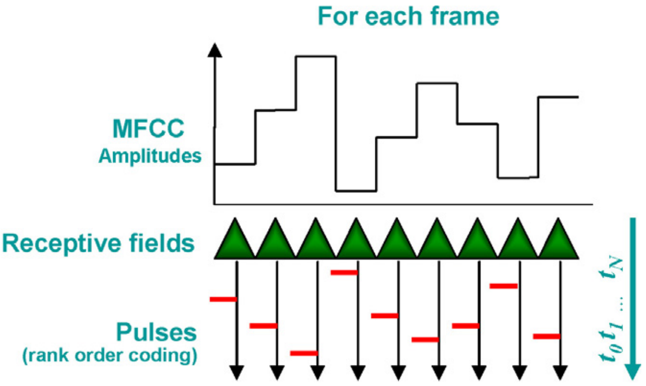

Fig. 7. MFCC encoded as spiking time with Rank Order Coding (Delorme et al., 2001). The higher the amplitude the shorter is the spike delay.

average as

$$W = \frac{W_{\text{new}} + N_{\text{Frames}} W}{1 + N_{\text{Frames}}} \tag{10}$$

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}:

$$\text{PSP}_{\text{Th}} = \frac{\text{PSP}_{\text{Thnew}} + N_{\text{Frames}} \text{PSP}_{\text{Th}}}{1 + N_{\text{Frames}}}. \tag{11}$$

Table 2 gives a summary of the main properties of the auditory SNN.

3.4. Experimental evaluation of the auditory SNN

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 his/her identity are some examples. For each of these scenarios, different and specifically tuned processing techniques seem to be the most effective. Here, we focus on the short-sentence text-independent problem, which is typically comprised of input utterances ranging from 3 s to 1 min. 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, Vector Quantization (VQ) (Burileanu, Moraru, Bojan, Puchiu, & Stan, 2002; Gray, 1984) and Gaussian Mixture Models (GMM) (Bimbot et al., 2004; Reynolds et al., 2000) became the standard approaches to tackle the text-independent speaker authentication problem. In our work, VQ is used for comparison purposes.

In order to test the classification properties of the system, i.e., neurons representing the speaker and background models, instead of the feature extraction based on spiking neurons, i.e. layers L1 and L2 in Fig. 5, Mel frequency cepstrum coefficients (MFCC) encoded in the time domain were used (Rabiner & Juang, 1993). The reason behind this simplification is to allow a better comparison of the learning procedure with previous works. Thus, in our shortcut implementation, each frame of the signal containing speech fragments generates an MFCC vector that is translated into spikes using Rank Order Coding. The process of encoding is schematically illustrated in Fig. 7. In the following experiments, one input neuron represents one MFCC vector. Fig. 8 shows network architecture used in the experiments.

The speech part of the VidTimit dataset (Sanderson & Paliwal, 2002) 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 (2002), which used Gaussian Mixture Model, the system was set to authenticate 35 individuals, each individual trained with 6 utterances. The remaining 4 utterances of each individual were 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

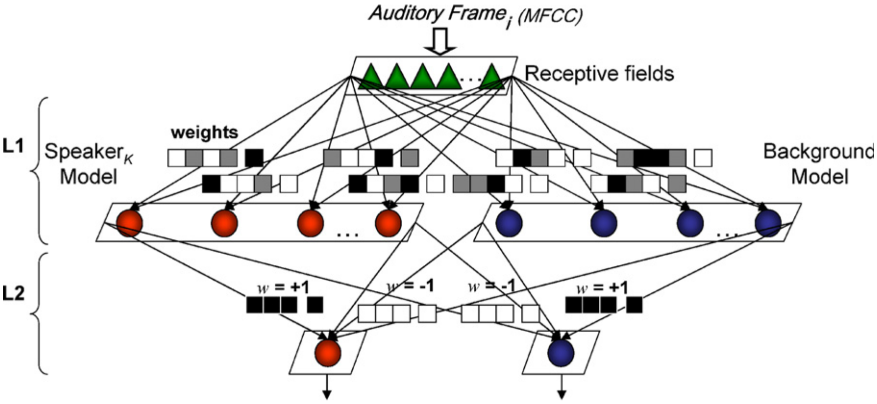

Fig. 8. Auditory SNN architecture. Frame-by-frame integration/accumulation of binary opinions.

that try to break into each model is (35 − 1 remaining user × 4 utterances) + (8 impostors × 4 utterances), which gives a total of 168 impostors. For all individual models of the entire dataset, there are (35 users × 4 utterances), totalling 140 true claimants and (35 users × 168 utterances) = 5880 impostors.

The speech signals were 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 was then encoded into spikes spread across 19 receptive field neurons. A specific background model for each speaker is trained. For the sake of simplicity, the background model of a speaker i was 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.

With respect to the SNN implementation, the number of neurons in the L3 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 L3 neurons L4 is composed of neurons with mod = 1. PSP_{Th} of L4 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 $40 \times 0.2 = 8$. The PSP_{Th} of L3 neurons were calculated as a proportion c of the maximum PSP obtained during the training procedure. The performance for $p = 0.2$ and the different values of c are shown in Fig. 9. The minimum TE reached was 31.1%. These results are worse than in Sanderson and Paliwal (2002), where with the same dataset, the authors reported total error TE = 22% using Gaussian Mixture Model. The standard vector quantization (VQ) algorithm (Burileanu et al., 2002) with k-means clustering was used for comparison and we obtained TE = 25% (Wysoski et al., 2007).

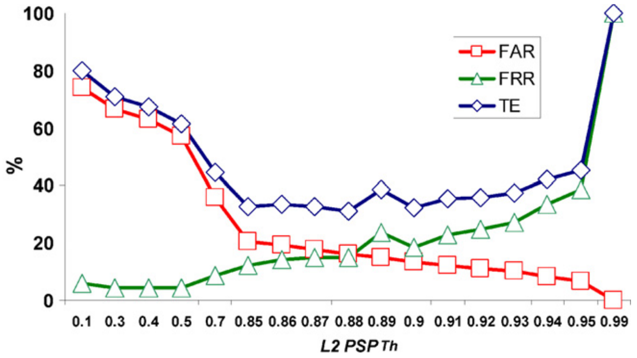

Fig. 9. Typical auditory SNN performance on VidTimit dataset for different values of c (proportion of the maximum PSP generated by a training sample). FAR is the false acceptance rate, FRR is the false rejection rate, and TE is total error (FAR + FRR).

4. Modular integration. Audiovisual and beyond

There is strong experimental evidence showing that integration of sensory information occurs in the brain (Calvert, 2001; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; von Kriegstein & Giraud, 2006; Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005; Stein & Meredith, 1993) 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, Jones, & Mosdell, 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).

Next section reviews several models (biologically realistic or not) of modality integration for the purpose of artificial pattern recognition. Several biologically realistic properties are then employed to describe a new integrative system for processing of multimodal information.

4.1. Short literature review

Brunelli and Falavigna (1995) presented a system where two classifiers are used to process speech signals and three others to recognize visual inputs. 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 & Soong, 1987). A local template matching approach at the pixel level, where particular areas of the face (eyes, nose, and mouth) are compared with a previously stored data, is used for face authentication. The results of these individual classifiers are connected to the input of a new integrative module based on HyperBF networks (Poggio & Girosi, 1990). Attempting to further improve the performance of the multimodal systems, several methods propose adaptation of the fusion mechanisms (Chibelushi, Deravi, & Mason, 1999; Sanderson & Paliwal, 2002). See Chibelushi, Deravi, and Mason (2002) for an extensive and comprehensive list.

Maciokas et al. (2002) applied brain-like approaches to tackle the problem of integrating the visual information of lip movements with the corresponding speech generated by it. The proposed system 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.

Mercier and Seguier (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 is 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) (Mozayyani, Baig, & Vaucher, 1998; Vaucher, 1998).

Chevallier, Paugam-Moisy, and Lemaitre (2005) present a system based on SNN to be used 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, Paugam-Moisy, Reynaud, and Puzenat (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, Hatef, Duin, and Matas (1998), after providing a review, tries to find a common basis for the problem of combining

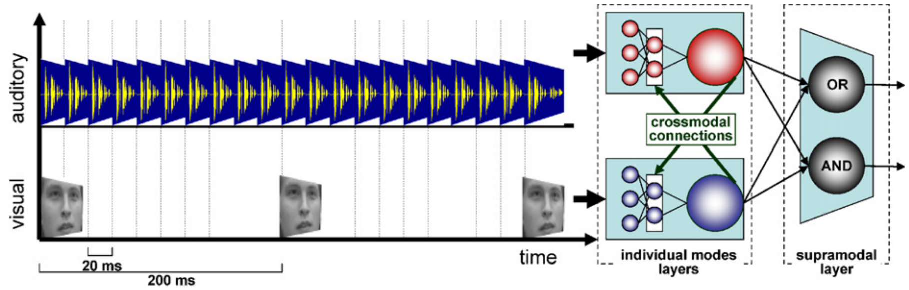

Fig. 10. Frame-based integration of modalities in a supramodal layer and by means of crossmodal connections. Unimodal and supramodal layers are implemented using spiking neurons.

Table 3
Main properties of the SNN-based integrative system.

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.
Learning	Online evolving procedures enable the ongoing learning of patterns extracted from 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.

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 et al. (2002).

In Yamauchi, Oota, and Ishii (1999); Yamauchi, Takama, Takeuchi, Sugiura, and Ishii (2001) a self-supervised learning procedure is described that learns categories automatically by integrating sensory information. After training, the system is able to control the priority of the sensors based on the characteristics of input signals. Particularly interesting in this design is the use of two networks for each sensor (forward and backward network). In the forward network, weights are adjusted for mapping incoming sensor signals (input) to the output, whereas in the backward network weights are adjusted for mapping the output vector to input signal. The result of the backward network is an estimation of the input signal, which is used during the recognition phase as a confidence measure of the sensory input. With this approach, priority between sensors can be established, which demonstrated to increase the learning ability of categories based on unlabelled datasets.

Among all the systems mentioned above, whether using traditional techniques or brain-like networks, none of them demonstrated performance degradation of multimodal systems. On the contrary, 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 bimodal 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, the main property of the system is the ability to receive and integrate information from different modules online, 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 that are based traditionally on the decomposition and consequent recombination of modalities. The illustration of a bimodal system for audiovisual processing with crossmodal connections is shown in Fig. 10.

4.2. SNN architecture for modality integration

The integration of modalities is implemented with spiking neurons. The same fast integrate-and-fire neuron described in the previous sections is used. Each individual modality has its own network of spiking neurons as described in the previous sections. In general, the output layer of each modality is composed of neurons that authenticate/not authenticate a class (i.e. user) they represent when output spikes are released. The modality integration is implemented attaching a new layer to the output of the individual modalities. This layer (supramodal layer) represents the supramodal region and contains neurons that are sensitive to more than one modality (Stein & Meredith, 1993). In the simplest case, the supramodal layer contains two spiking neurons for each class label (i.e. for each user). Each neuron of these two neurons, 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.

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. Thus, both excitatory and inhibitory connections are implemented for the crossmodal coupling. With this configuration, the output of a given class C in one modality excites the class C neuronal maps in other modalities and inhibits all other classes $\hat{C} \neq C$ in other modalities. Table 3 summarises the

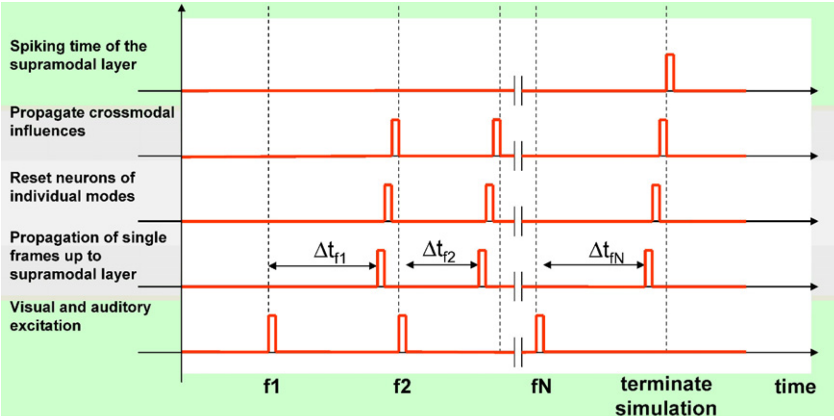
Spiking time of the supramodal layer
Propagate crossmodal influences
Reset neurons of individual modes
Propagation of single frames up to supramodal layer
Visual and auditory excitation
Δt_f1
Δt_f2
Δt_fN
f1
f2
fN
terminate simulation
time

Fig. 11. Typical behaviour of the integrated SNN architecture over time. The visual and auditory excitations (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.

main properties of the integrative system in terms of information processing units, information processing pathways and learning ability.

Fig. 11 illustrates the behaviour of the bimodal network over time. The dynamic behaviour is described as follows: each frame of the visual and auditory excitation (frames f1, f2, …, fN) are propagated through their corresponding modality 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 & Meredith, 1993).

In the following section, we evaluate the supra- and crossmodal concepts applied to the case of audiovisual integration in a person authentication problem based on face and speech information. A more detailed explanation of the implementation is also given.

4.3. *Experimental evaluation*

The integration of audiovisual modalities with a network of spiking neurons is exemplified with the VidTimit dataset (Sanderson & Paliwal, 2002). In this particular setup, 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 & Jones, 2001) implemented in the OpenCV library. Faces are converted into greyscale, normalized in size (height = 60 × width = 40), convolved with an elliptical mask, and encoded into spikes using Rank Order Coding. SNN does not require illumination normalization as demonstrated in 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 3×3 Gaussian grid with $\sigma = 0.9$ and scale 2 uses a 5×5 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 5×5 grid with a wavelength of $\lambda = 5$ and $\sigma = 2.5$ is used and in scale 2 a 7×7 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 (Fig. 7). 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 numbers of neurons in the auditory L1 neuronal maps for the speaker and background model are defined *a priori* to be 50 neurons each. In previous testing it was 80. The modulation factor for auditory neurons is set to 0.9.

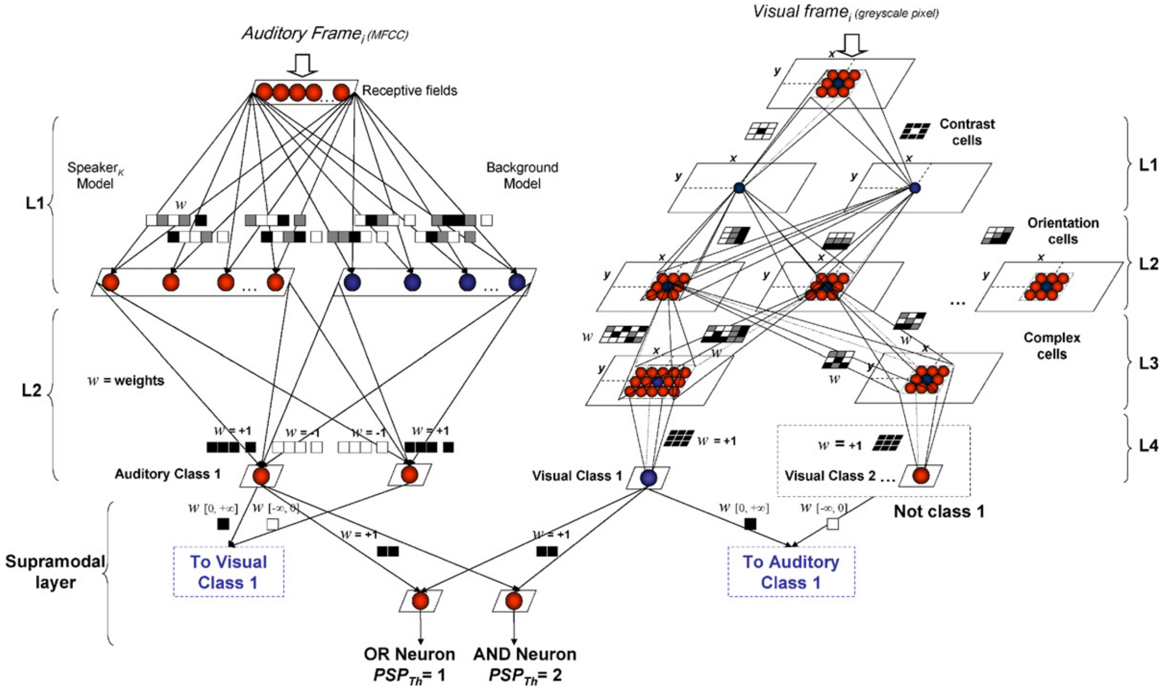

Fig. 12. Audiovisual crossmodal integration architecture. The supramodal layer integrates incoming sensory information from individual modalities and crossmodal connections influence of one modality upon the other.

- *Integration*: The crossmodal influence is parameterized as previously described in Wysoski et al. (2008b) and set as: CM_{AVexc}(audio to video excitation) $=$ CM_{VAexc}(video to audio excitation) $=$ 0.1 and CM_{AVinh} (audio to video inhibition) $=$ CM_{VAinh}(video to audio inhibition) $=$ 0. Results that do not take into account the crossmodal coupling are also presented, i.e., $CM_{\text{AVexc}} = CM_{\text{VAexc}} = CM_{\text{AVinh}} = CM_{\text{VAinh}} = 0$, 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 × 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. These settings are a little different from those used for testing individual modalities in order to manage effectively the joint bimodal system.

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. 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. Fig. 10 shows 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). The detailed audiovisual crossmodal integration architecture is shown in Fig. 12. The bottom part of Fig. 12 shows two neurons (OR and AND) representing the supramodal layer. 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. 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). In Fig. 13 is shown the performance obtained on each individual modality in this testing scenario for different values of L3 PSP_{Th} in the visual system and L1 PSP_{Th} in the auditory system. In this dataset setup, while the best total error (TE) for the face authentication was 21% (about the same as for 3 frames in Fig. 4), the auditory authentication module reached TE $\approx$ 38%.

The supramodal layer and the crossmodal coupling are updated when an individual modality outputs a spike, which may occur once in every frame. In reality, it is known that auditory stimuli are processed faster than visual (difference of approximately 40 to 60 ms Stein & Meredith, 1993). Here, we employ a simplification that the processing time for one frame is the same, regardless of the modality. Fig. 14 shows the performance of the system considering the type of integration held in the supramodal layer in terms of FAR, FRR and TE. 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 $CM_{\text{AVexc}} = CM_{\text{VAexc}} = 0.1$ and $CM_{\text{AVinh}} = CM_{\text{VAinh}} = 0$. The same parameters are used for individual modalities in this experiment, i.e., auditory parameters (L3 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 visual and auditory L3 PSP_{Th}.

In Fig. 14(a), we can see that the OR integration has the worst performance. It has very high false acceptance rate, because it relies effectively on one or the other positive recognition. TE reaches the minimum of 20% only for one combination of audiovisual parameters. As can be expected, AND integration lowers FAR and consequently TE, as both modalities have to agree on the result of authentication (Fig. 14(b)). The best overall performance has been achieved with the crossmodal AND integration (Fig. 14(c)). This type of bimodal integration resulted in the lowest TE $\approx$

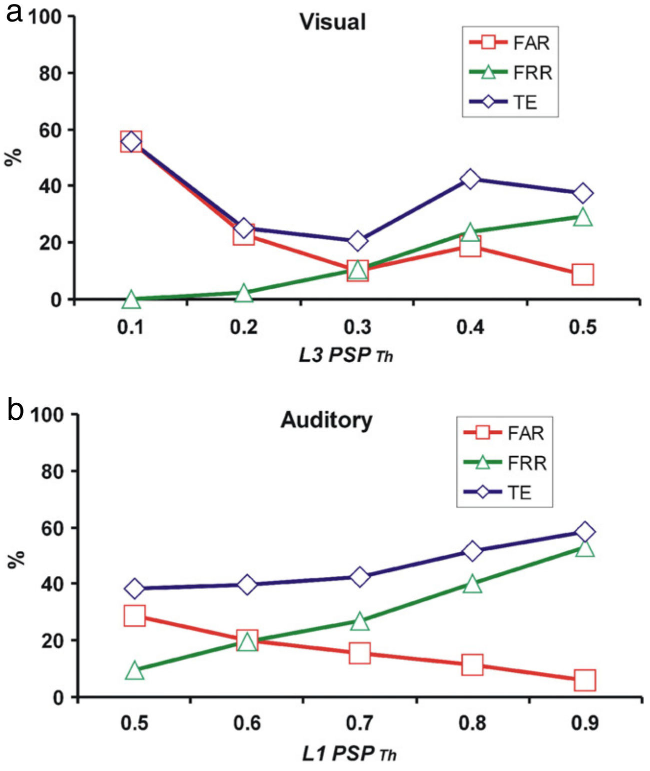

Fig. 13. Performance of individual modalities for different values of visual (L3 PSP_{Th}) and auditory parameters (L1 PSP_{Th}). (a) Visual system (b) auditory system. FAR is the false acceptance rate, FRR is the false rejection rate and TE is the total error (FAR + FRR).

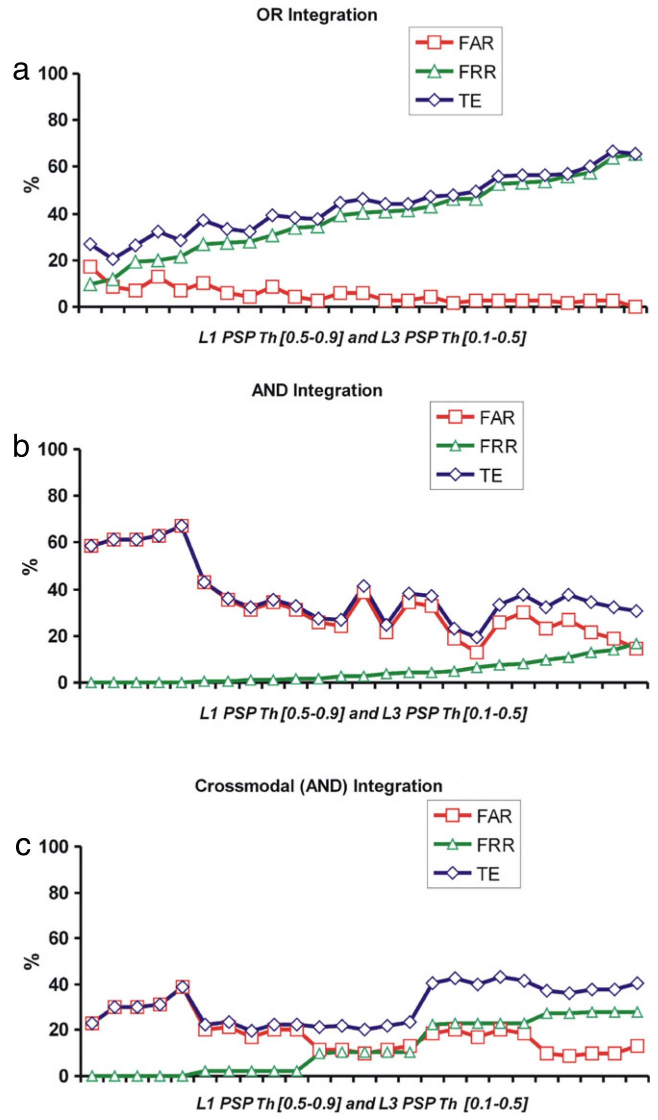

Fig. 14. Performance of the (a) OR and (b) AND integration of modalities with a supramodal layer of spiking neurons. (c) "Crossmodal AND" when excitatory crossmodal influences are activated (for auditory L1 PSP_{Th} and visual L3 PSP_{Th} ranging from [0.5, 0.9] and [0.1, 0.5], respectively).

20% for 10 combinations of audiovisual parameters. Thus, the crossmodal AND integration is the most robust type of bimodal integration. In summary, in this latter scenario the integrated performance is better than the performance of any of the systems taken individually.

Fig. 15 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. In addition this comparison clearly demonstrates that not well tuned parameters can deteriorate the results of the bimodal system to such an extent that in some cases the bimodal system has a lower accuracy when compared to a single modality. This highlights the importance of having parameters well tuned not only for individual modalities but also for an integration module.

5. Discussion

In terms of information pathways, neuroscientists have been drawing very accurate and detailed maps of the pathways taken by sensory information. In this work an integrated biologically inspired audiovisual pattern recognition system was designed and implemented, which contains a very simplified version of the major levels of processing. For the visual system, the functional behaviour of retina cells, orientationally selective cells and complex cells are implemented with a two-dimensional grid of spiking neurons. By complex cells in this context we mean neurons that are selective to the combination of orientations. Only feed-forward connections are used and no adaptation at lower levels is applied. With 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 more elaborate 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. The system was applied to the person authentication problem.

The main results of our bimodal system are:

1. *Visual system*. An 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.

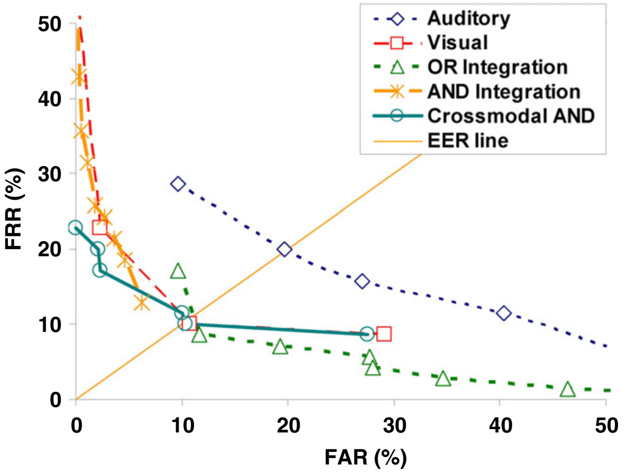

Fig. 15. Comparison between individual modes (auditory and visual) and the corresponding integration. Overall, the integration presents better performance than individual modes. OR, AND, "Crossmodal AND" alternate in the best position for different operating points. On the other hand, with certain combination of parameters the integration can provide lower accuracy than an individual modality alone. EER is the equal error rate (where FAR = FRR).

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. An SNN architecture was proposed, which achieved similar levels of performance when compared to a traditional Vector Quantization (VQ) 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.

The experiments and results presented in this paper do not explicitly consider the effects of noise on the performance. However, in a previous work, the behaviour of ensembles of the same "integrate-and-fire" neuron models and rank order information encoding used here have been thoroughly investigated (Delorme & Thorpe, 2001). The analysis demonstrated high invariance with respect to the illumination levels when processing visual information, mainly because the system is based on the relative contrasts due to Rank Order Coding. In addition, the system retained high accuracy (98% correct responses) even with input patterns corrupted with 30% of noise (Delorme & Thorpe, 2001). As the same principles of information processing remain in our system, we assume that a similar resistance to noise can be achieved.

Basically in our system, each modality is trained independently. During the recognition process, a supramodal layer integrates the result of individual modalities. In addition, if one single modality finalizes its process before others, crossmodal connections influence the decision of other modalities. In both works of Yamauchi et al. (1999, 2001), two networks for each sensory modality (forward and backward network) are proposed. In the forward network, weights are adjusted for mapping income sensor signals (input) to the output, whereas in the backward network weights are adjusted for mapping the output to input signal. The result of the backward network is an estimation of the input signal, which is used during the recognition phase as a confidence measure of the sensory modality. With this approach, priority between sensors can be established, which demonstrated to increase performance. This backward mechanism is not considered in our system. However, in our system the sense of most trusted modality is obtained by faster processing time, i.e., when a signal is very similar to a previously learnt pattern, neurons are more activated. Consequently the output signal of a given modality is generated sooner. The sooner a given modality generates an output, the sooner other modalities receive the crossmodal influence.

Under the multimodal pattern recognition perspective, our results show that the integration of modes enhances the performance in several operating points of the system. Sanderson and Paliwal (2002) explored the integration of modalities using a combination of mathematical and statistical methods on the same VidTimit dataset. The auditory system alone, using MFCC features and Gaussian Mixture Model in a noise-free setup, reached TE (total error) ≈22%. The visual system is reported to have TE ≈ 8% with features extracted using PCA and SVM for classification. After testing several adaptive and nonadaptive systems to perform integration, the best performance was 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 our experimental setup the results shown in this paper are clearly not as good as the results in Sanderson and Paliwal (2002). In our work, we have set up values of parameters experimentally. To extract the best performance from the system more elaborate optimization procedures can be incorporated. The computation with spiking neurons enables optimization of parameters and connection weights according to three different criteria: accuracy, energy (in the form of, e.g., minimum number of spikes), and processing time (e.g., minimum time needed to process). However even if a more sophisticated parameter optimization procedure was employed for our present model, there are some fundamental facts, more important than parameter optimization, which need to be taken into account when trying to improve our system further. In the following we will discuss these fundamental issues in detail.

Striving to achieve simplicity yet biological plausibility, we have employed many simplifications and abstractions in our system. From the lower levels of sensory processing to the higher levels of cognition, a simple model of spiking neurons was used. SNN inherently enables 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. However, in reality, 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. Alternatively, the buffer of working memory synchronizes the multimodal 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 & Meredith, 1993). 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 neuron models, perhaps even to simulate neurons at the level of ionic channels, and also to use neurons with different input/output function for different stages or tasks of processing. In addition, as the biological pathways are more and more clearly understood, a more detailed description of the biological pathways should be incorporated into the model, e.g., addition of new layers that represent hierarchically higher areas in the visual system (V2, V3, V4, IT, etc.). These hierarchically higher areas are all collapsed into one layer L3 in our present system.

Crucially important is to explore other information coding schemes. In our design, only one information coding mechanism is evaluated, that is one spike per neuron where the highest importance is given to the first spike to arrive. Spiking time theory (in opposition to the spiking rate theory, used for instance, in perceptrons) 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 was used that enabled the occurrence of only one spike per neuron. Based on these assumptions, concrete models were implemented and validated. The same spiking neurons when used on the integration layer provide a way to combine modalities and enable easy implementation of crossmodal influences between individual modes. A comparison with another connectionist-based system described in Sanderson and Paliwal (2002) indicate that our new design achieve comparable performance with parameters tuned by hand. However, it must be noted that it is difficult to assure, which neuron model and information encoding performs better on a given task without having a more systematic and autonomous way to optimize parameters.

An extension to our current design can be the reproduction of more natural patterns of spiking activity and other coding schemes. Although 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 & Kistler, 2002; Mazurek & Shadlen, 2002). This theory is proving gradually not to be universally valid, as several independent neurophysiological experiments demonstrate the existence of spike-timing patterns in both single and in ensembles of neurons. For instance, in Vaucher (1998), *in vivo* measurements enabled prediction of rat's behaviour responses through the analysis of spatio-temporal patterns of neuronal activity. 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 or asynchronous spiking activity behaviour. 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 & Gat, 2001) and computational models explored the storage and learning capabilities of this theory (Bienenstock, 1995; Gutig & Sompolinsky, 2006). One spike matters coding is based on experiments on ultra-fast perceptual classification (Thorpe et al., 1996). Person authentication studied in our model may be the task, which is beyond this type of information coding. 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. Extending our design to be able to model synchronicity in processing will entail exploration of feedback connections and stability problems.

Finally, with respect to learning 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 unsupervised 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. In the future, the learning procedure can be further elaborated to reproduce memory consolidation and forgetting. Further in this direction, it is necessary to define learning rules for integrative modules as well as a systematic procedure to train crossmodal connections.

Acknowledgements

The research was supported by the Tertiary Education Commission of New Zealand (SGW) and by the NERF grant AUTX02001 funded by FRST (LB and NK).

References

Abeles, M. (1982). *Local cortical circuits: an electrophysiological study*. Berlin: Springer.

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

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

Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, *4*, 430–451.

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

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

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

Chevallier, S., Paugam-Moisy, H., & Lemaitre, F. (2005). Distributed processing for modelling real-time multimodal perception in a virtual robot. In *International multi-conference parallel and distributed computing and networks* (pp. 393–398).

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

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

Crepet, A., Paugam-Moisy, H., Reynaud, E., & Puzenat, D. (2000). A modular neural model for binding several modalities. In *International conference on artificial intelligence* (pp. 921–928).

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

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

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

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

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

Eriksson, J. L., & Villa, A. E. P. (2006b). Artificial neural networks simulation of learning of auditory equivalence classes for vowels. In *International joint conference on neural networks* (pp. 1253–1260).

Fukushima, K., & Miyake, S. (1982). Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In S. Amari, & M. A. Arbib (Eds.), *Lecture notes in biomathematics*, *Competition and cooperation in neural nets* (pp. 267–285). Berlin, Heidelberg: Springer-Verlag.

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

Ganchev, T. (2005). Speaker recognition. *Ph.D. thesis*. Dept. of Electrical and Computer Engineering, University of Patras, Greece.

Gerstner, W., & Kistler, W. M. (2002). *Spiking neuron models*. Cambridge, MA: Cambridge University Press.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, *25*, 5004–5012.

Ghitza, O. (1988). 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*, 109–124.

Gray, R. M. (1984). Vector quantization. *IEEE ASSP Magazine*, 4–28.

Gutig, R., & Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing-based decisions. *Nature Neuroscience*, *9*, 420–429.

Holmberg, M., Gelbart, D., Ramacher, U., & Hemmert, W. (2005). Automatic speech recognition with neural spike trains. In *Interspeech* (pp. 1253–1256).

Hopfield, J. J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature*, *376*(6535), 33–36.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Iwasa, K., Inoue, H., Kugler, M., Kuroyanagi, S., & Iwata, A. (2007). Separation and recognition of multiple sound source using pulsed neuron model. In *Lecture notes in computer science*: *Vol. 4669*. *ICANN* (pp. 748–757).

Izhikevich, E. M. (2006). Polychronization: computation with spikes. *Neural Computation*, *18*(2), 245–282.

Kasabov, N. (2007). *Evolving connectionist systems*. Springer-Verlag.

Kiang, N. Y.-S., Watanabe, T., Thomas, E. C., & Clark, L. F. (1965). *Discharge patterns of single fibers in the cat's auditory nerve*. Cambridge: MIT Press.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

von Kriegstein, K., & Giraud, A. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), 1809–1820.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376.

Kruger, N., Lappe, M., & Worgotter, F. (2004). Biologically motivated multi-modal processing of visual primitives. *Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour, AISB*, *15*, 417–428.

Kuroyanagi, S., & Iwata, A. (1994). Auditory pulse neural network model to extract the inter-aural time and level difference for sound localization. *Transactions of IEICE*, *E77-D*(4), 466–474.

Loiselle, S., Rouat, J., Pressnitzer, D., & Thorpe, S. (2005). Exploration of Rank Order Coding with spiking neural networks for speech recognition. In *International joint conference on neural networks* (pp. 2076–2080).

Maciokas, J., Goodman, P. H., & Harris, F. C. Jr. (2002). Large-scale spike-timing dependent-plasticity model of bimodal (audio/visual) processing. In *Technical report of brain computation laboratory*. University of Nevada, Reno.

Matsugu, M., Mori, K., Ishii, M., & Mitarai, Y. (2002). Convolutional spiking neural network model for robust face detection. In *International conference on neural information processing* (pp. 660–664).

Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, *16*, 555–559.

Mazurek, M. E., & Shadlen, M. N. (2002). Limits to the temporal fidelity of cortical spike rate signals. *Nature Neuroscience*, *5*, 463–471.

McLennan, S., & Hockema, S. (2001). Spike-V: an adaptive mechanism for speech-rate independent timing. IULC working papers online 02-01.

Mel, B. W. (1998). SEEMORE: Combining colour, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, *9*, 777–804.

Mercier, D., & Seguier, R. (2002). Spiking neurons (STANNs) in speech recognition. In *3rd WSES international conference on neural networks and applications*.

Movellan, J. R. (1995). Visual speech recognition with stochastic networks. In G. Tesauro, D. Toruetzky, & T. Leen (Eds.), *Advances in neural information processing systems*: *Vol. 7* (pp. 851–858).

Mozayyani, N., Baig, A.R., & Vaucher, G. (1998). A fully neural solution for on-line handwritten character recognition. In *International joint conference on neural networks* (pp. 160–164).

Natschlager, T., & Ruf, B. (1998). Spatial and temporal pattern analysis via spiking neurons. *Network: Computation in Neural Systems*, *9*(3), 319–338.

Natschlager, T., & Ruf, B. (1999). Pattern analysis with spiking neurons using delay coding. *Neurocomputing*, *26–27*, 463–469.

Perrinet, L., & Samuelides, M. (2002). Sparse image coding using an asynchronous spiking neural network. *European Symposium on Artificial Neural Networks*, 313–318.

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*, 978–982.

Rabiner, L., & Juang, B. (1993). *Fundamentals of speech recognition*. New Jersey: Prentice Hall.

Reece, M. (2001). Encoding information in neuronal activity. In W. Maass, & C. Bishop (Eds.), *Pulsed neural networks*. MIT Press.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Robert, A., & Eriksson, J. L. (1999). A composite model of the auditory periphery for simulating responses to complex sounds. *Journal of the Acoustical Society of America*, *106*(4), 1852–1864.

Rosenberg, A. E., & Soong, F. K. (1987). Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. *Computer Speech and Language*, *2*(3–4), 143–157.

Rouat, J., Pichevar, R., & Loiselle, S. (2005). Perceptive, non-linear speech processing and spiking neural networks. In G. Chollet, et al., (Eds.), *Lecture notes on artificial intelligence*: *Vol. 3445*. *Nonlinear speech modelling* (pp. 317–337). Berlin, Heidelberg: Springer-Verlag.

Sanderson, C., & Paliwal, K. K. (2002). Identity verification using speech and face information. *Digital Signal Processing*, *14*, 449–480.

Seguier, R., & Mercier, D. (2001). A generic pretreatment for spiking neuron. Application on lipreading with STANN (Spatio-Temporal Artificial Neural Networks). In *5th international conference on artificial neural networks and genetic algorithms*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Shamma, S. A., Chadwick, R. S., Wilbur, W. J., Morrish, K. A., & Rinzel, J. (1986). A biophysical model of cochlear processing: intensity dependence of pure tone responses. *Journal of the Acoustical Society of America*, *78*, 1612–1621.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. MIT Press.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

Tikovic, P., Vöros, M., & Durackova, D. (2001). Implementation of a learning synapse and a neuron for pulse-coupled neural networks. *Journal of Electrical Engineering*, *52*(3–4), 68–73.

Vaucher, G. (1998). An algebraic interpretation of PSP composition. *Biosystems*, *48*, 241–246.

Viola, P., & Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition, Vol. 1* (pp. 511–518).

Wysoski, S. G., Benuskova, L., & Kasabov, N. (2006). On-line learning with structural adaptation in a network of spiking neurons for visual pattern recognition. In *Lecture notes in computer science*: *Vol. 4131*. *ICANN* (pp. 61–70). Berlin: Springer-Verlag.

Wysoski, S. G., Benuskova, L., & Kasabov, N. (2007). Text-independent speaker authentication with spiking neural networks. In *Lecture notes in computer science*: *Vol. 4669*. *ICANN* (pp. 758–767). New York: Springer-Verlag.

Wysoski, S. G., Benuskova, L., & Kasabov, N. (2008a). Fast and adaptive network of spiking neurons for multi-view visual pattern recognition. *Neurocomputing*, *71*(13–15), 2563–2575.

Wysoski, S. G., Benuskova, L., & Kasabov, N. (2008b). Adaptive spiking neural networks for audiovisual pattern recognition. In *Lecture notes in computer science*: *vol. 4985*. *ICONIP* (pp. 406–415). Berlin: Springer-Verlag.

Yamauchi, K., Oota, M., & Ishii, N. (1999). A self-supervised learning system for pattern recognition by sensory integration. *Neural Networks*, *12*(10), 1347–1358.

Yamauchi, K., Takama, J., Takeuchi, H., Sugiura, S., & Ishii, N. (2001). Sensory integrating neural network with selective attention architecture for autonomous robots. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *5*(3), 142–154.