Investigating Multisensory Integration in Emotion Recognition through Bio-inspired Computational Models

Esma Mansouri Benssassi and Juan Ye

Abstract—Emotion understanding represents a core aspect of human communication. Our social behaviours are closely linked to expressing our emotions and understanding others' emotional and mental states through social signals. The majority of the existing work proceeds by extracting meaningful features from each modality and applying fusion techniques either at a feature level or decision level. However, these techniques are incapable of translating the constant talk and feedback between different modalities. Such constant talk is particularly important in continuous emotion recognition, where one modality can predict, enhance and complement the other. This paper proposes three multisensory integration models, based on different pathways of multisensory integration in the brain; that is, integration by convergence, early cross-modal enhancement, and integration through neural synchrony. The proposed models are designed and implemented using third-generation neural networks, Spiking Neural Networks (SNN). The models are evaluated using widely adopted, third-party datasets and compared to state-of-the-art multimodal fusion techniques, such as early, late and deep learning fusion. Evaluation results show that the three proposed models have achieved comparable results to the state-of-the-art supervised learning techniques. More importantly, this paper demonstrates plausible ways to translate constant talk between modalities during the training phase, which also brings advantages in generalisation and robustness to noise.

Index Terms—Spiking neural network, mutilsensory integration, emotion recognition, neural synchrony, graph neural network

1 Introduction

HUMANS perceive emotions in a multisensory manner, where information from different sensory modalities such as facial expression, verbal, non-verbal signals, and body languages expresses our emotional states. Multisensory emotional percept is driven through a constant cross-talk between various sensory modalities.

Understanding emotions from multiple modalities is crucial for human-computer interaction (HCI) and affective computing with various applications such as gaming, mental health or car driving. Multisensory social signals of emotion recognition do not only provide more effective and efficient human-computer interaction but also facilitate the enhancement and efficiency of assistive technologies or social robots for individuals facing challenges in interpreting complex and subtle social cues; for example, in the area of autism, schizophrenia and dementia [1], [2]. Therefore, it is crucial to analyse and focus on the multisensory relationship between different modalities to get more accurate interpretation of emotions.

State-of-the-art multisensory fusion approaches offer a wide range of abilities. Recently, with the advances of deep learning techniques, research has turned towards applying deep learning architectures in social signals or emotions and social interaction recognition [3], [4], [5] for both unisensory and multisensory recognition tasks. However, they focus on feature extraction and are often combined with data fusion techniques such as feature concatenation or decision level fusion [6], [7]. No sufficient attention has been paid to translating constant interaction, cross-modal prediction [8], and full integration of multisensory social signals as well as how it occurs in the human brain [9], where prediction, interaction and integration play a significant role in translating multisensory information.

Recently bio-inspired approaches have started to emerge in the artificial intelligence field in general and machine learning in particular. Applying bio-inspired architectures in multisensory integration of social signals of emotions can represent a potential alternative to more classical data fusion techniques. These new methods can answer the key challenges faced by existing systems, such as the generalisation and robustness to noise [10]. They help not only in the fusion of information but in a more practical perceptual understanding of emotions.

This paper aims to explore novel biologically inspired architectures for multisensory integration. These novel methods are directly inspired by neuro-computational models and recent studies in neuroscience for multisensory integration in the brain [11]. To do so, we design and implement three multisensory integration models based on different pathways of multisensory integration in the brain: integration by convergence [12] (named *Convergence*), early cross-modal enhancement [13], [14] (named *Enhancement*), and integration through neural synchrony [15], [16] (named *Synchrony*). These three represent the main theories of multisensory integration in neuronsicence. Our key contributions and novelty are listed as follows.

1) We have designed and implemented three bio-inspired approaches that not only model social signals in visual and audio modalities but also model their interaction

• *Benssassi and Ye are with the School of Computer Science, University of St Andrews, UK, KY16 9SX.*
E-mail: juan.ye@st-andrews.ac.uk

Manuscript received April 15, 2020; revised XX XX, 2020.

and integration to enable more biologically plausible signal integration[1]. These approaches can perform emotion recognition in both unsupervised and semi-supervised manners.
2) We have evaluated these architectures on two real-world, public datasets and demonstrated the effectiveness of these architectures over the state-of-the-art techniques in emotion recognition.
3) We have shown that these approaches exhibit better generalisation capability; that is, they can maintain high recognition accuracy when trained on one dataset and tested on the other completely different dataset.
4) We have run the sensitive-to-noise experiments, where different types of noise have been injected to signals. The evaluation results have shown that the bio-inspired approaches are robust to noise and can still achieve high recognition accuracy.

The rest of the paper is organised in the following. Section 2 reviews the state of the art of multisensory integration models in early fusion, late fusion, and hybrid fusion, and identifies the limitation of the existing work. Section 3 briefly introduces the neuroscience theory on multisensory integration and Section 4 describes the design and implementation of the three main pathways of multisensory integration happening in the brain. Section 5 describes the setup and configuration of our models, base on which Section 6 presents and discusses the experiments. Section 7 summarises the findings and points out the future work.

2 Related Work

Multisensory emotion recognition consists of evaluating emotional states from various modalities such as facial expression, body gesture, verbal and non verbal speech. This section will briefly introduce different types of fusion techniques.

2.1 Early Feature Fusion

Early fusion or feature level fusion is one of the most straightforward methods for fusing features extracted from each modality. It works by concatenating extracted features together into one vector, then feeding them to classifiers for estimation and recognition. This fusion method often results in a high dimensional feature vector, to which dimension reduction techniques such as autoencoder are often applied. Feature level fusion remains the most adopted technique for data fusion in multisensory emotion recognition.

Liu et al. [17] have designed deep learning approaches for multimodal feature extraction in physiological data. They employ a Restricted Boltzmann Machine (RBM) [18] to extract features from EEG and eye movement data. They then extract intermediate features from the hidden layers, which are concatenated and fed to a supervised Support Vector Machine (SVM) classifier.

Lingenfelser et al. [19] have combined features extracted from audiovisual data to the Long Short-Term Memory (LSTM) network for continuous emotion recognition. They use short-timed events through a vector of space. Similarly, Chao et al. [20] also use LSTM for temporal feature extraction on both audio and video. Then they concatenate the feature vectors and feed them to a SVM model for the final emotion recognition. Zhang et al. [21] and Ma et al. [22] have used Convolutional Neural Network (CNN) and 3D-CNN [23] to extract meaningful features from audio and visual modalities. Then features are concatenated using a Deep Belief Network (DBN) and then fed to a linear SVM model for the final emotion classification.

These early fusion techniques are useful when data from different modalities are completely synchronised; that is with no temporal overlap, or delay. This is particularly hard for audio-visual data, as usually visual information is perceived earlier [24]. Another disadvantage of early fusion is that the correlation between features from different modalities is ignored and it is very challenging and difficult to learn any relation among modalities [25].

2.2 Late Decision Fusion

Late fusion, referred to as decision level fusion, is vastly used in multimodal emotion recognition, as it provides some answers some of the early fusion challenges by emphasising the uniqueness and individuality of each modality. In these fusion techniques, emotion classification is performed on individual modality first and then the classification results will be combined to form a final decision.

One of the most used decision level fusion techniques is Kalman filter [26]; that is, video is considered as a time series problem, and scores from individual classifiers are fused. The algorithm is based on Markov model, with the goal to reduce noise by taking several measurements and each step's estimation into account.

Felipe et al. [27] have proposed a real-time multimodal system based on decision level approaches. The primary system consists of two parallel models for facial and speech recognition respectively. The outcome of the two subsystems is then integrated in a Dynamic Bayesian Network.

Schels et al. [28] have created a classifier that fuses decisions inferred from video and physiological EEG data. A classifier for each module is created to learn features. Then a final classifier is built using different weights according to the model performance on each modality. The authors have applied more weights to the audio and physiological data, as they have produced higher accuracy individually. Sun et al. [29] have adopted a weighted product rule for fusing results from audio and visual modalities. SVM is applied for classification in each modality. Fusion is achieved by multiplying the weights in the fusion network by the posterior probabilities obtained on each class from each modality.

Nojavanasghari et al. [30] have compared late and early fusion for the prediction of persuasiveness in multimedia data where data from multiple modalities are used to predict a person's persuasiveness. They have explored two techniques for late fusion, averaging confidence scores obtained from each classifier. They have also experimented deep fusion where they have used these confidence scores as an input for a deep network classifier.

Duan et al. [31] have developed a novel approach based on kernel extreme learning (ELM) [32] for classification of multi-modal physiological and audio visual data. The kernel

1. The implementation will be accessible at: https://github.com/esmam-ai/MultisensoryEmotions.

ELM consists of one hidden layer feed-forward network, where the hidden layer does not need to be tuned and the kernel ELM is applied for each classier. Then a final Kernel ELM is applied on the result from each classifier.

Decision level approaches work by fusing decisions on each modality based on their confidence, while the main challenge is the lack of correlation between modalities. They assume complete independence between modalities' features [33]. This can result in loosing crucial information about the interdependence and interaction between modality such as audio and visual in emotion recognition.

2.3 Hybrid Fusion

Hybrid fusion consists of combining both feature and decision level fusion. A hybrid approach has been designed for multi modal emotion recognition for E-learning environment [34], where a decision level fusion is applied and features are extracted from each modality.

Wolmer et al. [35] have proposed a hybrid technique for sentiment analysis from Youtube videos dataset. Audio and visual features are extracted from video, and a bidirectional long short term memory (BLSTM) [36] is used to fuse data at feature level. Text data is classified with a SVM. Results from BLSTM and SVM are combined based on a decision level fusion approach for estimating sentiments. More recently, Amer et al. [37] have proposed a novel hybrid fusion approach for multimedia data fusion. They first apply a Discriminative Continuous Restrictive Botlzmann Machine (DCRBM) [38] to account for the temporal dimension for each modality. Then a Multimedia DCRBM is applied for fusing multiple DCRBMs combining multiple modalities.

More recently deep learning techniques have also been applied to fusion tasks, not only in feature extraction but also for multimodal or multisensory learning. Zhang et al. [39] use Deep Convolutional Neural Network (DCNN) for multimodal emotion recognition. They design two DCNNs to extract features from visual and auditory modalities. They then integrate the obtained features in a fusion network to obtain a multimodal feature representation. Poria et al. [40] introduce a CNN for sentiment and emotion prediction in visual, audio and text data. They use features extracted from all modalities and input them in a Multiple Kernel Learning [41] classifier. Ghaleb et al. [42] have used deep metric learning and fused visual and audio data with a gating mechanism.

Nguyen et al. [43] have proposed a novel approach using 3D convolutional neural network (C3D) [44] to model spatio-temporal video information, along with Deep Belief Network (DBNs) representing audio and video streams. Bhandar et al. [45] employ a modified stacked autoencoder in addition to a multilayer perceptron-based regression model. Experiments are conducted on the RECOLA dataset [46] by comparing unisensory against multimodal models. Ortega et al. [47] have proposed a novel DNN architecture by integrating three modalities: audio, visual and text. First, the network extracts features on individual modalities from hidden layers. Then extracted features are merged, followed by a fully connected layer and a regression layer.

The above hybrid fusion techniques are presented as specialised architectures for different modalities of interest. We are looking for generic integration models that simulate pathways in the brain when combining signals in decision making.

3 Theory of Bio-inspired Multisensory Integration

Social signals of emotions processing, understanding and perception involve various areas of the brain and a complex network [48]. Human brain proceeds by parsing inputs from different sensory modalities through segmentation, and then works on constructing meaningful representations through integration [49]. There exist four main steps in assessing and integrating social signals:

- Attention: The brain uses attention to select the emotional information for observation.
- Detection: This stage involves sensory modality-specific detection, where information is processed through different brain regions. All essential features from each modality are extracted in early sensory regions such as visual or audio cortices.
- Integration: A new percept is created, comprising multisensory features. Integration is not only achieved by fusing extracted sensory features but through a more elaborate mechanism. At this state, each modality is in constant interaction with others. Integration happens mainly in the Superior Temporal Sulcus (STS) of the brain, which includes a sub-region for each modality. In this region, there exists an overlapping sub-region as well as linking modality-specific regions.
- Evaluation: This final stage involves the evaluation of the affective state in the Inferior Frontal Gyrus region of the brain. Decisions are made on the interpretation of the social signal and emotional states.

Literature identifies three main pathways of multisensory integration happening at various areas in the brain [50]. These pathways start as soon as the brain receives sensory information. It starts by an early cross-modal integration and enhancement between modalities. Then an integration happens in higher order areas such as STS, which contain multisensory neuron groups facilitating integration. Multisensory integration is also driven through neural synchrony, where information is driven by synchronised spikes.

3.1 Integration Through Convergence

The most classical theory for multisensory integration is through convergence in higher order areas such as STS. Multisensory integration through convergence develops hierarchically through a progressive convergence of different sensory signals. Sensory signals get integrated in higher order areas such as Superior Colliculus (SC) [51]. This kind of area includes a higher number of multisensory neurons, which are usually seen as a way to multisensory integration.

3.2 Early Cross-Modal Enhancement

Early cross-modal enhancement describes the interaction between visual and auditory cortices; that is, activity in the auditory cortex is closely affected by visual information. It

represents one possibility of cross-modal prediction and interaction, especially for audio and visual pathways in emotion processing [52]. Visual information usually precedes auditory information, leading to a facilitation of auditory processing by visual information [48].

3.3 Neural Synchrony

Neural synchrony is defined as the simultaneous neural oscillations of different neuron groups in various brain cortical regions connected by synapses. It is considered as the main means of transferring information in the brain and drives the perception of multisensory emotions from auditory and visual stimuli. Audiovisual stimuli without delay provokes oscillatory activity changes during multisensory emotion processing, where the integration of facial and voice information is achieved through the increase in activity within the alpha and theta frequency band within the STS area [53].

3.4 Summary

TABLE 1: Bio-inspired multisensory integration models

Model	Brain Pathway	Characteristics
Integration through *convergence*	Superior Colliculus (SC)	Unisensory modalities converging into one multisensory area
Early cross-modal *enhancement*	Auditory and Visual cortex	Visual modality connected directly to the auditory one
Neural *synchrony*	Auditory and Visual cortex	Constant cross-talk between modalities, decentralised, temporal coherence, stimulus driven

Table 1 summarises the three pathways of multisensory integration. Multisensory information is gathered following specific rules such as temporal alignment, spatial and semantic congruence. Studies have identified various regions where multisensory integration happens, such as the temporal frontal and primary sensory areas [54]. They are not exclusive from each other, but happening at different stages [55]. Multisensory integration through convergence relies on firing rate changes in different cortical regions in a hierarchical and progressive manner. The integration happens in a convergence manner, where the response to multisensory information is compared to the sum of response to each unisensory input. However, multisensory integration does not solely happen in a convergence way [55], but can also occur through a constant cross-modal talk between various unisensory areas including at an early level [56]. Neural synchrony refers to simultaneous neural oscillations of different neuron groups in various brain cortical regions connected by synapses. It is considered as the main means of transferring information in the brain. In the next section, we will illustrate our design and implementation of these three models.

4 Design and Implementation of Bio-inspired Multisensory Integration Models

All three integration models are implemented on top of spiking neural networks (SNN), which are composed of spiking neurons inspired by biological neurons behaviours. In the following, we will first briefly introduce SNN (in Section 4.1), and then the design and implementation of three integration models, namely *Convergence* (in Section 4.2), *Enhancement* (in Section 4.3) and *Synchrony* (in Section 4.4).

4.1 Spiking Neural Network

Spiking neural network represents the third generation of neural networks and is an attempt to model how the brain processes information [57] [58]. Information in the brain is transmitted between neurons using action potentials via synapses. When a membrane potential reaches a certain threshold a spike is generated [59]. The computation of SNNs is based on the timing of spikes rather than their shape, where spikes that fire together get a stronger connection. SNNs have been extensively used for translating neuro-computational processes in the brain and successfully applied to machine vision tasks and lately for speech signals [60]. We have identified SNN as the best candidate to simulate and translate bio-inspired models for multisensory integration [10], [61], [62]. Neurons communicate through a series of spikes, which defines the unique patterns to distinguish different emotional states. To model the interaction between modalities, SNN consists of three main layers:

1) An input layer receives unisensory signals in both visual and auditory modalities;
2) An excitatory layer comprising two excitatory neuron groups translates information from auditory and visual inputs into spike patterns;
3) An inhibitory layer with two neuron groups linked to the excitatory layer for each modality with a lateral inhibition; that is, a neuron in the inhibitory layer is connected to all neurons in the excitatory layer apart from the one it receives signal from.

4.1.1 Interactions and Dynamics of Neurons

Neurons in a SNN communicate through spikes, enabling them to learn specific features at the excitatory layer. Each neuron behaviour is modelled through Leaky-Integrate-and-Fire (LIF) [63], as defined in the following equation:

$$\tau \frac{dV}{dt} = (E_{rest} - V) + g_e(E_e - V) + g_i(E_i - V). \quad (1)$$

V is the membrane voltage and E_{rest} represents the membrane potential in the resting phase. E_i and E_e represent the equilibrium potential for both inhibitory and excitatory synapses. g_e and g_i are the conductance value of synapses at the excitatory and inhibitory layers respectively.

Neurons fire when they reach a certain threshold and then enter a resting phase E_{rest} for an interval of 5ms. At this moment neurons cannot spike as they are in a refractory phase. τ is a time constant representing the time a synapse reaches its potential. This is set at 200ms and 100ms for excitatory and inhibitory neurons respectively. This delay between the excitatory and inhibitory layer is motivated by the learning process happening mainly in the brain.

In order to have a more stable network, *homeostasis* [63] is often applied through an adaptive membrane threshold to refrain some neurons from spiking for all the inputs [64]. At the inhibitory layer, all neurons are inhibited apart from the one they receive information, referred to as *lateral inhibition*.

This is used to encourage competition between neurons. That is, synapses conductance increases when pre-synaptic reaches the synapse before post-synaptic; otherwise, they decrease exponentially. The dynamics is ruled by a time constant as defined in the following equation.

$$\tau_{g_e} \frac{dg_e}{dt} = -g_e \tag{2}$$

where τ_{g_e} is a time constant of post-synaptic potential. The time constant is set to 1ms for the inhibitory conductance and to 2ms for the excitatory conductance.

4.1.2 Unsupervised Learning Through Spike Timing Dependent Plasticity

Learning in SNN is achieved in an unsupervised manner through Spike Timing Dependent Plasticity (STDP) [63]. STDP has been successfully used in facial expression recognition [61] and speech emotion recognition tasks [65]. It is a form of Hebbian learning, where connections between neurons are created and strengthened when they fire at the same time. The main learning is influenced by the time of spiking of pre-synaptic and post-synaptic neurons. Weights are updated by the following equation:

$$\Delta w = \eta(x_{pre} - x_{tar})(w_{max} - w)^{\mu} \tag{3}$$

η is the learning rate. w_{max} is the maximum weight and x_{tar} is the target value of the pre-synaptic trace when the post-synaptic spike fires. This is used to enable the disconnection of neurons that seldom lead to firing, when the post-synaptic neuron is rarely active. μ determines the dependence of updates on previous weight. x_{pre} is the pre-synaptic trace left every time pre-synaptic spike reaches a synapse. That is, weights are increased if pre-synaptic spikes fire prior to post-synaptic spikes. Otherwise, they decrease. The change of weights in STDP learning is computed by a function tracking differences in timing between pre-synaptic and post-synaptic spikes. STDP learning proves to be a simple and advantageous method compared to classical supervised learning such as back-propagation [66].

4.2 The *Convergence* Model

We design the integration through convergence model (named *Convergence*) by simulating the process of passing information from lower sensory areas to higher-order multisensory areas for integration. As depicted in Figure 1 (a), the convergence model consists of three layers:

- **Input** layer, which receives features from each modality. For bimodal integration, the network comprises two distinct neuron groups representing input from each modality. After feature extraction from each modality, spike trains will be generated from the features.
- **Excitatory** layer, where groups of neurons with excitatory ability are created. The layer comprises three main groups for bimodal integration. The first two groups define modalities such as audio and visual. The final group represents a higher-order multisensory region. The whole learning occurs in the excitatory layer. Excitatory neuron groups receive input from the input layer for each modality. The multisensory group receives information from each excitatory modality group. The recurrent connections between the unisensory neurons groups and the multisensory neuron group permit learning of distinctive patterns features for each class label.
- **Inhibitory** layer, which enables the network stability. The inhibitory layer comprises three main neuron groups representing unisensory modalities and a multisensory area. Network stability is achieved through lateral connections where each neuron in the inhibitory layer is connected to all other neurons, apart from the ones that receive input.

The main characteristic of the convergence model is the simulation of higher-order multisensory regions, where unisensory information converges to a multisensory area. Learning of multisensory patterns happens in two main stages. Firstly, unisensory excitatory neuron groups receive information from the input layer. Each group starts learning unisensory patterns where neurons spike for the same class label. Then, multisensory neuron group receives information through connection from both unisensory excitatory groups. Learning in the convergence group happens through STDP where neurons spiking for the same class label get a stronger connection. The connection between these neurons happens regardless of signals' origin. Training in the convergence model happens by presenting inputs from each modality with a delay, which simulates biologically realistic delay between visual and auditory sensory information reaching the brain.

4.3 The *Enhancement* Model

We use SNN to translate the cross-modal enhancement (named *Enhancement*) where spiking patterns in the visual modality affect the auditory part. This translates early multisensory integration in the brain; that is, influencing auditory processing with visual neurons spikes. Figure 1 (b) describes the workflow. The auditory excitatory layer receives input from both the auditory input layer and the visual excitatory layer. Following the same pattern in the brain, visual information precedes by few milliseconds the auditory processing. It is different from the recent cross-model learning [6] where a cross-modal transfer from the visual to auditory data is applied. The *Enhancement* model is more biologically plausible, where the auditory part does not use prediction from the visual part but learns from the spiking patterns. This represents a multisensory learning, which helps propagate spikes from the visual group to the auditory group [67].

4.4 The *Synchrony* Model

Neural synchrony (named *Synchrony*) allows cross-talk between modalities by setting recurrent connections at the excitatory layer between audio and visual neuron groups. This is achieved by connecting neurons that spike together between both modalities with the same temporal window, as shown in Figure 1 (c). This facilitates the integration of information from different sensory sources [50]; that is, learning and extracting relevant and crucial features from sensory inputs such as heterogeneous neuronal populations [68].

Different from the previous two models, neural synchrony focuses on cross-talk between neuron groups. SNN

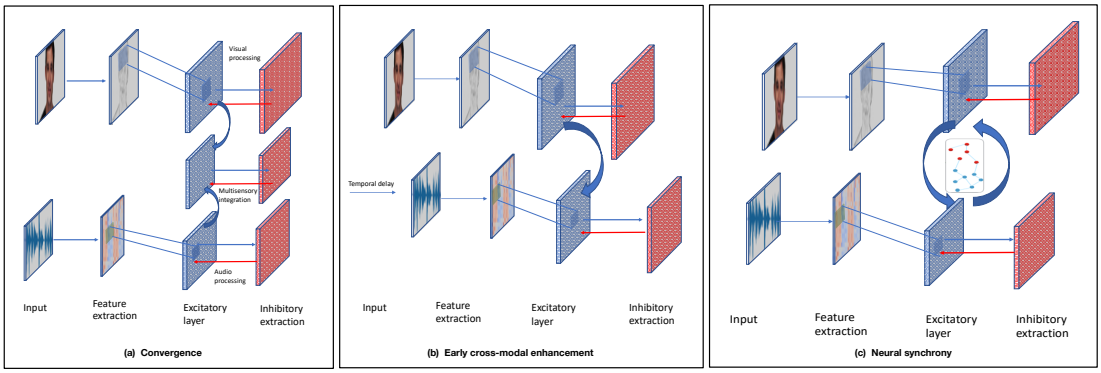

Fig. 1: Computational models and workflow of three integration models: *Convergence*, *Enhancement*, and *Synchrony*. Their workflow all starts with pre-processing both visual and audio inputs, extracting features, and feeding features to spiking neural networks (SNN). Each SNN is composed of an input layer, a feature layer, an excitatory layer and an inhibitory layer.

itself is not sufficient and therefore we propose to learn these complex patterns through a graph convolutional network (GCN). A neural synchrony graph network is defined as an un-directed graph: $G = (V, E)$, where V is a set of nodes representing neurons and E defines edges of relations between nodes. The edges include two types of relations: temporal and stimuli based. Edges are added between nodes which spike within a temporal window of integration.

We define emotion recognition as a subgraph classification problem; that is, assigning a class label to each subgraph. It stacks up multiple convolution layers. We use a deeper architecture compared to the one introduced in [69] by adding a hidden layer. Having a deeper network helps aggregate and translate the complex relationship between nodes to sub-graphs.

At each layer a GCN produces an output in the form of a feature matrix $Z_{N\times D}$, where D represents the dimension of output features for each graph and N is the number of nodes. Each layer can be represented by:

$$H^{(l+1)} = f(H^{(l)}, A), \tag{4}$$

$H^{(l)}$ represents the activation matrix at the lth layer and the activation matrix for the first layer is the feature matrix X. f is the propagation function that aggregates features at the lth layer with the adjacency matrix A, leading to features at the subsequent layer $l + 1$.

Spectral graph convolution is applied to the graphs by applying Eigen-decomposition of the graph Laplacian. The spectral convolutions are defined by the multiplication of graph signal $x \in R^N$ (which is a scalar value for every node) with a filter $g_\theta = diag(\theta)$ where $\theta \in R^N$ is in the Fourier domain [69]. The spectral convolution can be translated by:

$$g_\theta * x = U_{g_\theta} U^T x \tag{5}$$

U represents the matrix of eigenvectors of the normalised graph Laplacian $L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \Lambda U^T$, where Λ is the diagonal matrix of the eigenvalues. g_θ is a function of the eigenvalues of L. $U^T x$ is the graph Fourier transform of the graph signal x.

The input to the network consists of multiple sub-graphs each representing neural activities of a video input. The network consists of three layers followed by a pooling layer over graph [70] in order to combine features from all sub-graphs and enable the classification of subgraph. The main learning model and propagation rule can be defined as follows:

$$Z = f(X, A) = \texttt{softmax}(\hat{A}\sigma(\hat{A}\sigma(\hat{A}W^{(0)})W^{(1)})W^{(2)}), \tag{6}$$

where weights are defined by weight matrices with $W^{(0)}$ representing the input to hidden layer weight matrix, $W^{(1)}$ is the weight matrix from hidden layer 1 to hidden layer 2 and $W^{(2)}$ is the hidden to output weight matrix. $\hat{A} = A + I_N$ is the adjacency matrix of the graph with added self connection and I_N is the identity matrix. The loss function is defined as the cross-entropy over labelled neurons:

$$\mathcal{L} = -\sum_{d \in y_D}^{.} \sum_{c=1}^{C} Y_{d,c} ln Z_{d,c} \tag{7}$$

y_D is a set of neurons that are labelled and C represents the dimension of the output classes; *i.e.*, six basic emotions. The networks weights $W^{(0)}$, $W^{(1)}$, and $W^{(2)}$ are trained with gradient descent, where the full training set is used in each iteration [69].

5 Experiment and Evaluation Methodology

This section introduces the datasets and the model configuration details for these datasets.

5.1 Datasets

We use two open-source datasets to evaluate our model. The first dataset is the eNTERFACE'05 dataset [71] with 42 participants composed of 81% male and 19% female participants. The video resolution is 720 × 576 and the audio is recorded at 48000HZ in 16 bit format. There are 1166 videos recorded and there are 6 emotional classes: 'Angry' (17%), 'Disgusted' (16%), 'Fear' (16%), 'Happy' (18%), 'Sad' (17%), and 'Surprised' (16%). The second dataset is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [72]. The dataset consists of a balanced gender with 24 participants. The participants are actors reading a sentence in 6 emotional states, which are the same as eNTERFACE'05. Each emotion has the same number of recordings from each participant, leading to 4320 (= 24 × 6 × 30) videos in total. The video resolution is 1920×1080 and the audio is recorded at 480000HZ in 16 bit format.

5.2 Input Configuration

First of all, we use SNN to extract neuron features for the integration models. Each integration model will take the learnt spiking neurons as input.

For each audio sequence, we adopt 128 mel-filter bands up to 8000 Hz to extract mel-scale spectrogram using Fast Fourier Transform (FFT) [73] with the FFT window length of 128. MFCCs are then extracted from the mel-scale spectrogram by applying logs of power which are calculated for each mel frequency. Then Discrete Cosine Transform is applied on the mel log powers. The log mel spectrum is then converted back to temporal signal. The Cseptral representation of the speech enables the identification of local spectral properties of the audio signal for each temporal frame. The number of energies of filter banks is set at 40. All audio features are unified to have a temporal length of 388. Poisson distribution is used to encode MFCC into spike train. In the end, the audio neuron input to SNN has a size of 40 $\times$ 388. The mel-scale features are computed using using Librosa python library [74].

For the visual input we extract frames at each segment and convert them to a grey scale. We then identify and crop the face area of each frame, and resize them to 100 $\times$ 100. We apply Laplacian of Gaussian (LoG) to extract contours and edges of facial expression. LoG is selected for use as it achieves higher precision [75] and is represented in Equation 8.

$$\nabla^2 G_\sigma(x,y) = \frac{\partial G_\sigma(x,y)}{\partial x^2} + \frac{\partial G_\sigma(x,y)}{\partial y^2} \tag{8}$$

where ∇^2 is the Laplacian operator, σ is the smoothing value, and $G_\sigma(x,y)$ is the Gaussian filter applied to the image, given by:

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{9}$$

Each filtered image is then encoded into a Poisson spike train where the firing rate is proportional to the intensity of each pixel. The Poisson spike train with a size of 100 $\times$ 100 will be input for a SNN for learning, which is described in Figure 2.

5.3 *Convergence* Model Setup

The input layer of the network architecture consists of two groups of neurons each representing a modality. The number of neurons for each input neuron group is proportional to the size of the input; that is, the size of the audio features and video frame features. We use 40$\times$388 and 100$\times$100 input neurons for the auditory and visual input respectively.

Each input is divided into convolution features where a stride window moves through the input. The convolution window in the audio modality moves along the temporal axis. Convolutional windows are applied separately to each modality. That is, visual and audio modalities have different configurations in terms of convolutional window and the number of features and the total excitatory neurons. We have experimented with various configurations on the validation data and have chosen the best performing ones which set the window size and the stride size for the auditory and the visual features as 10. The number of features is set to 60 for the auditory modality and 60 for the visual modality. For the SNN network of each modality, the number of neurons for both excitatory and inhibitory layer is set as 4000.

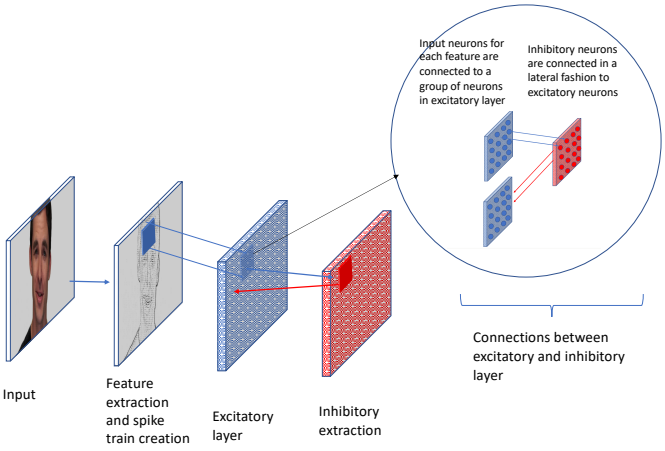

Fig. 2: SNN workflow for FER: LoG filters are applied to raw input, then the input is processed to create Poisson spikes train, as an input to the excitatory layer of SNN. The neurons are connected in a lateral fashion between the excitatory and inhibitory layers.

The excitatory layer comprises three distinct neuron groups. The first two groups correspond to each modality. The third is a multisensory group where integration happens. The inhibitory layer contains three distinct neuron groups. Two neuron groups are connected laterally to each excitatory group for each modality. A third set of connections is set between excitatory neurons in each modality to the multisensory excitatory group. There is no direct link between neurons from unisensory modalities. The main learning happens in the multisensory convergence area, receiving inputs from both modalities. We adopt the same parameter settings for SNN as [63], including the input firing rates, membrane threshold, and the resting phase duration.

5.4 *Enhancement* Model Setup

The input layer is similarly set as the *Convergence* model and then connected to a convolution excitatory layer which is connected to an inhibitory layer with a lateral inhibition, where neurons are connected to all neurons in the excitatory layer apart from the one receiving information from. After processing the visual frames, the audio input is fed to the network. Both visual and audio layers are connected through their excitatory layers through a recurrent connection. Speech features, visual features and cross-modal connections are learned using STDP unsupervised learning.

5.5 *Synchrony* Model Setup

The input layer is also similarly set as the *Convergence* model. The only difference is that the convolution parameters are set differently; that is, for each modality, the window size is 40 and the number of features is 20. Although setting feature number to a higher value and smaller convolutional window would increase the accuracy,

we have chosen the above setting due to computational power limitations.

The audio input is fed to the network after a 5ms delay. This is to model the natural temporal lag between visual and auditory sensory inputs in the brain [76]. Recurrent connections between modalities are applied at the excitatory layer. This enables the cross-talk between audio and visual modalities and help simulate multisensory interaction where modalities influence each other during the learning process. For example in Figure 3, we can see that excitatory neurons for visual and auditory modalities that spike together within the same temporal window are connected and thus their recurrent connection will be strengthened.

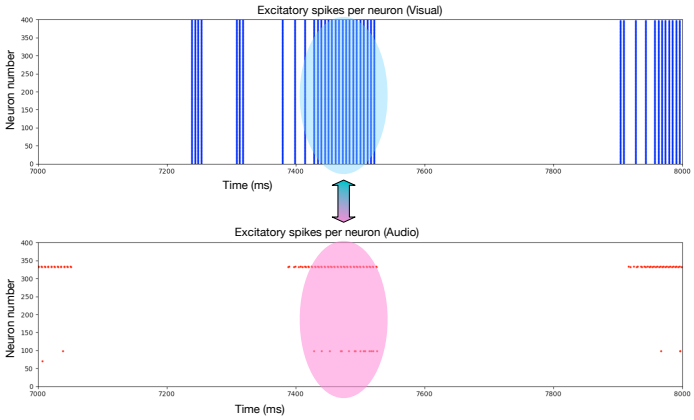

Fig. 3: Example of neuron output response and spike plot, showing cross-talk interaction between excitatory neurons of visual and auditory modalities

Taking the excitatory neurons with their location, time of spiking and modality, we construct a neural synchrony graph. The constructed graph on RAVDESS dataset consists of 814 sub-graphs and 130008 nodes in total. On the eNTERFACE'0 5 dataset we have obtained 1260 sub-graphs and 201600 nodes in total. After obtaining the basic structure for each graph we prepare the input for the GCN. We have trained a three-layer GCN with semi-supervised learning and have initialised the weights randomly [69]. We use Adam optimisation and a learning rate of 0.0001. These hyper-parameters are chosen after experimenting with various learning rate starting from 0.01. We use hidden layers of 64 units in the second and third layer. We train the network for 500 epochs with a dropout rate of 0.5.

There exists various evaluation methodologies such as stratified train-test data split [77], n-fold cross validation [78], and leave-one-user-out [39]. In this work, we opt for the first approach; that is, randomly shuffling and splitting the data into 60% for training, 20% for validation and 20% for testing. We run the process 10 times and report the averaged accuracy.

6 Results and Discussion

In this section, we seek to answer the following three questions:

- Which integration model is most effective for multisensory integration?
- Does the bio-inspired multisensory integration exhibit better *generalisation* capability compared to the state-of-the-art machine learning and/or deep learning techniques; that is, training on one dataset and test on another dataset?
- Does the bio-inspired multisensory integration present better *robustness* to noise; that is, the accuracy of recognising emotions will not be compromised when noise is introduced in each signal modality?

6.1 Effectiveness in Multisensory Emotion Recognition

Our first experiment is to assess the effectiveness of our three integration models in multisensory emotion recognition. First a baseline CNN is designed for visual and audio signal respectively. Each network is designed with three convolution layers, followed by a max pooling layer. Then we develop both *early* and *late* fusion models. In the early fusion model, features from both modalities are concatenated and represented as an input to a three-layer multi-perceptron classifier using a simple cross-entropy loss. In the late fusion model, the decisions (i.e, the confidence outputs from each modality) are used as input to a fully connected layer for prediction in a stacked manner [79].

TABLE 2: Comparison of accuracy on emotion recognition on the RAVDESS dataset.

Model	Feature Extraction	Fusion	Accuracy (%)
Synchrony	LoG, MFCC, SNN	Synchrony with GCN	**98.3**
Enhancement	LoG, MFCC, SNN	Cross-modal enhancement with SNN	73.3
Convergence	LoG, MFCC, SNN	Convergence with SNN	81.3
CNN (Early)	CNN	CNN	81.0
CNN (Late)	CNN	Majority voting at decision level	81.0

TABLE 3: Comparison of accuracy on emotion recognition on the eNTERFACE'05 dataset.

Model	Feature Extraction	Fusion	Accuracy (%)
Synchrony	LoG, MFCC, SNN	Synchrony with GCN	**96.8**
Enhancement	LoG, MFCC, SNN	Cross-modal enhancement with SNN	83.3
Convergence	LoG, MFCC, SNN	Convergence with SNN	83.3
CNN (Early)	CNN	CNN	79.0
CNN (Late)	CNN	Majority voting at decision level	83.0

Table 2 and 3 compare the accuracy on emotion recognition between *convergence*, *enhancement*, and *synchrony* models with the above two CNN baselines. On the RAVDESS dataset, the *Synchrony*, *Enhancement*, and *Convergence* models have achieved an overall accuracy of 98.3%, 73.3%, and 81.3% respectively. GCN has demonstrated as a powerful tool of learning synchrony patterns of neuron activities, which has resulted in the highest accuracy. Even as an unsupervised learning technique, the convergence and enhancement has achieved higher and comparable accuracy to the other supervised learning techniques. For example, Ghaleb et al. [42] employed the 10-fold cross validation and reported an accuracy of 67.7% with a sequence-based classification.

On the eNTERFACE'05 dataset, the *Synchrony*, *Enhancement*, and *Convergence* models have achieved an accuracy of 96.8%, 86.3% and 80.1%. Zhang et al. [21] employed the leave-one-subject-out evaluation and achieved an accuracy of 85.9% on a frame-based classification. Their approach enables feature learning in a multisensory way and quickly

translates the non-linear relationship between both modalities. However, the input data is segmented into several temporal intervals, which could lead to missing information. Noroozi et al. [80] reported 99.9% on the eNTERFACE'05 dataset, however, the accuracy is only measured on a small number of selected representative frames and thus their result is not comparable with ours and other techniques presented. They use supervised learning approach to perform late fusion on the decisions and confidence scores obtained from the visual and audio modalities.

In summary, we consider the three proposed integration models perform better and comparable to the state-of-the-art techniques, as they capture the dynamics and interactions between different modalities. The *Synchrony* model works best as it is supervised learning and also more importantly, it captures the most complex interactions between signals. Each modality influences the other during the learning process using connections between them. SNN enables capturing of multisensory learning through connections between audio and visual neuron groups. The *Convergence* and *Enhancement* models produce similar accuracy, and both are unsupervised learning and do not focus on learning interactions. With the help of GCN, *Synchrony* is able to model and learn synchrony patterns of neuron groups and enable multisensory emotion recognition across them. However, different from *Convergence*, *Enhancement*, and the other state-of-the-art models [39] that employ end-to-end training, *Synchrony* decouples the feature extraction via SNN from the classification via a GCN. This is necessary in our current design in that it would be difficult to build and converge the graph when both visual and auditory features are keeping updating. However, end-to-end training might help improve the accuracy further and we will look into ways to do so.

6.2 Cross-Dataset Generalisation Experiments

To assess the *generalisation* capability, we run cross-dataset experiments; that is, we train on one dataset and test on the other. Table 4 compares the accuracy between 3 integration models and a CNN baseline on 4 settings: train on RAVDESS and test on RAVDESS and eNTERFACE'05 respectively; and train on eNTERFACE'05 and test on eNTERFACE'05 and RAVDESS respectively. When the train and test datasets are the same, we adopt the random stratified train/test split evaluation methodology and use the results reported in Table 2 and 3.

TABLE 4: Generalisation results (accuracy in %) on cross-dataset experiments

Train	RAVDESS		eNTERFACE'05	
Test	RAVDESS	eNTERFACE'05	eNTERFACE'05	RAVDESS
Synchrony	**98.3**	**90.0**	**96.8**	**77.8**
Enhancement	83.3	43.2	86.3	65.7
Convergency	81.3	44.9	80.1	77.4
CNN (Early)	81.0	19.0	79.0	20.0
CNN (Late)	81.7	18.6	83.0	18.0

Neural synchrony exhibits a superior generalisation capability than the other models. When training on RAVDESS and testing on eNTERFACE'05, the *Synchrony* model achieves the highest accuracy of 90%, in comparison to the *Enhancement* model (43.2%), the *Convergence* model (44.9%), and the CNN baselines (19% and 18.6%). When training on eNTERFACE'05 and testing on RAVDESS, the *Synchrony* model also achieves the highest accuracy of 77.8%, while the *Enhancement* model (65.7%), the *Convergence* model (77.4%), and the CNN baselines (20.0% and 18.0%). The big difference between the *Synchrony* model and the baseline models shows that having constant cross-talk drives better performance compared to the CNN baselines with feature concatenation and decision integration. The baseline models are unable to generalise learnt features to a completely different dataset. The *Synchrony* model performs very well in generalisation tasks compared to the other presented models. Exploiting constant cross-talk, temporal synchrony and semantic similarity enables better feature learning.

Figure 4 presents the confusion matrices on the *Synchrony*, *Enhancement*, *Convergence*, and CNN models when trained on RAVDESS and tested on eNTERFACE'05. The performance of the CNN baselines drop significantly. For example, the highest accuracy achieved by the early fusion model is only 33.6% on the class 'surprised', and most classes are misclassified as 'surprised'. A potential reason could be that this class exhibits more distinctive facial features than the other emotional states. The accuracy on the *Enhancement* model degrades as well as they tend to bias towards the 'angry' class. Both *Convergence* and *Synchrony* models behave more stably. In particular, the *Synchrony* model can still maintain 100% accuracy on 3 classes and less diverse false positive rates on the other classes.

In summary, the bio-inspired integration models exhibit good generalisation capability through these cross-dataset evaluation. Here each dataset presents different subjects, ethnic groups, facial dimensions and characteristics; for example, different shapes and sizes of key facial regions like eyes or mouth or even data acquisition conditions [81]. The proposed models can learn inherent features and in particular the *Synchrony* model on characterising cross-talk between modalities enhances the generalisation capabilities to learn more robust features. The performance of cross-dataset experiments might be further improved via transfer learning; that is, fine-tuning the model previously trained on on dataset with a small number of data in the other dataset.

6.3 Robustness Experiments

Here we have experimented the sensitivity and robustness of the proposed three models compared to the state-of-the-art models to add noise for both audio and visual data.

Various types of noise have been used in the literature to assess the sensitivity of models for image recognition tasks, including colours changing, salt and pepper noise and Gaussian noise [82]. Noise degradation is also used to assess the sensitivity of different CNN models (ALexNet, VGG, and Googlenet) [82]. We have experimented with different intensity parameters of salt and pepper noise degradation ranging from 0 to 0.5. Salt and pepper noise represent intensity and sparse disturbances to an image where original pixels are randomly replaced with black and white pixels. After 0.5 noise intensity, we have noticed that the image is completely covered, thus not useful to get more insight of

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100	0	0	0	0	0
Disgusted	0	100	0	0	0	0
Fearful	0	0	42.9	0	0	57.1
Happy	0	0	0	60	0	40
Sad	0	0	0	0	77.8	22.2
Surprised	0	0	0	0	0	100

(a) Synchrony

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	83.7	0	4.7	4.7	4.7	2.3
Disgusted	42.9	26.2	4.8	9.5	11.9	4.8
Fearful	33.3	0	40.5	4.8	21.4	0
Happy	22	0	2.4	58.5	14.6	2.4
Sad	26.2	0	2.4	7.1	64.3	0
Surprised	46.3	0	4.9	17.1	9.8	22

(b) Enhancement

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	69.8	0	4.7	7	11.6	7
Disgusted	9.5	42.9	19	7.1	16.7	4.8
Fearful	2.4	0	59.5	11.9	23.8	2.4
Happy	4.9	0	2.4	68.3	22	2.4
Sad	7.1	2.4	2.4	0	85.7	2.4
Surprised	2.4	2.4	7.3	22	14.6	51.2

(c) Convergence

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	21.9	4.2	9.3	7.9	21.4	35.3
Disgusted	14.1	15.3	10.9	21	12.1	26.6
Fearful	8.6	10.1	28.1	13.5	8.6	31.1
Happy	11.4	9	25.6	23.7	8.5	21.8
Sad	10.2	10.6	16.1	25.2	14.2	23.7
Surprised	12.1	7.3	12.8	17.6	16.6	33.6

(d) CNN (Early)

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	10.2	17.9	17.7	18.5	17.7	18
Disgusted	10.2	22.8	18.3	17.1	15	16.6
Fearful	8.8	19	30.3	15.1	11.3	15.5
Happy	8.4	9.1	26.7	29	11.9	14.9
Sad	7.9	13.1	16.1	23.1	20.4	19.4
Surprised	4	12.3	10.9	15.6	19	38.2

(e) CNN (Late)

Fig. 4: Comparison of confusion matrices on the *Synchrony*, *Enhancement*, *Convergence*, and CNN models when trained on RAVDESS and tested on eNTERFACE'05

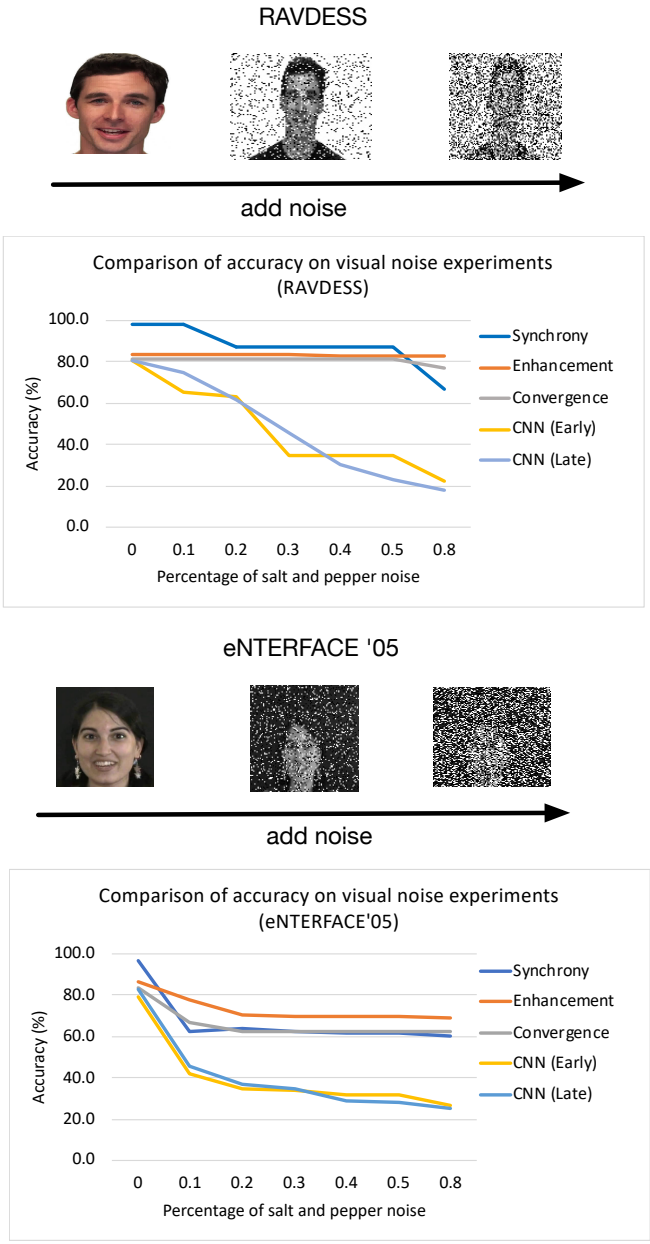

Fig. 5: Comparison of accuracy on emotion recognition in video noise experiments

the performance. However, we did not see much change on the recognition accuracy from 0.3 to 0.5 therefore a higher level of noise is chosen to demonstrate and evaluate the robustness to visual noise.

Figure 5 compares the accuracy of emotion recognition in video noise experiments. The CNN baselines obtain the lowest accuracy for all degrees of noise with the lowest accuracy for the noise probability of 0.8 with only 22% and 18% on RAVDESS and 27% and 25% on eNTERFACE'05.

The *Convergence* model has a drop in accuracy compared to the *Enhancement* and *Synchrony* models. Although the drop in accuracy is not as big as in the baseline models, it still performs less than the other two models. *Enhancement* is less affected by visual noise than the other models, as its accuracy remained stable with the largest drop only being 0.6% on RAVDESS and 10% on eNTERFACE'05. This is because of the architecture type of the model and type of connections between the auditory and visual neurons group. Connection from visual to auditory are set at an early level. The noise applied on visual modality alone did not affect the overall accuracy and the network, as the classification decision relied mainly on the auditory part.

For the audio data, we experiment three levels of audio noise with different power spectrum noises such as white, pink and brown noise. The white noise is characterised by a flat frequency spectrum, where the noise has an equal power spectrum. Thus the white noise represents a flat power. The pink noise has equal power in bands that are proportionally wide and the brown noise has higher energy at lower frequencies. These three levels of noise are used in speech recognition tasks to test the effect of noise in real-word error rate [83]. Figure 6 compares the accuracy of emotion recognition in audio noise experiments. All model accuracy decreases when applying white, pink and brown noise to the test data. The pink noise triggers more degradation than the other two types of noise, where the CNN baselines experience the worst drop (63%) and the integration models follow the same pattern. This is inline with the other findings [84].

Concerning individual labels, confusion matrices in Figure 7 show a sample of individual class accuracy when Brown noise is applied. Again the CNN baselines are significantly affected by the noise; *i.e.*, the best accuracy on the early fusion model that can be achieved is 48% and on the 'surprised' class. In comparison, the three proposed models can still maintain high accuracy, especially the *Convergence* and *Enhancement* models that achieve the averaged accuracy 78.1% and 76.6% respectively. The *Synchrony* model is biased towards 'surprised' class and fails to recognise 'sad'.

In addition, we experiment on a mixed noise experiment.

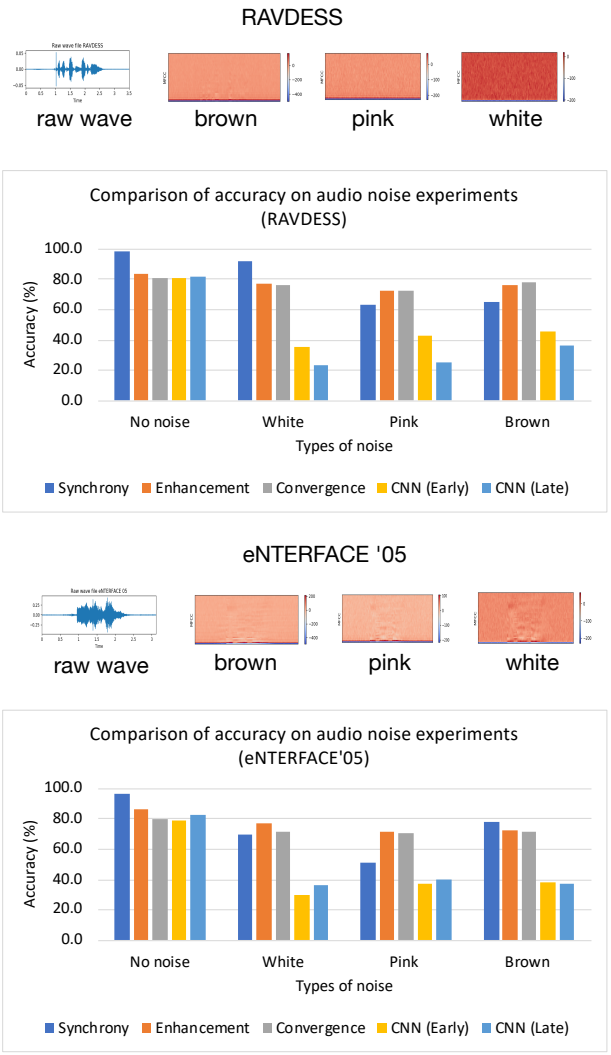

Fig. 6: Comparison of accuracy on emotion recognition in audio noise experiments

Figure 8 presents the accuracy when 0.2 salt and pepper noise is added on image frames and white noise is added on audio. The result is consistent with the previous results that our models outperform the CNN baseline models and the accuracy of *Synchrony* can degrade more than the other two integration models.

In summary, multisensory integration through neural synchrony does not produce the best accuracy for the three types of noise applications with a shallow drop in overall accuracy compared to the other models. This is mainly due to the nature of its implementation and being based in graph and adjacency matrix. Testing on noisy data is computationally costly as a new graph architecture is created for each new type of noise dataset. Creating a new graph when adding new subgraphs or nodes is due to the limitation of graph network with spectral learning, where it is needed to reload the whole graph when adding new subgraphs.

7 Conclusion and Future Work

This paper describes novel bio-inspired architectures and methods for multisensory integration with applications in audio-visual social signals of emotions. The evaluation results show that by adopting bio-inspired models with unsupervised and semi-supervised learning, we can achieve more accurate multisensory integration. Translating the interactions between modalities facilitates the interpretation of multisensory integration, thus producing better accuracy, generalisation and robustness to noise.

The proposed models focus on emotional state classification in an unsupervised learning manner. It would be beneficial to extend them to predict numerical emotional states; i.e., a score in the circumplex model [85]. We will explore ways to design the models for the regression problem. In the future, we will also look into how to combine these three models in a similar way that the brain does. This operation is particularly useful for integrating various sensory modalities as opposed to bimodal integration. Also the model can be extended to include other modalities such as body gesture or verbal speech information. Then more complex permutations of interaction will need to be explored. The high complexity has been a concern for using SNN in real-time applications [10], [86], and we will explore solutions to improve the computational efficiency.

References

[1] C. Chevallier, G. Kohls, V. Troiani, E. Brodkin, and R. Schultz, "The social motivation theory of autism," *Trends in Cognitive Sciences*, pp. 231–239, 2012.

[2] E. M. Benssassi, J.-C. Gomez, L. E. Boyd, G. R. Hayes, and J. Ye, "Wearable assistive technologies for autism: opportunities and challenges," *IEEE Pervasive Computing*, vol. 17, no. 2, pp. 11–21, 2018.

[3] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," in *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, 2015, pp. 179.1–179.12.

[4] M. R. Amer, B. Siddiquie, A. Tamrakar, D. A. Salter, B. Lande, D. Mehri, and A. Divakaran, "Human social interaction modeling using temporal deep networks," *CoRR*, vol. abs/1505.02137, 2015.

[5] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1971–1980.

[6] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 292–301. [Online]. Available: https://doi.org/10.1145/3240508.3240578

[7] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[8] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *WACV 2016*, March 2016, pp. 1–9.

[9] P. Garrido-Vásquez, M. D. Pell, S. Paulmann, and S. A. Kotz, "Dynamic facial expressions prime the processing of emotional prosody," *Frontiers in human neuroscience*, vol. 12, p. 244, 2018.

[10] E. Mansouri-Benssassi and J. Ye, "Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural networks," *Soft Computing*, vol. 25, pp. 1717–1730, 2021.

[11] C. Cappe, E. M. Rouiller, and P. Barone, "Cortical and thalamic pathways for multisensory and sensorimotor interplay," in *The neural bases of multisensory processes*, 2012.

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	86.4	0	0	0	0	13.6
Disgusted	0	74.1	3.7	0	0	22.2
Fearful	0	0	85	0	0	15
Happy	0	0	0	55.6	0	44.4
Sad	0	0	0	0	0	100
Surprised	0	0	0	0	0	100

(a) Synchrony

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	81	0	0	0	4.7	14.3
Disgusted	0	48.5	0	0	3	48.5
Fearful	0	0	81.2	0	0	18.8
Happy	0	0	0	76.5	5.9	17.6
Sad	0	0	0	0	83.3	16.7
Surprised	0	0	0	0	3.1	96.9

(b) Enhancement

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	85.7	0	0	0	9.5	4.8
Disgusted	0	81.8	0	0	12.1	6.1
Fearful	0	6.2	81.2	0	0	12.6
Happy	0	0	0	88.2	11.8	0
Sad	0	11.1	0	0	83.3	5.6
Surprised	0	3.1	3.2	0	12.5	81.2

(c) Convergence

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	34	11.2	12.6	8.7	14	19.5
Disgusted	4.4	39.1	14.5	13.8	12.5	15.7
Fearful	4.9	13.1	40.8	14.2	8.3	18.7
Happy	4.7	7.2	22.3	39.8	12.3	13.7
Sad	8.8	10.9	9.9	21.8	34.7	13.9
Surprised	6.4	6.7	11.3	11.3	16.2	48.1

(d) CNN (Early)

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	33.3	8.3	14.3	12.7	14.3	17.1
Disgusted	8.3	39.2	11.4	15.1	16.4	9.6
Fearful	12.3	6.6	47.5	11	13.2	9.4
Happy	10.4	15.5	15.1	38.1	11.5	9.4
Sad	19.7	11	15.9	13.6	29.9	9.8
Surprised	17.4	18.9	14.4	11.7	8.7	28.8

(e) CNN (Late)

Fig. 7: Confusion matrices of the three integration models for RAVDESS dataset with Brown noise

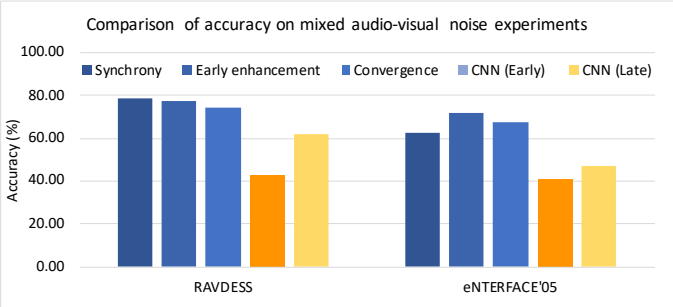

Fig. 8: Comparison of accuracy of *Synchrony*, *Enhancement*, *Convergence* and CNN baseline models on a mixed noise experiment (0.2 salt and pepper noise and white audio noise)

[12] C. Cuppini, M. Ursino, E. Magosso, B. A. Rowland, and B. E. Stein, "An emergent model of multisensory integration in superior colliculus neurons," *Frontiers in integrative neuroscience*, vol. 4, p. 6, 2010.

[13] A. Barutchu, C. Spence, and G. W. Humphreys, "Multisensory enhancement elicited by unconscious visual stimuli," *Experimental Brain Research*, vol. 236, no. 2, pp. 409–417, Feb 2018.

[14] H. Atilgan, S. M. Town, K. C. Wood, G. P. Jones, R. K. Maddox, A. K. Lee, and J. K. Bizley, "Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding," *Neuron*, vol. 97, no. 3, pp. 640–655, 2018.

[15] A. E. Symons, W. El-Deredy, M. Schwartze, and S. A. Kotz, "The functional role of neural oscillations in non-verbal emotional communication," *Frontiers in Human Neuroscience*, vol. 10, p. 239, 2016.

[16] J. Keil and D. Senkowski, "Neural oscillations orchestrate multisensory processing," *The Neuroscientist*, vol. 24, no. 6, pp. 609–626, 2018.

[17] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *International conference on neural information processing*. Springer, 2016, pp. 521–529.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran, "An event driven fusion approach for enjoyment recognition in real-time," in *MM '14*, 2014, pp. 377–386.

[20] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based encoding method for emotion recognition in video," in *ICASSP '16*, 2016, pp. 2752–2756.

[21] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[22] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV 2015*, 2015.

[24] M. Mukeshimana, X. Ban, N. Karani, and R. Liu, "Multimodal emotion recognition for human-computer interaction: A survey," *System*, vol. 9, p. 10.

[25] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[26] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, and F. Schwenker, *Kalman Filter Based Classifier Fusion for Affective State Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 85–94.

[27] C. Felipe, M. Luis J, and N. Pedro, "A novel multimodal emotion recognition approach for affective human robot interaction," in *IEEE/RSJ IROS '15*, 2015.

[28] M. Schels, M. Glodek, S. Meudt, S. Scherer, M. Schmidt, G. Layher, S. Tschechne, T. Brosch, D. Hrabal, S. Walter, H. C. Traue, G. Palm, F. Schwenker, M. Rojc, and N. Campbell, "Multi-modal classifier -fusion for the recognition of emotions," in *Converbal Synchrony in Human-Machine Interaction*. CRC Press, sep 2013, pp. 73–97.

[29] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *ICMI 2015*, 2015, pp. 497–502.

[30] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *ICMI 2016*, 2016, pp. 284–288.

[31] L. Duan, H. Ge, Z. Yang, and J. Chen, *Multimodal Fusion Using Kernel-Based ELM for Video Emotion Recognition*. Springer International Publishing, 2016, pp. 371–381.

[32] A. Iosifidis, A. Tefas, and I. Pitas, "On the kernel extreme learning machine classifier," *Pattern Recognition Letters*, vol. 54, pp. 11–17, 2015.

[33] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[34] K. Bahreini, R. Nadolski, and W. Westera, "Data fusion for real-time multimodal emotion recognition through webcams and microphones in e-learning," *International Journal of Human Computer Interaction*, vol. 32, no. 5, pp. 415–430, 2016.

[35] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, May 2013.

[36] M. WöLlmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vision Comput.*, vol. 31, no. 2, p. 153–163, Feb. 2013.

[37] M. R. Amer, T. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, and S. Chai, "Deep multimodal fusion: A hybrid approach," *International Journal of Computer Vision*, Feb 2017.

[38] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Two distributed-state models for generating high-dimensional time series," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 1025–1068, Jul. 2011.

[39] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *ICMR 2016*, 2016, pp. 281–284.

[40] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *ICDM '16*. IEEE, 2016, pp. 439–448.

[41] N. Subrahmanya and Y. C. Shin, "Sparse multiple kernel learning for signal processing applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 788–798, 2010.

[42] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *ACII 2019*, 2019.

[43] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *WACV '17*. IEEE, 2017, pp. 1215–1223.

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV '15*, 2015, pp. 4489–4497.

[45] D. Bhandari, S. Paul, and A. Narayan, "Multimodal data fusion and prediction of emotional dimensions using deep neural

network," in *Computational Intelligence: Theories, Applications and Future Directions-Volume II*. Springer, 2019, pp. 215–228.
[46] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *FG 2013*, 2013, pp. 1–8.
[47] J. D. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," *arXiv preprint arXiv:1907.03196*, 2019.
[48] S. Jessen and S. A. Kotz, "On the role of crossmodal prediction in audiovisual emotion perception," *Frontiers in Human Neuroscience*, vol. 7, p. 369, 2013.
[49] A. Shaked and G. L. Clore, "Breaking the world to make it whole again: Attribution in the construction of emotion," *Emotion Review*, vol. 9, no. 1, pp. 27–35, 2017.
[50] B. E. Stein, *The new handbook of multisensory processing*. The MIT Press, 2012.
[51] B. E. Stein and M. A. Meredith, *The merging of the senses.* The MIT Press, 1993.
[52] S. Molholm, W. Ritter, M. M. Murray, D. C. Javitt, C. E. Schroeder, and J. J. Foxe, "Multisensory auditory visual interactions during early sensory processing in humans: a high-density electrical mapping study," *Cognitive Brain Research*, vol. 14, no. 1, pp. 115 – 128, 2002, multisensory Proceedings.
[53] A. E. Symons, "Examining the role of temporal prediction in multisensory emotion perception," Ph.D. dissertation, The University of Manchester (United Kingdom), 2018.
[54] E. Tsilionis and A. Vatakis, "Multisensory binding: is the contribution of synchrony and semantic congruency obligatory?" *Current Opinion in Behavioral Sciences*, vol. 8, pp. 7–13, 2016.
[55] M. M. Murray and M. T. Wallace, *The neural bases of multisensory processes*. CRC Press, 2011.
[56] C. Kayser and N. K. Logothetis, "Do early sensory cortices integrate cross-modal information?" *Brain structure and function*, vol. 212, no. 2, pp. 121–132, 2007.
[57] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659 – 1671, 1997.
[58] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Bulletin of Mathematical Biology*, vol. 52, pp. 25–71, 1990.
[59] J. T. Jose, J. Amudha, and G. Sanjay, "A survey on spiking neural networks in image processing," in *Advances in Intelligent Informatics*, E.-S. M. El-Alfy, S. M. Thampi, H. Takagi, S. Piramuthu, and T. Hanne, Eds., 2015, pp. 107–115.
[60] J. P. Dominguez-Morales, Q. Liu, R. James, D. Gutierrez-Galan, A. Jimenez-Fernandez, S. Davidson, and S. Furber, "Deep spiking neural network model for time-variant signals classification: a real-time speech recognition approach," in *IJCNN 2018*, July 2018.
[61] E. Mansouri-Benssassi and J. Ye, "Bio-inspired spiking neural networks for facial expression recognition: Generalisation investigation," in *International Conference on Theory and Practice of Natural Computing*. Springer, 2018, pp. 426–437.
[62] ——, "Synch-graph: Multisensory emotion recognition through neural synchrony via graph convolutional networks," in *2020 AAAI*. AAAI, 2020.
[63] P. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 9, p. 99, 2015.
[64] N. Rathi and K. Roy, "Stdp-based unsupervised multimodal learning with cross-modal processing in spiking neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2018.
[65] E. Mansouri-Benssassi and J. Ye, "Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks," in *IJCNN '19*. IEEE, 2019, pp. 1–8.
[66] H. Hazan, D. Saunders, D. T. Sanghavi, H. Siegelmann, and R. Kozma, "Unsupervised learning with self-organizing spiking neural networks," in *IJCNN 2018*. IEEE, 2018, pp. 1–6.
[67] K. Strelnikov, J. Foxton, M. Marx, and P. Barone, "Brain prediction of auditory emphasis by facial expressions during audiovisual continuous speech," *Brain topography*, vol. 28, no. 3, pp. 494–505, 2015.
[68] R. Brette, "Computing with neural synchrony," *PLoS computational biology*, vol. 8, no. 6, p. e1002561, 2012.
[69] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR 2017*, 2017.
[70] D. K. Duvenaud and et al, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224–2232.
[71] I. Pitas, I. Kotsia, O. Martin, and B. Macq, "The eNTERFACE'05 audio-visual emotion database," in *ICDEW'06*, 2006.
[72] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS ONE*, vol. 13, no. 5, p. e0196391, 2018.
[73] K. Tarunika, R. B. Pradeeba, and P. Aruna, "Applying machine learning techniques for speech emotion recognition," in *ICCCNT '18*, July 2018, pp. 1–5.
[74] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa : Audio and music signal analysis in python," 2015.
[75] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London Series B*, vol. 23, pp. 187–217, 1980.
[76] N. Kilian-Hütten, E. Formisano, and J. Vroomen, "Multisensory integration in speech processing: Neural mechanisms of cross-modal aftereffects," in *Neural Mechanisms of Language*, M. Mody, Ed., 2017, pp. 105–127.
[77] R. D. Fonnegra and G. M. Díaz, "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model," in *Human-Computer Interaction. Theories, Methods, and Human Issues*, M. Kurosu, Ed., 2018, pp. 385–396.
[78] E. Di Nardo, A. Petrosino, and I. Ullah, "Emop3d: A brain like pyramidal deep neural network for emotion recognition," in *ECCV*, 2018.
[79] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Fusion of classifier predictions for audio-visual emotion recognition," in *ICPR '16*, 2016, pp. 61–66.
[80] ——, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 06 2017.
[81] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, no. Supplement C, pp. 610 – 628, 2017.
[82] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep cnn-based face recognition?" in *BIOSIG '16*. IEEE, 2016, pp. 1–5.
[83] M. Coto-Jiménez, J. Goddard-Close, and F. Martínez-Licona, "Improving automatic speech recognition containing additive noise using deep denoising autoencoders of lstm networks," in *SPECOM '16*, 2016, pp. 354–361.
[84] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan, "A spiking neural network framework for robust sound classification," *Frontiers in neuroscience*, vol. 12, 2018.
[85] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
[86] J. Kwisthout and N. Donselaar, "On the computational power and complexity of spiking neural networks," in *NICE '20*, 2020.

Esma Mansouri Benssassi is a PhD candidate in the School of Computer Science at the University of St Andrews. Her primary research areas are designing and developing bio-inspired models for multisensory integration. Contact her at emb24@st-andrews.ac.uk.

Juan Ye is a senior lecturer in the School of Computer Science at the University of St Andrews. Her research interests centre around adaptive pervasive systems, specialising in sensor-based human activity recognition, sensor fusion, context awareness, ontologies, and uncertainty reasoning. She has a PhD in computer science from University College Dublin. Contact her at juan.ye@st-andrews.ac.uk.