Cognitive Audio-Visual Associative Memories using Brain-inspired Spiking Neural Networks with Case Studies on Moving Object Recognition

Journal:	*IEEE Transactions on Cognitive and Developmental Systems*
Manuscript ID	Draft
Manuscript Type:	Regular
Date Submitted by the Author:	n/a
Complete List of Authors:	Kasabov, Nikola; Auckland University of Technology, School of Computing and Mathematical Sciences Bhattacharya , Basabdatta; BITS Pilani - KK Birla Goa Campus Rohitkumar, Patel ; BITS Pilani - KK Birla Goa Campus Aggarwal, Naman ; BITS Pilani - KK Birla Goa Campus Bankar, Tanmay ; BITS Pilani - KK Birla Goa Campus Abouhassan, Iman; TU-Sofia
Keywords:	cognitive computing, spiking neural networks, associative memory, audio-visual data, moving object recognition

Cognitive Audio-Visual Associative Memories using Brain-inspired Spiking Neural Networks with Case Studies on Moving Object Recognition

Nikola Kasabov, *Life Fellow, IEEE*, Basabdatta Sen Bhattacharya, *Senior Member, IEEE*, Dharmik Patel, Naman Aggarwal, Tanmay Bankar, Iman AbouHassan

***Abstract*—The paper introduces for the first time a framework for** cogntive **audio-visual associative memories based on brain-inspired spiking neural networks. The work is inspired by the associative memory ability of the human brain to integrated both audio and visual information for a better performance on classification, identification or prediction tasks. When using only one of the modalities for a recall of such a model** its **performance** can **still** be **better than using a single modality for these tasks. Our framework is based on the NeuCube brain-inspired SNN architecture, where audio- and visual streaming data are entered in a synchronised way into the areas of 3D SNN model that correspond to the auditory and the visual cortex according to a predefined brain template. The framework is manifested on two case study data of moving object recognition, incorporating video and audio information, and recalling the models on only one of the modalities. The results demonstrate the potential of the introduced** cognitive **framework and its applicability for predictive modeling in wider areas of** autonomous robots and vehicles, cognitive studies; assistive aids for blind and deaf people, security systems and other**.** The models can be incrementally developed to learn and recognise new audio-visual categories.

***Index Terms*—**cogntive computing; s**piking neural networks, audio-visual associative memory, moving object recognition.**

I. Introduction

THE brain integrates both sounds and images in a natural way to achieve better and faster recognition and location in the environment [1]. The auditory system in the brain not only deals with sound recognition but also performs source localization within a frequency range. The visual system allows us to perceive the environment. Both auditory and visual systems play a critical role in survival and motion detection.

N.K. Kasabov* (nkasabov@aut.ac.nz) is a Professor with the School of Engineering, Computer and Mathematical Science, Auckland University of Technology, AUT WZ building, St. Paul, Auckland,1010, New Zealand. He is also Chair Professor with the Intelligent Systems Research Centre at University of Ulster UK, and Visiting Professor with the Institute for Information and Communication technologies (IICT) Bulgarian Academy of Sciences and with Dalian University, China. He is Honorary Professor of the University of Auckland, NZ and Teesside University UK.

B. Sen Bhattacharya (basabdattab@goa.bits-pilani.ac.in) is an Associate Professor with the Department of Computer Science and Information Systems, Birla Institute of Technology and Science (BITS) Pilani, Goa Campus, Goa, India.

D. Patel, N. Aggarwal and T. Bankar are undergraduate students with the Department of Computer Science and Information Systems, BITS Pilani, Goa Campus, Goa, India.

I. AbouHassan (iabouhassan@tu-sofia.bg) is a Ph.D. candidate at the Technical University of Sofia, Bulgaria, and holds a senior position at Lebanon's Central Bank, Beirut, Lebanon (iabouhassan@bdl.gov.lb).

The image on the retina is transmitted via the optic nerve to the occipital lobe which is the seat of the visual cortex. It is here that the 'what' (recognition) and 'where' (localization) tracts are separated. It is still a fundamental question as to how the eyes and the ears work in synchrony for audio-visual (AV) signal integration and perception, but there are several mathematical and engineering approaches to integrate both modalities [2] [3]. Today, brain-inspired artificial intelligence is a much sought-after field [1]. In this regard, one critical question that needs answering is how we can utilize the efficient multi-modal information processing techniques of the brain to develop efficient machine learning algorithms. It is important that such systems are non-greedy in terms of power consumption, and can be easily adapted to run on green energy [4-13]. This project develops the world's first cognitive audio-visual associative memory based on brain-inspired spiking neural networks, called here CAViAM. It uses the NeuCube SNN architecture [14] [1] that is suitable for neuromorphic implementation [11] [17]. The latter is known for massively parallel and fast information processing and for low power consumption. As a feasibility study, we demonstrate the applicability of our method in two case studies, where audio-visual information is used to train a system for object recognition and then only one of the modalities is used to recall the system without sacrificing the recognition accuracy. Multi-sensory information processing architectures have been around for some time [15,16]. Previous works have investigated this area using a frame-based representation of audio and visual data using fuzzy neural networks [2] and evolving Spiking Neural Networks (SNN) [3]. SNNs use spikes at times for infoprmtion representation that allows for event-based signal processing. However, there are no existing methods for cognitive audio-visual associative memories based on brain-inspired SNN.

Our proposed CAViAM architecture will be the first using multi-modal spatio-temporal streaming data allowing for implementaion on low energy consuming neuromorphic hadware platforms. SNN are the 'next generation' neural networks that use biologically inspired events, commonly termed as spikes, to communicate amongst themselves, but in a 'need-based' manner. In doing so, they implement sparse computation and save energy. The challenge is to implement the architecture in a real-time. Our proposed architecture will be demonstrated for object recognition on-line in response to streaming multi-modal data, but its applications span across areas of robotics, cognitive studies, security, environment, unmanned vehicles, etc.

The paper is organized as follows. Section two presents some main principles of spiking neural networks and the brain-inspired NeuCube architecture. Section 3 reveals the proposed CAViAM model. Section 4 presents experimental results with two audio-visual data sets for object recognition. Section 5 presents some notions related to implementing CAViAM models on neuromorphic platforms and section 6 is the discussion and future work part.

II. Spiking neural networks and the brain-inspired NeuCube architecture

A. Spiking Neural Networks (SNN)

Spiking neural networks (SNN) are biologically inspired ANN where information is represented as binary events (spikes), similar to the event potentials in the brain, and learning is also inspired by principles in the brain. SNNs are also universal computational mechanisms. Learning in SNN relates to changes in the connection weights between two spiking neurons. Many learning paradigms, such as STDP, are inspired by the Hebbian learning principle [1, 22], in which the synaptic weights are adjusted based on the tem-poral order of the incoming spike (pre-synaptic) and the output spike (post-synaptic). This synaptic weight adjustment determines synaptic potentiation known as long-term potential (LTP) if the synaptic weight is increasing (positive change) and synaptic depression known as long-term depression (LTD) if the synaptic weight is decreasing (negative change). A particular connection is said to potentiate if a pre-synaptic spike arrives before a post-synaptic spike and is said to depress if it arrives after a post-synaptic spike. STDP is expressed in terms of STDP learning window W $(t_{pre}-t_{post})$ in which the difference between arrival time of the pre-synaptic spike and the arrival time of the post-synaptic spike will determine the synaptic weight (Equation 1). In the equation, τ_+ and τ_- refer to the pre-synaptic and post-synaptic time interval, and A_+ and A_- refer to the maximum fraction of synaptic adjustment if $t_{pre}<t_{post}$ approaches to zero.

$$W(t_{pre}-t_{post}) = \begin{cases} A_+ exp(\frac{t_{pre}-t_{post}}{\tau_+}), & \text{if } t_{pre} < t_{post} \\ A_- exp(-\frac{t_{pre}-t_{post}}{\tau_-}), & \text{if } t_{pre} > t_{post} \end{cases} \quad (1)$$

Izhikevich [22] has shown that similar activation patterns (called 'polychronous waves') can be generated in a SNN reservoir with recurrent connections to represent short term memory. This is a further extension of the 'synfire chain' theory by Abeles [19]. When using STDP learning, connection weights change to form LTP or LTD, which constitute long-term memory. Learned chains of connections can be 'stitched together' when additional data is used for further training. These previous results suggest that a SNN architecture can be explored for learning long (spatio-) temporal patterns and to be used as associative memory.

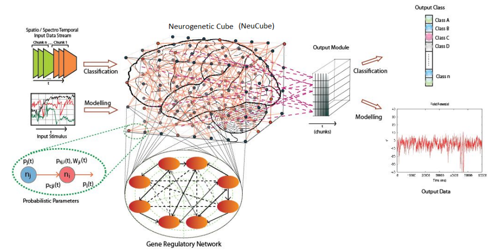


Fig. 1. The NeuCube brain-inspired SNN architecture (from [14]).

B. NeuCube

The NeuCube architecture is depicted in Fig.1. It consists of the following functional modules [14]:

- Input data encoding module;
- 3D SNN reservoir module (SNNc);
- Output function (classification) module;
- Gene regulatory network (GRN) module (optional);
- Parameter optimization module (optional).

The table below describes the functionality of the NeuCube (from [14] [1] [34]).

1) Temporal inputs (features) are converted into spike trains ([1] [20]).
2) Inputs are mapped spatially (brain-like) into a 3D SNNcube that consists of spiking neurons spatially organized in a topological 3D map. For modelling brain data the SNNcube is built with the use of a brain template (e.g. [27-29]).
3) Output classifier/regressor SNN is connected to neurons from the SNNcube, e.g. deSNN ([1]).
4) SNNcube structure is organized as small world connectivity 3D structure of spiking neurons.
5) Unsupervised learning is performed in the SNNcube using STDP.
6) Supervised learning is performed in the output SNN module, e.g. deSNN for classification.
7) Adaptive, deep learning of complex spatio-temporal patterns is performed in the SNNcube.
8) The BI-SNN operates in a fast, incremental learning mode.
9) The learned connectivity patterns in the SNNcube can be interpreted as deep knowledge representing deep spatio-temporal patterns in the data.
10) Learned connectivity patterns in the eSNN output module can be interpreted for rule extraction related to outputs.

The brain-inspired NeuCube SNN architecture has been used for brain-data modelling along with other applications [1] . NeuCube has been used for retinotopic mapping and recognition of moving digits [9] and for tonotpic mappig and recognition of musical signals [23] [1] . Recently, NeuCube has been used for the creation of spatio-temporal memories (STAM-SNN) [35], used for neuro-imaging data [36] and for the prediction of time series [37].

III. The proposed cAViAM Framework

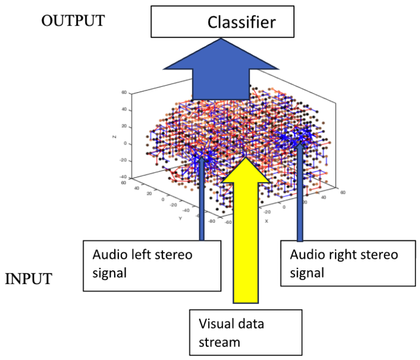


Fig. 2. A schematic diagram of the proposed CAViAM framework

MODEL BASED ON NEUCUBE

A. *A general architecture of CAViAM*

Here we propose a general method for the realization of CAViAM for classification of AV data on the NeuCube brain-inspired SNN. The proposed CAViAM architecture consists of the following modules (Fig.2.) :

1) Input audio and visual modules
2) Brain Inspired SNN
3) An output activation module.

The CAViAM follows how the human brain integrates audio and visual information and is based on the NeuCube architecture, as this architecture enables brain-like information processing. Each spiking unit simulates tightly packed meso-scale neuronal populations, and both short- and long-distance brain connectivity can be simulated.

B. *Mapping integrated visual and audio streaming data into CAViAM-SNN model*

The projection of visual spike data to a SNNcube is inspired by the proportion of projections that are made in the brain by the retina and cochlea to the visual and auditory cortex respectively, both signals traveling via the thalamus. Also, the retinal inputs are filtered by center-surround receptive fields in the retina and the thalamus and also taking into account peripheral vision, in addition to foveal processing, that will complement the sound recordings from the environment. An important characteristic of human vision is the very fast and simultaneous movement of both eyes, called saccades. Saccades help to scan a broader part of the visual field with the fovea and integrate this information into a detailed map. These mechanisms for eye movement are implemented in the spike encoding algorithm by changing the coordinates for the pooling of the visual pixels for each time step, thereby virtually moving the center of the visual field (see Fig.3 (from [9] [1]).

In terms of mapping the encoded into spikes audio information, synchronized with the video one, we use the same principle as shown in Figs. 4 [23, 1]. The brain forms deep neuronal structures when perceives audio information (Fig.5). The advantage of using NeuCube is that its Talairach brain template [29] allows a mapping of the transformed stereo-auditory signals to their corresponding brain regions. The experimental results from

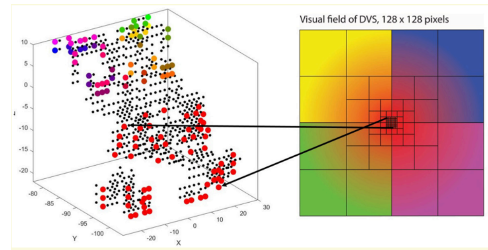


Fig. 3. Mapping visual information in the SNNcube area that corresponds to the visual cortex using the Talairach human brain template (from [9]).

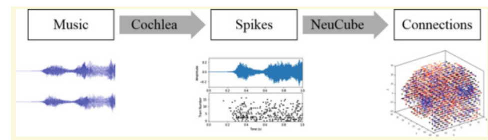


Fig. 4. The flowchart of an exemplified mapping of sound into CAViAM through cochlear simulator (from [23] [1])

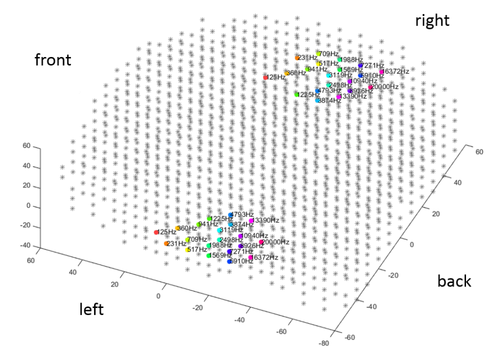


Fig. 5. Tonotopic stereo mapping of sound through cochleograms as part of the SNNcube (from [23] [1]).

previous works confirm that using such a setup with stereo and tonotopic mapping of sound is a promising approach that still needs to be explored further for multi-sensory integration such as the proposed here - Fig.5.

After mapping AV signals, the NeuCube model will be trained to learn the association of these signals in relation to a defined output (e.g. moving object recognition) as explained below.

C. *Learning and classification of integrated AV data in a CAViAM model*

*Our **first hypothesis** is that, similar to how the human brain works, using two AV modalities, if we train an CAViAM classification system* it will result *in a better accuracy when compared with a system developed with the use of a single modality.*

Spatio-temporal patterns from data can be learned in a SNNcube. Connections are created and strengthened. Once data is learned, the SNN retains the connections as long-term memory. Since the SNNcube learns functional pathways of spiking activities represented as structural pathways of connections, when only a small initial part of input data is entered the SNN will 'synfire' and ' chain-fire' learned connection pathways to reproduce learned functional pathways. This is the rationale for using SNN to realize a CAViAM.

This is demonstrated in experimental Sec. IV case studies.

D. Using one of the modalities to recall a trained CAViAM model on both modalities

In [22] polychronous neuronal groups are studied. Since the number of such groups of synchronously activated neurons is very large, this brings a new perspectives of developing CAViAM. In principle, CAViAM is a system that is trained for classification on all available AV data and recalled on new data that contains less number of variables or even missing a hole modality. Learning in the SNNcube is spatio-temporal, i.e. AV data is first spatially mapped into a brain-template structured SNNcube as shown above, and then brain-inspired learning rule is applied to the encoded into spike sequences data, that changes the connectivity of the SNNcube in space and time.

Based on this unique feature of NeuCube, our ***second hypotheses*** *is that once an CAViAM model is trained on full scale AV data for classification, it can generalize on using only one of the two modalities (A or V) subject to proper settings and proper data recordings.*

The rationale behind the hypothesis is the following. The SNNcube is structured, so that it accommodates the structural (spatial) information from the data using a suitable brain template (e.g. Talairach [29], MNI, MRI, etc) and the AV variables are mapped in this structure according to their spatial co-ordinates per the used template (auditory cortex and visual cortex correspondingly). After that the SNNcube is trained with a STDP rule, so that the connection weights capture temporal associations between the structurally distributed neurons, thus reflecting on the spatio/spectro-temporal associations in the AV data. Once a SNNcube is trained on AV data of time length T, the model can be recalled only on one of the modalities and a shorter time length, $T_1 < T$, as the already created connections during learning can be activated even when some input variables are missing in the recall procedure following the principles of *polychronisation* and *synfire* described above. And these connections can be activated even for a shorter time of AV variables, rather than for the full time used in the training of the full model.

This is demonstrated in the experimental section IV on two case studies.

IV. Experimental results with two AV data sets for moving object recognition

Two small scale case studies are presented here to demonstrate the ability of the proposed CAViAM in terms of:

- Integration of audio and visual information for object recognition;
- Using only one of the modalities to recall a system trained on both modalities.

In Sec. IV-A, we present our first case with recorded data of passing aeroplane as a small object, far away from the viewer who is recording the sound in a noisy environment. In Sec. IV-B, our second case is with data of close objects exemplified by a moving train.

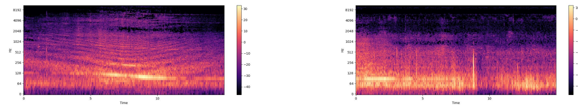


Fig. 6. Short-time-Fourier-Transform (STFT) on each 15 s audio segment containing audio (a) with aeroplane and (b) without aeroplane demonstrates the time-frequency representation of the recorded signals.

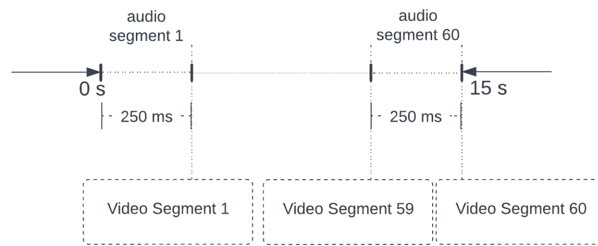


Fig. 7. Synchronised extraction of audio and visual segment from video clips.

A. Case Study I - Recognition of a small moving object, far away from the viewer, and the sound is recorded in a noisy environment

1) Data Collection: The data consists of a 60 seconds video of a far away passing plane (a small object) collected by the authors (DP and NA) at the BITS Pilani, K K Birla Goa Campus, that is close to the Goa Dabolim airport. Thus, the video is of one airplane approaching landing. An iPhone 11 Pro mobile phone camera (12 mega pixels, 1/2.55-inch sensor with 1.4μm pixels, 26mm f/1.8-aperture lens) was used for the recording; no tripod was used i.e. the phone was hand-held that could have added a spatial 'noise' during recording.

2) Data pre-processing: An audio-visual clip of 15 seconds (0:00 to 0:15 s) duration is extracted from the video clip — this will constitute the data that we refer to as 'with plane'. Another clip of the same duration (0:18 to 0:33 s) was extracted when the plane had passed and there was only environmental (background) noise i.e. frames that had the sounds of leaving (the camera frame) aeroplane were discarded; this we refer to as data 'without plane'. The extraction was done using the librosa library in Python that encodes the audio in .mp3 format with the default data sampling rate as 22KHz (22050 data points per second). A time-frequency representation of both 15 s audio segments with plane and without plane are shown in Fig. 6. Next, we extracted independent audio and visual data segments from each of the two above-mentioned 'with plane' and 'without plane' audio-visual clips such that the data are synchronised with one another in time, see Fig. 7. Each audio segment extracted from the 15 s audio-visual clip is of 250 ms duration. Thus, we have 60 audio segments for each of the two types of audio-visual clip. Sixty visual segments

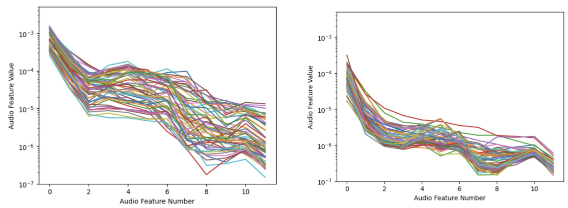


Fig. 8. Twelve features extracted from each of the 250 ms audio segment obtained from the 15 second video clip (a) with airplane and (b) without airplane. The y-axis is logarithmic scale for easy visualisation. Clearly, the overall power is increased during the presence of the plain, in addition to a wider frequency band occupancy.

were also extracted where each segment was a snapshot of the scene at every 250 ms interval, starting from the first 250 ms.

3) Feature Extraction for Audio input: Each of the 60 audio segments had 5513 data points (at the above-mentioned sampling rate of 22 KHz). We can think of the processed audio data as a 60×5513 matrix. Thus we had two audio-data matrices corresponding to 'with plane' (M_p) and 'without plane' (M_{np}). Each matrix was passed through a Fast Fourier Transform (fft) algorithm using Python Scipy library with sampling frequency 22KHz and frequency resolution 0.25 Hz. This operation mapped $M_p^{60\times5513}$ ($M_{np}^{60\times5513}$) in the time domain to $F_p^{60\times5513}$ ($F_{np}^{60\times5513}$) in the frequency domain.

We extracted 12 features from each audio data segment and mapped them to the input nodes in the 3D SNN that correspond to the closest auditory areas from the Talairach Daemon software [27, 28, 29]. This is of course an approximate mapping and it is one of many possible mappings, not necessarily the optimal one in terms of classification results. Towards this, each row r of the matrix F_p (and F_{np}) is mapped to a vector $I_r \in \mathcal{R}^{12} \ni$ each element of I_r,

$$I_{rc} = \frac{1}{N}\sum_{j=1}^{N} F_p^{r,j}$$

where $c \in \{1, 12\}, r \in \{1, 60\}, N = 460$. The resultant matrix of feature vectors will be $A_p^{60\times12}$ ($A_{np}^{60\times12}$), where each row represents a 250 ms audio segment from the respective 15 s clipped video. The 60 plots with 12 features each for each of the two cases are shown in Fig. 8.

4) Feature Extraction for Visual input: The visual 'snapshots' (see Fig. 7) extracted at every 250 ms from the video clips were downsampled to 1080×1080 frames. Thus, we had 60 visual frames of each data type viz. with plane ($D_p^{1080\times1080\times60}$) and without plane ($D_{np}^{1080\times1080\times60}$). For homogeneity, we normalise these tensors such that every pixel value lies between 0 and 1. Our design is to extract 16 features from each of these 60 frames for projecting to the input nodes corresponding to the visual cortex on NeuCube. Towards this, we mapped each adjacent 270×270 blocks of a frame to $\mathcal{R}^1$ by averaging all the pixels. The resulting 4×4 matrix is flattened to a 1×16 vector. These operations result in two feature matrices corresponding to with plane ($V_p^{60\times16}$) and without plane ($V_{np}^{60\times16}$), where each row of V_p (V_{np}) represent the 16 visual features in the scene extracted after a 250 ms window. The 16 featuires are mapped into the 3D SNN nmeurons that correspond to the visual cortex. This mapping is approximate and not necessarily the optimal on for a brain locations or from classification results points of view. It is one of many possible mappings and it is used here for the purpose of AViAM framework illustration.

5) AViAM Model simulation: We refer to each dataset as belonging to either Class 1, where the airplane is present, and Class 2 otherwise. Training was performed with three different datasets viz. audio, visual and audio-visual. All simulation parameters were maintained at the default values set on the NeuCube simulator [1] [14]. To load the data on to NeuCube, we did the following: from the Talairach Daemon atlas, we obtained the details of the "nearest gray matter" region in the brain corresponding to the co-ordinates of each input neuron on NeuCube [27]–[29]. From the total 1485 co-ordinates that were thus mapped, we identified all the 209 and 101 co-ordinates corresponding to the temporal and the occipital lobes respectively. This was based on our experimental design where we assume that: all visual and auditory data are being processed at the primary visual and auditory cortices respectively, the locations of which are at the occipital and temporal lobes respectively of the brain. Furthermore, we considered 12 feature vectors for the audio signals that were mapped onto 12 of the available 209 temporal lobe co-ordinates. This was done by visual inspection such that the selected nodes were spatially spread across the region of the NeuCube that represented the temporal lobe. The details of these 12 selected input neuron co-ordinates on NeuCube and their mapping to regions indicated by the Talairach Daemon brain atlas are shown in Table I. Similar mapping was done for the visual features and the details are shown in Table II. Thus initialised, the NeuCube is now trained and tested with the experimental data and results are shown in Table III.

Three sets of training and testing were done: (a) only audio feature vectors; (b) only visual features; (c) with both audio-visual features. The training with audio-visual data was tested for a recall under three conditions: (i) only audio data, (ii) only video data and (iii) the audio-visual data. All data were manually labelled (Truth values 1 and 2 for Classes 1 and 2 respectively) and the test-train split was 50%. There were 20 trials for every type of train-test, and the accuracy extracted for each trial. esults are discussed in Sec. IV-A6.

6) Results of Case Study I: This is a difficult task, as the object is small and far away and the recordings of are made in a noisy environment. The accuracy obtained demonstrates a clear agreement with our Hypotheses 1 and 2. Class 1 accuracy in Table III correspond to true positives i.e. presence of the small object. We see that training and testing a CAViAM with both audio and visual data results in a higher test accuracy compared to just training with audio or visual data. We note that the visual data accuracy is closer to, but not greater than when both modalities are used. When a CAViAM is trained on both the audio-visual data, the recall accuracy with partial data is significantly higher in the case of audio data than those obtained by training with single audio modality. This feature of the model may be used for early warning of a far away moving object when there is no visibility or for blind people.

TABLE I
COORDINATES OF THE 12 Cognitive AUDIO FEATURES IN THE 3D SNN AND THEIR CORRESPONDING "NEAREST GREY MATTER" REGION FROM THE TALAIRACH BRAIN DAEMON ATLAS [27-29].THIS IS ONLY AN EXEMPLAR TONOTOPIC MAPPING OF SMALL NUMER OF COGNITIVE AUDIO FEATURES.

Feature neuron indx	Record no.	X coor	Y coor	Z coor	Level 1	Level 2	Level 3	Level 4	Level 5
a1	1	-60	-60	0	Left Cerebrum	Temporal Lobe	Middle Temporal Gyrus	Gray Matter	Brodmann area 21
a2	18	-60	-30	10	Left Cerebrum	Temporal Lobe	Superior Temporal Gyrus	White Matter	*
a3	32	-60	-10	10	Left Cerebrum	Temporal Lobe	Transverse Temporal Gyrus	Gray Matter	Brodmann area 42
a4	67	-50	-60	-10	Left Cerebrum	Temporal Lobe	Inferior Temporal Gyrus	White Matter	*
a5	83	-50	-40	-10	Left Cerebrum	Temporal Lobe	Middle Temporal Gyrus	White Matter	*
a6	100	-50	-20	-10	Left Cerebrum	Temporal Lobe	Sub-Gyral	White Matter	*
a7	118	-50	0	-10	Left Cerebrum	Temporal Lobe	Superior Temporal Gyrus	White Matter	*
a8	168	-40	-70	30	Left Cerebrum	Temporal Lobe	Angular Gyrus	White Matter	*
a9	201	-40	-30	-10	Left Cerebrum	Temporal Lobe	Sub-Gyral	White Matter	*
a10	238	-40	10	-30	Left Cerebrum	Temporal Lobe	Superior Temporal Gyrus	Gray Matter	Brodmann area 38
a11	304	-30	-50	0	Left Cerebrum	Temporal Lobe	Sub-Gyral	White Matter	*
a12	425	-20	-60	20	Left Cerebrum	Temporal Lobe	Sub-Gyral	White Matter	*

TABLE II
COORDINATES OF THE 16 Cognitive VISUAL FEATURES IN THE 3D SNN AND THEIR CORRESPONDING "NEAREST GREY MATTER" REGIONS FROM THE TALAIRACH BRAIN DAEMON ATLAS.THIS IS ONLY AN EXEMPLAR RETINOTOPIC MAPPING OF A SMALL NUMBER OF FEATURES.

Feature neuron indx	Record No.	X coor	Y coor	Z coor	Level 1	Level 2	Level 3	Level 4	Level 5
v1	54	-50	-80	0	Left Cerebrum	Occipital Lobe	Middle Occipital Gyrus	Gray Matter	Brodmann area 19
v2	60	-50	-70	0	Left Cerebrum	Occipital Lobe	Inferior Temporal Gyrus	White Matter	*
v3	158	-40	-80	10	Left Cerebrum	Occipital Lobe	Middle Occipital Gyrus	White Matter	*
v4	165	-40	-70	0	Left Cerebrum	Occipital Lobe	Inferior Temporal Gyrus	White Matter	*
v5	183	-40	-50	-10	Left Cerebrum	Occipital Lobe	Sub-Gyral	White Matter	*
v6	277	-30	-80	10	Left Cerebrum	Occipital Lobe	Middle Occipital Gyrus	White Matter	*
v7	285	-30	-70	0	Left Cerebrum	Occipital Lobe	Sub-Gyral	White Matter	*
v8	405	-20	-80	10	Left Cerebrum	Occipital Lobe	Cuneus	Gray Matter	Brodmann area 17
v9	414	-20	-70	10	Left Cerebrum	Occipital Lobe	Cuneus	Gray Matter	Brodmann area 30
v10	422	-20	-60	-10	Left Cerebrum	Occipital Lobe	Fusiform Gyrus	*	*
v11	538	-10	-80	10	Left Cerebrum	Occipital Lobe	Cuneus	Gray Matter	Brodmann area 17
v12	547	-10	-70	10	Left Cerebrum	Occipital Lobe	Cuneus	Gray Matter	Brodmann area 30
v13	670	0	-80	20	Left Cerebrum	Occipital Lobe	Cuneus	*	*
v14	798	10	-80	0	Right Cerebrum	Occipital Lobe	Lingual Gyrus	White Matter	*
v15	809	10	-70	20	Right Cerebrum	Occipital Lobe	Cuneus	*	*
v16	930	20	-80	0	Right Cerebrum	Occipital Lobe	Lingual Gyrus	White Matter	*

TABLE III
MODEL ACCURACY OBTAINED FROM THE CAViAM MODEL WHEN TRAINED AND TESTED WITH AUDIO-VISUAL DATA OF A FAR AWAY AIRPLANE FLYING PAST ON ITS APPROACH TO LANDING. THE SOUND IS MEASURED IN A NOISY ENVIRONMENT.

Training	Testing/Recall	Classifier Accuracy (mean, std)		
		Class 1	Class 2	Overall
audio	audio	45, 31.455	98, 0.	72 15.72
visual	visual	50, 20	80 15.18	65 13.56
audio-visual	visual (no audio)	56, 13.61	86, 19.57	71
	audio (no visual)	90, 10.25	99, 4.47	94
	audio-visual	66, 8.20	94, 9.40	80 5.12

is tested only on visual data due to a noisy environment during the sound recordings.

B. Case Study II - Recognition of a larger and closer to the subject moving object with the sound recorded in a noisy environment

1) Data Acquisition and Processing: The data consists of a 53 seconds video of a train arriving at a platform that was downloaded from YouTube (https://www.youtube.com/watch?v=p7o1mpwyfa0&ab_channel=DutchRailwayExplorer). An audio-visual clip of 15 seconds (26:00 to 41:00 s) duration is extracted from the video clip — this will constitute the data with train and is referred to as Class 1. Another clip of 15 seconds (00:00 to 15:00 s) duration is extracted that did not contain the train and neither approaching sound, and is referred to as Class 2.

2) Feature Extraction and Model Simulation: The segmentation of the 15 second video clips were done for 100 ms for both audio and visual features. Thus there were a total of 150 frames per class of each type. Also, the resolution of the visual frames here was 1080×1920.

The spectrograms of the audio data with and without train is shown in Fig 9 (top row). There is a presence of higher frequency components in the case of 'with train'. The logarithmic power plot for each of the twelve features agree with the spectrogram and are shown in the Fig 9 (bottom row). Training and validation results are shown in Table IV.

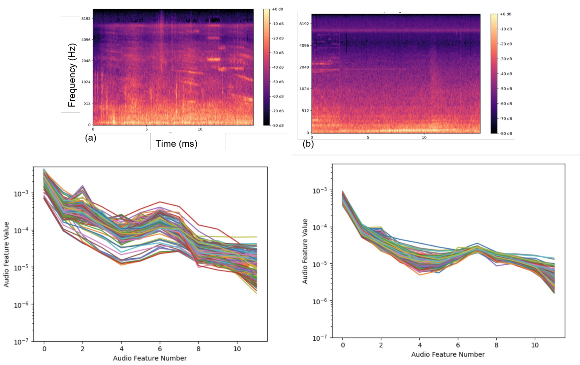


Fig. 9. (Top row) Short-time-Fourier-Transform (STFT) on each 15s audio segment containing audio (a) with train and (b) without train, demonstrate the time-frequency representation of the recorded signals. (Bottom row) Twelve features extracted from each of the 250 ms audio segment obtained from the 15 second video clip (a) with train and (b) without train. The y-axis is logarithmic scale for easy visualisation. Similar to that for case study 1, the overall power is increased during the presence of the train, in addition to a wider frequency band occupancy.

TABLE IV
MODEL ACCURACY OBTAINED by the AViAM-SNN SIMULATOR WHEN TRAINED AND TESTED WITH AUDIO-VISUAL DATA OF AN APPROACHING TRAIN.

Training	Testing	Classifier Accuracy (mean, std)		
		Class 1	Class 2	Overall
audio	audio	85.38, 8.24	99, 0	92 4.12
visual	visual	80.38, 9.49	98.84, 2.81	89.61, 5.59
audio-visual	visual (no audio)	40 13.86	98 0	69
	audio (no visual)	99 0	89 5.73	94
	audio-visual	89 6.35	99 0	94

3) Results of Case Study II: The accuracy obtained for this dataset are presented in Table IV. These results prove our Hypothesis 1 and Hypothesis 2. Interestingly, the highest accuracy was obtained when a CAViAM model was trained on both audio and visual data and afterwards recalled only on audio data.

C. *Feature Interaction in the A-, V-, and the AV Data*

To further understand our observations presented in Tables III and IV, we have generated directed graphs that indicate the level of spike interactions between the input features during the unsupervised training process in the 3D SNN. We can interpret these interactions in a brain-inspired way due to the SNN structure being based on a brain template and the AV input variables being mapped into this structure accordingly.

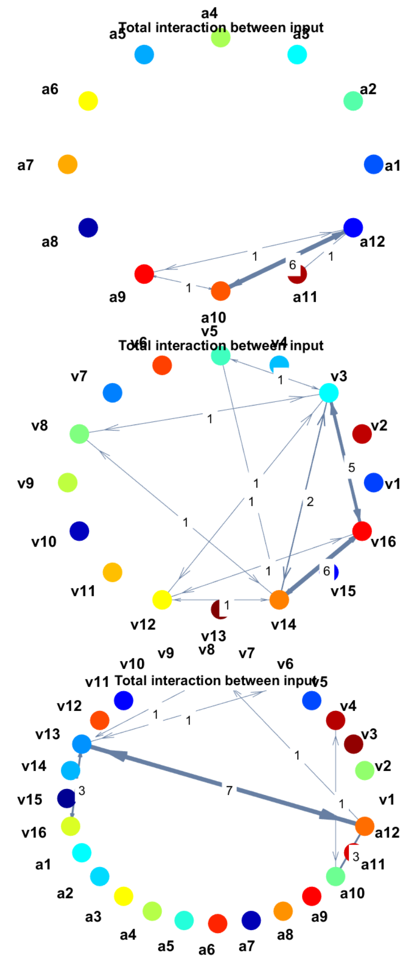


Fig. 10. Directed Feature Interaction Graph recorded on the 3D SNN during training of an CAViAM model with (a) audio, (b) visual and (c) audio-visual modalities for the airplane dataset in our case study I.

The interpretation given below suggests that the proposed AViAM framework can be used for cognitive studies in the future.

Figure 10 shows the directed feature interaction graphs for our case study I. The strongest interaction when trained with just the audio dataset in Fig. 10 (a) is shown to be between features/neurons $a10$ and $a12$, that is identified from Table I as the superior temporal gyrus (gray matter) and the sub-gyral (white matter) in the temporal lobe of the left cerebrum. In Fig. 10 (b), the strongest interaction when trained with just the visual modality are two bidirectional links between $(v14, v16)$ and $(v3, v16)$ with connection strengths 6 and 5 respectively; from Table II, we identify the former to be between the lingual gyrus white matter of the right cerebrum at different co-ordinates, and the latter to be between the middle occipital gyrus and the lingual gyrus white matters of the left and right cerebrums respectively. When trained with both audio-visual modality, the strongest bidirectionally linked pair of nodes is $(a12, v13)$ with connection strength 7 and representing the sub-gyral white matter of the temporal lobe and the cuneus of the occipital lobe, both in the left cerebrum.

Overall for case study I, we note that the audio node $a10$ corresponding to the temporal lobe of the left cerebrum is identified as an important input feature when training with both

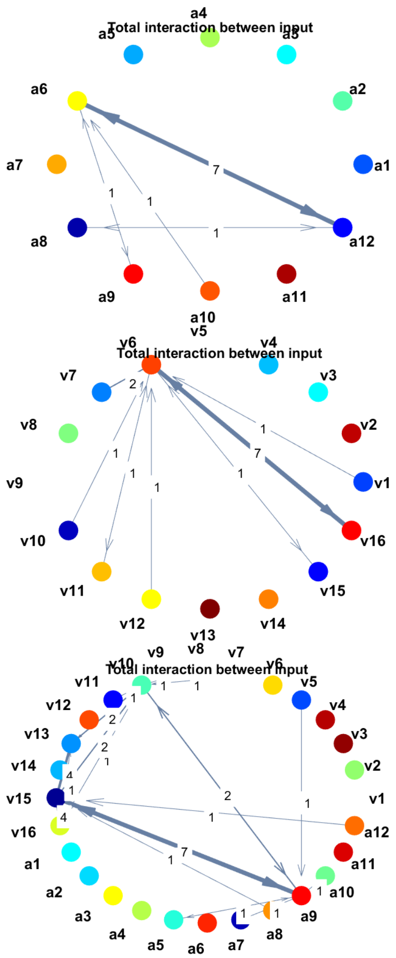


Fig. 11. Directed Feature Interaction Graph recorded from the 3D SNN of a CAViAM simulator during training with (a) audio, (b) visual and (c) audio-visual modalities for the train dataset in our case study II.

audio and audio-visual modalites. For the visual input features, the node $v13$ in the left cerebrum seems important in forming a multimodal audio-visual representation. During independent training with visual modality though, the node $v16$ in the right cerebrum is identified as a strong input feature.

The dynamic feature interaction graph for our case study II on the train dataset is shown in Fig. 11. The nodal pairs with strongest connectivity when trained with audio-visual modality (Fig. 11 (c)) is $(a9, v15)$ that correspond to the sub-gyral white matter of the left cerebrum temporal lobe, and the cuneus of the right cerebrum occipital lobe. Note that this is aligned with our observation for case I, where the nodal pairs were the sub-gyral white matter and the cuneus, albeit in different cerebral hemispheres and co-ordinates.

During training with the audio modality shown in Fig. 11 (a), the $(a6, a12)$ nodes shows the strongest bidirectional connection representing (Table I) sub-gyral white matters of the left cerebrum temporal lobe at different co-ordinates. In Fig. 11 (b), the $(v6, v16)$ pair has the strongest bidirectional connectivity representing (Table IV-A5) middle occipital gyrus and lingual gyrus white matters of the left and right cerebrum.

Overall, for both case studies, the nodes $a12$ and $v16$ corresponding to the sub-gyral and lingual gyral white matters in the left and right cerebrum respectively. The middle occipital gyrus and cuneus in the occipital lobe are also identified as important input feature nodes.

Our work demonstrates the potential of making our learning system explainable by identifying potential connectivity pathways during audio-visual perception when more precise brain data and experimental settings are used. That said, this is only a preliminary experimentation along these lines. Future work will look into more brain-related selection of 16-of-101 and 12-of-209 (or many more) features to identify the combination that is closest to biological explainability. Besides, and as mentioned before, we plan to work on a larger dataset to verify the consistency of our results.

V. Implementing the AViAM-SNN on Neuromorphic Platforms

The proposed CAViAM is suitable for a realization on neuromorphic chips and platforms. One scenario is described here. Streaming visual data will be via a software module or a hardware module, such as Dynamic Vision Sensor (DVS) [8]. The principle is to convert a dynamic and continuous environmental video data to spatio-temporal spike patterns on a 'need-based' manner, i.e. only a movement in the scene is recorded. This is unlike conventional cameras where each frame stores the same details regardless of change, thus introducing redundancy in stored data, and consuming needless resources, time and energy. In terms of sound, the streaming audio data will be encoded into spike trains via a software simulated cochlea method (for example, see https://www.phon.ucl.ac.uk/resource/cochsim/) or a hardware device. In the latter case there are several groups producing such devices , such as INI/ETH/UZH in Zurich and a group in the University of Seville that has an FPGA implementation of a Spiking AUdio Sensor which is open source (http://www.t-cober.es/pdf/).

Along with the mentioned above DVS and Cochleogram, an CAViAM implementation can be based on a neuromorphic hardware, such as SpiNNaker [10]. The hardware interconnection of the DVS and the SpiNNaker is possible in real time for high-speed real-time processing. SpiNNaker is a brain inspired (neuromorphic) hardware to implement flexible spiking neural network architectures in real time 10-12]. The technology allows parallel and event-based computing and is built with very low power ARM processors. Implementation of NeuCube on SpiNNaker has already been made available [11] which can be used for future implementations of CAViAM systems.

Other neuromorphic platforms, can also be explored for efficient implementation [13, 18].

VI. Discussions and Future Work

While on-line encoding of video and audio streaming data is not a novelty[30-33,39], neither is the notion of associative memories [24, 25, 35] or the notion of fuzzy inference [26, 1], the integration of audio and visual streaming information in a brain-inspired SNN, to make a spatio-temporal associative memory, is a novel approach in cognitive and developmental systems and also in neural networks, learning systems and AI. In this respect, the proposed here CAViAM is a world-first framework to be explored further in terms of new methods developed and their wide range of applications. The presented here CAViAM framework is a starting point. It is illustrated here on two small AV data sets. The mapping of the variables in to the SNN model, being brain-inspired, is still on a small scale, mainly to illustrate the potential of this approach. Several directions of study can be followed: new brain-inspired methods for associative learning (e.g.[40]); new methods for mapping AV data into a SNN model; using other modalities in one model, along with audio and visual (e.g. [38]). Potential applications of CAViAM include:

- Autonomous robots and unmanned vehicles;
- Video or/and audio information retrieval from large repositories;
- Human-machine communication systems;
- On-line robotic control systems;
- Cognitive study systems;
- Assistive aids for visually or audio- impaired;
- Security systems.

References

[1] N. Kasabov, *Time-Space, Spiking Neural Networks and Brain-Inspired Artificial Intelligence*. Springer, 2018. [Online]. Available: https://www.springer.com/gp/book/9783662577134

[2] N. Kasabov, E. Postma, and J. van den Herik, "Avis: a connectionist-based framework for integrated auditory and visual information processing," *Information Sciences*, vol. 123, no. 1, pp. 127–148, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025599001140

[3] S. G. Wysoski, L. Benuskova, and N. Kasabov, "Evolving spiking neural networks for audiovisual information processing," *Neural Netw.*, vol. 23, no. 7, p. 819–835, sep 2010. [Online]. Available: https://doi.org/10.1016/j.neunet.2010.04.009

[4] B. Sen-Bhattacharya, T. Serrano-Gotarredona, L. Balassa, A. Bhattacharya, A. B. Stokes, A. Rowley, I. Sugiarto, and S. B. Furber, "A spiking neural network model of the lateral geniculate nucleus on the spinnaker machine," *Frontiers in Neuroscience*, vol. 11, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:34036126

[5] C. Chiplunkar, N. Gautam, I. Mediratta, A. Gait, S. Thomas, A. Rowley, T. Serrano-Gotarredona, and B. Sen-Bhattacharya, "A reduced-scale cortical network with izhikevich's neurons on spinnaker," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[6] R. James, J. Garside, M. Hopkins, L. A. Plana, S. Temple, S. Davidson, and S. Furber, "Parallel distribution of an inner hair cell and auditory nerve model for real-time application," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2017, pp. 1–4.

[7] K. Meier, "Special report : Can we copy the brain? - the brain as computer," *IEEE Spectrum*, vol. 54, no. 6, pp. 28–33, 2017.

[8] T. Serrano-Gotarredona and B. Linares-Barranco, "A 128 $\times$ 128 1.5% contrast sensitivity 0.9% fpn 3 µs latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 827–838, 2013.

[9] L. Paulun, A. Wendt, and N. K. Kasabov, "A retinotopic spiking neural network system for accurate recognition of moving objects using neucube and dynamic vision sensors," *Frontiers in Computational Neuroscience*, vol. 12, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:47020057

[10] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[11] J. Behrenbeck, Z. Tayeb, C. Bhiri, C. Richter, O. Rhodes, N. Kasabov, J. Espinosa-Ramos, S. Furber, G. Cheng, and J. Conradt, "Classification and regression of spatio-temporal signals using neucube and its realization on spinnaker neuromorphic hardware," *Journal of Neural Engineering*, vol. 16, no. 2, 2019, publisher Copyright: © 2019 IOP Publishing Ltd. Copyright: Copyright 2019 Elsevier B.V., All rights reserved.

[12] J. Humble, S. Denham, and T. Wennekers, "Spatio-temporal pattern recognizers using spiking neurons and spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 6, 2012. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2012.00084

[13] E. Neftci, E. Chicca, G. Indiveri, and R. Douglas, "A systematic method for configuring vlsi networks of spiking neurons," *Neural Computation*, vol. 23, no. 10, pp. 2457–2497, 2011.

[14] N. K. Kasabov, "Neucube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data," *Neural Networks*, vol. 52, pp. 62–76, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014000070

[15] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 828–836, 2003.

[16] Q. Yue, X. Wu, and J. Gao, "Audio-visual event localization based on cross-modal interacting guidance," in *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2021, pp. 104–107.

[17] J. Behrenbeck, "Neucube spiking neural network algorithm for eeg signals decoding on spinnaker neuromorphic hardware," 2018. [Online]. Available: https://github.com/behrenbeck/NeuCube_SpiNNaker

[18] [Online]. Available: https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html

[19] M. Abeles, *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge University Press, 1991.

[20] N. Nuntalid, K. Dhoble, and N. Kasabov, "Eeg classification with bsa spike encoding algorithm and evolving probabilistic spiking neural network," in *Neural Information Processing*, B.-L. Lu, L. Zhang, and J. Kwok, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 451–460.

[21] I.-C. Chern, K.-H. Hung, Y.-T. Chen, T. Hussain, M. Gogate, A. Hussain, Y. Tsao, and J.-C. Hou, "Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023, pp. 1–5.

[22] E. M. Izhikevich, "Polychronization: Computation with spikes," *Neural Comput.*, vol. 18, no. 2, p. 245–282, feb 2006. [Online]. Available: https://doi.org/10.1162/089976606775093882

[23] G. Saraceno, "Deep learning and memorizing of spectro-temporal data (music) in the spatio-temporal brain," Master's thesis, University of Trento, 2017.

[24] H. Kwan and C. Lee, "Temporal associative memories using cascade and ring architectures," in *International 1989 Joint Conference on Neural Networks*, 1989, pp. 570 vol.2–.

[25] H. Zhang, B. Zhang, W. Huang, and Q. Tian, "Gabor wavelet associative memory for face recognition," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 275–278, 2005.

[26] Q. Song and N. Kasabov, "Nfi: a neuro-fuzzy inference method for transductive reasoning," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 6, pp. 799–808, 2005.

[27] J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, and P. T. Fox, "Automated talairach atlas labels for functional brain mapping," *Human Brain Mapping*, vol. 10, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:19025072

[28] J. L. Lancaster, L. Rainey, J. Summerlin, C. S. Freitas, P. T. Fox, A. C. Evans, A. W. Toga, and J. C. Mazziotta, "Automated labeling of the human brain: A preliminary report on the development and evaluation of a forward-transform method," *Human Brain Mapping*, vol. 5, pp. 238–242, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:15242390

[29] J. Talairach, P. Tournoux, and M. Rayport, "Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: An approach to cerebral imaging," *The Journal of Laryngology & Otology*, vol. 104, pp. 72 – 73, 1988. [Online]. Available: https://api.semanticscholar.org/CorpusID:222247765

[30] S. Liu, W. Quan, Y. Liu, and D.-M. Yan, "Bi-directional modality fusion network for audio-visual event localization," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4868–4872.

[31] L. Lacheze, Y. Guo, R. Benosman, B. Gas, and C. Couverture, "Audio/video fusion for objects recognition," in *2009 IEEE/RSJ Interna-*

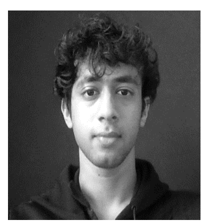
tional Conference on Intelligent Robots and Systems, 2009, pp. 652–657.
[32] R. Su, L. Wang, and X. Liu, “Multimodal learning using 3d audio-visual data for audio-visual speech recognition,” in *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 40–43.
[33] X. Zheng and Y. Wei, “Audio-visual event and sound source localization based on spatial-channel feature fusion,” in *2022 7th International Conference on Signal and Image Processing (ICSIP)*, 2022, pp. 106–110.
[34] N. Kasabov, “Neucube evospike architecture for spatio-temporal modelling and pattern recognition of brain signals,” in *Artificial Neural Networks in Pattern Recognition*, N. Mana, F. Schwenker, and E. Trentin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 225–243. [Online]. Available: https://doi.org/10.1007/978-3-642-33212-8_21
[35] N. K. Kasabov, “Spatio-temporal associative memories in brain-inspired spiking neural networks: Concepts and perspectives,” *TechRxiv*, 2023, preprint. [Online]. Available: https://doi.org/10.36227/techrxiv.23723208.v1
[36] N. Kasabov, H. Bahrami, M. Doborjeh, and A. Wang, “Brain inspired spatio-temporal associative memories for neuroimaging data: EEG and fMRI,” MDPI, Bioengineering, *Preprints*, 2023. [Online]. Available: https: //doi.org/10.20944/preprints202308.0333.v1
[37] I. AbouHassan, N. Kasabov, T. Bankar, R. Garg, and B. S. Bhattacharya, “Pamet-snn: Predictive associative memory for multiple time series based on spiking neural networks with case studies in economics and finance,” *TechRxiv*, 9 2023, preprint. [Online]. Available: https://doi.org/10.36227/techrxiv.24063975.v1
[38] I. AbouHassan, N. K. Kasabov, V. G. Jagtap, and P. Kulkarni, “Spiking neural networks for predictive and explainable modelling of multimodal streaming data on the case study of financial time series data and on-line news,” *Scientific Reports*, vol. 13, no. 1, p. 18367, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-42605-0
[39] C. Chen, Z. Al-Halah, and K. Grauman, “Semantic audio-visual navigation,” 2021.
[40] Y. Imai and T. Taniguchi, “Associative memory by virtual oscillator network based on single spin-torque oscillator,” *Scientific Reports*, vol. 13, p. 15809, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-42951-z

Nikola K. Kasabov (IEEE M’93, SM’98, F’10, LF’22) received his MSc and Ph.D from the Technical University of Sofia, Bulgaria. He is the Founding Director of KEDRI and Professor of Knowledge Engineering at Auckland University of Technology, New Zealand. He holds also professorial positions at the University of Ulster UK, IICT Bulgarian Academy of Sciences and Dalian University, China. He is Director of the knowledgeengineering.ai. His research areas are computational intelligence, neuroinformatics, knowledge discovery, spiking neural networks, with more than 320 journal publications.

Basabdatta Sen Bhattacharya (IEEE M 2009, SM 2015) received her PhD from the School of Computer Science, University of Manchester, UK (2008), Masters in Engineering from Jadavpur University, Kolkata, India (2002), and Bachelors of Engineering from National Institute of Technology (NIT), Silchar, Assam, India (1992). She was born (1969) and raised in Silchar. She has worked as a Researcher with the University of Manchester (UK), University of Bordeaux (France) and the University of Ulster (Northern Ireland, UK). She has worked as an Academic at the University of Lincoln (UK) and as a Lecturer at the NIT, Rourkela, Odisha, India. She has worked as a Manager (Technical Category) with the Steel Authority of India Ltd. Her primary research area is Brain-inspired Neural Networks with applications to Computational Neuroscience and Artificial Intelligence.

Dharmik Patel is a final-year undergraduate student at BITS Pilani, Goa (2024), majoring in Computer Science and Engineering (BE. Honors) with a minor in Data Science. He is also an undergraduate Researcher in the Sivak Group at Simon Fraser University (SFU), Burnaby, Canada. He worked as a summer intern at Sprinklr, India (2023) in the AI team. He is also a member of the Brain Inspired Neural Network (BINN) lab at BITS Pilani, Goa. His research interests are machine learning and its applications in different areas .

Naman Aggarwal is a final year undergraduate student at BITS Pilani, Goa, India. He is majoring in Computer Science and Engineering (BE. Honors). He has previously worked as Software Engineering Intern at Palo Alto Networks and will be joining Phonepe as Software Developer Engineer.



Tanmay T. Bankar is a final year undergraduate student at BITS Pilani, Goa, India. He is majoring in Computer Science and Engineering (BE. Honors) and is also a visiting researcher at the KEDRI Lab at Auckland University of Technology. His research areas are machine learning, spiking neural networks, and multi-modal predictive time-series modeling.

Iman AbouHassan is a Ph.D. candidate at the Technical University of Sofia / Faculty of Computer Systems and Technology, and a Researcher at knowledgeengineering.ai. Her research focuses on the practical applications and advancements of artificial intelligence and neural network technologies in the analysis of economic and financial data. Apart from her academic pursuits, she holds a master’s degree in computer engineering from TU-Sofia. She is the Head of the External Sector Division at Banque du Liban. Her expertise is acknowledged by the International Monetary Fund (IMF) and the World Bank.