Elsevier Editorial System(tm) for Neural Networks
Manuscript Draft

Manuscript Number:

Title: Evolving Spiking Neural Networks for Audio-visual Information Processing

Article Type: Contributed Article

Section/Category: Engineering and Design

Keywords: Spiking neural network; audio and visual pattern recognition; face recognition; speaker authentication; on-line classification

Corresponding Author: Mr. Simei Gomes Wysoski, M.E.

Corresponding Author's Institution: Knowledge Engineering and Discovery Research Institute - Auckland University of Technology

First Author: Simei G Wysoski, Master Course

Order of Authors: Simei G Wysoski, Master Course; Lubica Benuskova, PhD; Nikola Kasabov, PhD

Abstract: This paper presents new modular and integrative information methods and systems inspired by the way the brain performs information processing, in particular, pattern recognition. Spiking neural networks are used to model human visual and auditory pathways, and train the system to perform the specific task of person authentication. The systems are individually tuned and trained to recognize facial information and to analyze sound signals from spoken sentences. The modelling of the integration of different sources of information (multisensory integration) using spiking neural network architectures and learning procedures trained in an evolvable and adaptive way are also subject of investigation.

Title

Evolving Spiking Neural Networks for Audio-visual Information Processing

Keywords

Spiking neural network, audio and visual pattern recognition, face recognition, speaker authentication, on-line classification

Authors and Affiliation

Simei Gomes Wysoski, Lubica Benuskova, and Nikola Kasabov

Knowledge Engineering and Discovery Research Institute (http://www.kedri.info)
Auckland University of Technology, 581-585 Great South Rd, Auckland, New Zealand
{swysoski, lbenusko, nkasabov}@aut.ac.nz

1) Introduction

A good number of systems have used the terms *biologically realistic* or *brain-like* to define a new generation of neural networks that attempts 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 struggles to deliver general and reliable solutions. Each application requires a uniquely tailored artificial system whereas the human brain effortlessly processes information, integrates sensory modalities, controls motor activities while taking care of vital involuntary functions using only a few kJoules of energy per hour.

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

Several models of visual system reused the Hubel and Wiesel model of the primary visual cortex with contrast, directionally selective and complex cells placed in an hierarchical pathway [28] for the purpose of pattern recognition [18, 45, 56]. Examples of auditory models can be found in [23, 63]. 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 [19] and [31] for extensive reviews on adaptive methods and procedures).

Thus, up to this point, there are *brain-like* models of network structures and *brain-like* ways to perform network connectivity and reconfiguration. Recently another factor has added to the momentum. The fact that neurons exchange information using spikes and processing mechanisms use action potentials is another addition to the *biologically realistic* realm [27]. In [21] this concept is properly clarified, stating that, in order to avoid any prior assumptions on neural computation, neurons need to process at the level of action potentials. Thus, spiking neurons and spiking neural networks (SNNs), historically used as a tool for neuroscientists to study the dynamics of single or ensembles of neuronal units, emerged as a new generation of neural network models for pattern recognition.

As the theory of spiking neurons is currently most commonly accepted as the most human-like manner of processing [21], they are the basis for all 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) Evaluation and further extension of adaptive learning procedures to perform audio-visual pattern recognition. A learning procedure that enables the system to change its structure, creating/merging neuronal maps of spiking neurons is presented and evaluated on a person authentication problem.

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

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

A simple illustration of the complexity of integrating the auditory and the visual senses is shown in Figure 1. Each sensory modality has mostly distinct pathways where information is processed. Within a sensory modality, information is decomposed, e.g., in the visual system, the information is divided into sub modalities (colour, shape and movement) that are independently processed in different pathways. In the auditory system, the ventral cochlear nucleous with mainly tonotopical organization of cells and dorsal cochlear nucleous (mainly non tonotopical) also define different pathways. In different modalities and sub modalities, it is reasonable to think that the speed of transduction and the speed of information propagation in different pathways is not the same. If this is true, afferent stimuli from different sensory modalities arrive at the cerebral cortex at different times. The separation and integration of pathways within a modality as well as the integration of pathways from different modalities (and all the synchronizations implied in it) constitute a complex network that cannot be accurately described and reproduced.

The visual and the auditory models, partially described in [71, 72, 73] are evaluated in Section 2 and 3. Section 4 explores the integration of modalities (also partially presented in [70], 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) 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 [18]. A new type of model for object recognition based on computational properties found in the brain cortex was described by Riesenhuber and Poggio [56]. This model uses hierarchical layers similar to the Neocognitron and processing units based on MAX-like operation, to define the postsynaptic response, which results in relative position- and scale-invariant features. This biologically motivated hierarchical method is carefully analyzed by Serre *et al* [62] on several real-world datasets, extracting shape and texture properties. The analysis encompassed invariance on single object recognition and recognition of multiple objects in complex visual scenes (e.g. leaves, cars, faces, airplanes, motorcycles). The method presented comparable performance with benchmark algorithms.

In the same way, Mel [45] 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 non rigid 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 [36]. 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 [27] proposed a model and learning algorithm for spiking neurons to realize Radial Basis Functions (RBFs) where spatial-temporal information is presented based on the timing of single spikes, i.e., not in a rate-based fashion. Natschlager and Ruf further extended the idea, by defining the pattern not only by the sequence of input spikes, but also by the exact firing time [49, 50]. 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 [40] goes down to the level of ionic channels to describe a model of an audiovisual system that reproduces the responses of the GABAergic cells. Audio features were extracted using Short Term Fourier Transform and represented in tonotopic maps. The visual information of lip movement was extracted using Gabor filters. The two main results described in his work are: a) the accurate model of diverse firing behaviours of GABAergic cells; and b) proof that a large-scale network of the cortical processing preserves information in audiovisual modalities using an entropy measure. No attempts to test the classification abilities of the network have been made.

Thorpe [65] suggests that in order to be coherent with the time measured in certain classes of behavioural experiments on perceptual activities, the information processing mechanisms can afford to have a single neuron exchanging only one or a few spikes. The time between information acquisition and the cognitive response is too short to have rate-based neuronal encoding, since the information needs to travel sequentially over several different compartments located in distinct brain areas. Thus, the information needs to be sparsely encoded and, highly complex cognitive activities are reached through a complex wiring system that connects neuronal units. As an output of this work, the authors proposed a multi-layer feed-forward network (SpikeNet) using fast integrate-and-fire neurons that can successfully track and recognize faces in real time [12, 14]. Coding of information in this model is based on the so-called rank order coding, where the first spike is the most important. It has been shown that using rank order coding and tuning the scale sensitivity according to

the statistics of the natural images can lead to a very efficient retina coding strategy, which compares to image processing standards like JPEG [51].

Matsugu *et al* utilized a different coding strategy in a hybrid of a convolutional and SNN architecture for face detection tasks [41] In this hierarchical network, local patterns defined by a set of primitive features are represented in the timing structure of pulse signals. The training mentioned in the work is for the bottom feature-detecting layer to use the standard error back-propagation algorithm. The model implements hierarchical pattern matching by temporal integration of structured pulse packets. The packet signal represents intermediate or complex visual features (like an eye, nose, corners, a pair of line segments) that constitute a face model. As a result of the spatio-temporal dynamics the authors achieved size and rotation invariant internal representation of objects. Endowed with a rule-based algorithm for facial expression classification, this hybrid architecture achieved robust facial expression recognition together with robust face detection [42].

Next section follows the conceptual approach described in [12, 14], from which the basic building blocks of the model are borrowed, e.g., the fast integrate-and-fire neuron model and its respective learning rule, and the network structure, which is formed from hierarchical layers composed of neurons grouped in neuronal maps.

2.2) SNN architecture for visual information processing

The system uses the neuronal model described in [14]. The network structure, where neurons are placed in two-dimensional grids forming neuronal maps and consequent layers of maps, also follows the same pattern. The neural network is composed of four layers of integrate-and-fire neurons (See Figure 2). In the first two layers (L1 and L2) there is no learning, they simply act as passive filters and time domain encoders. In the third layer (L3), where the learning takes place, maps are trained to be sensitive to incoming excitation of more complex patterns. Neuronal maps are created or merged during learning, according to the online learning procedure described in [71]. There are lateral inhibitory connections between neuronal maps in the third layer, so that when a neuron fires in a certain map, other maps receive inhibitory pulses in an area centred in the same spatial position. An input pattern belongs to a certain class if a neuron in the corresponding neuronal map spikes first.

Layer 4 (L4), an addition to the original model described in details in [71], has one neuronal map containing a single neuron for each pattern class. The L4 neuron of a given class is connected to the corresponding L3 neuronal maps. There are excitatory connections (typically $w = +1$) between the L4 neuron and the neurons located close to the centre of L3 maps. Thus, L4 combines the results of a sequence of visual patterns, i.e. accumulates opinions from several frames.

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 used. With the inclusion of this simple procedure, it is possible to assess how many positive opinions from different frames are required to recognize a pattern successfully.

2.3) Visual patterns learning procedure

The learning procedure presented in details in [71] can be summarized in four sequential steps:

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

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

$$\Delta w_{j,i} = \mathrm{mod}^{order(a_j)} \quad \textbf{(1)}$$

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

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

$$PSP_{threshold} = c \max(PSP) \quad \textbf{(2)}$$

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).

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

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

$$W = \frac{W_{Map_{C(k)}} + N_{samples} W_{Map_{C(Ksimilar)}}}{1 + N_{samples}} \quad \textbf{(3)}$$

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

$$PSP_{Th} = \frac{PSP_{Map_{C(k)}} + N_{samples} PSP_{Map_{C(Ksimilar)}}}{1 + N_{samples}} \quad \textbf{(4)}$$

Notice that the learning procedure updates W and PSP_{Th} as well as enabling map merging for each incoming sample during training. For this reason, presenting the samples to the network in a different order can potentially lead to different network structures as well as different resultant W and PSP_{Th}. In other words, samples presented in a different order could potentially form slightly different clusters (different numbers of output maps for a given class), which can in turn affect the performance of the network.

The system has the properties summarized in Table 1.

2.4) Experimental evaluation

The system has been extensively evaluated in our previous experiments [71]. Here we present some additional results in order obtain a fair comparison with the integrated setup as described in the following sections. The system is trained to recognize 35 individuals. For testing, all 43 individuals are used, so 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 evaluated, particularly in which concerns to the adaptive addition of neuronal maps within a class to accommodate several training samples (views). For this, different numbers of samples (1, 3 and 5) were used to train on the 35 users. The training samples were chosen from different video streams (using the first frame from each video stream). The similarity threshold for merging neuronal maps was kept at a high level in order to inhibit 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. Figure 3 shows the results on test frames for different numbers of training samples 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, in all instances. 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.

3) The auditory model

3.1) Literature review

Robert and Eriksson [57] proposed a model of the auditory periphery to simulate the response to complex sounds. The model basically reproduces the filtering executed by the outer/middle ear, basilar membrane, inner hair cells, and auditory nerve fibers. The purpose of 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 output of the inner hair cells and auditory nerve fibers are properly represented with trains of spikes. This model has been used in [16] to simulate the learning of synthetic vowels by rats reported in [17]. 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 [37] and in sound source separation and source recognition in [29]. In [44] 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 [26] the importance of temporal and spectral characteristics of sound signals is described. The spectral properties are inherently represented with "rate-place code" during the transduction of the inner hair cells. Temporal information, on the other hand, provides additional cues, such as amplitude modulation and onset time. In the same work a multi-layer

auditory model is presented, which emulates inner ear filtering, compression and transduction. The work mainly concentrates on using spiking neurons to model octopus neurons, which are neurons located at the cochlear nucleus. Octopus neurons enhance the amplitude modulations of speech signals and are sensitive to signal onsets. Preliminary experiments showed that the system performs in much the same way as Mel Frequency Cepstral Coefficients (MFCC) [53].

Rouat *et al* [59] envisage the advantages of merging perceptual speech characteristics and biologically realistic neural networks. After a description of the perceptual properties of the auditory system and non-linear processing realized by spiking neural networks, a biologically inspired system to perform source separation on auditory signals is proposed. In the same work and in [38], a preliminary evaluation used SNN for recognition of spoken numbers.

Mercier and Seguier [46] 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 [47]. STANNs were initially proposed to process visual information [61].

The next section presents the design of a new network architecture based on fast spiking neurons [14] 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

The systemic behaviour of the ensemble of inner hair cells is simulated with biologically inspired basic processing units (spiking neurons) to be used in artificial speech processing systems. Note that, this design does not aim to accurately reproduce the activity of the inner hair cells.

Sound signals are described with spectral characteristics. Cochlear fibers are sharply tuned to specific frequencies [32], which are commonly modelled with the Short Term Fourier Transform (STFT) or wavelets. STFT as a discrete mathematical method has the intrinsic characteristic of being able to provide high spectral resolution of low frequency signals and low spectral resolution at high frequencies. This property does not affect the extraction of speech features for speech recognition. The Mel scale that forms the Mel filter banks also has sharply tuned filters at low frequencies and broadly tuned filters at higher frequencies.

Nonetheless, as described in [53] and the main object of research for [20], Mel filter banks and consequently MFCC, extract features particularly suitable for speech recognition. MFCC is also used successfully for speaker authentication, but it may occlude other features that can facilitate a unique description of a speaker. Ganchev [20] 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 [14] 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 [13]. One neuron represents each frame position. From hereafter the processing is done through spikes;
4) Layer 1 (L1) neurons (see Figure 4) 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:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \quad \textbf{(5)}$$

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 [73] and summarized 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 (see Figure 4) using spiking neurons includes of two techniques that have already demonstrated themselves to be efficient in traditional methods [24, 55]. They are:

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

Layer 3 (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/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 [4].

Similar to L3, 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 are reset to the rest potential 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.

Thus, 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 Figure 5. 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 [4]. 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 patterns learning procedure

Training is done in the synapses connecting L2 and L3 neurons. To update weights during training, the simple rule described in Eq. 1 and Eq. 2 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 [73] 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 Eq. 1 and Eq. 2
    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
Eq. 6 and Eq. 7
```

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

$$W = \frac{W_{new} + N_{Frames} W}{1 + N_{Frames}} \quad \textbf{(6)}$$

where N_{Frames} is the number of frames previously used to update the neuron in question.

Similarly, the average is also computed to update the corresponding PSP_{Th}:

$$PSP_{Th} = \frac{PSP_{Thnew} + N_{Frames} PSP_{Th}}{1 + N_{Frames}} \quad \textbf{(7)}$$

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

3.4) Experimental evaluation

Computer-based speaker authentication presents a number of possible scenarios. Text-dependent, text-independent, long sentences, single words, speaker willing to be recognized, speaker trying to hide their identity are some examples. For each of these scenarios, different and specifically tuned processing techniques seem to be the most effective. Here, we focus is on the short-sentence text-independent problem, which is typically comprised of input utterances ranging from 3 seconds to 1 minute. In this scenario, a speaker being authenticated does not necessarily need to present the same word or sentence used during training. Moreover, due to the short length of the signal, it is not possible to acquire long-term dependencies of features that could eventually supply additional information that would enhance performance. Thus, state machines to detect phonemes, words, and bigrams cannot be setup at full strength. Based on these properties, in recent years, Vector Quantization (VQ) [6, 24] and Gaussian Mixture Models (GMM) [4, 55] 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, Mel frequency cepstrum coefficients (MFCC) [53] encoded in the time domain were used. Each frame of the signal containing speech fragments generates an MFCC vector that is translated into spikes using Rank Order Coding [13] (Figure 6). In the following experiments, one input neuron represents one MFCC vector. The network setup is represented in Figure 7.

The speech part of the VidTimit dataset [60] was used for performance evaluation. VidTimit contains 10 utterances from 43 different speakers. In order to make a comparison with the experiments described in [60], 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 that try to break into each model is (35 - 1 remaining user x 4 utterances) + (8 impostors x 4 utterances), which gives a

total of 168 impostors. For all individual models of the entire dataset, there are (35 users x 4 utterances), totalling 140 true claimants and (35 users x 168 utterances) = 5880 impostors.

The speech signals 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.

The standard vector quantization (VQ) algorithm [6] with *k-means* clustering was used for comparison. Training of VQ was done with the same features (19 MFCCs) and the same strategy for selecting background models was applied. Figure 8 reports the best performance of the VQ algorithm obtained with 32 prototypes for speakers and 32 prototypes for the background model. These results are comparable with the work presented by [60], where, with the same dataset, the authors reported total error TE = false acceptance rate (FAR) + false rejection rate (FRR) = 22 % in slightly different setup conditions using Gaussian Mixture Model. The VQ implementation presented here obtained TE = 25 %.

In respect to the SNN implementation, the number of neurons in the 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 x 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 Figure 9. The minimum TE reached was 31.1%.

4) Modular integration. Audiovisual and beyond

There is strong experimental evidence showing that integration of sensory information occurs in the brain [7, 22, 34, 35, 64] 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 [15]. 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 [7, 22].

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) Literature review

Brunelli and Falavigna [5] presented a system where two classifiers are used to process speech signals and three others to recognize visual inputs. The results of these individual

classifiers are connected to the input of a new integrative module based on HyperBF networks [52]. MFCC and the corresponding derivatives are used as features, and each speaker is represented by a set of vectors based on Vector Quantization (VQ) [58]. A local template matching approach at the pixel level, where particular areas of the face (eyes, nose, mouth) are compared with a previously stored data, is used for face authentication.

Attempting to further improve the performance of the multimodal systems, several methods propose adaptation of the fusion mechanisms [10 ,60]. Chibelushi *et al* [9] provides an extensive and comprehensive list.

Maciokas *et al* [40], based on brain-like approaches, tackle the problem of integrating the visual information of lip movements with the corresponding speech generated by it. It uses a biologically realistic spiking neural network with 25.000 neurons placed in 10 columns and several layers. Tonotopic maps fed from Short Term Fourier Transform (STFT) with a neural architecture that resembles MEL scale filters are used for converting audio signals to spikes. Gabor Filters extract the lip movements. The encoding of three distinct sentences in three distinct spiking patterns was demonstrated. In addition, after using the Hebbian rule for training the output spiking patterns were also distinguishable from each other.

Seguier and Mercier [46] 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) [48, 67].

Chevallier *et al* [8] present a system based on SNN to be use in a robot capable of processing audiovisual sensory information in a prey-predator environment. In reality, the system is composed of several neural networks (prototype-based incremental classifier), one for each sensorial modality. A centralized compartment for data integration is implemented as a bidirectional associative memory. A network (also incremental) is used to perform the final classification (This architecture is described in detail in [11]). Particularly interesting in the prey-predator implementation is the spike-based bidirectional associative memory used. As properly suggested by the authors, the implementation using spikes enables the flow of information over time. The integration of these streams of incoming data is also processed on the fly as soon as the data from different modalities are made available. Furthermore, the bidirectional associative memory implemented with the spiking mechanism enables the simulation of crossmodal interaction.

Kittler *et al* [33], after providing a review, tries to find a common basis for the problem of combining classifiers through a theoretical framework. It is argued that most of the methods proposed so far can be roughly classified in one of the following types: product rule, sum rule, min rule, max rule, median rule and majority voting. After performing error sensitivity analysis on several combined systems, it is further suggested that the sum rule outperforms the other combination procedures. A more specific review of the speech-based audiovisual integration problem (speech and speaker recognition) is provided in [9].

Among all the systems mentioned before, whether using traditional techniques or brain-like networks, none of them demonstrated a degradation of performance of multimodal systems. The integration, in a synergistic way, achieves higher accuracy levels when compared with single modalities alone. The next section presents a simple attempt to process multimodal sensory information with a new architecture of fast spiking neurons. Besides the inherent ability of the neurons to process information in a simple and fast way [12], the main property of the system is the ability to receive and integrate information from several different modules on the fly, as the information becomes available. Because the entire system is based on the same principle of computation (spiking units) and the processing time of the information is also meaningful, back and forth connections as well as connections that emulate crossmodal influences are able to be simulated in a more biologically realistic manner. The crossmodal connections enrich the architecture of the current multimodal systems that are based traditionally on the decomposition and consequent recombination of modalities. The illustration of multimodal systems with crossmodal connections is shown in Figure 10. In particular, the system tackles the person authentication problem with the integration of audiovisual cues.

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. In general, the output layer of each modality is composed of neurons that authenticate/not authenticate a class they represent when output spikes are released. The integration is implemented attaching a new layer onto the output of the individual modes. This layer (supramodal layer) represents the supramodal region and contains neurons that are sensitive to more than one modality [64]. In the simplest case, the supramodal layer contains two spiking neurons for each class label. Each neuron representing a given class C in the supramodal layer has incoming excitatory connections from the output of class C neurons of each individual modality. The two neurons have the same dynamics, yet different thresholds for spike generation (PSP_{Th}). For one neuron, the PSP_{Th} is set in such a way that an output spike is generated after receiving incoming spikes from any single modality (effectively it is a spike-based implementation of an OR gate). The other neuron has PSP_{Th} set so that incoming spikes from all individual modalities are necessary to trigger an output spike (AND gate). AND neuron maximizes the accuracy and OR neuron maximizes the recall.

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.

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

Table 3 summarises the main properties of the integrative system in terms of information processing units, information processing pathways and learning ability.

In the following section, the supra/cross modal concepts are applied to the case of audiovisual integration in a person authentication problem based on face and speech information. The implementation of the visual model follows the description given in Section 2 and the auditory model uses the architecture described in Section 3. 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 [60]. 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 [69] implemented in the OpenCV library. Faces are converted into greyscale, normalized in size (height = 60 x width = 40), convolved with an elliptical mask, and encoded into spikes using rank order coding [13]. SNN does not require illumination normalization [14]. There are two scales of On/Off cells (4 L1 neuronal maps). In scale 1, the retina filters are implemented using a 3 x 3 Gaussian grid with $\sigma = 0.9$ and scale 2 uses a 5 x 5 grid with $\sigma = 1.5$. In L2, there are eight different directions in each frequency scale with a total of 16 neuronal maps. The directionally selective filters are implemented using Gabor functions with aspect ratio $\gamma = 0.5$ and phase offset $\varphi = \pi/2$. In scale 1 a 5 x 5 grid with a wavelength of $\lambda = 5$ and $\sigma = 2.5$ is used and in scale 2 a 7 x 7 grid with λ and σ set to 7 and 3.5, respectively. The modulation factor for the visual neurons was set to 0.995.
- **Auditory**: Speech signals are sampled at 16 kHz, and features extracted using standard MFCC with 19 MEL filter sub-bands ranging from 200 Hz to 7 kHz. Each MFCC is then encoded into spikes using rank order coding. One receptive field neuron is used to represent each MFCC (19 input receptive fields). A specific background model is trained for each speaker model. For the sake of simplicity, the following procedure is applied: the background model of a speaker *i* is trained using the same amount of utterances used to train the speaker model. The utterances are randomly chosen from the remaining training speakers. For the experiments, the number of neurons in the auditory L1 neuronal maps for the speaker and background model are defined *a priori* (50 neurons each). The modulation factor for auditory neurons is set to 0.9.
- **Integration**: The crossmodal influence is parameterized as described in [70] 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., $CM_{AVexc} = CM_{VAexc} = CM_{AVinh} = CM_{VAinh} = 0$, which effectively correspond to AND or OR integration.

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

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

The test is carried out frame-by-frame keeping the time correspondence between speech and visual frames. However, to speed up the computational simulations, the visual frames are downsampled. Five visual frames per second are used whereas the speech samples have a rate of 50 frames per second. The downsampling of the visual frames does not affect the performance, as for a period lower than 200 ms no substantial differences between one facial posture and another can be noticed in the VidTimit dataset. Figure 11 shows typical input streams to the SNN-based audiovisual person authentication system, where frames of detected faces are sampled at 200 ms (5 frames/second) and 19 MFCC extracted from the detected speech parts are processed every 20 ms (50 frames/second).

The supramodal layer and the crossmodal coupling are updated when an individual modality outputs a spike, which may occur once in every frame. Here, it is assumed that the processing time for one frame is the same, regardless of the modality, although it is well known that auditory stimuli are processed faster than visual (difference of approximately 40 to 60 ms [64].

For the speech mode, the number of opinions to validate a person is set proportionally to the size of a given utterance (20% of the total number of frames in an utterance is used). For the visual mode, the number of opinions to authenticate a person is set to two (two frames). Figure 12 shows the best performance obtained on each individual modality. While the best total error (TE) for the face authentication is 21%, the auditory authentication is TE ≈ 38% (varying values of L1 PSP_{Th} in the auditory system and L3 PSP_{Th} in the visual system).

Figure 13 shows the best performance of the system considering the type of integration held in the supramodal layer. First, the crossmodal coupling parameters are set to zero, simulating only the OR and AND integration of individual modalities done by the supramodal layer. Then, the crossmodal coupling is made active ("Crossmodal AND"), setting $CM_{AVexc} = CM_{VAexc} = 0.1$ and $CM_{AVinh} = CM_{VAinh} = 0$. The same parameters are used for individual modalities in this experiment, i.e., auditory parameters (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} ordered according to the performance.

Under the pattern recognition perspective, the results suggest that the integration of modes enhances the performance in several operating points of the system when the learning is done with the same training examples. For a comparative analysis, in [60], the integration of modalities is explored with the VidTimit dataset using a combination of mathematical and statistical methods. The auditory system alone, using MFCC features and GMM in a noise-free setup, reached TE (total error) = FAR (false acceptance rate) + FRR (false rejection rate) ≈ 22%. The visual system is reported to have TE ≈ 8% with features extracted using PCA

(principal component analysis) and SVM (support vector machine) for classification. After testing several adaptive and non adaptive systems to perform integration, the best performance is obtained with a new approach that builds the decision boundaries for integration with consideration of how the distribution of opinions are likely to change under noisy conditions. The accuracy with the integration reached TE ≈ 6% involving 35 users for training and 8 users acting as impostors. Despite some differences between the experiments setup when compared to [60], the results shown in Figure 13 are clearly not as good. Nonetheless, to extract the best performance from the system and evaluate the crossmodal influence specifically on the pattern discrimination ability, an optimization mechanism needs to be incorporated. Similarly important is to explore different information coding schemes.

5) Conclusion

In this work an integrated biologically inspired audiovisual pattern recognition system was designed and implemented. The system was applied to the person authentication problem. Striving to be closer to the biological way of processing, this work integrated several stages of information processing with a single type of processing unit. From the lower levels of sensory processing to the higher levels of cognition, a simple model of spiking neurons was used. Additional to the biological appeal, SNNs enable a close integration of feature extraction and decision-making modules as well as the integration of multiple modalities. This close integration is mainly possible because the processing time has a meaning in spiking neuron systems. In other words, with spiking neurons, the time a spike takes to travel from one neuron to another can be explicitly set up. The generation of postsynaptic potential also occurs in time, set up through the excitatory/inhibitory time constants of a neuron (τ). These values can be set in accordance with biological measurements. Having the processing time of single units and the time spent in communication between units, the time taken by an area for processing can also be defined. This process can ultimately lead towards the simulation of an entire pathway where the information flows in a relevant time scale. The implication of achieving information processing where the time matters for pattern recognition is that it breaks the existing hard separation between feature extraction and classification. Features are propagated as soon as they are processed and they can arrive at different times in areas where classification is undertaken. Similarly, processing time in different modalities vary. Thus, the individual modalities asynchronously feed a global decision-making process. Computation with real processing time also enables the implementation of crossmodal connections between modalities, where one modality can influence others according to its partial opinions in time. This phenomena can effectively increase overall accuracy (as proved to be the case in the human brain) or make the decision-making process faster [64].

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

In terms of information pathways, neuroscientists have been drawing very accurate and detailed maps of the pathways taken by sensory information. In this research, a very simplified version of the major levels of processing is implemented. For the visual system, the functional behaviour of retina cells, directionally selective cells and complex cells are implemented with a two-dimensional grid of spiking neurons. Only feed-forward connections are used and no adaptation at lower levels is applied. In respect to the auditory speaker recognition process, features extracted from a functional model that resembles the characteristics of the human ear (MFCC) are used during the design and evaluation of the decision-making process for speech signals. A subsequent design using tonotopic organization of the spiking neurons (wavelet-based) is proposed that amounts to the entire processing of sound signals being undertaken with spiking neurons. The integration of modalities is also accomplished with spiking neurons. Supramodal layers of spiking neurons as well as crossmodal connections were implemented.

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

Finally, in 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 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. On another front, it is necessary to define learning rules for integrative modules as well as a systematic procedure to train crossmodal connections.

In that which is concerned with quantitative analysis, the main results are:

1. **Visual system**. A SNN-based multi-view face authentication system demonstrated:

a) the ability to adaptively learn from multiple frames. More frames for training of a class increased the accuracy. A peak in performance is reached after five frames.

b) the ability of the system to accumulate opinions from several frames for decision-making. More test frames increased accuracy. The accuracy level flattens after five frames.

2. **Auditory system**. In the text-independent speaker authentication scenario using SNNs, the adaptive learning procedure was used to create speaker codebooks. Neuronal maps representing background models were also introduced to achieve similarity normalization. A SNN architecture was proposed, which achieved similar levels of performance when compared with a traditional Linear Vector Quantization (LVQ) model to authenticate 43 users uttering short-sentences.
3. **Audiovisual system**. A supramodal area as well as crossmodal connections were used to process audiovisual features for person authentication. Different

configurations of the integrated system clearly outperformed individual modalities.

Spiking time theory (in opposition to the spiking rate theory) was used in this work for the conceptual design and implementation of algorithms. In particular, a spiking neuron model was used that privileges early spikes and a constraint that enabled the occurrence of only one spike per neuron [60]. Based on these assumptions, concrete models were implemented and validated. However, as coding schemes utilized by the brain are still not clearly understood other spike-based coding mechanisms can be evaluated. A good introduction to the issues related to the encoding of information in neuronal activity can be found in [54]. A traditional theory, suggests that information is transmitted by firing rates (see [21, 43]). This theory is proving gradually not to be completely right, as several independent neurophysiological experiments demonstrate the existence of spike-timing patterns in both single and in ensembles of neurons. For instance, in [68], *in vivo* measurements enabled to the prediction of rat's behaviour responses through the analysis of spatio-temporal patterns of neuronal activity. Izhikevich [30] created the term "polychronization" to define the spatio-temporal behaviour of a group of neurons that are "time-locked" to each other, a term to distinguish it from synchronous, asynchronous or polysynchronous spiking activity behaviour. Abeles, in 1982 [2], 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 [1]) and computational models explored the storage and learning capabilities of this theory [3, 25]. From all these theories, it is also reasonable to believe that different areas in the brain can utilize different coding schemes. If this is the case, combined approaches would be needed to better represent a given information pathway.

6) References

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

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

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

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

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

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] 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.

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

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

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

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

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

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

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

[20] Ganchev, T. (2005). Speaker Recognition, PhD Thesis, Dept. of Electrical and Computer Engineering, University of Patras, Greece.

[21] Gerstner, W., & Kistler, W. M. (2002). Spiking Neuron Models. Cambridge University Press, Cambridge, MA.

[22] 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.

[23] 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.

[24] Gray, R. M. (1984). Vector quantization. IEEE Acoustics, Speech, and Signal Magazine, 4-28.

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

[26] Holmberg, M., Gelbart, D., Ramacher, U., & Hemmert, W. (2005). Automatic speech recognition with neural spike trains. Interspeech, 1253-1256.

[27] Hopfield, J. J. (1995). Pattern recognition computation using action potential timing for stimulus representation. Nature, 376 (6535), 33-36.

[28] 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.

[29] Iwasa, K., Inoue, H., Kugler, M., Kuroyanagi, S., & Iwata, A. (2007). Separation and recognition of multiple sound source using pulsed neuron model. ICANN, Lecture Notes in Computer Science, 4669, 748–757.

[30] Izhikevich, E. M. (2006). Polychronization: computation with spikes. Neural Computation, 18 2, 245-282.

[31] Kasabov, N. (2007). Evolving Connectionist Systems. Springer-Verlag.

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

[33] 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.

[34] Kriegstein, K. von, & Giraud, A. (2006). Implicit multisensory associations influence voice recognition. PLoS Biology, 4 (10), 1809-1820.

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

[36] Kruger, N., Lappe, M., & Worgotter, F. (2004). Biologically motivated multi-modal processing of visual primitives. Interdisciplinary Journal of Artificial Intelligence the Simulation of Behaviors, AISB, 15, 417-428.

[37] 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.

[38] Loiselle, S., Rouat, J., Pressnitzer, D., & Thorpe, S. (2005). Exploration of Rank Order Coding with spiking neural networks for speech recognition, IJCNN. 2076-2080, Montreal.

[39] Maciokas, J. B. (2003). Towards an understanding of the synergistic properties of cortical processing: a neuronal computational modeling approach. PhD Thesis, University of Nevada.

[40] Maciokas, J., Goodman, P. H., & Harris Jr., F. C. (2002). Large-scale spike-timing dependent-plasticity model of bimodal (audio/visual) processing. Technical Report of Brain Computation Laboratory, University of Nevada, Reno.

[41] Matsugu, M., Mori, K., Ishii, M., & Mitarai, Y. (2002). Convolutional spiking neural network model for robust face detection. International Conference on Neural Information Processing, ICONIP, 660-664.

[42] 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.

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

[44] McLennan, S., & Hockema, S. (2001) Spike-V: an adaptive mechanism for speech-rate independent timing. IULC Working Papers Online 02-01.

[45] 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.

[46] Mercier, D., & Seguier, R. (2002). Spiking neurons (STANNs) in speech recognition. 3rd WSES International Conference on Neural Networks and Applications. Interlaken.

[47] Movellan J. R. (1995). Visual speech recognition with stochastic networks. In: Tesauro, G., Toruetzky, D., & Leen, T. (eds), Advances in Neural Information Processing Systems. 7, 851-858.

[48] Mozayyani, N., Baig, A. R., & Vaucher, G. (1998). A fully neural solution for on-line handwritten character recognition. International Joint Conference on Neural Networks, IJCNN, (pp. 160-164). Alaska.

[49] Natschlager, T., & Ruf, B. (1999). Pattern analysis with spiking neurons using delay coding. Neurocomputing, 26-27, 463-469.

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

[51] Perrinet, L., & Samuelides, M. (2002). Sparse image coding using an asynchronous spiking neural network. European Symposium on Artificial Neural Networks. (pp. 313-318). Bruges.

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

[53] Rabiner, L., & Juang, B. (1993). Fundamentals of Speech Recognition. Prentice Hall, New Jersey.

[54] Reece, M. (2001). Encoding information in neuronal activity. In: Maass, W., & Bishop, C. (eds). Pulsed Neural Networks. MIT Press.

[55] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing, 10, 19-41.

[56] Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2 (11), 1019-1025.

[57] Robert, A., & Eriksson, J. L. (1999). A composite model of the auditory periphery for simulating responses to complex sounds. Journal of Acoustics Society of America, 106 (4), 1852-1864.

[58] 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.

[59] Rouat, J., Pichevar, R., & Loiselle, S. (2005). Perceptive, non-linear speech processing and spiking neural networks. In: Chollet, G. *et al* (eds), Nonlinear speech modelling. Lecture Notes on Artificial Intelligence, 3445, 317-337.

[60] Sanderson, C., & Paliwal, K. K. (2002). Identity verification using speech and face information. Digital Signal Processing, 14, 449-480.

[61] Seguier, R., & Mercier, D. (2001). A generic pretreatment for spiking neuron. application on lipreading with STANN (Spatio-Temporal Artificial Neural Networks). 5th International Conference on Artificial Neural Networks and Genetic Algorithms.

[62] 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.

[63] 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.

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

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

[66] Thorpe, S. (1990). Spike arrival times: a highly efficient coding scheme for neural networks. In: Eckmiller, R., Hartman, G., & Hauske, G. (eds). Parallel processing in neural systems, Elsevier. 91-94.

[67] Vaucher, G. (1998). An algebraic interpretation of PSP composition. Biosystems, 48, 241-246.

[68] Villa, A. E., Tetko, I. V., Hyland, B., & Najem, A. (1999). Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioned task. Proceedings of the National Academy of Sciences (pp. 1106-1111). USA.

[69] Viola, P., & Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 511-518.

[70] Wysoski, S. G., Benuskova, L., & Kasabov, N. (2008). Adaptive spiking neural networks for audiovisual pattern recognition. ICONIP, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 4985, 406-415.

[71] Wysoski, S. G., Benuskova, L., & Kasabov, N. (2008). Fast and adaptive network of spiking neurons for multi-view visual pattern recognition. Neurocomputing. doi: 10.1016/j.neucom.2007.12.038.

[72] Wysoski, S. G., Benuskova, L., & Kasabov, N. (2006). On-line learning with structural adaptation in a network of spiking neurons for visual pattern recognition. ICANN, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 4131, 61-70.

[73] Wysoski, S. G., Benuskova, L., & Kasabov, N. (2007). Text-independent speaker authentication with spiking neural networks. ICANN, Lecture Notes in Computer Science, Springer-Verlag, New York, 4669, 758-767.

List of Tables

List of Figures

Table1

Processing Units	A fast and computationally inexpensive version of spiking neuron is used as processing unit in all stages of visual information processing (Delorme and 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, direction selective cells, complex cells).
Learning	The online evolving procedure enables the learning of external stimuli through synaptic plasticity and structural adaptation. The addition of new classes is done in a supervised way whereas the system adapts in an unsupervised fashion when new samples of a class are presented (see (Wysoski *et al*, 2008) for details).

Table2

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 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 learning of external stimuli through synaptic plasticity and structural adaptation. The addition of new classes is done in a supervised way. The adaptive learning creates/merges neurons that respond optimally to different speech phones in a supervised or unsupervised fashion when new utterances of a class are presented.

Table3

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 learning of external stimuli through synaptic plasticity and structural adaptation separately for each modality. Algorithms to train the strength of crossmodal connections and weights of the supramodal layer still need to be designed.

Figure1

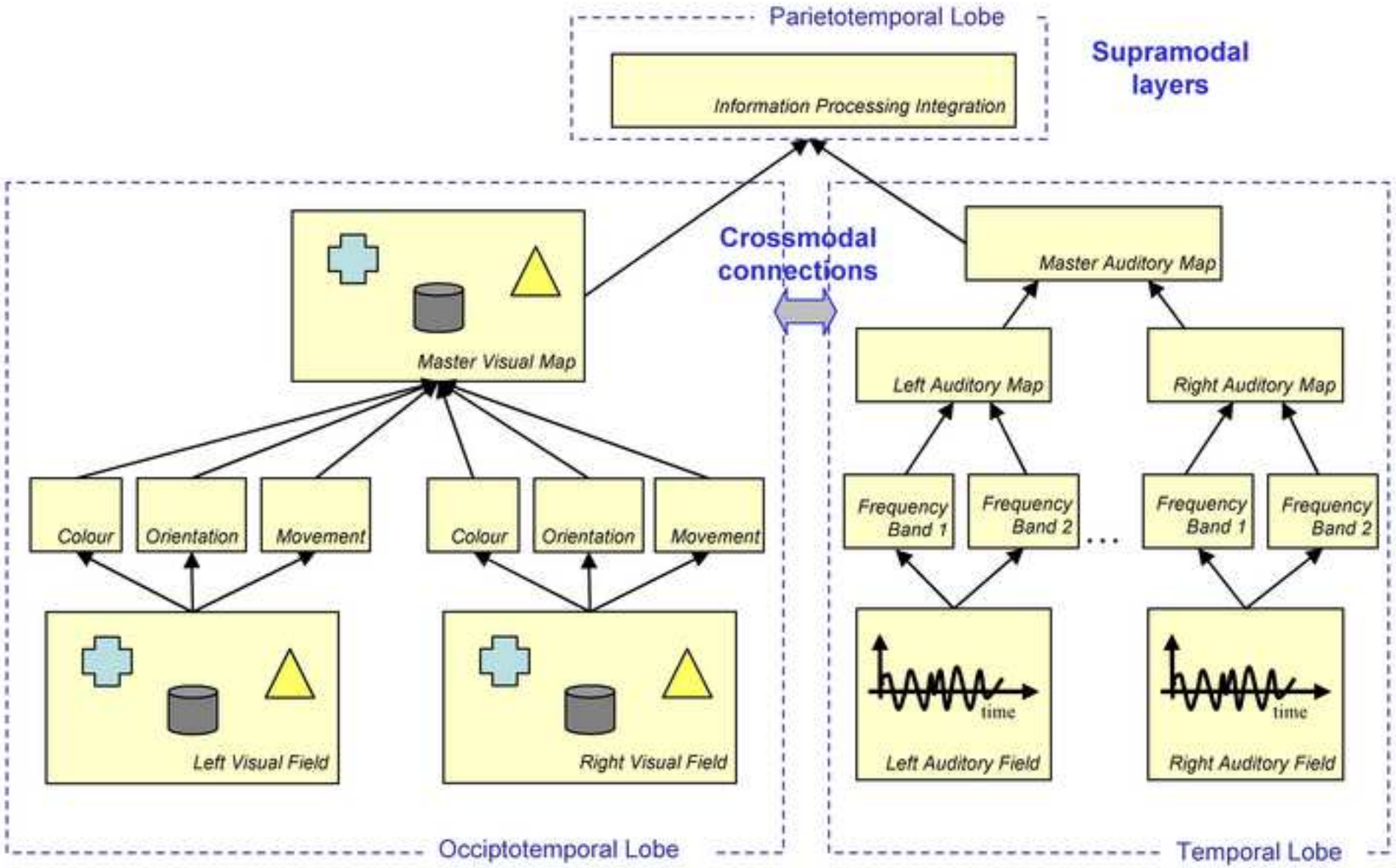
Parietotemporal Lobe
Information Processing Integration
Supramodal layers
Crossmodal connections
Master Visual Map
Master Auditory Map
Left Auditory Map
Right Auditory Map
Colour
Orientation
Movement
Colour
Orientation
Movement
Frequency Band 1
Frequency Band 2
Frequency Band 1
Frequency Band 2
time
time
Left Visual Field
Right Visual Field
Left Auditory Field
Right Auditory Field
Occiptotemporal Lobe
Temporal Lobe

Figure2

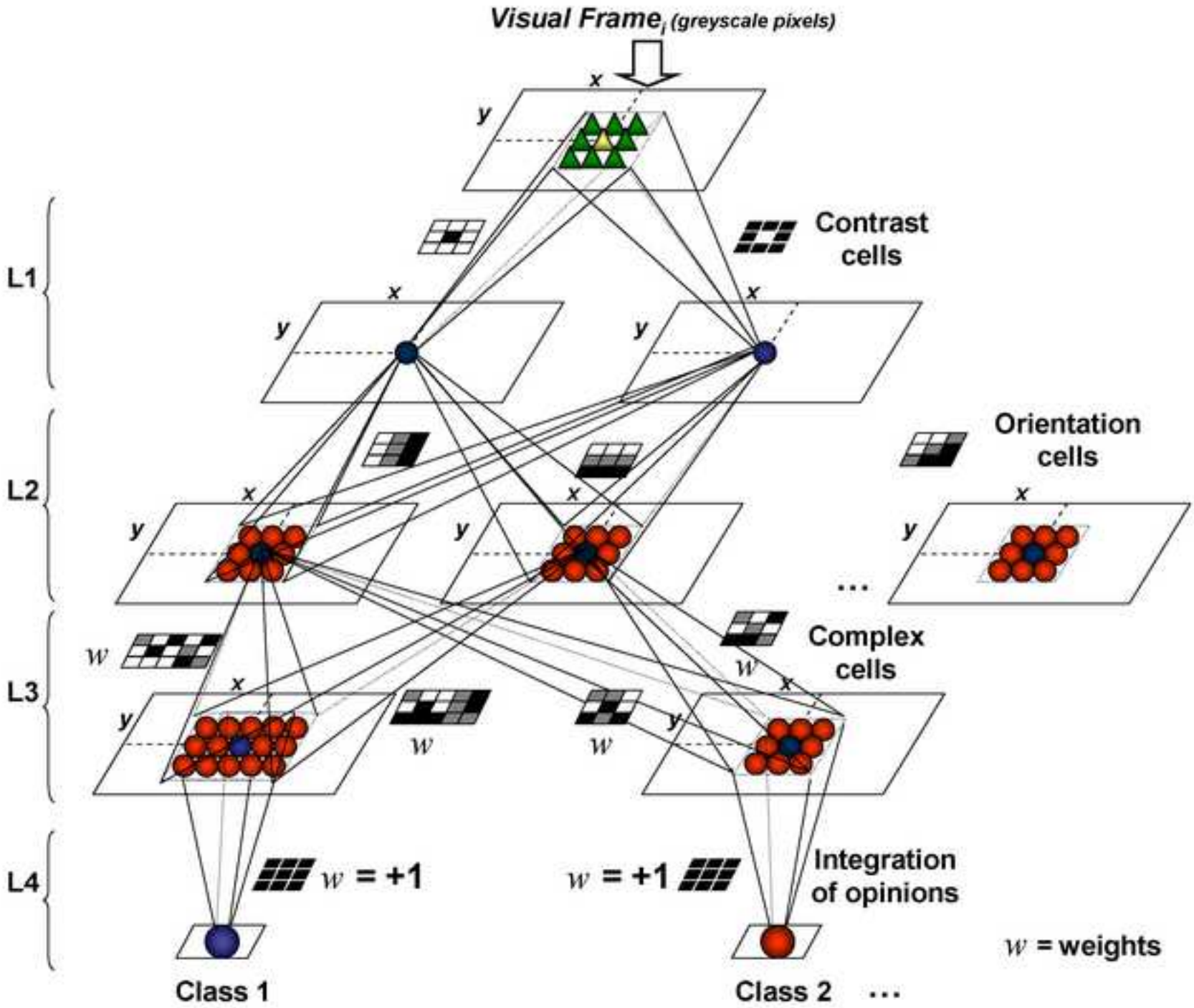
Visual Frame$_i$ (greyscale pixels)
x
y
L1
Contrast cells
L2
Orientation cells
...
L3
Complex cells
w
L4
w = +1
w = +1
Integration of opinions
w = weights
Class 1
Class 2 ...

Figure3

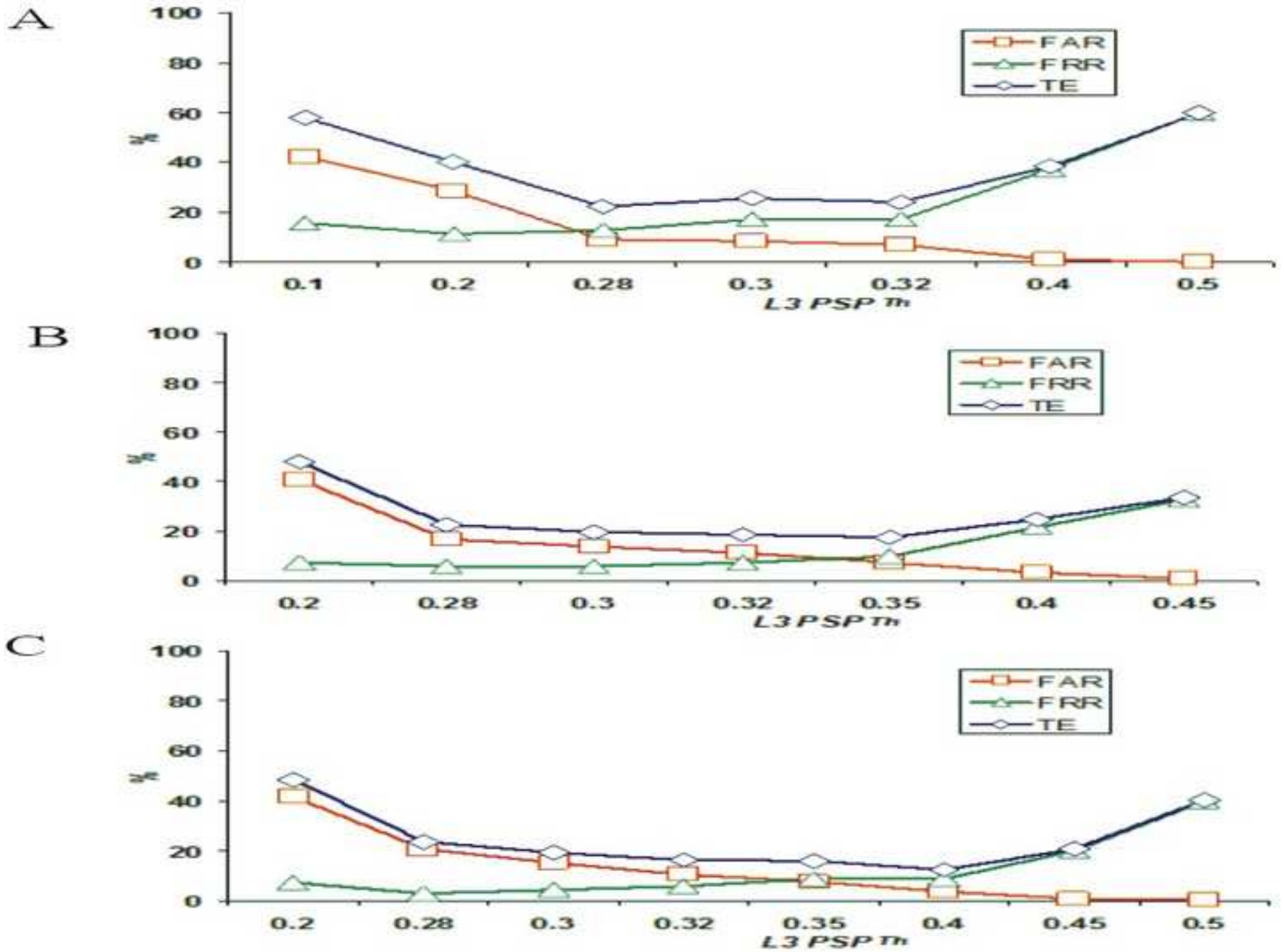
A
100
80
60
40
20
0
%
FAR
FRR
TE
0.1
0.2
0.28
0.3
0.32
0.4
0.5
L3 PSP Th
B
100
80
60
40
20
0
%
FAR
FRR
TE
0.2
0.28
0.3
0.32
0.35
0.4
0.45
L3 PSP Th
C
100
80
60
40
20
0
%
FAR
FRR
TE
0.2
0.28
0.3
0.32
0.35
0.4
0.45
0.5
L3 PSP Th

Figure4

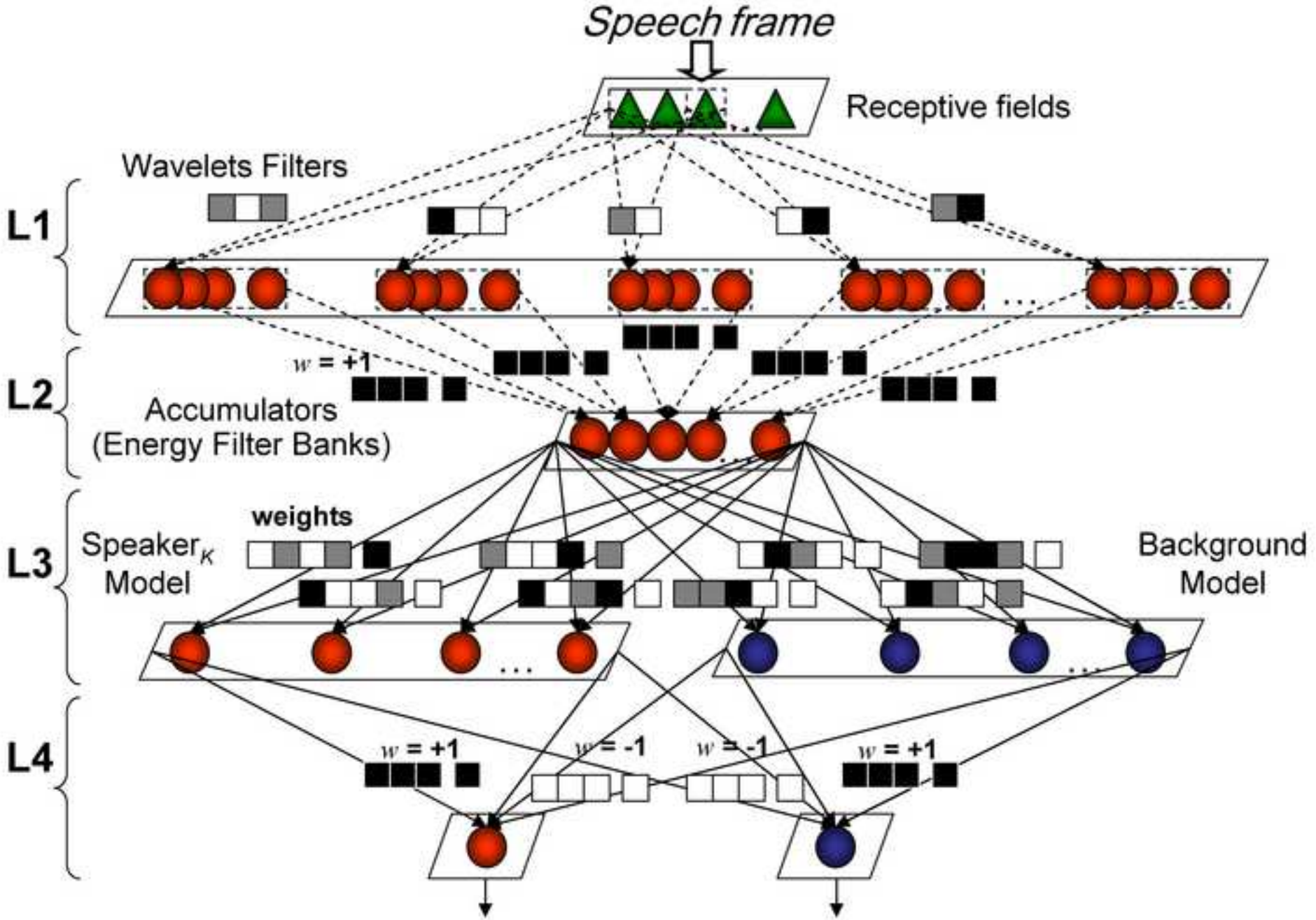
Speech frame
Receptive fields
Wavelets Filters
L1
L2
w = +1
Accumulators
(Energy Filter Banks)
weights
L3
Speaker$_K$
Model
Background
Model
L4
w = +1
w = -1
w = -1
w = +1

Figure5

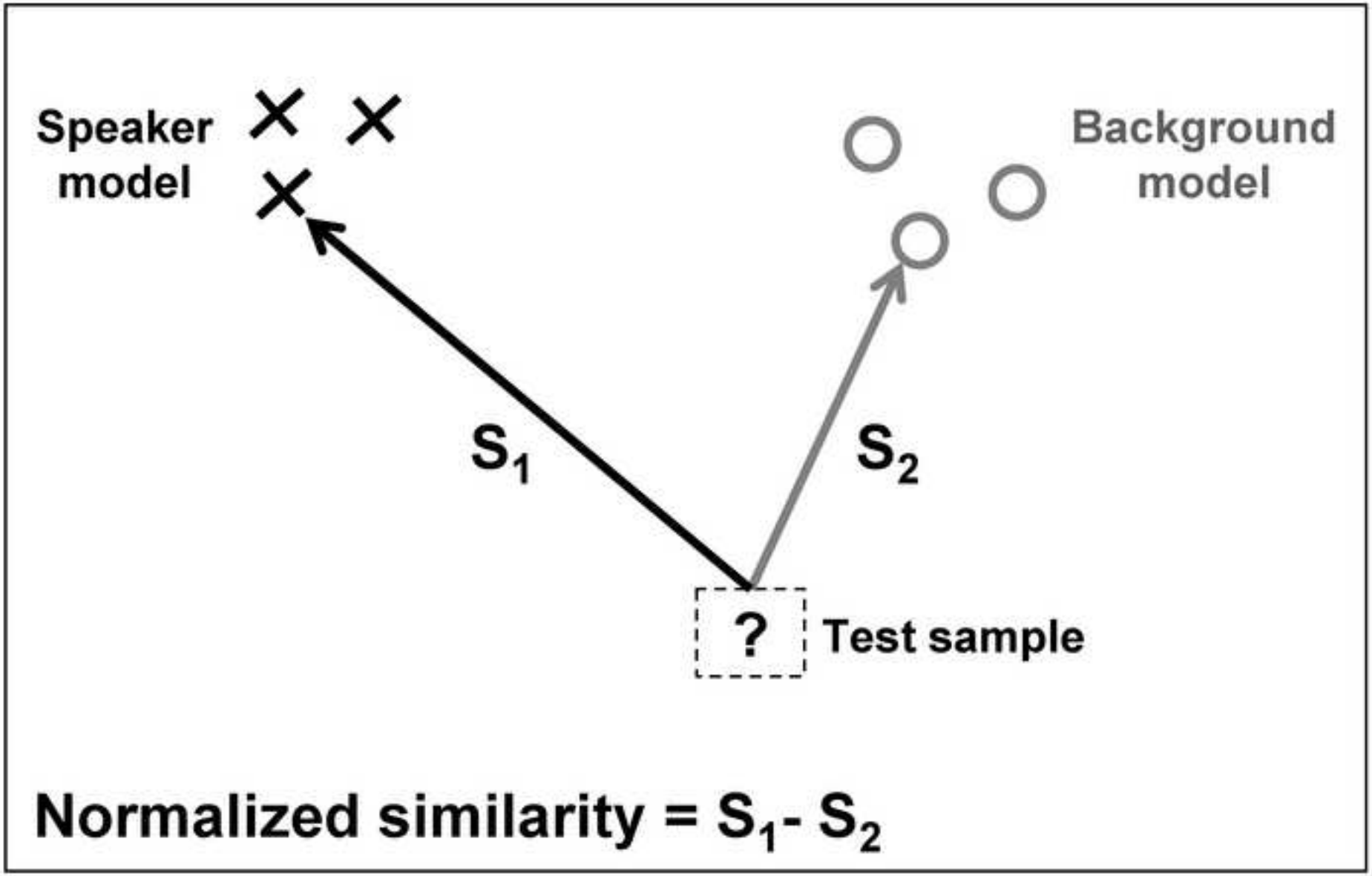
Speaker model
Background model
S_1
S_2
?
Test sample
Normalized similarity = S_1- S_2

Figure6

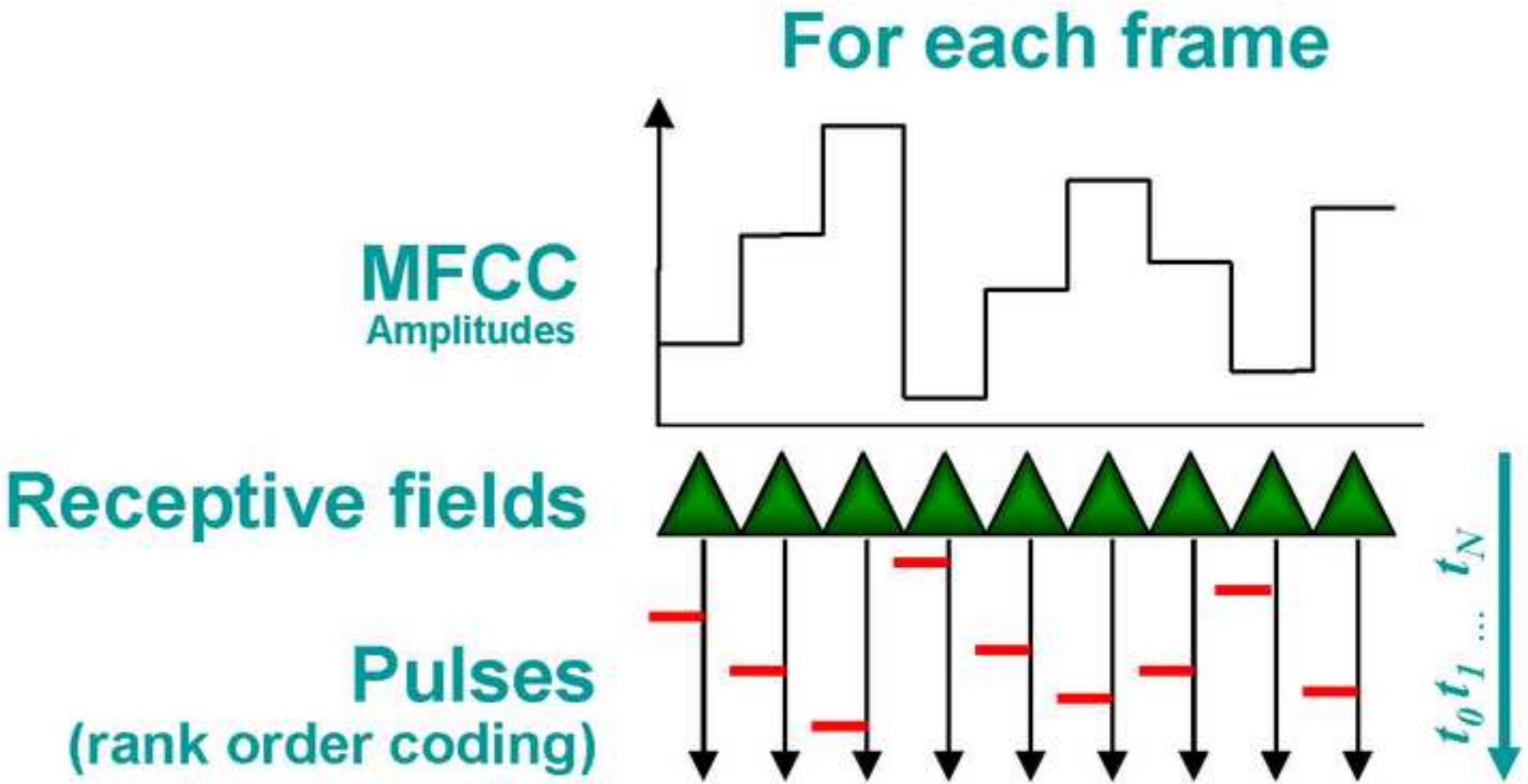
For each frame
MFCC
Amplitudes
Receptive fields
Pulses
(rank order coding)
$t_0 t_1 \ldots t_N$

Figure7

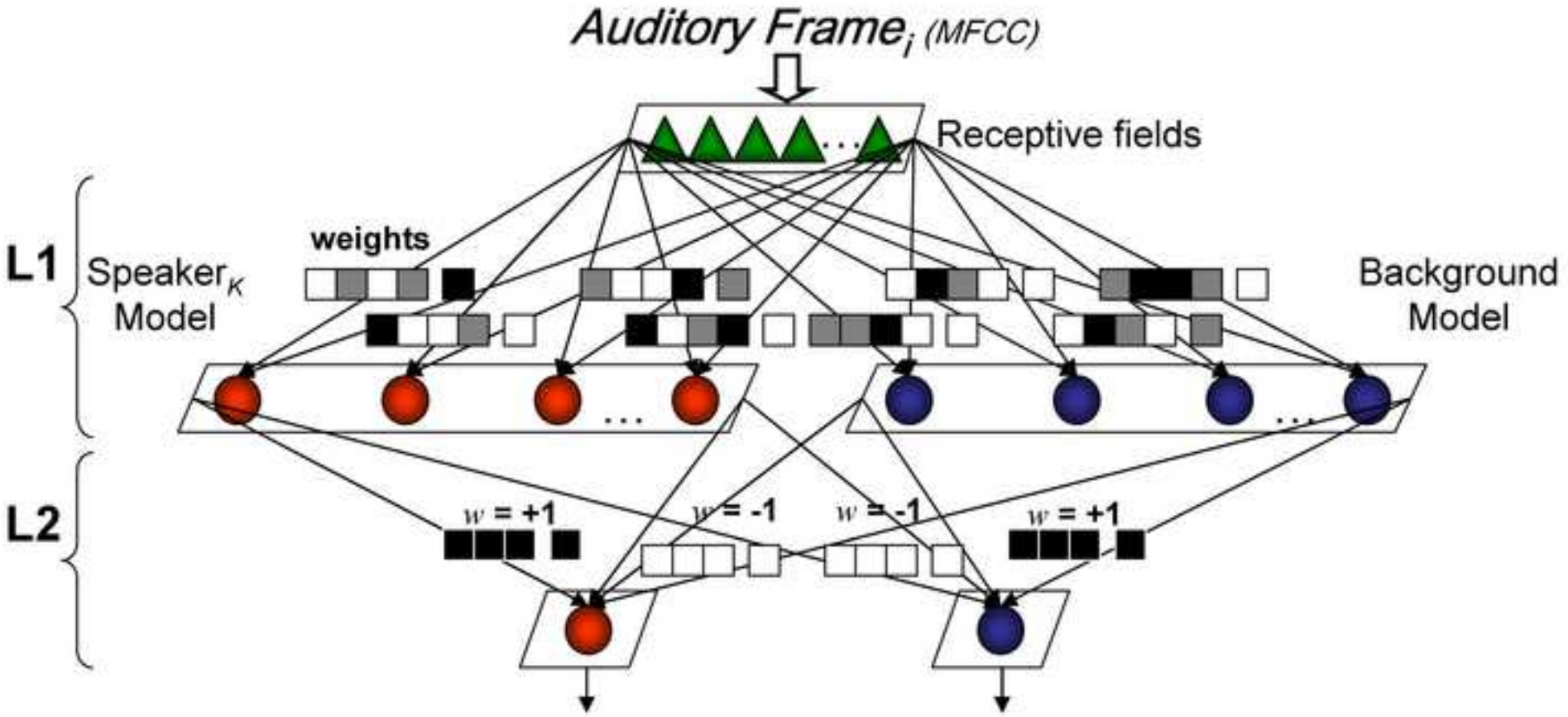
Auditory Frame$_i$ (MFCC)
Receptive fields
L1
Speaker$_K$ Model
weights
Background Model
L2
w = +1
w = -1
w = -1
w = +1

Figure8

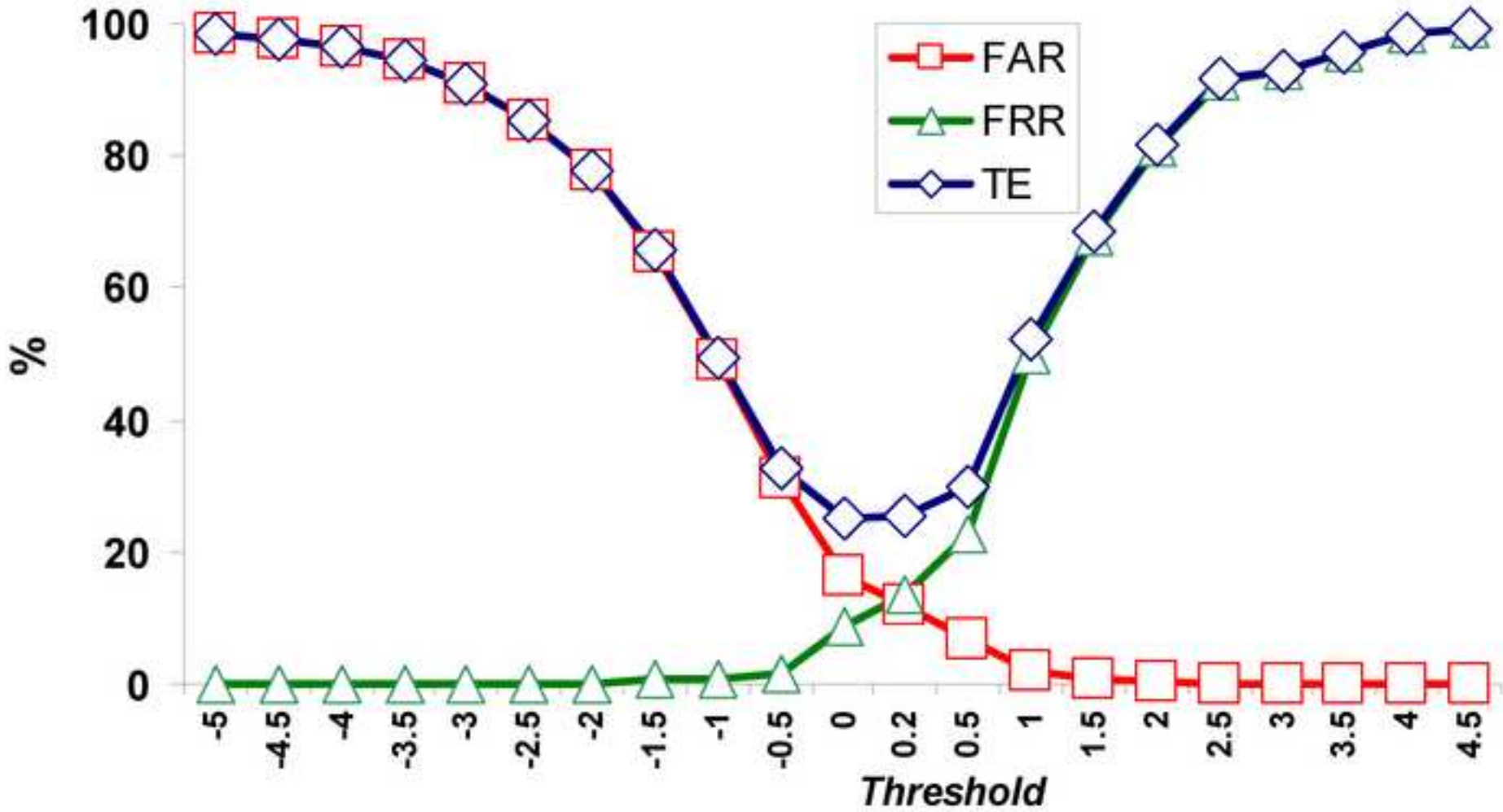
100
80
60
40
20
0
%
FAR
FRR
TE
-5
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.2
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Threshold

Figure9

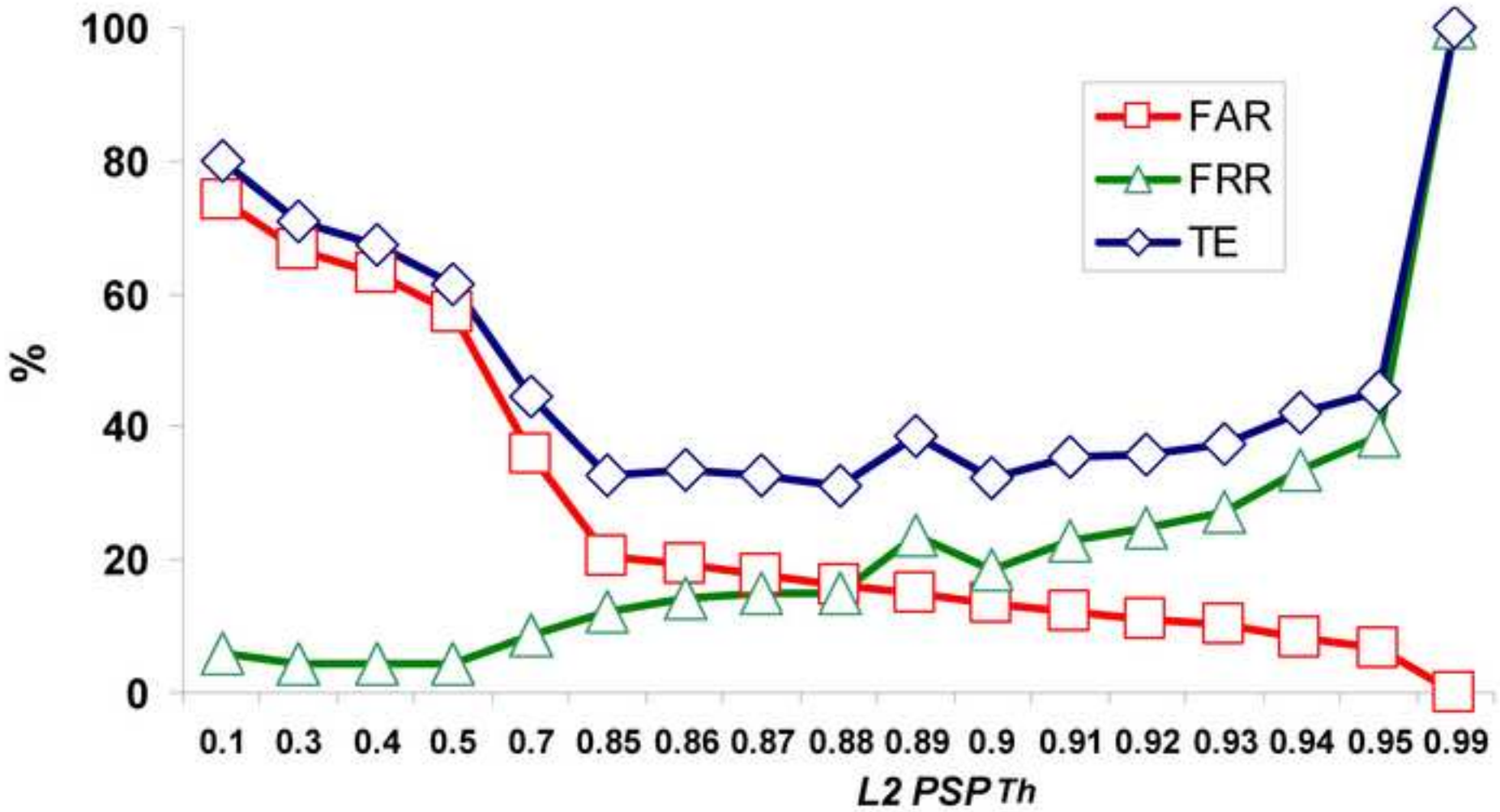
100
80
60
40
20
0
%
FAR
FRR
TE
0.1
0.3
0.4
0.5
0.7
0.85
0.86
0.87
0.88
0.89
0.9
0.91
0.92
0.93
0.94
0.95
0.99
L2 PSP Th

Figure10

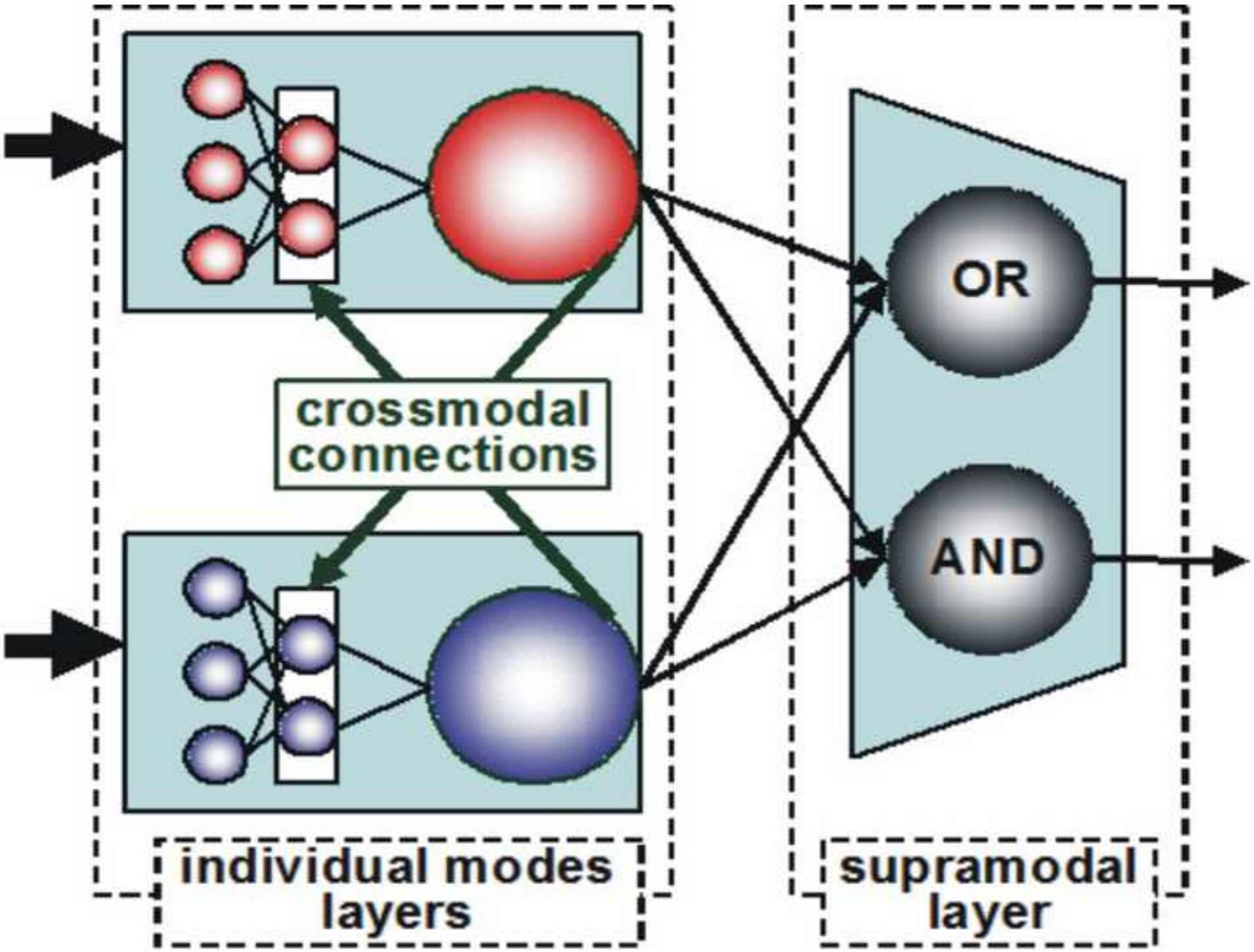
OR
AND
crossmodal
connections
individual modes
layers
supramodal
layer

Figure11

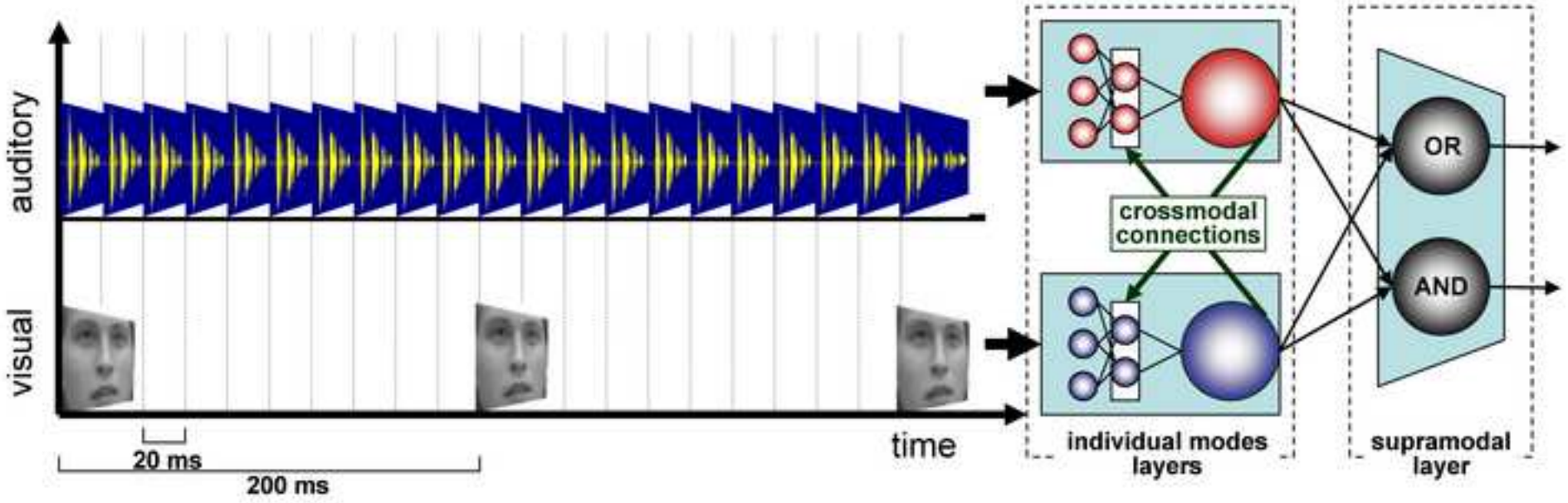
auditory
visual
20 ms
200 ms
time
crossmodal
connections
OR
AND
individual modes
layers
supramodal
layer

Figure12Top

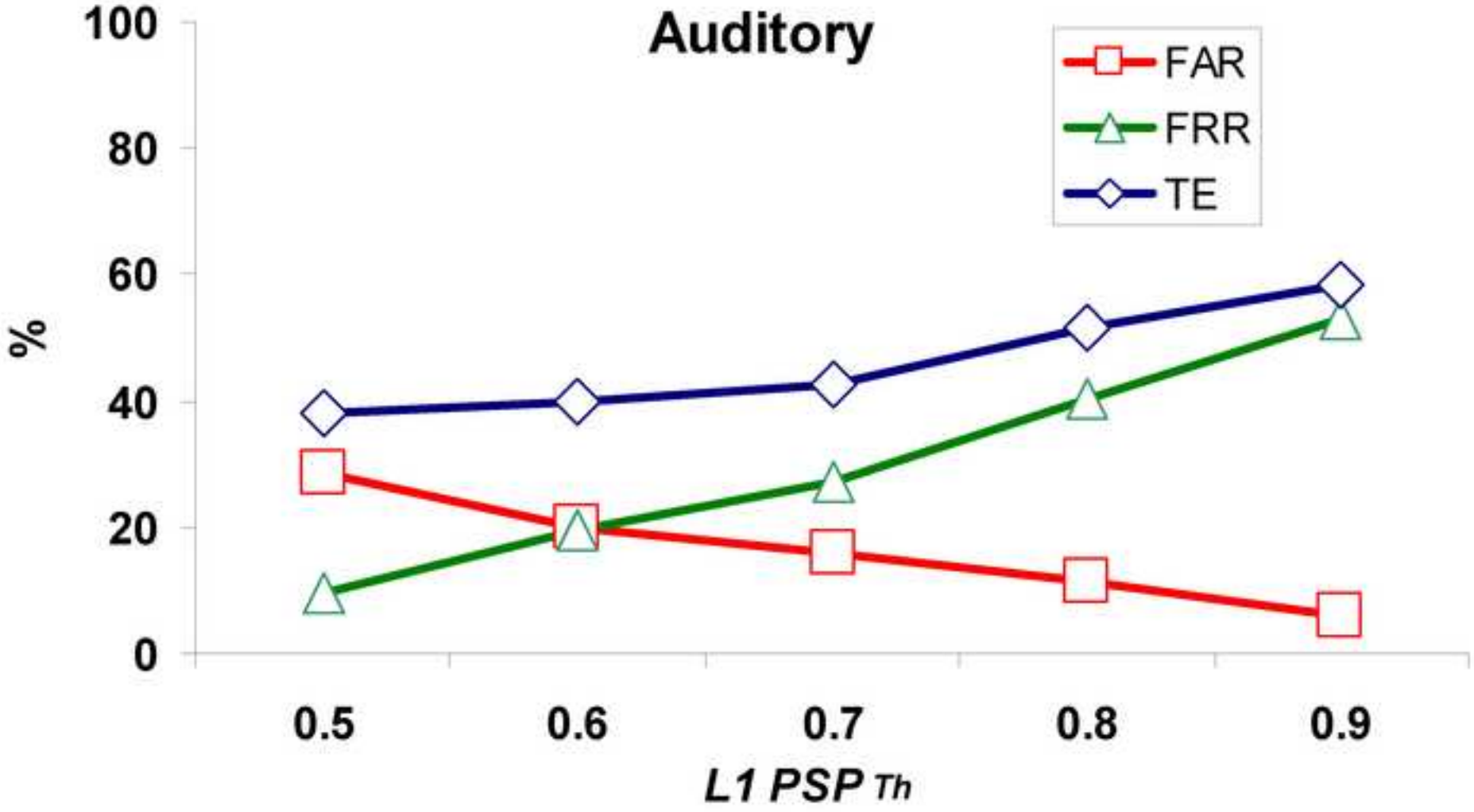

Figure12Bottom

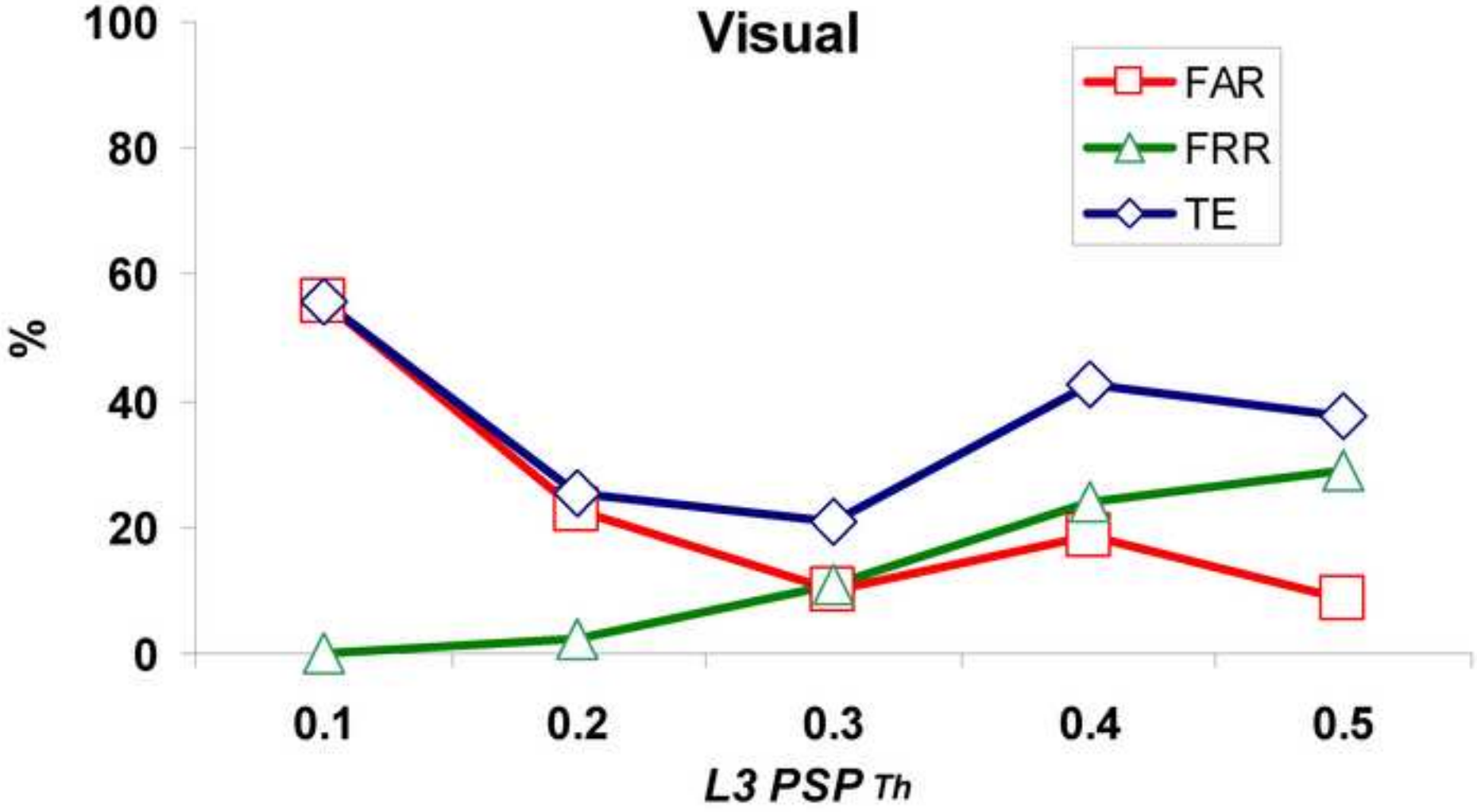

Figure13a

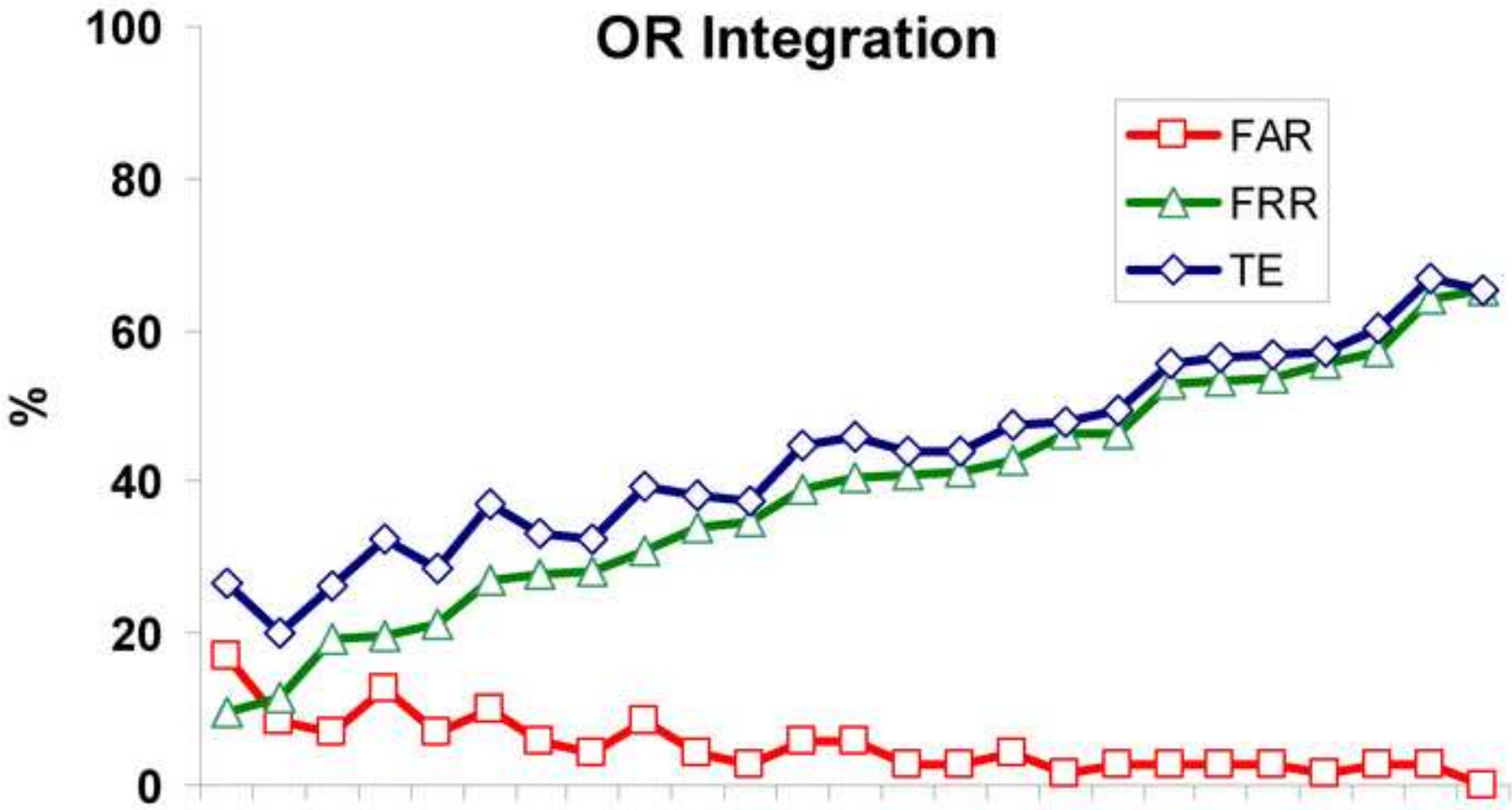

Figure13b

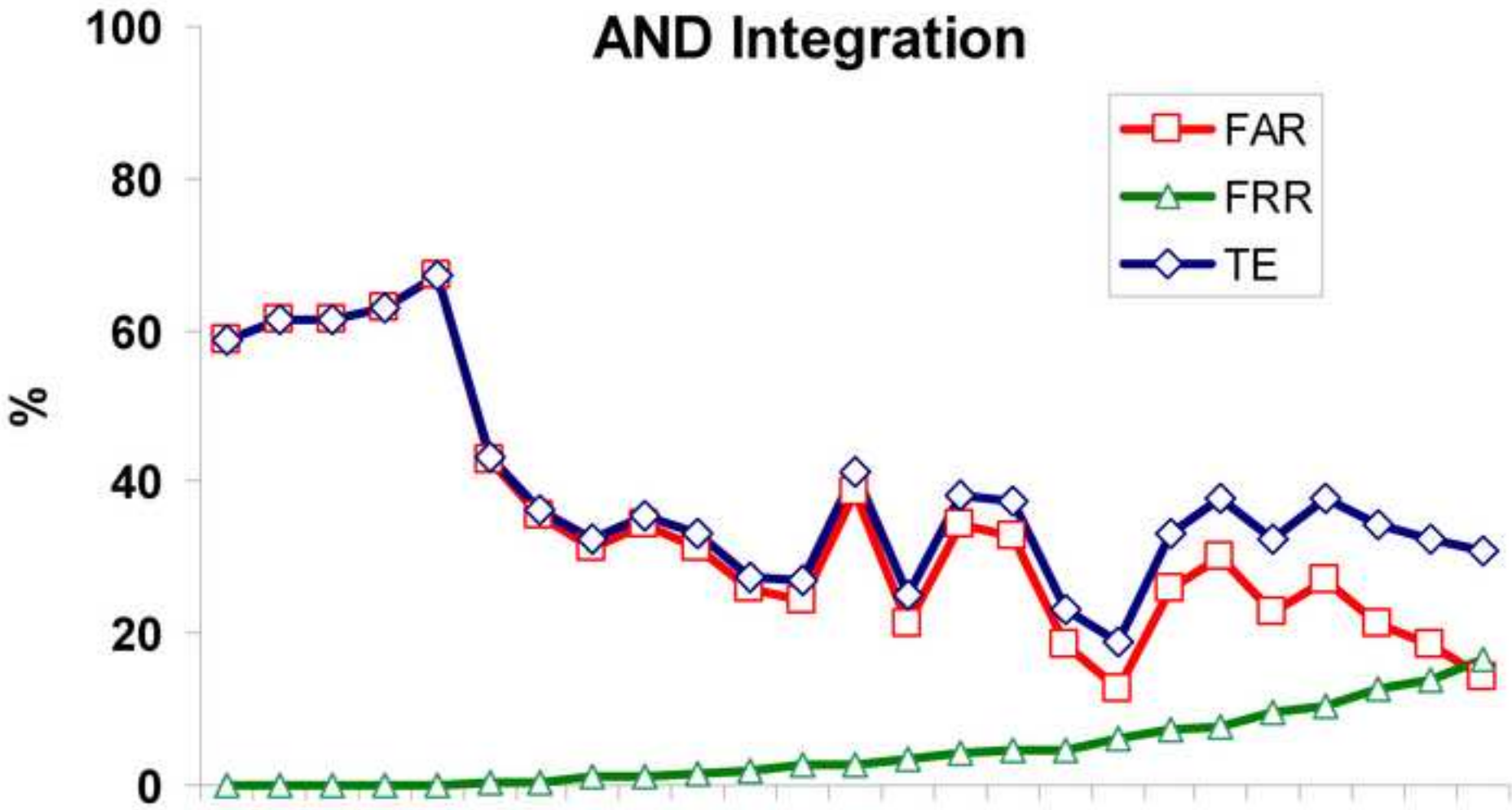

Figure13c