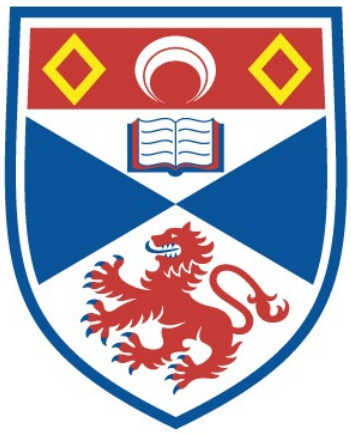
BIO-INSPIRED MULTISENSORY INTEGRATION OF SOCIAL SIGNS

Esma Mansouri Benssassi

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2020

Bio-inspired Multisensory Integration of Social Signals

Esma Mansouri Benssassi

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

February 2020

Abstract

Emotions 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. Emotions are expressed in a multisensory manner, where humans use social signals from different sensory modalities such as facial expression, vocal changes, or body language. The human brain integrates all relevant information to create a new multisensory percept and derives emotional meaning.

There exists a great interest for emotions recognition in various fields such as HCI, gaming, marketing, and assistive technologies. This demand is driving an increase in research on multisensory emotion recognition. The majority of 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 ineffective in translating the constant talk and feedback between different modalities. Such constant talk is particularly crucial in continuous emotion recognition, where one modality can predict, enhance and complete the other.

This thesis proposes novel architectures for multisensory emotions recognition inspired by multisensory integration in the brain. First, we explore the use of bio-inspired unsupervised learning for unisensory emotion recognition for audio and visual modalities. Then we propose three multisensory integration models, based on different pathways for 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) with unsupervised learning. 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 achieve comparable results to state-of-the-art supervised

learning techniques. More importantly, this thesis shows models that can translate a constant talk between modalities during the training phase. Each modality can predict, complement and enhance the other using constant feedback. The cross-talk between modalities adds an insight into emotions compared to traditional fusion techniques.

Acknowledgements

I would like to express all my thanks to my supervisor Dr. Juan Ye for guiding me well throughout the research work from the beginning to the final results. Her motivation and energy have given me more power and spirit to excel in this research work. I would like to thank her for being always available and answering all my questions.

I also won't forget to express my gratitude to the administration, and support teams at the school of computer science.

Last but not least, I am very grateful to my husband Abdel without whom I would not have been able to achieve this goal, and would like to thank him for being always supportive and patient with me for the past four years. I would also like to thank my daughters Nadia and Sofia who never stopped cheering me up. They gave me enough moral support, encouragement and motivation to accomplish this work. I would also like to thank my mother and sisters who always kept me in their prayers.

Candidate's declaration

I, Esma Mansouri Benssassi, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 50,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2016.

I confirm that no funding was received for this work.

Date 27.05.2020 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 27.05.2020 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Esma Mansouri Benssassi, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 27.05.2020 Signature of candidate

Date Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Esma Mansouri Benssassi, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 27.05.2020 Signature of candidate

Publications

Some of the work presented in this thesis has been published in the following papers:

1. E. Mansouri Benssassi and J. Ye. Synch-graph: multisensory emotion recognition through neural synchrony via graph convolutional networks. In Procceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020). AAAI Press, 2020.

2. E. Mansouri Benssassi and J. Ye (2019). Generalisation and robustness investigation for facial and speech emotion recognition on bio-inspired spiking neural networks. Under Review.

3. E. Mansouri Benssassi and J. Ye (2019). Speech Emotion Recognition With Early Visual Cross-Modal Enhancement Using Spiking Neural Networks. Accepted by IJCNN '19: the international conference on neural networks. July 14-19, 2019, Budapest, Hungary.

4. E. Mansouri Benssassi, J. Gomez, L. E. Boyd, G. R. Hayes, and J. Ye (2018), Wearable assistive technologies for autism: opportunities and challenges. IEEE Pervasive Magazine's special issue on augmenting humans. 2018.

5. E. Mansouri Benssassi and J. Ye (2018). Bio-Inspired Spiking Neural Networks for Facial Expression Recognition: Generalisation Investigation. TPNC'18: the 7th international conference on the theory and practice of natural computing. December 12-14, 2018, Dublin, Ireland.

6. E. Mansouri Benssassi (2017). A decentralised multimodal integration of social signals: a bio-inspired approach. ICMI 2017 Doctoral Consortium, 3rd August 2017.

Contents

List of Figures

List of Tables

Acronyms

HCI Human Computer Interaction

SNN Spiking Neural Network

FER Facial Rxpression recognition

SER Speech Emotion Recognition

SC Superior Colliculus

STDP Spike Timing Dependent Plasticity

GCN Graph Convolution Neural Network

STS Superior Temporal Sulcus

MEG Magnetoencephalography

EEG Electroencephalogram

fMRI Functional Magnetic Resonance Imaging

rpSTS Right Posterior Superior Temporal Sulcus

HRI Human Robot Interaction

CNN Convolution Neural Network

AU Action Units

LBP Local Binary Pattern

SVM Support Vector Machines

ANN Artificial Neural Network

PCA Principal Component Analysis

HOG Histograms of Oriented Gradients

DCNN Descriminative Convolutional Neural Network

MFCCs Mel Frequency Cepstral Coefficients

HMM Hidden-Markov Model

GMM Gaussian Mixture Models

LSTM Long Short Term Memory

RNN Recurrent Neural Networks

LSM Liquid State Machines

SVR Support Vector Regression

RBM Restricted Boltzmann Machine

DBN Deep Belief Network

LIF Leaky Integrate And Fire

LoG Laplacian Of Gaussian

Chapter 1

Introduction

1.1 Motivation

Humans perceive emotions in a multisensory manner, where information from different sensory modalities such as facial expression, verbal, non-verbal speech signals, and body language translate our emotional states. Multisensory emotional processing is driven by a constant cross-talk between various sensory modalities.

Understanding emotions from multiple sensory modalities is crucial for Human Computer Interaction (HCI) and affective computing with various applications such as gaming, mental healthcare or car driving [24]. Multisensory emotions recognition does not only provide more effective and efficient Human-Computer Interaction (HCI) but also facilitates the enhancement and efficiency of assistive technologies or social robots for individuals facing challenges in interpreting complex and subtle social cues [50, 24]. Therefore, it is crucial to analyse and focus on multisensory relationship between different modalities to get a more accurate meaning and a better interpretation of emotions.

There exists a high interest in developing effective multisensory systems for emotion recognition in HCI and affective computing fields. Classical techniques have been used in feature extraction, fusion and classification of multisensory signals. However, current data fusion techniques are generally not able to translate the constant cross-talk and complementarity between modalities. Moreover, current data fusion techniques applied in emotion recognition and affective computing do not translate the multisensory precept, where information is delivered from

different modalities through a constant talk and feedback between multiple sensory modalities.

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 address various challenges faced by existing systems and can help not only in the fusion of information but in a more practical perceptual understanding of emotions by modelling the learning and interaction between modalities.

1.2 Challenges

Nowadays, with the progress in sensing and intelligent technologies, social signals of emotions from various sensory modalities can be captured through various means such as portable or wearable devices in real-time. However, designing a practical integration approach is challenging, since social signals from different modalities can come at different time onsets with noisy environments. These can affect their reliability and contribution to the final inference. Systems need to discern, extract, and process signals in order to derive meaningful interpretation of social signals. With the current advances in machine learning, computer vision and human-computer interactions techniques, there is growing interest in developing techniques for better interpretation of multisensory social signals 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 and social interaction recognition [63, 10, 116] for both unisensory and multisensory recognition tasks. However, they only focus on features extraction and often combine with conventional data fusion techniques such as feature concatenation or decision level fusion [7, 15].

Current techniques often lack in accurately converting signals into new multisensory precepts and integrating information from different modalities. They also fail to translate the relationship and constant cross-talk between different sensory modalities, where each modality receives input throughout the learning phase.

Current multisensory integration techniques face various challenges including inconsistency

of one or more signals, lack of interaction between different modalities or timing, and asynchrony. They do not translate cross-modal prediction [227]. The prediction, interaction and integration play a significant role in translating multisensory information. This is how the integration of multisensory social signals occurs in the human brain [82].

1.3 Aims and Objectives

Research presented in this thesis aims to explore and propose novel biologically inspired architectures for multisensory integration. These novel methods are directly inspired by neuro-computational models and studies in neuroscience on multisensory integration [40].

The main objectives of this research are summarised as follows:

1. Carry out a thorough literature review on existing multisensory data integration techniques used in multisensory emotion recognition and identify their main challenges.

2. Investigative current research trends on neuroscience models and theories for multisensory integration pathways in the brain.

3. Propose, create and design novel models for multisensory integration with application in social signals of emotions.

4. Evaluate the proposed models on various multisensory emotions datasets, explore their strength, and identify their limitations.

1.4 Thesis Hypothesis

The main hypothesis of this thesis is: *"Bio-inspired architectures enable better multisensory integration and more accurate translation of constant interaction between different modalities"*

1.5 Research Questions

The main questions to be investigated in this thesis are:

1. Are bio-inspired architecture effective for unisensory emotions recognition tasks?
2. Does applying bio-inspired models in multisensory integration increase the accuracy of multisensory recognition systems?
3. Do bio-inspired models present better generalisation capacity compared to state-of-the-art?
4. Are bio-inspired architectures robust to signal noise?

In order to answer the presented questions, this thesis proceeds to the following tasks:

- Create bio-inspired models for unisensory emotion recognition for audio and visual data.
- Explore, investigate and create novel bio-inspired models for multisensory integration of social signals of emotions.
- Evaluate the robustness to noise and generalisation capacity of such models.

1.6 Main Contributions

This thesis proposes novel bio-inspired approaches to not only model social signals of each modality but also model their interaction and integration to enable more biologically plausible signal integration and achieve better performance. This work is novel for affective computing in general and multisensory integration in particular by using models inspired by the brain. The main contributions can be summarised in the following sections.

1.6.1 Bio-inspired Model – Spiking Neural Network

Applying bio-inspired models in multisensory integration helps alleviate some of the challenges faced by the current data fusion techniques. Preserving spatio-temporal relation between multiple modalities is essential, as is the constant interaction between them. The implementation of the proposed models is achieved through Spiking Neural Network (SNN). SNNs represents the third generation of neural networks, and have been mainly used for implementing neuro-computational models [176] [110]. The main difference from artificial neural networks relies in the way SNNs process information based on spikes, where neurons communicate through series of spikes by

firing spikes when they reach a certain threshold [221]. The computation in SNNs is based on timing of spikes, where spikes that fire together get a stronger connection. SNNs are becoming more popular with the advances in Neuromorphic computing research. They include various applications such as audio signal processing, pattern recognition.

The first contribution in this research work consists of adapting SNNs with unsupervised learning in a novel way for unisensory emotion recognition for both audio and visual data and extract essential features that can be generalised across datasets and robust to noise degradation [179].

1.6.2 Multisensory Social Signal Integration

The most significant contribution is to propose three novel models for multisensory social signals of emotions recognition, which focus on applying bio-inspired techniques derived from three different pathways of multisensory integration in the brain.

Humans and animals perceive events in a multisensory manner, where information enters the brain through various sensory modalities. Information is integrated following specific rules such as temporal alignment and spatial and semantic congruence. Multisensory integration represents the process of combining inputs from different modalities, such as visual and auditory. The brain responds to multisensory information by first, processing information from each sensory modality, and then integrating data to form a new multisensory percept.

The proposed models are detailed as follows:

- *Integration through convergence* Multisensory integration is the first proposed integration model. This model applies a classical theory in multisensory integration which happens in higher-order areas of the brain; that is, convergence of various information from unisensory areas into one multisensory area [253]. First, information is processed in each unisensory area. Then features converge in a higher-order multisensory regions. This method mainly relies on firing rate changes in different cortical regions through a hierarchical and progressive manner. In this approach, the integration happens in a convergence manner, where the response to multisensory information is compared to the sum of response to unisensory input.

- *Integration through cross-modal enhancement* The second porposed model is integration through early cross-modal enhancement. Early cross-modal enhancement is a pathway in multisensory integration in the brain and is derived from the work detailed in [265]. Studies suggest that multisensory areas such as Superior Colliculus (SC) use a Spike Timing Dependent Plasticity (STDP) learning at a neural level for the interaction between different unisensory modalities [265]. It has also been described that unisensory areas interact at early sensory levels [14] during multisensory integration. Auditory and visual areas interconnect with recurrent connections. This idea of early sensory interaction represents a possibility of cross-modal prediction and interaction especially for audio-visual pathway in emotions processing [20, 14] [191] [119].

  The model proposed in this thesis enhances speech emotion recognition through visual information and achieves better performance compared to most commonly used state-of-the-art data fusion approaches in multisensory emotion recognition [310]. This model is more simplified and computationally advantageous. Also, rather than a simple fusion, this model promotes loose coupling between multiple signal modalities, which can be more flexible and robust. For example, where one modality fails or is very noisy, it will less affect the overall recognition accuracy.

- *Integration through neural synchrony* Integration through neural synchrony is the third proposed model. Studies have identified various regions where multisensory integration is achieved, such as the temporal frontal and primary sensory areas [283].

  Neural synchrony represents one of the most recent views for multisensory integration [133]. It is derived from various experiments on humans and animals. It is defined as the simultaneous neural oscillations of different neuron groups in various brain cortical regions connected by synapses. It is considered as the primary means of transferring information in the brain. Numerous studies have been conducted in order to define the exact role of neural synchrony in multisensory integration [284]. Neural synchrony is defined as the synchronisation of different brain oscillations in different frequencies. Each frequency band drives a specific type of information such as cognitive functions. Multisensory integration through neural synchrony is modelled using SNNs and (Graph Convolution

Neural Network (GCN))

The three models present three distinctive pathways in multisensory integration in the brain. Multisensory integration in the brain has various pathways from early sensory areas to higher-order areas. This thesis aims at representing the main pathways for multisensory integration from early sensory interaction to higher-order multisensory areas.

1.7 Organisation of the Thesis

This thesis is organised in eight main chapters as follows:

- **Chapter 2** introduces the background and the motivation of this thesis. It describes an overview of the nature of emotions in humans. Then it gives a summary of state-of-the-art emotion recognition techniques both in unisensory and multisensory tasks. It focuses on emotion recognition from videos with audio-visual integration. It also outlines the main challenges in current multisensory emotion recognition techniques.

- **Chapter 3** describes how biological neurons function and how signals are transmitted in the brain. It provides mathematical models about neurons communication. It introduces SNNs, and describes their different possible architectures and learning methods.

- **Chapter 4** proposes two bio-inspired models for unisensory emotion recognition. It details the application of bio-inspired models for Facial Expression Recognition FER and Speech Emotion Recognition SER tasks.

- **Chapter 5** outlines the major contribution to this thesis. It details the design of three models for multisensory integration. It details the architecture and design of multisensory integration through convergence, early cross-modal enhancement and finally through neural synchrony.

- **Chapter 6** details all experimental setup and implementation details of various models proposed in this thesis.

- **Chapter 7** details experiments to evaluation the bio-inspired methods for unisensory emotion recognition in FER and SER. It describes all experimental setups, tools, baseline models. It also details all experimental results and discussion.
- **Chapter 8** describes the evaluation of the three proposed models for multisensory integration. It describes experimental setups, results for each models and a comparison of models in generalisation and robustness to noise evaluations.
- **Chapter 9** concludes the thesis and summarises the main contribution, evaluations and results. It also discusses the limitations of the current work and points out the future direction.

Chapter 2

Background and Literature Review

2.1 Introduction

This chapter provides an overview of existing methods for multisensory emotion recognition focusing on visual and auditory sensory modalities. It consists of three parts:

1) First, it outlines the nature of human emotions and emotion perception in psychology and neuroscience. It starts by illustrating the importance of emotions for human communication. Then, it explains different models for emotions in the literature. After that, it describes the nature of emotions and social signals in the human brain and how humans perceive them.

2) The second part consists of a thorough review of emotion recognition in machine learning and affective computing fields. Although the review describes various modalities such as body language, text, facial expression or non-verbal speech, it focuses on two primary unisensory modalities; that is, audio and visual. These are also the modalities that this thesis focuses on.

3) This chapter then outlines state-of-the-art for multisensory emotion recognition by listing the most common multisensory integration and data fusion techniques used in emotion recognition.

Finally, it identifies the main challenges faced by the recent multisensory emotion recognition techniques, which will be tackled by this thesis.

2.2 The Nature of Human Social Signals of Emotions

Humans interact through multiple social signals, translating different mental states and feelings. Emotions represent the way we communicate our internal mental states, resulting from reactions to external stimuli. Emotions are driven by rewards and punishment motivations [234]. They are considered as states elicited through reinforcement behaviours. According to [233], emotions can have the following functions:

- Physiological changes: Physiological response elicitation such as changes in heart rate.
- Reinforcement: Changes in emotional response to external reinforcement stimuli, where the brain responds to two types of reinforcement, either reward or punishment.
- Communication: Humans and primates can communicate through different means, including facial expressions. Facial expression can translate various emotional states through deformation of different facial muscles. There are special areas in the brain for processing facial expressions.
- Social bonding: Emotions are linked to attachment in humans and primates, such as parents and their children.
- Motivation: Emotions are motivating. For example, they can elicit reactions through stimulus reinforcement associations in the case of fear.
- Effect on cognitive functions and memory: Emotions can help elicit and store events in memory. Memories such as episodic memory can be facilitated through emotions. One way for emotions to elicit memories is by triggering perceptual representation in the brain.
- Direction to behaviours: Emotions can change behaviours in both humans and primates, such as fear that drives a change in behaviour.

Emotions have been investigated thoroughly throughout history. The first documented work on human emotions dates back to 1872. Darwin [58] was the first to provide a theory of emotions and facial expressions. According to Darwin's studies, emotion expressions and affective states in humans and other primates happen over time through deformation of various facial muscles. He

has also claimed that some emotional expressions were universal and shared the same expression through cultures and ethnic groups. Humans or animals from the same species react in the same manner when presented with the same stimuli or facing the same situation in those universal expression of emotions.

Emotions are categorised by their type, intensity and other parameters such as context [106]. These parameters make up emotion models. Emotion models consist of defining emotions based on scores, dimensions and ranks. Existing emotion models are based on intensity, dynamic change or even appraisal elicitation or behavioural change.

Research in psychology is dominated by two main theories for emotions: ***dimensional*** or ***categorical***. Categorical models describe emotions by category which are entirely distinct. Ekman [70] has proposed one of the most popular categorical models. Ekman has researched facial expression of emotions and studied the importance of facial expression in sharing one's mental states. He also identified six basic emotions shared across cultures (anger, happiness, sadness, fear, disgust and surprise). Other studies have produced different models [280] as to the definition of basic emotions.

On the other hand, dimensional emotions models as defined by Russel [239] define emotions by dimensions of some predefined parameters. Dimensions comprise valence (*indicating if the emotion is positive or negative*) and arousal (*defining the intensity of the emotion*). A third dimension consists of dominance, indicating the level of control. Plutchik [220] proposed a hybrid emotion model with eight basic emotions (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy) with three-dimensional levels.

Figures 2.1, 2.2 and 2.3 show an example of Ekman's categorical emotion model, Russel's dimensional model and the categorical dimensional model from Plutchik's model. The categorical model presented by Ekman as shown in Figure 2.1 shows emotions as discrete categories. On the other hand, dimensional models such as Russel's model as shown in 2.2 define emotions as distributed in two dimensional space with the x-axis defining the valence and the y-axis the arousal dimension. The hybrid model presented by Plutchik as shown in Figure 2.3 represent emotions in three main dimensions representing a hybrid between the discrete basic emotions and dimensional model. This emotion model results in several emotions with different intensities. The main difference between categorical models such as Ekman's and dimensional models such

Figure 2.1: Ekmann's emotion model (Categorical) with six basic emotions (surprise, sad, happy, angry, fearful and disgust

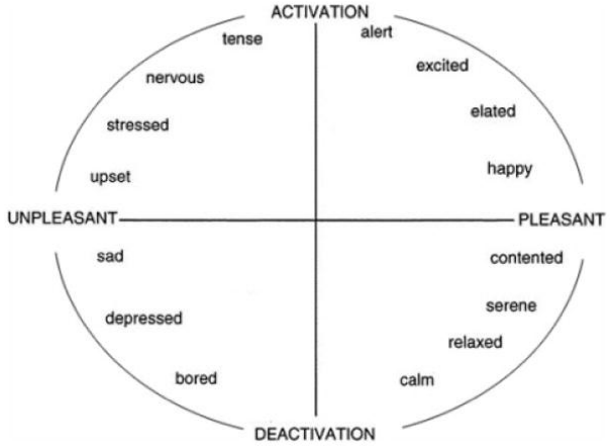


Figure 2.2: Russel's emotion model (Dimentional) [239]

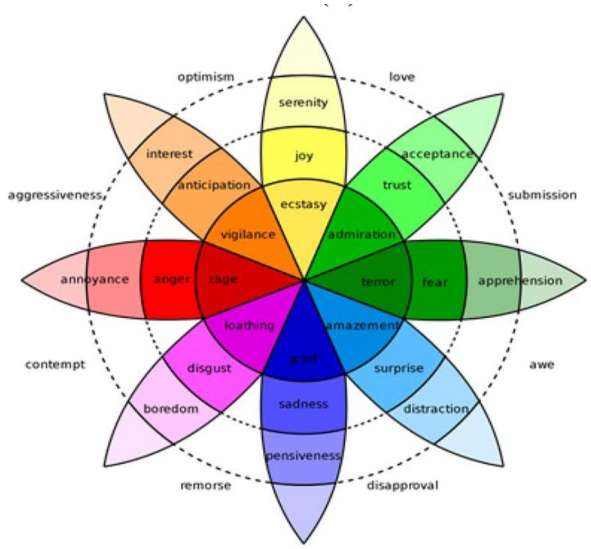


Figure 2.3: Pultchik's emotion model (Hybrid) [220] comprises concentric circles where inner circle represent basic emotions and outer circles more complex.

as Russel's or Plutchnik is that they present emotions in different dimensions in contrast to discreet emotions proposed in categorical models.

In addition to facial expressions, there are other different ways of expressing emotions, such as body gesture, verbal and non-verbals speech signals. This group of manifestations constitute social signals. Social signals represent a significant part of humans interactions and are crucial in understanding communications. They also influence behaviours, reaction and internal mental states of individuals [187]. Methu et al. [187] have shown that humans' mental states are closely influenced by watching and hearing emotional stimuli such as laughter, cries, threatening gestures or voices.

Social signals and emotions are interpreted differently depending on contexts such as cultural differences, place or time [106]. For example, a facial expression can be interpreted in various ways depending on a cultural context. Social signals cues can also play a role of context. A tone of voice can affect and influence our understanding of facial expressions; this is primarily present in complex emotions such as sarcasm. Thus, the ability to comprehend the whole picture of emotions manifestation is crucial in humans communication.

Understanding others' emotional state involves integrating various social signals such as tone of voice, facial expression and body gesture. The inability to integrate these social signals to understand others' mental states and emotions can be missing. Individuals with disorders and conditions such as autism, schizophrenia, or dementia find it very hard being in social situations and making meaning from different social signals. This barrier creates challenges in communicating with others. Therefore the integration of social signals is very crucial in understanding the whole picture of emotional states and is referred to multisensory integration.

Social signals of emotions processing, understanding and perception involves various areas of the brain and a complex network [120]. The human brain starts by parsing input from different senses through segmentation and then works on constructing meaningful models through integration [254]. These two processes are always active in the brain, constructing meaningful interpretation of the present from past events. For example, in order to conclude a speech sound, the brain needs to segment all possible auditory inputs from noisy environments to identify which sound corresponds to the speech of a person. It is achieved by looking at temporally and spatially adjacent sensory information from both visual and auditory inputs such

as facial expressions and non-emotional verbal sounds.

2.3 Multisensory Integration of Emotions in the Brain

The brain goes through three main steps for integrating emotions social signals. It first uses attention to select the emotional information for the observed data. It then integrates the affective information where a new multisensory precept is created. In this step, all unisensory modalities interact constantly. Finally, an evaluation and interpretation is made based on the new percept.

Multisensory integration represents the process by which information from different sensory modalities are gathered and integrated to form an overall emotional precept. Multisensory integration have been studied widely in neuroscience and psychology fields [23]. It comprises various domains such as cognitive tasks, motor tasks or emotion recognition. Research includes both behavioural experiments or imaging studies using Functional Magnetic Resonance Imaging (fMRI), Electroencephalogram (EEG) or Magnetoencephalography (MEG) [60]. Most of recent studies in multisensory integration focus on the interaction between faces and voices, body and faces or body and voices [219].

Multisensory integration in emotions follows a different process in the brain [59] compared to more general multisensory integration. In fact Davies et al. [59] have found that the brain processes emotional information in a different manner, where brain regions such as right posterior superior temporal sulcus (Right Posterior Superior Temporal Sulcus (rpSTS)) respond only to multisensory emotional information, when the brain is presented with both audio and visual emotional information.

The interaction and integration between faces and voice represents the most investigated area in multisensory integration of emotions. Experiments range between studying static and dynamic facial expressions. Static facial expression based experiments consist of employing static images defining a emotional state. Dynamic facial expression based experiments, refers to using dynamic expression through videos showing continuous change of facial expressions.

Experiments conducted by [60] have focused on static emotion recognition by presenting participants with static facial expression ranging from happy to sad in addition to short spoken sentences were added with either a happy or sad emotional voice tone. Participants were always

biased and influenced by the voice when judging facial expression even if instructed to ignore the voice stimuli.

Garrido et al. [82] have focused on dynamic facial expression and voices in emotion perception. They have investigated early cross-modal interaction between visual and auditory modalities. They have presented participants with dynamic facial expressions such as angry, happy and neutral as primes to auditory emotional tones happy and sad as targets. They have investigated how visual information would affect early auditory processing and how cross-modal prediction occurs. They have found that facial expressions affect auditory emotion processing at an early stage. When information between visual and auditory information is incongruent, there is an additional processing in the brain.

Multisensory integration of emotion have also been investigated through the interaction of body gestures and voices. Studies show that body gestures do influence the perceptions of emotions in voices [121].

Emotions are naturally multisensory, where each sensory modality influences, enhances and complements the others. Having effective multisensory integration is crucial in understanding emotions. Multisensory integration is not only crucial for affective perception but also major cognitive processes. Studies show that individuals encountering difficulties in emotional understanding such as dementia, schizophrenia or autism may have a multisensory integration impairment [236], [33] ,[74].

Most studies show an altered multisensory integration for emotional stimuli. Some research hypothesise that this can be due to some defective neurofunctional mechanisms in some brain areas such as the superior temporal cortex [282]. Therapy methods are focusing on assistive technologies to help individuals with emotional understanding. HCI, Human Robot Interaction (HRI) and affective computing areas have witnessed a surge in applications in assistive technologies or social robots [50] [24]. Applications in assistive technologies can alleviate some difficulties faced by individuals with impaired emotions understanding. Therefore, it is imperative to capture, interpret, and deliver the right social signals in assistive technologies.

Nowadays, following advances in sensing and intelligent technologies, signals from different modalities can be captured through portable or wearable devices in real-time. Each modality provides one aspect of social interaction and often needs to be integrated to derive a more

robust and comprehensive meaning of a social situation. Creating a new percept or accurate interpretation of data from different modalities can be very challenging, as they can come at different timing onsets and need to be included in a noisy environment. Systems need to discern, extract and process the right inputs to have a meaningful output and interpretation of social signals.

The following sections outline the current research trends in both unisensory and multisensory emotion recognition in affective computing and machine learning, focusing on facial expressions and non-verbal speech modalities. First, the section describes state-of-the-art review of unisensory emotion recognition focusing on the most popular ones; that is; speech and facial expression tasks. Then, it details current trends in multisensory emotion recognition in the literature. Finally, an outline of some critical challenges faced by current methods in multisensory emotion recognition is given.

2.4 The Importance of Emotions Datasets

One of the most challenging steps in multisensory emotion recognition process is collecting multisensory data reflecting various types of emotions from multiple subjects. Each dataset is prepared using different conditions. Therefore, it can be challenging to compare models using different datasets due to variation of subjects, data collection context or data dimension. Most existing methods for multisensory emotion recognition are based on publicly available datasets, where data are either acted or induced, or natural. However, there are very few datasets publicly available. Acted or posed datasets are prepared by asking subjects to show predefined emotional states. Induced emotions dataset, on the other hand, are usually prepared by putting subjects in certain emotional situations. Natural emotions dataset are usually recorded in real-life situation, where emotions expression are genuine and define the most natural emotional states. However, the natural emotions datasets often contain other factors and noise beyond facial expression, so state-of-the-art methods usually use acted or induced dataset emotion for evaluation and models training.

2.5 Facial Emotion Recognition

Facial expression recognition and classification represents one of the fast-growing and significant research areas in recent years in the computer vision field. While the main goal is to facilitate human-computer interaction, it forms the basis of affective computing where research is still mainly concentrated on FER [112].

Facial expressions represent a crucial non-verbal means of expressing emotions and mental states and are defined by the deformation of multiple muscles in the face. Distinct combinations of these deformations form a representation of different emotions. The study of facial expression goes back to Darwin and later to Ekman et al. [70] where they described that some expressions are universal and have defined six basic expressions namely, sadness, happiness, disgust, fear, anger and surprise.

Automatic facial expression recognition has developed significantly in recent years with the development of computer vision, machine learning and image processing techniques. However having accurate systems is still a challenging task in naturalistic and non-controlled environments, due to changes in face dimensions, head pose or facial features. Most research report different accuracy depending on datasets and methods used.

One can distinguish three main methods for facial expression recognition in the literature: handcrafted features, model-based and deep learning. The majority of conventional approaches consist of two primary methods; that is features and appearance-based. Feature-based approaches aim at extracting facial features such as nose, brows eyes or mouth and assessing changes in the geometrical features of these areas of interest. They usually rely on subtle changes in each facial features and are very sensitive to noise.

Appearance-based approach, on the other hand, handle images in a holistic approach. These methods apply spatial image analysis to the whole image and not only some regions of interest. Deep neural networks (DNN) have been introduced recently and proved very useful in facial expression recognition task, especially the convolution neural network (Convolution Neural Network (CNN)). Studies have showed that CNN could distinguish several Action Units (AU) features [139]. Action Units represent different deformation of facial muscles, defining different facial expressions. FER tasks follow various steps such as follows :

1. Image normalisation and noise reduction
2. Facial area detection
3. Facial features extraction
4. Training a model
5. Facial expression identification

Some models architectures follow an end-to-end approach such as deep learning models, whereas other follow a pipeline approach such as methods using handcrafted feature extraction.

The following sections describe the most common state-of-the-art methods in facial expression recognition tasks.

2.5.1 Conventional FER Approaches

Extracting meaningful features from images in facial expression recognition represents a crucial step in the classification process. There exists various methods to extract meaningful features for facial expression recognition. Facial expressions changes are represented by subtle or minor deformation of different facial parts and muscles such as brows, eyes or nose [243]. There are two distinct methods for facial features extraction; that is, geometrical based and appearance-based. Most methods are either based on geometrical differences or facial muscle deformation in action units Action Units (AU), or appearance where various filters are applied to detect textures differences.

2.5.1.1 Geometrical features

Geometrical based extraction techniques consist of extracting significant features from geometrical information such as AU changes. They consist of selecting regions of interests, such as the eyes or the mouth and detecting changes in the muscle. Geometrical-based features are one of the most used features extraction techniques. Majumber et al. [177] have used geometrical and appearance features in addition to a deep learning model based features fusion for automatic facial expression recognition. The method consists of detecting different regions of the face (eyes, nose and mouth). The geometrical features extracted represent the ratio of horizontal and

vertical projection of each region in a sequence of frames. For each point, specific points are extracted, such as the left corner and centre of eyes or brows. In total, there are 22 features on each frame. The algorithm assumes that the first frame represents a neutral position of a face and the features extracted from the successive frames represent the differences between the reference (neutral) and the other frames. The algorithm also includes Local binary pattern (Local Binary Pattern (LBP)) features which are detailed in the next section. The extracted geometrical and appearance features are then fused using deep learning algorithm.

[87] have presented different techniques by using appearance model (AM) to track changes in facial landmarks. The authors have experimented on two methods; that is, AdaBoost and Support Vector Machines (SVM). The geometrical features are based on tracking changes in facial expressions. Geometrical features are also used in [68], by using facial landmarks for facial expression recognition. They use 18 facial points identifying regions of interested such as eyes and mouth areas. Then they calculated Euclidean distances between all pointed and used an Artificial Neural Network (ANN) classifier to infer facial expressions. Other techniques have been used, such as curvelet local features such as in [71], where salient keys are extracted from the face region.

2.5.1.2 Appearance Features

Appearance features are a set of image features based on the change of the texture [189]. LBP is one of the most popular methods introduced by Ojala et al. [205]. They have been first used in texture analysis and later applied in facial expression recognition. The basic algorithm works on assuming that image texture has two complementary aspects which are pattern and the strength.

Liu et al. [165] have used LBP in a combination of grey pixel values. Then (Principal Component Analysis (PCA)) is used for dimensionality reduction of the obtained features. The algorithm utilises active facial patches using region of interests (ROI) where significant changes occur in facial expressions. Ahonen et al. [4] also used LBP based algorithms, where each face is divided into several regions of interest where LBP was applied. The operation resulted in a histogram representation of the image. Histograms are concatenated and fed to a nearest neighbour classifier for recognition.

Another popular appearance-based method consists on applying histograms of oriented

gradients (Histograms of Oriented Gradients (HOG)). HOG descriptors are based on constructing a histogram feature vector by computing the accumulation of gradient direction over each pixel of each small region. It was first successfully used in human detection [57]. Carcagni et al. [211] conducted a comprehensive study on using HOG feature for facial expression recognition. The authors test various HOG parameters in terms of cell size and number of orientation bins.

2.5.2 Model Based Approach

Model-based facial expression recognition methods is the process of reconstructing a model of the face in order to track facial muscles deformation.

Tie et al. [279] propose a 3D deformable facial expression model with 26 fiducial points that are tracked through video frames using multiple particle filters. They then used a discriminative Isomap-based classification to classify the tracked facial deformation into a facial expression of emotion. Gilani et al. [90] used 3D face model to compute the correspondence between different constructed 3D models. The correspondence is achieved by morphing the model to new faces. Chen et al. [48] have used 3D facial point-clouds on a CNN model. Their method has proved accurate in term of speed of feature extraction and tolerance to pose changes. They have achieved an accuracy of 86.67% using BU-3DFE dataset.

2.5.3 Deep Learning Based Approaches

Recently and with the advance in research turned toward using deep learning for FER.

Kim et al. [141] have used deep learning for facial expression recognition in the wild. They use a descriminative convolutional neural network (Descriminative Convolutional Neural Network (DCNN)) to fuse aligned and non-aligned facial frame input. The DCNN is also used to align non-alignable faces in video input. Their proposed method achieved an accuracy of 73.73 % for the FER 2013 dataset.

CNN were also used in [167]. The authors used the CNN with specific data pre-processing approach in order to overcome small datasets training. The authors added eye localisation , rotation correction and intensity normalisation before feeding their training data to the CNN network. They have achieved an accuracy of 96.76% using CK+ dataset. [192] also used deep

leaning for facial expression recognition. The author propose a novel architecture for a CNN with two convolution layers where each is followed by a max pooling and four Inception layers. Using Inception layers gives more depths and width to the network without affecting the computational cost. Their model acheived an accuracy of 92.3% on CK+ and 66.4% on FER2013 datasets.

Zeng et al. [308] presents a framework for facial expression based on facial geometric and appearance features. The have used these features along deep sparse auto-encoder (DSAE) for classification they achieved an accuracy of 95.79% on the extended Cohen–Kanade (CK+).

2.6 Speech Emotion Recognition

Similar to facial expression recognition, SER represents an essential aspect in understanding human emotions as it contains salient features. In addition to linguistic content, non-verbal components play an important role in emotion comprehension. These features have been successfully used in various research for developing SER systems. Prosodic features, for example, represent the intonation or music of the voice and can be represented by energy and pitch [251]. Other features are represented by phonetic features which can give a better insight on emotions.

There are various steps to follow in order to achieve emotion recognition through non-verbal speech. The most crucial step in SER tasks consists of extracting and learning features translating differences of various emotional states in speech. Audio features can represent both speech and non-speech. Classification methods are then applied in order to get the classification of emotion.

This section describes state-of-the-art methods in audio features extraction in SER, then details some of the most used method for SER classification techniques.

2.6.1 Features in Speech Emotion Recognition

Feature selection and extraction is the first and most crucial step in SER tasks. It is a challenging task as any classification depends on it. Humans can distinguish automatically between different vocal features, either linguistic or para-linguistic, and can distinguish between different features. Voice features comprise various types, including excitation source, vocal tract, continuous speech, global and local. The primary step preceding features extraction is to segment speech signal

into temporal windows. In the following, we list some of the most popular features and their extraction techniques used in SER.

- **Mel Frequency Cepstral Coefficients (Mel Frequency Cepstral Coefficients (MFCCs))**: MFCCs are the most biologically plausible method and mimics how human process sound [96]. They are one of the most common feature extraction method for vocal tract features, along with LPCCs (Linear pre-sceptral coefficients). They are derived from the cepstral domain representing the vocal tract. They are also referred to system or segments features. They are based on vocal tract shapes for each temporal segment. They are used in various applications of speech recognition, especially emotion recognition. They are computed as a linear cosine transform of log power spectrum representing short-term power of signals.

- **Spectral centroid**: Spectral centroid represents the centre mass of the spectrum magnitude indicating quick changes in the audio signal [277]. They are computed with the centre mass of the magnitude of spectrum. They have been successfully used along with convolution neural network [54].

- **Pitch**: Pitch represents the nature of a tone, either being high or low. It consists of the quality of the audio signal computed by the vibration rate [214].

- **Energy**: Energy denotes the presence of a signal at a given temporal interval. Energy of an audio signal is calculated by measuring the occurrence of an audio signal in a small time window interval. [45] used energy along with volume, MFCCs, Zero Crossing Rate, Formants and Spectral Centroid as predictors for emotion classification through speech. Energy is usually calculated from small-time intervals and consists of finding the presence of a signal through a temporal interval.

2.6.2 Conventional SER Approaches

Most SER models use MFCCs features to extract the best features for this particular tasks [96]. Another popular feature extraction for speech emotion recognition is spectral power, which mainly represents the brightness of an audio signal [277]. Other forms of features are either used

individually or in combination with the above features such as the energy or the pitch of the signal.

Research mainly focus on two distinct areas in SER; that is, dynamic processing or static processing. Dynamic processing or frame-based processing partitions auditory signals into frames and focuses on learning temporal relationships between frames in emotion recognition [11]. Static processing, on the other hand, aims at the recognition of emotion through the whole utterance of the emotion through the audio signal features [73]. For both aspects, the essential steps reside in extracting meaningful features, which can translate the right emotion. Traditional SER model use handcrafted features sited above and rely on classical machine learning algorithm such as SVM in [46] [306]. Other classifiers are widely used such as Hidden-Markov Model (Hidden-Markov Model (HMM)), Gaussian Mixture Models (Gaussian Mixture Models (GMM)) and more recently deep neural network. For example, Yang et al. have fed the above features to SVM to recognise different emotional states [306]. Anagnostopoulos et al. [11]. have used HMM in dynamic learning.

2.6.3 Deep Learning

Deep learning methods for SER have produced more accurate results compared to classical methods [154]. The proposed deep learning on SER tasks apply deep learning architectures to hand-crafted features such as MFCCs, or Spectrogram power. There have been few recent work on applying deep learning to raw audio signal for SER tasks. Hand-crafted feature are considered to represent the audio signal with a global level acoustic feature, where once extracted, they tend to lose the dynamic relation in the temporal dimension. Most of deep learning work using hand-crafted features take the feature input as a whole regardless of dynamic relationships within time.

Niu et al. [201] and Satt et al. [245] both use spectrogram features as an input for a deep learning network. Nio et al. [201] have proposed the application of Deep Retinal Convolution Neural Network (DRCNN). This novel approach consists of two parts. The first is the data augmentation step where the principal of retina and convex lens imaging is used on the set of spectrogram features for each input. The second step involves applying a Deep convolution Neural network on the spectrogram feature to extract high-level features. They obtained an overall

accuracy of 48.8% on IEMOCAP dataset. Satt et al. [245] have investigated two types of network on spectrogram features for SER tasks. They first trained a CNN on the extracted spectrogram data, with experimentation on different network topologies. They then train a CNN in addition to a Long Short Term Memory (LSTM) layer. Adding a LSTM layer proved beneficial to the overall accuracy where it reaches 68% compared to 62% with CNN layers alone on IEMOCAP dataset. Other deep learning techniques have been investigated by Lee et al. [158] where they used a Recurrent Neural Networks (RNN) to draw feature representation of audio signals. Their model acheived an accuracy of 63.89% on IEMOCAP dataset.

2.6.4 Bio-inspired Approaches

Bio-inspired approaches are under-explored in the literature for SER tasks. An early attempt has been made in [36] where they use a spiking neural network for speech emotion recognition. The primary approach consists of applying Spiking Neural Networks on raw speech data. The SER task is applied on the linguistic part of the speech by decomposing each sentence into different parts for each vowel occurrence. For each part, MFCCs feature are extracted. The features are encoded into spike trains using average rate coding. The network is trained using reinforcement learning algorithm. Another biologically-inspired approach is investigated in [168]. The author used raw speech signal as an input and used Liquid State Machines (Liquid State Machines (LSM)) for classification. LSMs are a type of reservoir computing [80] which reservoir represents a Spiking Neural Network. The speech input goes through several pre-processing techniques, where linear prediction analysis is applied to audio signal. The overall classification tasks has an accuracy of 82.35%, which is comparable to state-of-the-art for the same datasets.

2.7 Multisensory Emotion Recognition Fusion Techniques

Humans express their feelings and emotions through various social signals such as facial expressions, body gestures and both verbal and non-verbal speech signals. The integration of these different signals makes up distinct emotions precepts and enables us to communicate effectively.

With the developments of various computing areas such as in computer vision, affective computing or human-computer interaction and having more accurate sensor technologies, it

becomes very crucial to have adequate systems that can draw human emotional states through multisensory integration approaches. The development of accurate real-time systems that can derive emotional states from different signals can play an essential role in enhancing many areas in human-computer interaction. It can have substantial positive effects on applications such as assistive technologies or behavioural analytics. Various attempts have been made, where research focus mainly on the development and enhancement of individual modalities recognition techniques rather than focusing on enhancing multisensory integration and fusion techniques.

Social signals from different modalities only make sense when integrated in a specific way. Multisensory social signals of emotions integration define the way we integrate information from different senses and create a new precept from it. This new precept constitutes emotional state interpretation [218]. Creating a new precept or concrete interpretation of information from different modalities can be very challenging, as sensory information can come at different timing onsets and need to be integrated in noisy environments. Systems need to discern, extract and process the right inputs in order to have a meaningful output and interpretation of emotions.

Multisensory integration for emotions recognition is essential in fields such as affective computing or assistive technologies [223]. Multisensory emotion recognition consists of evaluating emotional states from various modalities such as facial expression, body gesture, verbal and non-verbal speech. Integrating information from different sensory modalities is very crucial as they provide additional information on subtle changes in unisensory modalities that can go undetected in unisensory systems, such as facial expression systems only. Current research trends aim at exploiting information from various modalities and focus on two integration techniques: early and late fusion. Late or score-level fusion techniques is one of the most popular techniques which works by fusing scores from individual modalities. Early feature level fusion, on the other hand, consists of concatenating inputs at an early level and proceeding the final scoring using the obtained features. Current state-of-the-art techniques in multisensory integration are described in the following sections.

2.7.1 Early 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 vectors. The high dimensionality is alleviated by using dimensionality reduction techniques. Feature level fusion remains the most adopted technique for data fusion in multisensory emotion recognition.

Kessous et al. [137] have presented a framework for multisensory emotion recognition from body gesture, facial expression and auditory speech information. The proposed method starts by extracting features from facial, speech and body gesture input data. Combining three modalities produces high accuracy of 78.3% compared to unisensory or bimodal emotion recognition with 48.3% and 62.5% for speech and facial modalities.

Schuller et al. [252] have concatenated audio and visual feature into one vector and then have used Support Vector Regression (SVR) for classification as a baseline for AVEC 2012 emotion recognition. The study uses four dimension measures for emotion recognition; that is, arousal, valence, expectation and power.

Lingenfelser et al. have combined features extracted from audiovisual data to LSTM for continuous emotion recognition [162]. They use short timed events through a vector of space. Chao et al. [44] also opted for early fusion; they first use LSTM-RNN for temporal feature extraction on both audio and video. Then they concatenate the features vectors and feed them to SVM for final emotion recognition. Liu et al. [164] have used deep learning approaches for multimodal feature extraction in physiological data. They implemented a Restricted Boltzmann Machine (Restricted Boltzmann Machine (RBM)) to extract features from EEG and eye movement data. They then obtained intermediate features in the hidden layers. These are concatenated and fed to a supervised SVM classifier.

Zhang et al. [312] have used CNN and 3D-CNN to extract meaningful features from audio and visual modalities. Then features are concatenated using a Deep Belief Network (Deep Belief Network (DBN)). Finally, a linear SVM is used for final classification of emotions. Ma et al. [175] also employed a similar approach. They used a 3D-CNN for visual feature extraction and a 2D-CNN for audio feature extraction. They then use a DBN for features concatenation and a SVM for multisensory emotion classification.

Early fusion techniques are more useful when data from different modalities is completely synchronised; that is, with no temporal overlap or delay. This is particularly difficult for audiovi-

sual data, as frequently visual information is perceived earlier [196]. Another limitation of early fusion is that it requires heavy pre-processing of different features due to the heterogeneity of data. The difference between features from modalities is ignored, and it is very challenging and difficult to learn any relation or relevance between modalities [224].

2.7.2 Late Fusion

Late fusion also referred to as decision level fusion is a commonly used technique in multimodal emotion recognition, as it answers some of the early fusion challenges by emphasising the uniqueness and individuality of each modality. In this fusion method, each modality is classified separately. Then a rule is chosen for combining the classification results from each modality. Considering that fusion is achieved using classification results of each modality, this fusion technique is more advantageous compared to early fusion as the data fused have the same dimension and format.

One of the most used decision level fusion techniques is Kalman filter as applied by Glodek et al. [91]; that is, video is considered as a time series problem, and scores from individual classifier are fused. The algorithm is mainly based on Markov model, with the primary goal to reduce noise by taking several measurements and each step's estimation into account. Glodek et al. have used Kalman filters to track estimations for each classifier.

In decision level fusion, a local decision is derived from each individual modality classifier. Then all decisions are combined to form a final score using various voting or classification techniques. This method has been applied in most multisensory social signals and emotion recognition [75]. For example, Felipe et al. [75] have proposed a real-time multi-modal system based on decision level classifiers. The primary system consists of two parallel models for facial and speech recognition respectively. The outcome of the two subsystems is then integrated using a Dynamic Bayesian Network.

Schels et al. have created a classifier that fuses decisions inferred from video and physiological EEG data [248]. Firstly a classifier for each module is created to classify different features. Then a final classifier is built using different weights according to the accuracy on individual modules. The authors have applied more weights to the audio and physiological data, as they have shown more accuracy individually. They have also tested a combination of different classifiers, and the

overall accuracy is around 60%.

Sun et al. [269] also have used late fusion by adopting 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 probabilities value of each class in each obtained feature. Values belonging to the same class are added. They have used values with greater probability for classification in each class. Huang et al. [113] used sum and production rule to combine classification results from visual modality with facial expression and EEG.

Other studies have also opted for the late fusion techniques. In [202] the authors 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 which is averaging the confidence level for each classifier. They have also experimented deep fusion where they have used the score for each classifier as an input for a deep network classifier. This fusion technique has also been used in [67] where the authors have developed a novel approach based on kernel extreme learning (ELM) for classification of multi-modal physiological and audio visual data. The main characteristic for the kernel ELM consists of one hidden layer feed-forward network, where the hidden layer doesn't 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 represent a promising method in multisensory data fusion. However, their main challenge lies the lack of connection between modalities. In fact, in decision level fusion, complete independence is assumed between modalities [163]. It can result in losing crucial information about the inter-dependability and interactions of modality such as audio and visual in emotion recognition. For example in emotion recognition, auditory information is predicted by preceding visual information, where for a example a deformation of mouth can predict the type of verbal sound produced.

2.7.3 Hybrid Fusion

Hybrid fusion consists of combining both feature and decision level fusion. A hybrid approach has been designed for multimodal emotion recognition for E-learning environment [17]. First

features are extracted from each modality and a decision level fusion is applied. Then feature fusion technique is applied to combine all features from multimodal dataset.

Wolmer et al. [297] 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) is used to fuse data at feature level. SVM is use to classify text data. Results from BLSTM and SVM are used in decision level fusion for estimating sentiments.

More recently, Amer et al. [9] have proposed a novel hybrid fusion approach for multimedia data fusion. They first apply a Discriminative Continuous Restrictive Botlzmann Machine (DCRBM) to account for the temporal dimension for each modality. Then a Multimedia DCRBM is applied for the fusion of multiple DCRBMs combining multiple modalities.

2.7.4 Deep Learning Based Fusion

More recently deep learning techniques have been applied to fusion tasks, not only in feature extraction but also for multisensory learning.

Zhang et al. [311] use CNN for multimodal emotion recognition. First, they use two CNNs to extract features from visual and auditory modalities. They then integrate the obtained feature in a fusion network to obtain a multimodal features representation.

Poria et al. [225] introduce a Convolution Neural Network (CNN) for sentiment and emotion prediction in visual, audio and text data feature extraction. They use features extracted from all modalities and input them in a Multiple Kernel Learning (MKL) classifier. Nguyen et al. [199] have proposed a novel approach using 3D convolutional neural network (C3D) to model spatio-temporal video information, along with DBNs representing audio and video streams. Bhandar et al. [26] employ a modified stacked autoencoders in addition to a multilayer perceptron-based regression model. Ortega et al. [208] propose a novel DNN architecture by integrating three modalities: audio, visual and text. First, the network extracts individual modalities' features from hidden layers. Then extracted features are merged in a merging layer, followed by a fully connected layer and a regression layer. The network is trained in an end to end fashion.

Table 2.1: Summary of State-of-the-art multisensory integration techniques

Fusion Techniques	Model	Dataset	Modality
Early	Bimodal Deep Autoencoder [164]	SEED, DEAP	EEG and Eye Signals
	Chao et al. [44]	Emotiw 2015	Audio and Visual
	Kessous et al. [137]	Own database	Audio and Visual
	Schuller et al. [252]	AVEC	Audio and Visual
	Linhenfelser et al. [162]	Belfast	Audio and Visual
Late	Goltek et al. [91]	AVEC	Audio and Visual
	Felipe et al. [75]	SAVEE	Audio and Visual
	Schels et al. [248]	AVEC	Audio, Visual
	Sun et al' [269]	AFEW	Audio, Visual
	DNN Nojavanasghari et al. [202]	POM	Audio, Visual and Text
	Duan et al. [67]	EEG	Audio, Visual and EEG
	Huang et al. [113]	EEG capture	Facial and EEG
Hybrid	Bahreini et al. [17]	SEMAINE dataset for training and a dataset with 12 participants for evaluation of the software	Audio and Visual
	Olmer et al. [297]	Youtube	Audio, Visual, Linguistic
	Amer et al. [9]	AVEC	Audio and Visual
Deep Learning	CNN Zhang et al. [311]	RML	Audio and Visual
	CNN Poria et al. [225]	IEMOCAP	Audio, Visual, and Text
	DBN Nguyen et al. [199]	eNTERFACE'05	Audio and Visual
	autoencoder Bhandar et al. [26]	RECOLA	Audio and Visual
	DNN Ortega et al. [208]	RECOLA	Audio, Visual and Text
	DBN Zhang et al. [312]	eNTERFACE'05	Audio and Video
	DBN Ma et al. [175]	eNTERFACE'05, RML	Audio and Visual

2.8 Challenges in Multisensory Emotion Recognition

Table 2.1 summarises the multisensory integration models described in the previous sections. Most of the presented work describe fusion techniques for audio and visual data with categorical emotions. Work presented in the literature displays varying accuracy results depending on multiple factors such as datasets quality, feature extraction techniques, accuracy metrics, experimental setup and mainly fusion techniques. For studies using categorical emotions with six basic emotions and overall accuracy as metrics, the best performing models are the ones using deep learning either as features extractors or fusion technique. Early research focused more on the use of early fusion and later on late fusion. Recently, the most popular researched topic is using deep learning models [223].

Although most cited techniques have promising overall accuracy results, they face various challenges when it comes to social signals of emotions data. Multisensory integration of emotions

have brought various challenges due to their complex types of interaction between modalities. State-of-the-art methods on multisensory social signals and emotion recognition rely only on applying conventional fusion techniques.

However, there is a significant difference in the multisensory processing of emotion data compared to any other type of data in the human brain as explained in section 2.2. In this sense, social signal of emotions integration should be different compared to general data integration. One can consider social signals integration more as a constant communication and interaction between sensory modalities rather than a simple fusion of features or classification results [55].

Current studies on social signal and emotions data fusion and recognition focus on the integration of audio and visual data. This includes many challenges, as identified in [131]. Katsaggelos et al. [131] have identified some challenges for audiovisual data fusion in general that are all applicable for general sensory data integration beyond social signal integration, which are summarised as below [19]:

- *Reliability*: the reliability of each modality varies in a noisy environment. Some modalities can be more reliable than others, where, for example, in a noisy auditory environment, visual input is more important.

- *Inconsistency*: each modality might produce conflicting or inconsistent data. For example, each data from each modality derives conflicting emotional states. There is a need to resolve their uncertainty and derive a commonly agreed conclusion.

- *Interaction*: there exists cross-modal interaction and prediction between different modalities. This occurs when signals from one modality can be used to predict signals in another. For example, visual signals can be used to predict an auditory sound. Current fusion techniques do not focus on the interaction between different modalities.

- *Integration timing*: integration time is particularly important for continuous and real-time recognition. The integration of data from different modalities needs to occur at a particular time after the onset of the sensory input. If the integration time is too long, data from different sources might not be able to be integrated.

- *Asynchrony*: time asynchrony where a visual input precedes auditory inputs is particularly not accounted for in feature level fusion. Hence time dimensionality is mostly ignored. This could result in errors or accuracy deficiency. Especially for social signals where time plays an important role in the evolution of emotions [12].

- *Real-time*: state of the art fusion techniques rarely account for real-time applications where a uni-modal classifier can fail, due to a defect of a sensor, for example. This can usually solved by relying on the classification and estimations of other non-failing classifiers [124].

Although existing techniques and state-of-the-art methods attempt to address some of the challenges, each method still have various drawbacks which can be summarised as follows :

- Early fusion techniques lack handling multisensory data when there is a difference of reliability between modalities

- The main challenge facing late or decision-based fusion is the discrepancies between results from different modalities

- Another main challenge in late fusion is a problem with time synchronisation, as results from different modalities can translate temporal delays.

Another problem faced by state-of-the-art multimodal fusion for emotion and affective computing is the dimensionality challenge. Affective data is highly dimensional, where extracted features from various modalities contain thousands of variables. Dimensionality issue can be alleviated using dimensionality reduction methods, which have been employed in order to reduce features space. Denis et al. [232] use principal component analysis along with linear discriminant analysis for dimensionality reduction.

Another major challenge from state of the art method, including deep learning methods is that these methods usually ignore the constant cross-talk and temporal relationship between different modalities, where modalities receive and send feedback to each other and information is integrated within a temporal window. The recent study in cognitive neuroscience on cross-modal modulation in emotion processing [82] has shown that cross-modal interaction is particularly crucial in emotion recognition, where signals from different modalities can complement each

other in learning and thus signals in one modality can used to predict the other. For example, dynamic facial expressions can influence vocal emotion processing.

This thesis aims at answering three main challenges of the existing multisensory fusion techniques in social signals of emotions recognition which are **interaction**, **reliability** and **asynchrony**. The work in this thesis addresses those challenges by using findings in neuroscience, to try to model the constant cross-modal interaction between modalities, more robustness to noise.

2.9 Summary

This chapter presents existing work on emotion recognition in general and multisensory integration for social signals of emotions in particular. It first starts by describing the nature of emotions in psychology. It describes the different emotion models as identified in psychology, such as categorical or dimensional emotions. It also provides an overview of multisensory integration and processing of emotions in the brain. Literature shows that emotions are mainly multisensory, where humans understand and process emotions principally in a multisensory way. Work also demonstrate that multisensory integration in the brain follow various paths, where various regions of the brain are involved in the processing, and integration including early sensory regions.

The chapter outlines the related work in unisensory emotion recognition focusing on audio and visual emotion recognition tasks. After that, it describes the current trends in multisensory emotion recognition and existing fusion techniques. Finally, this chapter presents the current challenges faced by current multisensory integration techniques and current limitation of these methods. Challenges includes many aspects such as synchrony, timing, reliability, consistency and more importantly translating the constant cross talk and interaction between modalities.

The next chapters describe novel biological architectures designed to address some of the most significant challenges faced by current fusion techniques in multisensory emotion recognition such as interaction, reliability, asynchrony and interaction.

Chapter 3

Introduction of Spiking Neural Networks

3.1 Introduction

Spiking Neural Network (SNN) represent the third generation of artificial neural networks. They are composed of spiking neurons inspired by biological neurons behaviours. They are different from artificial neural networks in various aspects, where they are more biologically plausible and follow different learning rules. SNNs represent the primary means for the implementation of the proposed biologically inspired models for both unisensory and multisensory emotion recognition in this thesis.

This chapter provides an introduction to SNNs. It first gives an overview of biological neural networks. Then, it describes how information is transmitted in the brain between various brain regions. It introduces the main components of SNNs and various steps in implementing them. These include:

1. Computing neural dynamics through various neuron models.
2. Overview of different SNN topologies and architectures.
3. Description of the existing learning algorithms for SNN.

3.2 Biological Neural Networks

Neurons are the primary component in the brain. They have different roles within brain functions, from cognition and memory to motor action. The human brain can contain around 100 billion neurons with billions of connections [169]. Neurons have a particular shape, compared to other cells in the body. Having a longer shape enables them to send signals very rapidly and with exceptional precision to other neurons.

Neurons are composed of three main parts: dendrites, axons and cell body (stoma) as detailed in Figure 3.1. Dendrites which resemble small branches help neurons to receive information and stimuli from linked neurons and transmit signals to other neurons' cell body. Cell body or **stoma** holds neurons' nucleus and behaves a non-linear processor that creates a spike if the input exceeds a certain threshold.

Neurons receive signals from dendrites and transmit information through axons. Many axons are covered with an insulating substance called **myelin**. **Myelin** is produced by **Scawnn cells**. There are gaps in the myelin refered to **node of Ranvier**. Their main function is to facilitate a quick conduction of nerve impluses.

Axon terminal represent the axon endings and are button shaped. They make synaptic connections with other nerve cells.

Communication between neurons happens at the synapse level, representing a junction or link between them. Pre-synaptic cell sends signals, and post-synaptic cell receives a signal. The brain is a very complex system, where signals are transmitted very rapidly. Connections between neurons change according to their spiking patterns. These modifications make up various cognitive patterns [114].

3.2.1 Information Transmission in the Brain

Information transmission in the brain goes through waves referred to as spikes. Spikes are high-speed electrical signals that propagate information from one neuron to another. Information is transmitted through junctions (synapses) when an axon terminal of pre-synaptic neurons is very close to the cell body of a post-synaptic neuron.

The transmission creates a small gap referred to as synaptic cleft, where a chemical neuro-

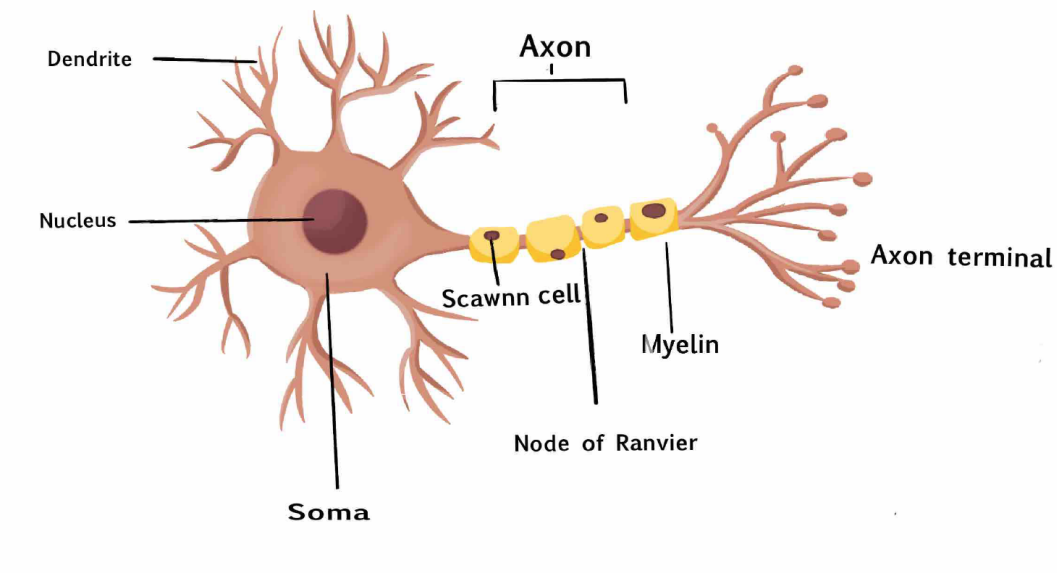


Figure 3.1: Components of neurons

transmitter is released following the action potential generated by the pre-synaptic neuron. This neurotransmitter is then detected by post-synaptic neurons, which permit an electrical current to be passed, as shown in Figure 3.2. The primary role of synapses is to transform electrical to chemical molecules.

The number of chemical neurotransmitters defines the strength of the electric current. Changes in chemical synapses govern the synaptic strength effectiveness. This operation sets synaptic plasticity. Synaptic plasticity describes changes of connection and modulate synaptic efficacy. It helps create neuronal learning and memory, which constitute synaptic plasticity [122].

3.3 Spiking Neural Networks

Spiking Neural Networks are the third generation of neural networks [176]. They aim to bridge the gap between artificial intelligence and neuroscience by using biologically inspired mathematical models to model neurons behaviours. The main difference between SNN and classical Artificial Neural Networks (ANN) remains in the way they process data based on spikes. Although classical ANNs get some inspiration from biological networks, they diverge in their implementations. ANNs use neurons with continuous variable outputs, and training usually happens using error backpropagation. Neurons in SNNs communicate through series of spikes by firing excitatory

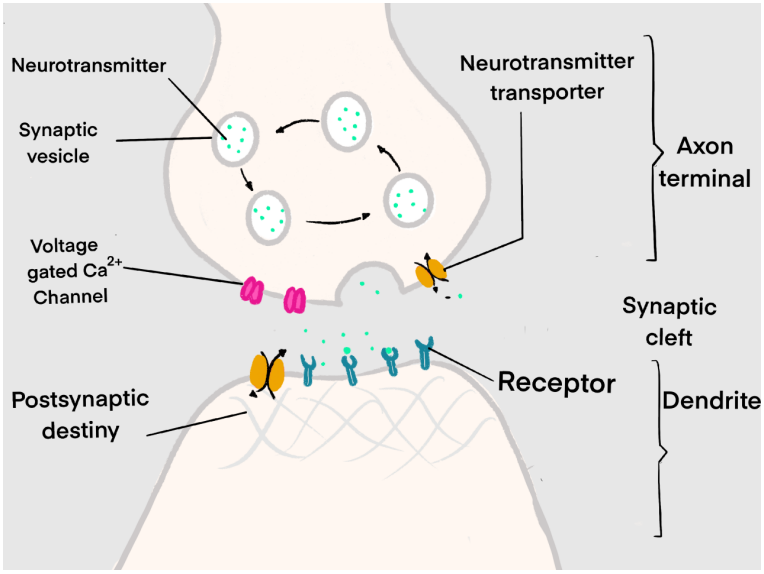


Figure 3.2: Transmission of information in the brain

inputs when they reach a certain threshold. Then, these spikes are decreased by inhibitory inputs [221]. Calculations are mainly achieved by differential equations representing various biological processes. The membrane potential of neurons represents the most critical computation aspects.

3.3.1 Neuron Models

Neurons models represent the computation of behaviours for each neuron in the brain. Neurons communicate by generating action potential representing electrical pulses [126], constituting the core of communication in spiking neural models. Spiking neural networks work by processing information from various inputs and produce one spiking output signal. They also operate by generating spikes which increase at excitatory inputs and decrease at inhibitory inputs. Spikes are fired when internal variables reach a certain threshold. Timing in spiking neural network plays a significant role in implementing neuron models [85]. This sequence of time-based firing information represent spike trains. Modelling spiking neurons is principal in two main aspects: computing the evolution of the membrane potential and setting a mechanism of spike generation.

3.3.1.1 Huxley-Hodgkin Model

Huxley and Hodgkin model is first presented in [110]. The model is based on modelling the electrochemical information transmission between neurons using electrical circuits containing

capacitors and **resistors**. The model can be translated in the following equations [215]:

$$C\frac{du}{dt} = g_{N_a}m^3h(u-E_{N_a}) - g_K n^4(u-E_K) - g_L(u-E_L) + I(t) \tag{3.1}$$

C represents the capacitance of a membrane, g_{N_a}, g_K, g_L represent the conductance parameters denoting the different ion channels for the neuron-transmitters for sodium and potassium and L represents leak conductance.

The equilibrium potentials for each ion channel are represented by E_K, E_L, E_{N_a} .

$$\tau_n\frac{dn}{dt} = -\Big[n-n_0(u))\Big], \tau_m\frac{dm}{dt} = -\Big[m-m_0(u))\Big], \tau_h\frac{dh}{dt} = -\Big[h-h_0(u))\Big] \tag{3.2}$$

m, h, and n are variables describe voltage-dependent channels opening and closing. τ_n, τ_m and τ_h represent membrane time constant for each voltage-dependent channel.

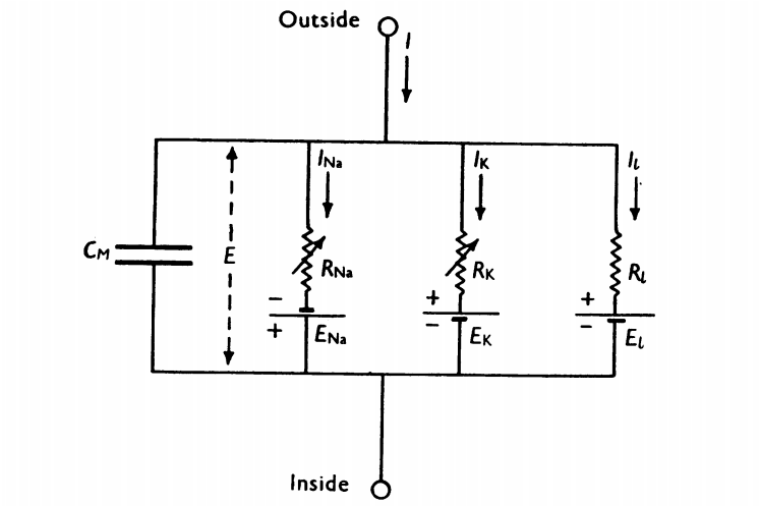


Figure 3.3: Electrical circuit representing membrane [110]

The model can be described by the circuit detailed in Figure 3.3

Hodgkin-Huxley model successfully models biologically realistic properties of membrane potentials, with realistic behaviours comparable to natural neurons. This is characterised by a sudden and large increase at firing time, which is followed by a refractory period where a neuron cannot spike again, followed by a time interval where the membrane is depolarised.

Although Hodgkin-Huxley model demonstrates to be very powerful to model neuronal behaviours realistically, its implementation is very complex for numerically solving the system

of differential equation using SNNs. Computing the temporal interaction in the HH model is also computationally very costly.

Large networks can be very difficult to model using the Hodgkin Huxley model [83], and it is not particularly suitable in more complex tasks such as pattern recognition.

3.3.1.2 Leaky Integrate and Fire Model

Integrate and Fire (IF) more specifically, Leaky Integrate And Fire (LIF) models are a simpler model compared to Hodgkin-Huxley neuron models. LIF models were introduced by Abbotts [2]. LIF models are a simplification of Hodgkin-Huxley models by considering every spike as a uniform event defined solely by the time of spiking. Besides, the shape of action potentials is neglected in LIF models. The general dynamics and evolution of membrane potential in LIF model neuron can be computed using a single first-order linear differential equation:

$$\tau_m \frac{du}{dt} = u_{rest} - u(t) + RI(t)) \tag{3.3}$$

where $\tau_m = R\,C$ represents a membrane time constant, C is the capacitor and R is the resistor. RI represents the circuit resistor. When a membrane reaches a threshold u_{thresh}, it is reset to a lower value than u_{rest}.

The membrane potential u is equivalent to the u_{rest} when in resting phase. When the current arrives to neurons, the capacitor is fed an electric current and is discharged through the resistor. The membrane potential returns to u_{rest} when the current stops by leaking through the resistance until it reaches the resting potential u_{rest}.

Figure 3.4 shows the LIF circuit. After receiving a stimulus, the membrane potential increased up to a threshold VT. After that, v(t) is reset to its rest value Vrest. At this point a spike is generated.

Compared to Hodgkin-Huxley models, LIF models are less biologically plausible. However, LIF is less computationally costly. LIF model is suitable when using SNN in machine learning and pattern recognition applications.

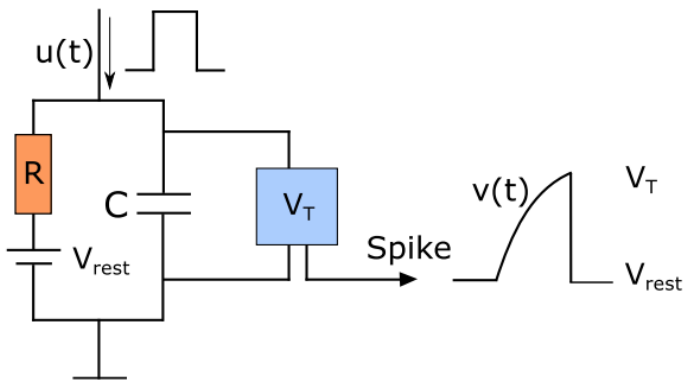


Figure 3.4: The leaky integrate-and-fire circuit. [198]

3.3.1.3 Izhikevich Model

The Izhikevich model is introduced in 2003 by Izhikevich [118] to alleviate some problems arising from the Hodgkin-Huxley model in terms of computational complexity and capability of LIF models. The simplification of Izhikevich model is achieved by reducing previous models to a two-dimensional system of ordinary differential equations. This model is particularly useful to simulate extensive brain models using real biological neurons. The model can be summarised through the following differential equations [118]:

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u - I(t) \tag{3.4}$$

$$\frac{du}{dt} = a(bv - u), \tag{3.5}$$

where v is the membrane potential and u represents the membrane recovery which negatively feeds back to v. The model uses auxiliary after spike resetting represented by :

$$\text{if } v \geq 30mV, \text{ then } \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases} \tag{3.6}$$

where a is the time scale of the variable u. b represents the sensitivity of u to the sub-threshold fluctuation of membrane potential v. c is a reset value of the membrane potential v. d describes the reset value after spikes of u. Neurons in various brain regions can exhibit different value

choices of the above parameters.

3.3.2 Synapse Models

Synapses represent the link by which neurons communicate and pass information as described in Section 3.2.1. Chemical reactions represent the primary means of information transmission, making pre-synaptic and post-synaptic neurons electrically coupled. Neurons model can also be applied for synaptic spikes transmission. There exist two types of synapses: conductance and current based models. When signals pass through synapses they provoke the following reactions at the post-synaptic level:

- The flow of post-synaptic current (PSC)
- Opening of neuron membrane on nearby ion channels.
- Changes in voltage of membrane by either increasing or decreasing. This is referred to as post-synaptic potential (PSP)

In synapses, the direction of the post-synaptic flow current and voltages depends on the nature of neurons. In excitatory neurons, PSC depolarises the membrane and triggers the excitatory post-synaptic potential (EPSP). For inhibitory neurons, PSC flow results in a membrane hyper-polarisation and an inhibitory post-synaptic potential (IPSP). It is directly linked to the strength of a synapse.

3.3.3 Architectures of Spiking Neural Networks

Similarly to classical Artificial Neural Networks (ANN), SNNs are designed using different topologies described as follows:

- Feedforward: In this topology, information flows in one direction with no feedback connection. These kind of topology are usually used in SNN to model low-level sensory systems, such as vision systems. They have also been used for binding tasks such as Spatio-temporal spikes or spike synchronisation [273]. They have been used as topology in [260], and [276].

- Recurrent: Neurons interact through feedback connections, where a dynamic temporal behaviour represents the network. Although this topology is harder to compute, it can have higher computational power. Recurrent architectures are particularly useful for modelling or analysing dynamic objects. However, it is computationally more challenging to apply supervised learning on this type of architecture [62]. Recurrent architectures can also be applied to investigate extensive population activities and analysing neuronal populations dynamics.

Feedforward topology is the most common topology for general pattern recognition as it mimics the hierarchical structure of visual cortex [5]. This topology represents the right candidate for tasks such as emotion recognition.

3.3.4 Learning in Spiking Neural Networks

The primary mechanism for memory and cognitive function in the brain is primarily governed by synaptic plasticity as described in Section 3.2. Computing synaptic plasticity can take various forms, where the only difference resides in the time scale. Some models rely on pulse paired facilitation while others use decay or even long term potentiation or depression.

3.3.4.1 Unsupervised Learning

Unsupervised learning in SNNs follows Hebb's law. Hebb was the first to introduce the theory on synapses modifying their weights to process and store data [103]. Hebb's formula stipulates that the change in weights affects synaptic coupling. Coupling between synapses strengthens whenever neurons fire together. The original Hebb's formula did not include the synaptic depression, which is later added along with potentiation by Stent et al. [266]. Automatic reorganisation of connection in the Hebbian learning permits the ability of unsupervised learning with various potential applications, such as clustering or pattern recognition. Unsupervised learning with Hebbian formula enables learning of distinct pattern without using classes labels or having a specific learning goal [109], [29], and [95].

Following neurophysiological studies, Markram et al. [183] have demonstrated that the timing of spikes closely influence the plasticity computed through Hebbian learning. Besides,

more experiments indicate that the order of pre-synaptic and post-synaptic spike creates different Hebbian processes. The order is particularly important as it either induces depression or potentiation. This phenomenon is described as Spike Timing Synaptic Plasticity (STDP) [39]. When pre-synaptic spikes precede post-synaptic spikes, it creates a potentiation which mathematically represents an increase in weights. Otherwise, if post-synaptic spikes precede pre-synaptic, this induces a depression or a decrease in weights. STDP process is a modified version of Hebbian learning [85]. STDP learning can be translated though the following general equation:

$$\frac{d}{dt}w_{ji}(t) = a_0 + a_1 S_i(t) + a_2 S_j(t) + a_3 S_i(t)\bar{S}_j(t) + a_4 \bar{S}_i(t) S_j(t) \tag{3.7}$$

$w_{ji}(t)$ represents the strength of the synaptic coupling from neuron i to neuron j. Pre-synaptic spike trains are represented by $S_i(t)$ and post-synaptic spike trains are represented by $S_j(t)$. Spike trains are represented by Dirac impulses for each firing time:

$$S(t) = \sum_f \delta(t - t^f), \tag{3.8}$$

where t^f represents firing times for spike trains.

$\bar{S}_i(t)$ and $\bar{S}_j(t)$ represent the low pass filter of pre-synaptic $S_i(t)$ and post-synaptic $S_j(t)$ respectively. The level of change in the synaptic efficacy (weights) are governed by the constant values a_0 ... a_4. The constants are related to weights and vary from 0 to $(w_{max} - w)$

One of the most popular STDP variant is the online STDP. The value for the pre-synatptic trace is increased by 1 every time a spike reaches a synapse. It then decays exponentially in any other cases. The weight modifications are computed using the synaptic trace using the following equation:

$$\Delta w = \eta(x_{pre} - x_{tar})(w_{max} - w)^{\mu} \tag{3.9}$$

where Δ w represents the weight change, η represents the learning rate, μ represents the weight dependence of the previous weight. , x_{tar} is the target value of the pre-synaptic trace and (w_{max} is the maximum weight.

3.3.4.2 Supervised Learning

The implementation of supervised learning in a Hebbian way for biologically inspired models is achieved by adding a supervision signal to reinforce firing at target times. Studies of motor learning and control have confirmed supervised learning in the brain, primarily in the cerebellum and the cerebellar cortex [147]. In the motor, cortex uses supervised learning to learning specific representations of the body and its environment [298], [66]. Some cognitive tasks may also be processed through supervised learning in the brain, such as language acquisition [151].

Supervised learning in SNNs is achieved through applying Hebbian learning. The supervision is done by spike-based Hebbian process by reinforcing the post-synaptic neuron in order to fire at preset timing and not spiking at other times. The reinforcement signal is transmitted through synaptic currents.

3.3.4.3 Reinforcement Learning

This kind of learning enables learning directly from the environment where SNN includes a rewarding signal spike. Reinforcement learning is directly inspired by how animals learn new instructions following constant feedback and reactions. Actions are mainly reinforced by positive reward, whereas undesired action receives by negative feedback. Reinforcement learning has been successfully applied, especially in machine learning [30]. Reinforcement learning in biological neural models have only been investigated recently [210]. Several models of reinforcement learning in SNN have been developed in the literature, such as the work from [72]. Farries and Fairhall [72] have combined STDP learning with reinforcement learning. Most proposed reinforcement learning models follow the following general equation [159]:

$$\frac{d}{dt}w_{ji}(t) = c_{ji}(t)d(t), \tag{3.10}$$

where w_{ji} represents the weight of a synapse from neuron i to neuron j. c_{ji} represents an eligibility trace collecting weights changes from the general STDP learning process. $d(t)$ is defined by:

$$d(t) = h(t) - h_0 \tag{3.11}$$

$h(t)$ is the neuro-modulatory signal and its mean value is represented by h_0. Reinforcement learning in SNNs can be both applied in feedforward or recurrent architectures [159]. There have also been applications in applying reinforcement learning in SNNs mainly in robotics [261] [27].

3.4 SNN Simulators

SNNs are implemented differently from traditional Neural Networks due to their nature. They rely on spike timing rather than rate; therefore, their implementation or simulation needs to take into account the precise timing of each spike firing. Thus, SNNs can be described as several timed spikes. Parallel computing can be beneficial for SNN implementation, due to spiking neurons not needing to receive weight values from each pre-synaptic neuron at each computation step [215].

There have been various attempts to create full simulations for SNNs both in computational neuroscience models and machine learning. Some first attempts such as NEURON [107] and GENESIS [31] have aimed at simulating biophysical models of individual neurons rather than whole networks of neurons. Hines et al. [107] have introduced a novelty in NEURON by implementing event-driven mechanism and parallel computing ability in [108]. This addition enabled more applications and computations using NEURON simulator. Verstraeten et al. [291] are the first to introduce a toolbox for reservoir computing. They create a toolbox for the implementation of three reservoir implementations and simulation: Backpropagation Decorrelation (BPDC) learning rule, Echo State Networks (ESNs) and Liquid State Machines (LSMs).

Mouraud et al. [194] applied parallel computing for SNN simulation with a parallel event-driven simulator. They present Distributed And Multi-threaded Neural Event-Driven DAMNED, a simulator that runs efficiently and uses multi-threaded programming for communication optimisation. They use an event-driven architecture. They implemented the simulator on a cluster computer.

Goodman et al. [92] have introduced BRIAN – a clock-based SNN simulator with the ability to have event-driven computation and was the first to create a simulator using Python. BRIAN enables the implementation of various neuron models by implementing differential equation in ordinary mathematical notations. They use vectorisation, which enables efficient simulations.

BRIAN is particularly useful for simulating bio-inspired architectures and includes easy-to-use syntax. BRIAN has applications in computational neuroscience and simulating machine learning algorithms.

More recently, Hazan et al. [102] have introduced BINDsNet a platform for applying SNNs to machine learning tasks. BINDsNet is implemented in Python with a wide range of functionality focused on machine learning and reinforcement learning. It is built using PyTorch library for deep learning. The implementation enables the implementation of SNNs in both CPU and GPU environment. Although this simulator is very promising, it fails to implement the clock-driven calculations in SNNs.

This thesis identifies BRIAN as the best simulator suitable for the implementation of bio-inspired models for multisensory emotion recognition. BRIAN simulation is more biologically plausible, and the use of clock-driven computing makes it more suitable for the models proposed in this thesis.

3.5 Applications of SNN

SNNs have been successfully used to simulate the brain processes for different tasks, including pattern recognition and image processing. Wu et al. have used an SNN architecture based on Integrate-and-Fire IF neurons [301]. The design mimics the visual cortex, which consists of various receptive fields for colour and orientation. Weights matrices are used for filtering different patterns or colours.

Iakymchuk et al. [115] have used an SNN for pattern classification where the STDP algorithm is applied. STDP represents a biologically inspired learning method in an unsupervised manner. The main spiking neural network used in this work is mainly created for feature extraction with applications in classification for embedded systems configuration. Experiments have been conducted on a handwriting digit dataset.

Spiking neural networks have also successfully been used in modelling spatio-temporal data such as EEG [129]. Kasabov et al. have created a framework using a SNN, called NeuroCube for modelling spatio-temporal interaction of various brain imaging techniques. Fu et al. have created a feedforward SNN architecture to recognise faces in a supervised learning manner [79].

They employ a hierarchical architecture to simulate the visual cortex using SNN with supervised learning for facial expression recognition.

More recently, there have been a surge in broader applications. Piotr et al. [207] applied SNNs for classifying Fashion-MNIST images. They have used BindsNET library to implement and SNN with LIF and Bernoulli distribution for input encoding. Al Zoubi et al. [6] have proposed the use of SNNs to build a highly adaptive supervised learning based on divide and conquer rule and hierarchical abstraction. Wu et al. [300] have proposed a model for automatic sound classification. They proposed a framework using Self Organising Map (SOM) implemented through SNN. It starts by extracting features from sound and feature representation learning. They used temporal SNN classification in a bio-inspired fashion.

Rathi et al. [229] have presented a cross-modal framework using SNN for the classification of digits. They create two unimodal ensemble using SNN and created connections between both ensembles. They have demonstrated that adding connections between the ensembles increased accuracy with 98% compared to 93.20% and 96% for visual and auditory unimodal, respectively. Although this work proposes a novel approach on multimodal pattern recognition, they fail to define a biologically plausible integration. The authors trained the unimodal ensembles first for a particular iteration amount than connected the two modalities. This method is not in line with multisensory integration pathway in the brain, where cross-modal interaction and learning happens from the onset of stimuli.

3.6 Summary

This chapter has introduced the structure and basis of information transmission in the brain. It also has detailed mathematical models translating the brain operations, which represent the essential components of SNNs. It has also introduced various learning models in the literature and described some successful applications of these networks. In addition it has presented various possible architectures for SNN implementation. This thesis focuses on applying feedforward architectures which represents the most biologically plausible one for visual, auditory and multisensory tasks modelling.

The next chapters 4 and 5 will describe the proposed models in this thesis for both unisensory

and multisensory emotion recognition. The proposed models are implemented using SNN as described in this chapter.

Chapter 4

Bio-inspired Unisensory Emotion Recognition

4.1 Introduction

The previous chapter has given an introduction and general background of spiking neural networks (SNNs), based on which this chapter will detail how to use SNN for unisensory emotion recognition. This also forms the foundation for multisensory integration models that will be described in Chapter 5.

As described in Chapter 3, there exists a variety of architectures, neuron models and learning techniques in SNN. The application of SNN for pattern recognition and classification often needs to identify the appropriate SNN architecture and topology for the task in hand. In unisensory emotion recognition, we will choose a hierarchical topology that has been successfully adopted for general pattern recognition in SNNs [65]. The topology follows the biological network topology in the auditory and visual cortex in the brain [290], and [182]. We use LIF neuron model, and apply STDP learning algorithms. The use of SNN will involve preparing input for SNN via feature extraction and input encoding, training a SNN, and perform classification.

After choosing the right architecture for the network, the implementation of the models goes through various steps, as shown in Figure 4.1.

1) Features extraction for each modality that is: audio and visual features;

2) Input encoding consisting of Spike-train generation for SNN;

3) Learning method for SNN;

4) Training phase;

5) Labelling phase.

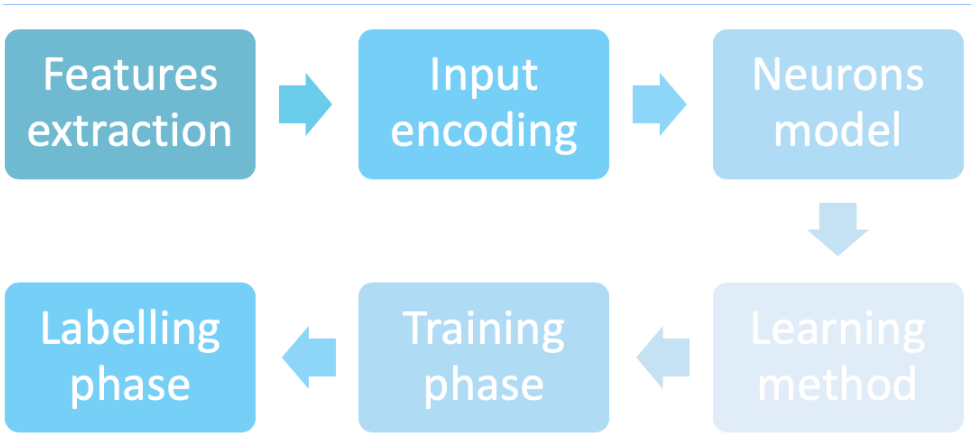


Figure 4.1: Process of applying SNN for unisensory emotion recognition

4.2 Architecture and Topology

This thesis selects a hierarchical model proposed in [65]. There are two reasons. First of all, this model has been successfully adopted for pattern recognition tasks. Secondly, the biological network topology in the auditory and visual cortex in the brain justifies the chosen hierarchical topology [290], and [182]. The architecture is composed of three layers: input, excitatory, and inhibitory layer. In the following, we will give a brief introduction of them.

1. **Input Layer**: The input layer is the first layer of SNN and its primary goal is to encode the input to be suitable for SNN network. This thesis focuses on two types of inputs: (1) visual input, consisting of facial expression images, and (2) audio input compromising raw acoustic signals. On these two types of input, feature extraction process will be run and the resulted features are encoded into Spike-trains to be fed to the next layers.

2. **Excitatory Layer**: The excitatory layer contains excitatory neurons groups and receives input in the form of spike-trains from the input layer. Excitatory neurons are excited by external stimuli such as audio or visual features. Excitatory layer neurons trigger a positive

increase in the membrane of post-synaptic neurons when they reach a threshold. In this layer weights of neurons are updated, and primary learning happens.

3. **Inhibitory Layer**: The inhibitory layer contains inhibitory neurons. These neurons trigger a negative change in the membrane of a post-synaptic connected neuron. The inhibitory layer contains neurons connected to the excitatory layer in a lateral fashion, where each neuron in the inhibitory layer is connected to all neurons in the excitatory layer, at the exception of the neurons it receives connection from them. The number of neurons in the excitatory layer are proportionate to the number of neurons in the input layer.

Figure 4.2 presents a general example of SNN topology and architecture for facial expression recognition task. The original architecture from [65] is modified to include convolution patches. Each input is divided into various convolution windows or patches. Each patch is connected to a number of neurons in the excitatory layer. Neurons in each patch of the input layer are connected in a all to all way to the corresponding excitatory neurons group. There is no connection between patches and each group of neurons receiving connections from each patch will learn a specific feature. The excitatory neurons are connected in a lateral fashion to the inhibitory layer. Each group of neuron in the inhibitory layer will inhibit all groups from the excitatory layer apart from the one it receives input from.

In this thesis, both unisensory and multisensory emotion recognition models are built using this basic SNN architecture with three main layers. In the following, we will introduce the setup of a SNN for unisensory emotion recognition on visual and audio input. More specifically, we will first introduce feature extraction (in Section 4.3), input encoding (in Section 4.4), neuron models (in Section 4.5), learning algorithms 4.6), training process (in Section 4.7), and prediction (in Section 4.8).

4.3 Feature Extraction

The first and most crucial step in applying SNNs to emotion recognition is to extract meaningful features depending on the type and nature of the input. This thesis focuses on audio and visual modalities consisting of non-verbal speech features and facial expression.

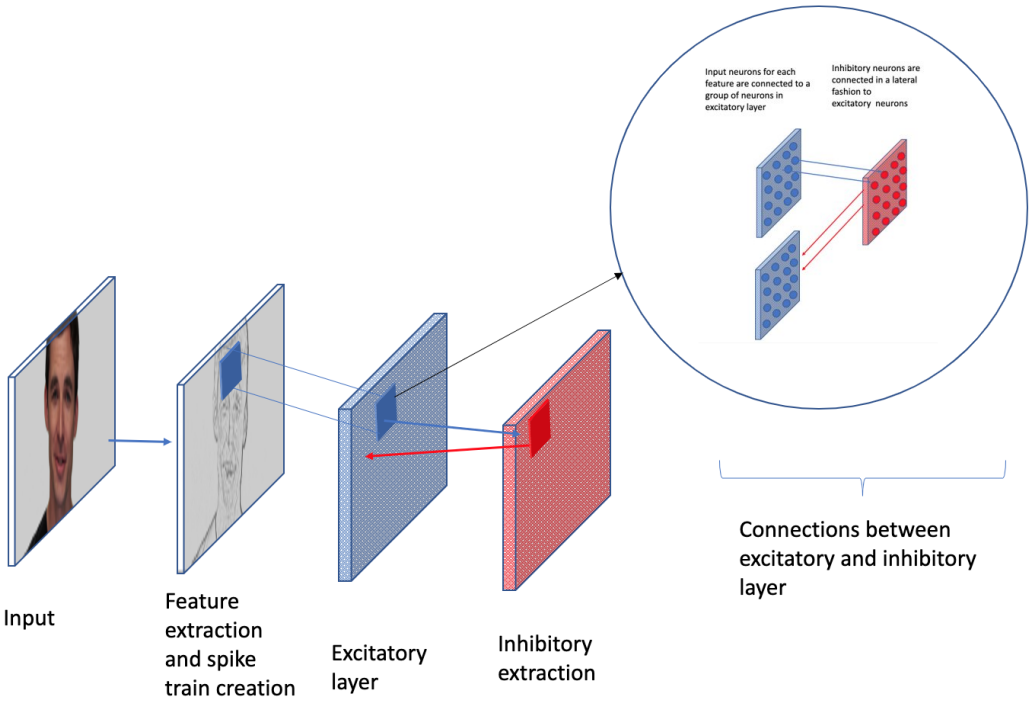


Figure 4.2: SNN for unisensory emotion recognition: Raw input goes through features extraction then input encoding. It is then fed to excitatory layer. The excitatory layer is connected to the inhibitory layer in a lateral fashion

4.3.1 Facial Features Extraction

Before proceeding to feature extraction, raw input goes through various pre-processing steps. First, all input images are resized to a uniform size and converted to greyscale. Then all input images are cropped to the face area. Feature extraction consists of defining essential facial features, distinguishing the main features that play a role in emotion recognition. Chapter 2 describes some of state-of-the-art methods in facial expression features extraction. This section describes feature extractions techniques chosen for the application of SNNs in facial expression recognition. The aim in the chosen features extraction technique is to distinguish between different facial features and prepare the input for SNN.

Feature extraction starts by defining contours of essential facial characteristics in input images. Spatial filters represent a beneficial way for contour detection in facial expression inputs. Filters such as Different of Gaussian (DoG) are successfully applied to pre-process input data and prepare it as input to SNNs. For example, DoG has been applied on pre-process handwriting images [138]. This thesis applies Laplacian of Gaussian (LoG) to extract contours and edges

of facial expression features on input images. Although LoG and DoG are quite similar, where DoG represents an approximation of the LoG. LoG is selected as it achieves higher precision [184] and is formulated in Equation 4.1.

$$\nabla^2 G_\sigma(x,y) = \frac{\partial G_\sigma(x,y)}{\partial x^2} + \frac{\partial G_\sigma(x,y)}{\partial y^2}, \tag{4.1}$$

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:

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.2}$$

Gaussian filters are first applied to remove noise and smooth the input image. Then Laplacian filters are then applied to detect, locate and extract all critical facial contours and features. Having well defined facial expression features enables input encoding and preparation for SNN.

4.3.2 Audio Feature Extraction

Audio signal processing goes through various steps such as analysis, feature extraction and prediction of behaviour. There exist various ways to extract features from audio signals in SER tasks. State-of-the-art methods in SER have used feature extraction techniques such as Mel-Scale Spectrogram, pitch, MFCCs or raw input [274]. They also have used a combination of approaches, such as prosody and spectral features. Rathi et al. [230] use raw audio data and generate Lyon's cochlear model. This thesis first explores several methods for features extraction in audio, such as raw features, Mel-scale Spectrogram and MFCCs. The main goal in applying feature extractions techniques to SER is to minimise the number of features, simplify and create distinctive features.

This thesis uses audio features that have achieved the best performance in state-of-the-art SER tasks [240]; *i.e.*, Mel-scale Spectrogram, and MFCCs. The choice of audio features for bio-inspired model using SNN in SER tasks is challenging because input needs to be encoded into binary data and needs apparent, distinctive and concise features.

Mel-scale Spectrogram

Mel-scale Spectrogram features are successfully applied in emotion recognition in the literature [246]. Mel-scale frequency represent the frequency perception in humans and represent the spectrum of frequency in audio signals [185]. For each audio sequence, Mel-scale spectrogram are extracted using Fast Fourier Transform (FFT) [274]. First, the magnitude spectrogram is calculated from the raw input signal. Then it is mapped onto the Mel scale with a power spectrum.

Mel Frequency Cepstral Coefficients (MFCCs)

MFCCs represent one of the most popular ways of extracting handcrafted features for emotion recognition in the SER tasks. The main characteristics of these features reside in mimicking speech processing in the human ears by applying Cepstral analysis [131]. MFCCs are extracted from the Mel-Scale Spectrogram by applying logs of power which are computed 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 audio signals for each temporal frame.

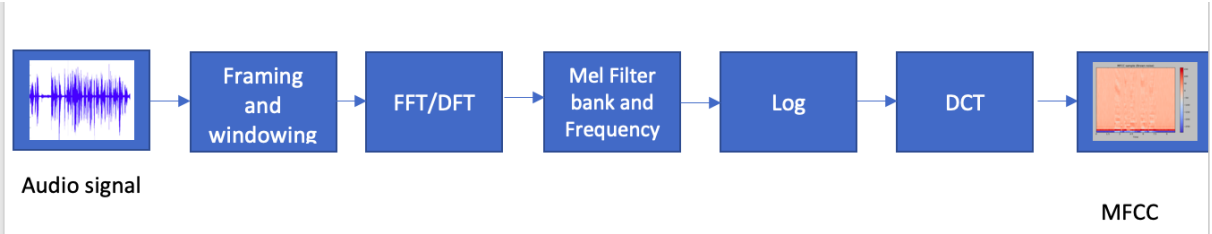


Figure 4.3: The process of extracting MFCCs features from raw audio signal

4.4 Input Encoding

Input encoding is an essential step in the application of SNNs for emotion recognition. Input encoding enables transforming input into a spike patterns. The process involved converting real values input into temporal Spike-train. There are various methods for input encoding, where the choice of methods depends on the nature of input.

Visual feature encoding Facial features encoding consists of transforming each pixel into a neuron. Thus the number of neurons is equivalent to the total number of pixels. Poisson distribution is the most popular method for input encoding applied to encode input into spike-trains [41]. Poisson distribution is used to compute features into spike-trains, which represent a

binary value of either true or false for each neuron. The Poisson distribution P is given by the following equation:

$$P(n) = \frac{(rt)^n}{n!} exp^{-rt} \tag{4.3}$$

where n is the number of spikes occurring in a time interval Δt and r is randomly generated in a small time interval where only one spike occurs. Each r has to be less than the firing rate in the Δt time interval.

Number of neurons in the Poisson group represented by spike-train are equivalent the size of the input. For example if an input image size is 100×100, the total number of neurons of the spike train is 10000. Rates of neurons in the Poisson group are proportionate to each pixel intensity in the input. The maximum rate is set to the maximum intensity of the input pixel. The main process of converting input into Poisson spike-train is described through the following algorithm:

Algorithm 1: Poisson spike train generation

Input: input features extracted for raw input
Output: Poisson spike-train
while *input presented to for a period of time* **do**
- 1- Set number of neurons to input size
- 2- Set firing threshold rates of neurons according to the corresponding input intensity
- 3- Set a time step
- 3- Sub-divide time into a group of temporal intervals using time step
- 4- Generate numbers x between 0 and 1 (number of x is number of neurons).
- 5- For each interval if x is higher than firing rate generate spike
- 4- Save spike train corresponding to the input

Figure 4.4 shows spike-train generation using Poisson distribution for an input image representing a facial expression.

Audio feature encoding

Similar to visual input, auditory input is encoded into Poisson groups of neurons with the number of neurons proportionate to the size of the audio features. Figure 4.5 shows an example of spike-train generation for audio features. Encoding temporal data, such as audio, is more challenging as the results need preserving the temporal information of audio features such as MFCCs or Sceptral features. There are two approaches in encoding audio features:

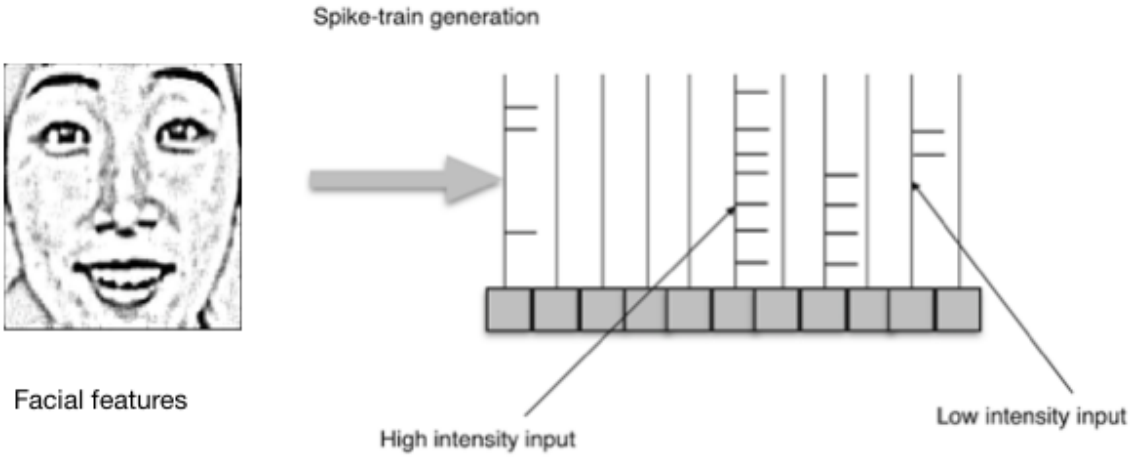


Figure 4.4: Spike-train generation using Poisson distribution: input represents facial features. Each pixel represents an input neuron. Rates of Poisson spike-train are proportionate to pixel intensity

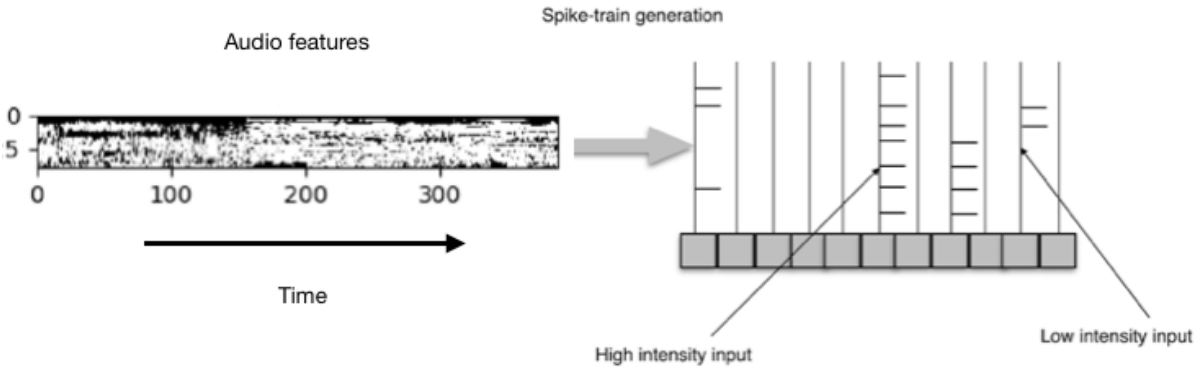


Figure 4.5: Spike-train generation using Poisson distribution: input represents audio features.

- **Temporal segments** where input features are divided in small temporal segments, and each segment is encoded separately,
- **Input features with whole temporal information** where features are encoded in one full temporal segment, and the features are encoded as one.

In this thesis, features are encoded as a whole temporal information mainly to preserve the whole temporal relationship and for computational efficiency in the training process.

4.5 Neuron Model

Chapter 3 describes numerous neuron models used in SNNs, LIF represents the most popular and most straightforward method to translate neurons dynamics. Similarly to work presented in [65], Leaky Integrate and Fire (LIF) model is identified as the most suitable model for emotion recognition neurons dynamics. Neurons communicate through a series of spikes, and thus neurons can learn unique features that distinguish different emotional states. The membrane voltages of neurons are translated by the following equation:

$$\tau\frac{dV}{dt} = (E_{rest} - V) + g_e(E_e - V) + g_i(E_i - V). \tag{4.4}$$

V is the membrane voltage, and E_{rest} represents the resting membrane potential. E_i and E_e represent the equilibrium potential for inhibitory and excitatory synapses, respectively. g_e and g_i represent the conductance of the synapses for the excitatory and inhibitory synapses. When a membrane reaches a certain threshold, the neuron fires spikes followed by a resting phase E_{rest} for a specific time interval (5ms). The temporal interval represents a refractory period where the neuron cannot spike. τ is a time constant representing the time a synapse reaches its potential, and it is longer for excitatory neurons.

In order to achieve better network balance and stability, *homoeostasis* is employed. Homeostasis is first proposed by [41] as a stabiliser of STDP learning in SNN networks. It is an adaptive membrane threshold V_{thresh} mechanism [65]. That is, V_{thresh} =V_{thresh} + θ, where V_{thresh} initial value is a constant and θ increases when a neuron fires and then decays exponentially when θ reaches the neuron's rate with a time constant of (5ms) which is the time of the refractory

period of the excitatory neurons. In this way, homeostasis prevents some neurons firing for all presented inputs and as well as avoids few neurons from dominating emotional patterns [230]. We also employ lateral inhibition encouraging competition between neurons.

Changes in conductance drive synapses models. Synapses conductance increase when pre-synaptic reaches a synapses; otherwise, the conductance decreases exponentially. The conductance dynamics are governed by a time constant of post-synaptic potential following the equation:

$$\tau_{g_e} \frac{dg_e}{dt} = -g_e, \tag{4.5}$$

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

4.6 Learning Algorithm

Learning in the brain is occurs mainly in an unsupervised way, rather than supervised learning [152]. Synapses in the neocortex are constantly influenced by changing patterns in sensory neurons such as in visual and auditory cortices. The sensory information provided does not have any supervised learning where information about the pattern is given in the learning phase.

As presented in Chapter 3, Spike Timing Dependent Plasticity (STDP) is one of the most popular learning methods for classification tasks [138], [195]. The algorithm is successfully employed in pattern recognition and image classification tasks [38]. This thesis proposes the application and adaptation of STPD learning for emotion recognition in both facial expression and speech tasks.

The difference between auditory and visual SNNs resides in the input pre-processing, feature extraction and input layer settings. Learning from the input layer to the excitatory layer is achieved through unsupervised STDP learning [65]; that is, learning distinctive features for each emotional class label in an unsupervised manner. STDP represents a spike-based type of Hebbian learning, where the connection between neurons strengthens when they fire together. The plasticity is influenced by the timing of the pre-synatptic and post-synaptic spikes. Post-synaptic weight are updated when a post-synaptic spike reaches a synapse, which is characterised

by the following equation:

$$\Delta w = \eta(x_{pre} - x_{tar})(w_{max} - w)^{\mu} \tag{4.6}$$

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 is the dependence of updates on previous weight. x_{pre} is the pre-synaptic trace left every time pre-synpatc spike reaches a synapse. That is, weights are increased by Δw if pre-synaptic spikes fire prior to post-synaptic spikes. Otherwise, they decrease. The changes of weights in STDP learning is computed by a function of difference between pre-synaptic and post-synaptic spike firing timing. Learning with STDP is advantageous compared to classical back-propagation as weights do not need to be learned through backward and forward pass [101]. In SNN, weights representing each synapse are updated independently according to time of spikes of their pre-synaptic and post-synaptic neurons.

4.7 Training Process

The training phase starts by presenting input one by one to the network for a specific amount of time with a delay after each input. This delay enables the network to reset all values and get ready for the next input. The network records all neurons' spikes where each group of neurons represents a feature in the input. The network updates weights after each training interval during the training phase. The training interval are decided dependent on the dataset size. After training, a class label is assigned to a neuron group based on their spiking behaviour. For example, various voting mechanisms can be identified for assigning neurons to class labels. Saunders et al. [247] have presented three different voting mechanisms:

- **all** where excitatory spikes are summed for each label class and the label with the highest sum is assigned;
- **most-spikes** where the neurons with the most spikes for a class label are assigned a class label;

- **top percent** where only a percentage of the most spiked neuron are used to identify a class label; and
- **correlation clustering** where at each training interval a vector of the most spiked neurons for each class label are recorded which are then compared to the testing phase vector and labels are identified.

In this thesis, we use the **most-spikes** mechanism: neurons with the highest spike response to a label class during a training phase are assigned to that class label. During the training phase, neurons are assigned label at each training interval.Input where for both facial and audio features are identified by distinctive patterns in both spatial and temporal locations in the input. The main training phase is described in the following algorithm.

Algorithm 2: SNN training

Input: Input features extracted from raw input
Output: learned weights
while *input presented to SNN for a period of time* **do**
    1- Encode input into Poisson spike trains;
    2- Set input neuron group rates to the encoded input.
    3- Unsupervised learning using STDP
    4- Update weights after each update interval

4.8 Prediction

After the learning phase, neurons are assigned a label according to their spiking behaviours. A neuron is assigned a label when it spikes more compared to other labels. After the learning stage weights are fixed.

We adopt the 'most-spiked' voting strategy, where neurons are assigned a label if they spike most when presented with such label. The excitatory layer of the network represents the voting layer where neurons are assigned labels following their spiking activity over time. First each neuron is assign a random class label. Then, each output class is represented by a group of neurons. When adding convolution layer to the network, each neuron group for each class comprises subgroups, each representing a feature within a class label. This voting mechanism enables the classification of new data according to neurons activation. In the prediction phase

input are classified by analysing spiking activities of all neuron groups. Each class label has specific spiking patterns where only a certain number of neurons have the highest spiking activity. The prediction process can be summarised through the following algorithm:

Algorithm 3: SNN prediction phase

Input: input features
Output: label prediction
while *input presented to SNN for a period of time* **do**
    1- Encode input into Poisson spike trains;
    2- Set input neuron group rates to the encoded input.
    3- Unsupervised learning using STDP
    4- Record spike activity for each neuron
    5- For each convolution patch count the spikes for neurons
    6 - Get the number of neurons that spiked the most
    7- Assign label when neurons spike most for a specific label

4.9 Summary

This chapter introduces the first contribution of this thesis; introducing novel methods of applying bio-inspired architecture in emotion recognition tasks for unisensory models for both visual and audio modalities. It starts with a SNN architecture. It describes techniques chosen to extract features from input data for both visual and audio modality. After that, it describes input encoding step, where audio and visual features are encoded into meaningful input for the SNN. The chapter also details neuron models used for representing neural dynamics in the network, explaining how individual neurons interact in the network. It also describes the learning algorithm chosen for emotion recognition tasks, consisting of STDP with unsupervised learning. Unsupervised learning represents the most biologically plausible learning method for emotion recognition tasks. Finally, it describes the training, testing and labelling process for the network, where the network if first trained in an unsupervised manner. Each group of neurons learns specific features from the input.

The next chapter 5 will detail the proposed models for multisensory integration. The chapter describes three main models based on different pathways of multisensory integration in the brain. All multisensory models are implemented using SNN with similar neurons models and

unsupervised STDP learning.

Chapter 5

Bio-inspired Multisensory Emotion Recognition

5.1 Introduction

As highlighted in Chapter 2, the key challenge in multisensory integration in emotion recognition tasks is the inability to translate the multisensory nature of emotions. Most commonly used fusion techniques such as early or late fusion, consider information from different modalities as independent. These methods rarely translate the constant cross-talk between modalities, where learning in each modality can be enhanced by feedback from other modalities. To directly tackle this challenge, this chapter introduces three bio-inspired architectures in multisensory integration that model three pathways in the brain of processing social signals.

This chapter starts by giving an overview of the nature of multisensory integration in the brain for audio-visual data. Then it describes the design of three proposed models inspired by three different pathways for multisensory integration; that is,

1) Multisensory integration through convergence in higher-order areas in the brain;

2) Multisensory integration through cross-modal enhancement happening in sensory areas such as visual and auditory;

3) Multisensory integration through neural synchrony, translating the importance of the role of temporal neural correlation in multisensory integration.

5.2 Background on Multisensory Integration of Social Signals

Social signals processing, understanding and perception involve various areas of the brain and a complex network [120]. The human brain starts by parsing inputs from different sensory modalities through segmentation then works on constructing meaningful representations through integration [254]. The two processes are always active in the brain to interpret the present from previous past events. For example, to process and understand a speech sound, the brain needs to segment all possible auditory inputs from noisy environments to identify which sound corresponds to which speech of a person. Segmentation is achieved by looking at temporally and spatially adjacent sensory information from both visual and auditory inputs such as facial expressions and non-verbal sounds.

The mechanisms of multisensory integration of social signals in emotions differs from multisensory integration of other tasks in the brain. Social signals and affective information automatically draw attention in the brain; that is, the brain automatically attends to emotional and social information as a priority. It automatically processes emotional faces in a background [272]. There exist four main steps in assessing and integrating social signals:

- Attention: The brain uses attention to select emotional information for observation.
- Detection: This first stage involves sensory modality-specific detection, where information is processed through different brain regions. In this phase, all essential features from each modality are extracted in early sensory regions such as visual or audio cortices.
- Integration: In this step, a new percept is created, comprising multisensory features. Integration is not only achieved by fusing the extracted sensory features but through a more elaborate mechanism. At this state, each modality is in constant interaction with others. Integration also happens 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. In this final stage, decisions are made on the interpretation of the social signal and emotional states.

Until quite recently, multisensory integration believed to be occurring only in high-level associative areas. However, other pathways are also discovered. Evidence also shows that multisensory integration happens with constant feedback between modalities. Multisensory integration also involves in primary sensory areas such as visual or auditory cortices.

Literature identifies three main pathways of multisensory integration happening at various areas in the brain. These pathways starts as soon as the brain receives sensory information. It starts with an early cross-modal integration and enhancement between modalities. Then an integration in higher-order areas such as STS. Higher-order areas contains multisensory neurons groups facilitating integration. Multisensory integration is also driven through neural synchrony, where synchronised spikes drive information. Multisensory integration follow three main principles: 1) spatial alignment, 2) temporal synchrony, and 3) inverse effectiveness where multisensory enhancement plays a role of increasing any sparse or noisy sensory modality.

The next sections describe the proposed models based on three different pathways for multisensory integration:

1) Multisensory integration through convergence simulating the integration in higher order multisensory areas.

2) Cross-model enhancement in early sensory areas where one modality precedes, predicts and enhances other modalities.

3) Integration through neural synchrony describing the integration through temporal and semantic synchrony of neurons groups from different modalities, where the synchrony of various neurons drive the integration of multisensory information.

5.3 Multisensory Integration through Convergence

This section describes the first model for multisensory emotion recognition based on multisensory integration by convergence as experimented in some multisensory brain areas.

5.3.1 Background on Multisensory Integration through Convergence

The most classical theory for multisensory integration is through convergence in areas such as Superior Temporal Sulcus (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 [264]. This kind of areas includes a higher number of multisensory neurons, which constitute a way to multisensory integration.

The common assumption is that multisensory areas, containing multisensory neurons, receive convergent inputs from various modalities such as auditory and visual. Thus, multisensory integration happens by changes in the firing rate of multisensory neuron populations [153]. Many experiments have been conducted to explain multisensory integration by convergence.

Davies et al. [59] have proposed a hierarchical model for emotion integration of faces and audio stimuli. They have conducted experiments to study the pathways of multisensory integration of emotions, ranging from presenting congruent facial and audio stimuli and incongruent stimuli. They also varied emotions stimuli from happy to fearful. They have found that the final integration happens by merging in the right posterior superior temporal sulcus (rpSTS). rpSTS responds more to bimodal information than unimodal facial stimuli. These findings are in conjunction with the notion of supra-additivity in multisensory integration. While the response to multisensory stimuli in some multisensory areas is more significant than responses to unisensory stimuli, it is also higher than the algebraic sum of neuronal responses to unisensory stimuli [237]. Morrow et al. [193] also proved that there were multisensory areas such as the amygdala in multisensory emotion recognition. They demonstrated that large portions of neurons in the amygdala are multisensory, receiving and responding more than one sensory modality.

5.3.2 Convergence model

Multisensory integration through convergence is the first model proposed in this thesis. This model represents the multisensory convergence pathway in the Superior Colliculus (SC) brain region. The SC region of the midbrain is a crucial multisensory area for audio-visual integration [153]. Various neurons in this area respond to both audio and visual modalities and are fully multisensory. There have been various studies on the computational principles of multisensory

integration within the SC brain region. Research has focused on the study of responses of multisensory neurons compared to unisensory modalities [55] [188]. The convergence model presented in this section is more simpler than models presented by [55], where sensory areas neurons are represented in one region for each modality and not in subregions. The simplification of the convergence model makes it more suitable for affective computing applications as opposed to creating neurocomputational models.

Integration through convergence represents the most common approach for multisensory integration in the brain, as described in Section 5.3.1. In this thesis, multisensory by convergence model architecture is designed by simulating the process of passing information from lower sensory areas to higher-order multisensory areas for integration. The unisensory model described in Chapter 4 is used to model unisensory lower-order areas such as visual facial expression or auditory speech.

5.3.3 Model Architecture

Multisensory integration by convergence model is designed using SNN which follows the same high-level design in terms of neuron models and learning as the unisensory models in Chapter 4. The high level description of the model is shown in Figure 5.1. The figure shows how information flows from both modalities and converges in a multisensory area. The network contains three layers as shown in Figure 5.2. The main input layers consisting of inputs from both audio and visual modalities connects to an excitatory layer. The excitatory layer contains three main neurons groups. These are connected in a lateral way to inhibitory layers. We descirbe each layer as follows:

- **Input** layer, 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 are generated from the extracted 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 unisensory 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 for

Figure 5.1: High level description of multisensory integration by convergence

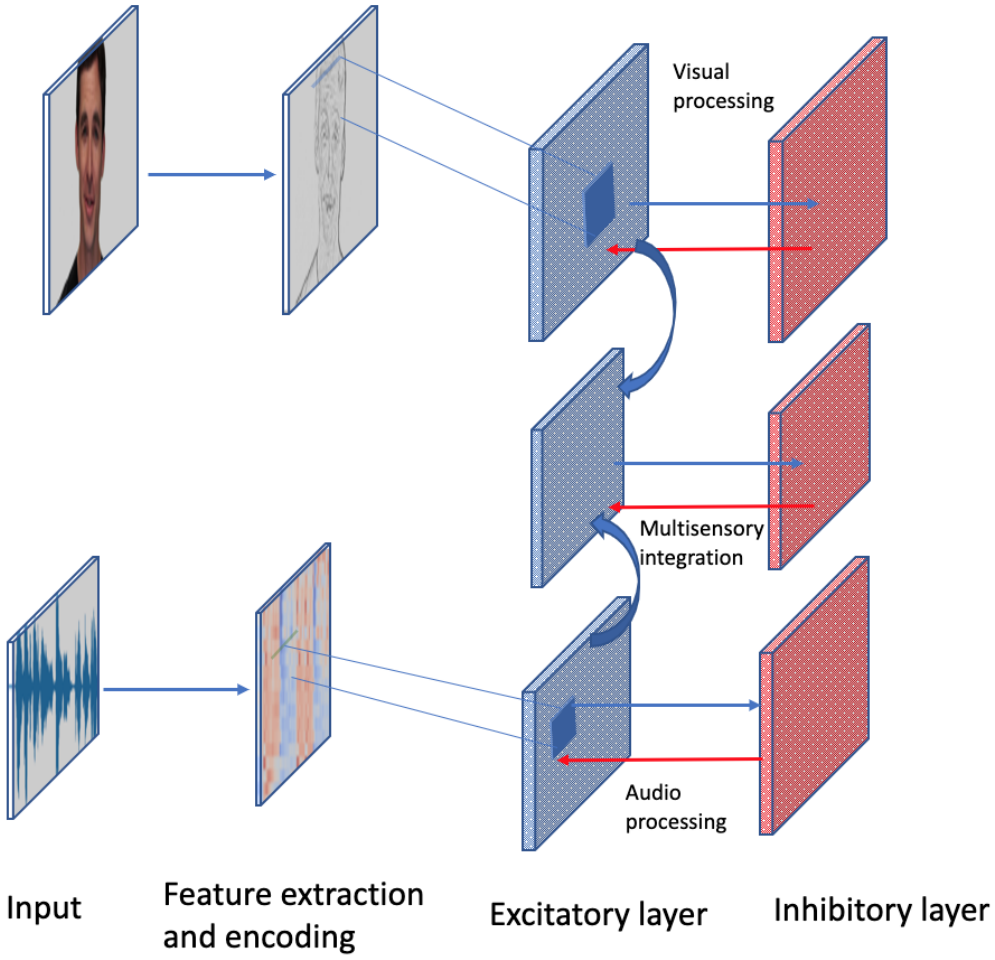


Figure 5.2: Multisensory integration by convergence model architecture

both modalities receive input from the input layer for each modality. The multisensory excitatory group receives information from each excitatory modality group, as shown in Figure 5.3. The recurrent connections between the unisensory neurons groups and the multisensory neuron group allow learning of distinctive patterns features for each class label. The connections from the multisensory group to each modality group enables feedback from the multisensory to unisensory groups [286].The number of neurons in the excitatory layer in each group depends on the choice of the number of features and convolution window size and input neurons.

- **Inhibitory** layer, 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

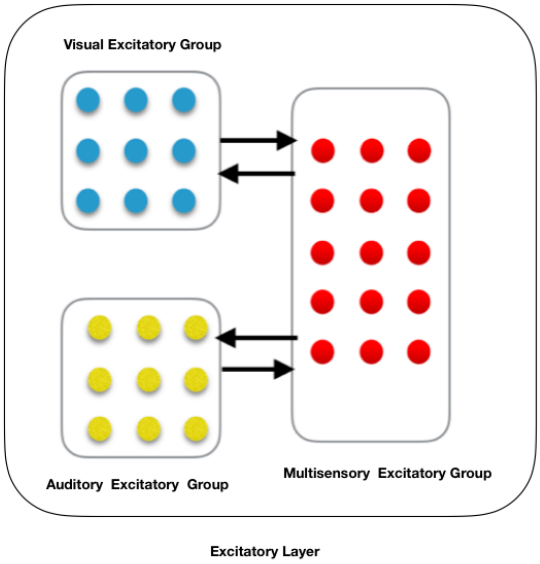


Figure 5.3: Recurrent connections at the excitatory layer

is connected to all other neurons, apart from the one it is receiving input. Each inhibitory group is connected in a lateral fashion to the corresponding excitatory group.

5.3.4 Model Learning and Training

Learning in the convergence model happens in the same order as unisensory models; learning is achieved through STDP. STDP is considered to be a powerful learning method as it enables learning through unsupervised methods, as described in Chapter 3.

The main component of the convergence model is the simulation of higher-order multisensory regions, where unisensory information converges to a multisensory area. Learning multisensory patterns happens in two main stages. First, unisensory excitatory neuron groups receive information from the input layer. Each group starts learning unisensory patterns where groups of neurons spike for the same class label. Then, multisensory neuron group receive information through connection from both unisensory excitatory groups. Learning in the convergence group happen in an unsupervised manner through STDP where neurons spiking for the same class label get a stronger connection. The connection between these neurons happens regardless of signals modality. Training in the convergence model happens by presenting inputs from each modality with a delay between input. This delay enables to mimic the biologically realistic small delay

between visual and auditory sensory information reaching the brain. Learning in the network happens in a temporal manner, where each input is presented for a certain temporal window.

5.4 Early Cross-Modal Enhancement

This section describes the second proposed model on applying bio-inspired architectures in multisensory emotion recognition. It proposes a model based on early cross-modal enhancement, which represents a different pathway of cross-modal interaction and multisensory integration in the brain. This section first outlines a background on cross-modal enhancement. Then it details the proposed model in terms of architecture and learning method.

5.4.1 Background on Cross-modal Enhancement

Early cross-modal enhancement denotes the interaction between visual and auditory cortices. Activity in the auditory cortex is closely affected by visual information. Literature reveals that there exists an early multisensory integration in the auditory cortex, where visual information enhances the perception of auditory information [20, 14].

Most of the existing experimental studies proceed by presenting subjects with video information where there is a single utterance of a syllable [117]. In these experiments, visual information is given earlier to investigate its influence on the auditory processing. Results show that auditory information is often predicted by visual information which affects the brain's response to auditory information. The prediction happens within 100ms on the onset of auditory stimuli [117]. Visual information is transmitted to the auditory cortex through multiple routes and influences the auditory process [206]. Altigan et al. [14] have also demonstrated that there is an early cross-modal enhancement of auditory processing using visual information.

Unisensory areas in the brain have a constant interaction at early sensory levels [265] during multisensory integration. This idea of early sensory enhancement represents one possibility of cross-modal prediction and interaction, especially for audio and visual pathways in emotion processing [191].

Arnal et al. [13] have identified two main pathways for early cross-modal enhancement in multisensory integration. The first pathway is defined by information from early visual sensory

area influencing the auditory area through a third area; that is, the superior temporal sulcus (STS). This indirect pathway helps predict the type of auditory stimuli. The second pathway has a direct connection between the visual and auditory areas without any involvement of an intermediatory area. Having a direct connection helps predict the onset of the auditory stimuli.

In emotion recognition, visual information usually precedes the auditory, leading to facilitation of auditory processing by visual information [119]. There exist various studies investigating the interaction between faces and voices in a social setting. Most studies lead to a similar conclusion; that is, auditory processing is influenced by visual information at very early stages of processing. Jessen et al. [119] have investigated the role of early cross-modal enhancement through various EEG studies. The experiments investigated the dynamic interaction between body gesture and audio modalities.

Garrido et al. [82] also investigated the role of cross-modal enhancement in emotion recognition. They study the dynamic interaction of visual and auditory emotional information. They proved that facial expressions influence the processing of auditory prosody processing at very early stages. They also discovered that any mismatch between facial expression and auditory information results in further processing in the brain. The additional processing is due to the violation of early cross-modal prediction from visual modality.

Lee et al. [156] studied the influence of visual information on auditory processing for emotion understanding. Visual information influences auditory processing at a very early stage and in the auditory area. Kokinous et al. [149] also investigated the role of cross-modal enhancement in multisensory emotion recognition. They focused on two expressions that are angry and neutral. They first experimented on auditory stimulus only as a baseline. They compare audiovisual responses for both congruent and non-congruent stimuli. Their findings converge to conclude that there exists an early cross-modal enhancement incongruent audiovisual information, which also suggests a direct connection between visual and auditory processing areas.

Similarly to multisensory integration by convergence, early cross-modal enhancement is simulated and implemented using SNN. The main difference resides in the network topology and architecture. Weights are learned using STDP as described in the previous sections.

5.4.2 Cross-modal Enhancement Model Architecture

Early cross-modal enhancement is modelled and implemented using a SNN with two separate early pathways, corresponding to visual and auditory modalities respectively [265]. The cross-modal enhancement model depicts the enhancement of the auditory modality with a preceding visual modality as shown in Figure 5.4. Cross-modal enhancement model is represented by adding connections from visual modality to the auditory modality at the excitatory layer level, as shown in Figure 5.5. Those connections translate early multisensory integration in the brain with early cross-modal enhancement; that is, influencing auditory processing with visual information proceeding and vice versa. The auditory excitatory layer receives input from both auditory input layer and visual excitatory layer, following the same pattern in the brain where visual information precedes auditory processing by a few milliseconds.

This model differs from classical data fusion techniques employed in multisensory emotion recognition from the state-of-the-art as detailed in Chapter 2. The existing fusion techniques consist either of concatenating features extracted from each modality while ignoring interactions between them, or late fusion where each modality is if first classified on its own, ignoring any interaction between modalities. Cross-modal enhancement is also different from the recent cross-model learning technique presented in [7] where a cross-modal transfer from the visual to auditory data is applied. This thesis presents an approach that is more biologically plausible, where the auditory part does not use only prediction from the visual modality but learns from their spiking patterns. This type of learning is multisensory, which improves the propagation of learning from the visual group to the auditory group [268].

Figure 5.6 shows the workflow of our approach, which mainly consists of two learning parts:

(1) Unimodal learning on visual and audio signals based on SNN;

(2) Early cross-modal interaction in the brain [20, 14] to enhance audio signals using visual stimuli. In the following, we will detail the main design.

5.4.3 Model Implementation and Learning

Similar to the unisensory model in Chapter 4 and the above multisensory convergence model, the proposed cross-modal enhancement model is implemented using SNN [65]. It follows the same

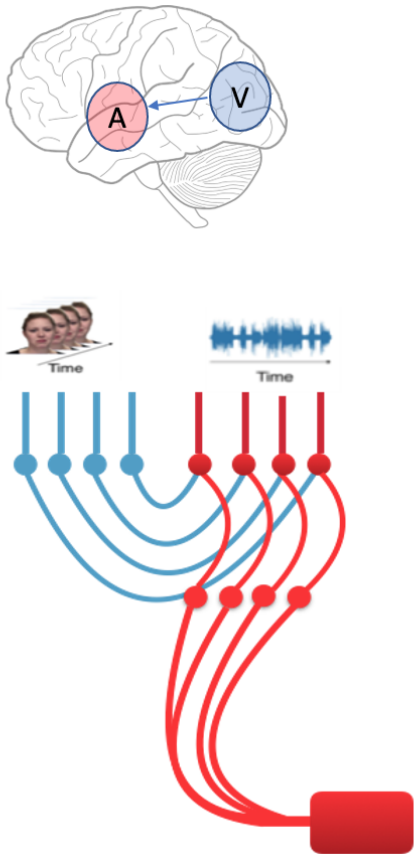


Figure 5.4: High level description of cross-modal enhancement model

process in feature extraction and input encoding as in Section 4.3 and 4.4, in order to process raw visual and audio signals and prepare inputs for the SNN.

Similar to the proposed architectures for unisensory models 4, a convolution layer is applied on input features in each modality at the excitatory level; that is, the input layer is connected to a convolution excitatory layer coupled with an inhibitory layer. Each input is divided into convolution features windows where a stride window moves through the input. The stride window moves along the temporal axis of the audio features and along adjacent features in visual input. Adding a convolution layer has demonstrated to be beneficial in improving the overall accuracy on unimodal learning from general image classification [247].

Poisson spike trains are generated from both visual and audio inputs. Then spike trains representing each modality are fed to both visual and audio areas to learn distinctive features of image and speech for each type of emotions. The input layer for each modality is recurrently connected to an excitatory layer that is in turn connected to an inhibitory layer in a one-to-one

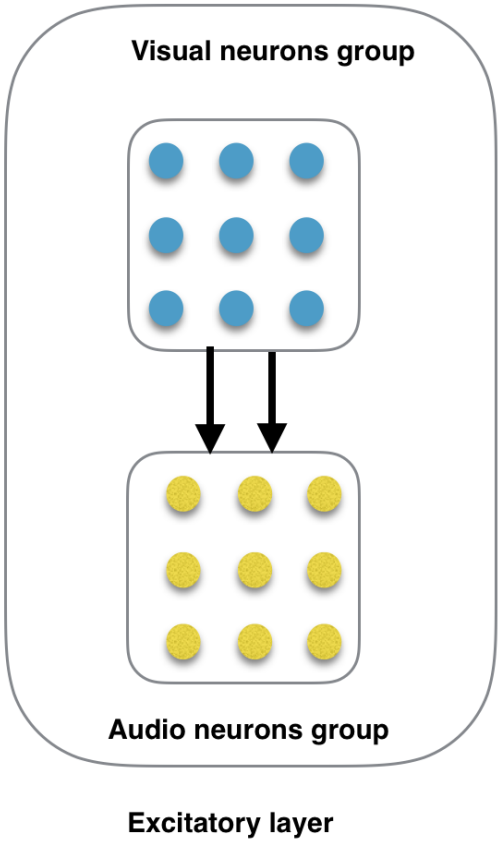


Figure 5.5: Early cross-modal connections from visual to auditory modality in the excitatory layer

aspect providing lateral inhibition. Neurons in the inhibitory layer are connected to neurons from all features in the excitatory layer apart from the ones it receives input.

Connection are set between the neurons in visual modality at the excitatory layer to audio neurons at the excitatory layer in the auditory modality. The choice of the direction of connection, i.e from visual to auditory is justified by the nature of the task. In fact for emotions processing and recognition the attention goes first to the face and facial expression information precedes the auditory by few milliseconds [119]. These connections activate new neurons in the audio modality. After receiving video frames input at the visual modality, the network runs and learns different spike patterns. After learning from the video input, the network enters a resting phase. The audio modality then learns from both the audio input and the visual spike patterns. Neurons spiking for audio modality play a multisensory role; accepting connections from the visual modality. Connections between the visual excitatory and the auditory neuron groups help the

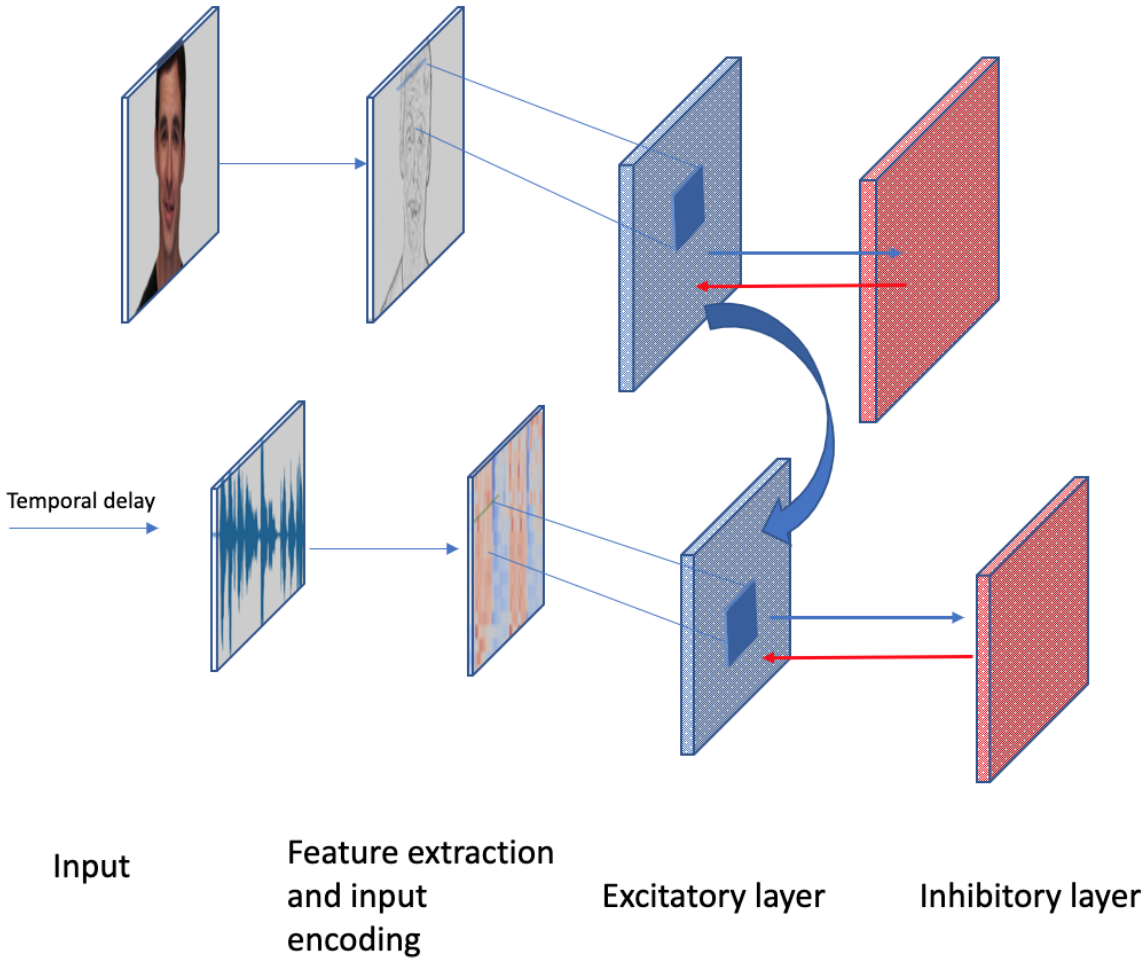


Figure 5.6: Main architecture of the cross-modal enhancement model

transfer of spikes from the visual to the auditory modality. The weights of these connections are initialised similarly to the connection weights from inputs to excitatory layers.

The visual-audio interaction is learned through STDP in the same fashion as unimodal modalities and convergence model. The evaluation of unsupervised STDP learning is achieved in two stages. During the training, weights are updated after each training interval, and spiking neurons for each feature at the audio excitatory layer are allocated a class label, according to which neurons spiked most for each feature. In the second stage and during the testing phase, classification is achieved by allocating the testing data with a class label for the most spiked neurons saved through the training phase.

5.5 Multisensory Integration through Neural Synchrony

The previous sections have introduced two biologically inspired models for multisensory integration. The two presented models translate two different pathways in the brain. Multisensory integration by convergence and early cross-modal enhancement are explored, modelled and simulated using Spiking Neural Networks (SNNs) with unsupervised online STDP learning.

This section introduces a third model based on the latest findings in neuroscience [133]. It presents a model based on neural synchrony. It is inspired by temporal synchrony of spiking neurons in various sensory regions corresponding to different types of stimuli. Temporal and stimuli based synchrony drives multisensory integration of social signals of emotions. This section first presents an overview and describes the importance of neural synchrony in multisensory integration. Then it details a novel method for modelling multisensory integration through neural synchrony based on SNNs and Graph Convolution Networks (GCN).

5.5.1 Background on Neural Synchrony

Humans perceive emotional events and social signals in a multisensory manner where information enters the brain through various sensory paths. 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 [283]. The brain creates connections that are feedforward, lateral and feedback for integrating information from several senses [140]. Multisensory integration is facilitated by signals being adjacent temporally, spatially or both. The temporal and spatial proximity of sensory stimuli makes the information likely to come from the same event. Multisensory integration also depends on the semantic congruence between signals, i.e. unisensory signals having come from the same stimuli. This section details the role of semantic and temporal congruence in multisensory integration.

Role of Stimuli Congruence in Multisensory Integration

Multisensory integration depends strictly on various aspects such as stimuli or semantic congruence between unisensory senses. Kim et al. [144] demonstrate that there is a difference of multisensory learning effect between congruent stimuli and incongruent one. Multisensory

learning with congruent audio and visual stimuli enhanced multisensory integration compared to incongruent audiovisual stimuli. Barutchu et al. [21] also show an enhancement in learning when presented with congruent stimuli for each modality. When stimulus from modalities in audiovisual integration are incongruent, the perception of auditory information changes, creating an illusional effect such as in McGurk effect [117].

Role of Timing in Multisensory Integration

Multisensory integration and interaction are closely affected by the timing of the onset of different modalities. Sanders et al. [242] show through experiments that the temporal relationship between audio and visual stimuli influences multisensory integration. In addition to the timing of the onset of stimulus from different modalities, timing is essential in the relation between different neuron groups. Timing synchrony between different cortical areas translates the coordination between these areas. Neural oscillations in different frequencies and their synchrony drive multisensory processing.

5.5.2 Neural Synchrony in Multisensory Integration

Neural synchrony, along with convergence and information binding, represents the main pathways for multisensory integration [197]. Multisensory integration through convergence rely on firing rate changes in different cortical region through 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, this theory experienced limitations in terms of multisensory precepts generalisation. Multisensory integration does not solely happen in a convergence way [197], but occurs through a constant crossmodal talk between various unisensory areas including at an early level [132].

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. Numerous studies have been conducted to define the exact role of neural synchrony in multisensory integration [284]. Neural synchrony is regarded as the synchronisation of different brain oscillations at different frequencies. Each frequency band drives certain types of information as defined below [134]:

- Delta waves: having functions in memory with a frequency lower than 1Hz;
- Beta waves: dealing with attention or cognitive tasks and having a frequency range between 13 and 30Hz;
- Gamma waves: running over 30Hz and driving stimulus processing and features binding;
- Alpha waves: ranging from 8 to 12Hz and driving attention and distributing information;
- Theta waves: ranging from 4 to 7Hz and driving attention memory and cognitive control.

Role of oscillation band frequencies is detailed in [272] for emotion and social signal processing. Symons et al. [272] have outlined a detailed study on the role of neural oscillations synchrony for the perception of emotions in both unisensory and multisensory. They have argued that neural synchrony drives the perception of multisensory emotions from auditory and visual stimuli. Experiments show that 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. On the other hand, experiments with induced temporal delays between visual and auditory stimuli show a cross-modal enhancement from visual to auditory areas.

Neural synchrony also drives general multisensory integration as detailed in [133]. Kiel et al. argued that the transfer of information between different brain area in primary sensory, frontal cortical or multisensory areas, is achieved through neural synchrony in different oscillatory frequencies. Distinct frequency ranges define the different aspects of multisensory integration. The importance of neural synchrony is also demonstrated in [285] where neural synchrony within different brain regions is altered in some conditions such as schizophrenia, epilepsy, or autism. Temporal coordination of neuronal response plays a vital role in the integration and transfer of information.

Various studies have been conducted to investigate and define the role of neural synchrony in multisensory integration [253]. Data show that coherent and synchronised neural signals, especially in the gamma band, drive the integration of multimodal information. More recently [242] defined the importance of temporal binding window (TBW) in multisensory integration. Temporal binding window is highly relevant in information integration from different modalities.

In audio-visual stimuli, information needs to be integrated within a specific temporal range. In humans, it ranges from 150 to 250ms [33]. Having a longer or shorter TBW can impair the perception of multisensory integration such as in autism of schizophrenia [267]. TBW can be plastic in human being; that is, it can be narrowed down by training such as for musicians [267] [33].

The proposed model for integration through neural synchrony is implemented using SNN and GCN, which will be described in the following sections.

5.5.3 Background on Graph Neural Network

Graph neural networks are gaining more and more attention in dealing with problems on unstructured data such as classification of social networks, representations of biological systems and chemical reactions. The notion of graph was first introduced by Gori et al. [93] to represent the learning of graph-structured data through nodes neighbour information propagation. They introduced the notion of Recurrent Graph Neural Networks (RGNN) where learning happen in an iterative way until the network reaches a stable fixed point. Computing with iterations is computationally costly especially for large networks with a large number of nodes. Other research has attempted optimising these findings; for example, Yujia et al. [161] introduced the use of gated recurrent units and optimisation for feature learning in Graph Neural Networks. Following the success of Convolutional Neural Networks, Bruna et al. are one of the first who have applied convolutional layers to graph neural network [35]. They employ a spectrum of graph Laplacian that translates convolutional properties into the Fourier domain resulting in a more straightforward representation of graph data.

Henaff et al. have applied graph convolutional network (GCN) and spectral learning to large classification problems such as ImageNet object recognition and bioinformatics [105]. They have designed unsupervised learning for graph estimation when the graph structure is unknown. Spatial convolutions are introduced to address the limitations of spectral methods for large graph, which allows learning functions by aggregating features between neighbouring nodes. They are particularly useful for node classification as they do not require to process the whole graph simultaneously as for spectral methods.

Kipf et al. have introduced a semi-supervised method using a localised first order approxima-

tion of spectral graph convolutions for node classification [145]. This helps in alleviating the complexity challenge of spectral convolutions on processing whole graph. They experiment on citation networks and the results have shown that the model can effectively learn hidden layer representations encoding local graph structure and features of individual nodes.

Hamilton et al. overcome the challenge of large graphs by introducing inductive node embedding [97], where node features are used to learn an embedding function generalising on unseen nodes. This is achieved by using the topological structure of local neighbours of each node. It trains on aggregator functions instead of feature vectors on each node. An unsupervised loss function is designed so as to enable training without using task-specific labels.

Gao et al. have used Learnable Graph Convolutional Layer (LGCL) to enable convolution operations on large graphs [81]. This works by transforming the graphs into 1-D format grid to make the use of convolutions easier and more accurate. They have developed subgraph training to reduce the computational complexity of the current training method that uses the whole adjacency matrix as an input.

Applications of GCN are starting to emerge in computer vision and affective computing recently. Nian et al. propose the use of GCN in facial feature recognition [200]. They have used GCN for defining facial attributes such as hair colour, eyes or brow shape. They first extract facial features using CNN, which are then transformed into the above attributes. These are used to construct a graph with nodes representing facial attributes and edges representing relations between them.

GCNs have been used for emotion recognition through EEG data [259]. Song et al. have proposed Dynamical Graph Convolutional Neural Network (DGCN) to model multi channel EEG features where each EEG channel represents a node in the graph. The adjacency matrix is learned in a dynamic way, where the matrix is updated at training time. This is the opposite of classical GCN where the adjacency matrix is often fixed at the beginning of the training.

Zhang et al. have used GCN to model context in emotion recognition [309]. They compute the relation between context information with a graph and one example of context is facial expression of the interlocutor. Then facial features are extracted with CNN and concatenated to context information.

GCNs have demonstrated promising results in various applications [303, 98] and play an

important role in the advancement of affective computing and emotion recognition. In this paper, we apply GCN in modelling neural synchrony to learn complex interaction patterns between synchronised neuron activities captured in a spiking neural network. To the best of our knowledge, we are the first to apply GCN to bio-inspired multisensory emotion recognition.

5.5.3.1 Notations and Definitions

This section introduces the main definitions and notations. It will also focus on graph convolution neural networks (GCN). Graphs are defined by two main attributes: nodes and edges. Graph are defined as follows: $G = (V, E)$, where V is a set of nodes and E defines edges of relations between nodes. Inputs in GCN are identified as follows:

- an input feature matrix X defined by $N \times D$, where N is the number of nodes and D is the number of input features for each node;
- an adjacency matrix A which defines the main structure of the graph network [145].

A GCN graph produces an output in the form of a feature matrix Z defined by $N \times F$. F represents the number of output features for each node, and N is the number of nodes. Generally all GCNs share the same main architecture where layers can be represented as follows:

$$H^{(l+1)} = f(H^{(l)}, A)) \tag{5.1}$$

The first layer $H^{(0)}$ represents the input feature X and f represents the propagation rule. f can be represented by a simple activation function such as ReLU. Each layer is represented by the number of nodes and the feature representation of nodes. For each layer, features are aggregated and represent the features in the subsequent layer. This is achieved using a propagation function f. In this way features become more abstract.

5.5.4 Modelling Neural Synchrony for Multimodal Emotion Recognition

The previous section provides an overview about the role of neural synchrony in multisensory integration. Neural synchrony drives multisensory integration through various network areas from primary cortical regions to multisensory regions. This characteritic can be exploited for

creating novel models for multisensory integration and data fusion. This will address some challenges in current data fusion techniques for social signals and emotion recognition. The temporal synchrony translates the temporal adjacency for multisensory information and can be applied for constant cross-modal interaction.

This section describes a bio-inspired approach to model multisensory emotion recognition using neural synchrony. It consists of three main components: (1) simulating and modelling multisensory integration and interaction via SNN; (2) modelling neural synchrony through a graph network; and (3) applying graph convolution network to multimodal emotion recognition. In the following, we will describe each of these components in details.

Neural dynamics and learning rules in this model follow the unisensory modality learning process described in Chapter 5. LIF model is used to model neurons dynamics and unsupervised STDP learning is used to to model the multisensory learning.

5.5.4.1 Modelling Neural Synchrony in Graph Network

To enable multisensory integration, recurrent connections are set at the excitatory layer between audio and visual neuron groups in order to allow cross-talk between modalities by connecting neurons that spike together between both modalities.

After training SNN, important information is obtained on neuron activities and multisensory interaction. This includes the spatial location of a neuron at the excitatory layer, time of spiking in milliseconds, and the modality type of each neuron, which together defines patterns for each type of emotional states.

Neural synchrony represents neurons spiking within the same temporal window. This facilitates the integration of information from different sensory sources [263]; that is, learning and extracting relevant and crucial features from sensory inputs such as heterogeneous neuronal populations [34]. This model proposes to model neural synchrony with a graph network. Neurons are modelled as nodes and their spiking synchrony as edges. In this way, complex patterns can be learnt between visual and auditory neuron groups through graph neural network to enable multisensory emotion recognition.

Neural synchrony graph network is defined as an undirected 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.

Node feature matrix $X_{N\times D}$ is defined, where N is the number of nodes and D is the dimension of input features on each node. Nodes represent neurons with features defining the type of neurons; that is, either audio or visual. Nodes are connected if they belong to the same video and spike within the temporal window of integration. The temporal window is set to 150ms to simulate a biologically realistic temporal window of integration in multisensory integration [18].

The adjacency matrix A is represented by a sparse matrix containing adjacency matrices for each subgraph that is constructed on a video input. The adjacency matrix is represented by two main aspects: *temporal coordination* between neuron spikes and *stimulus* based relations, where neurons belonging to the same subgraph and class type are linked together.

5.5.4.2 Multisensory Interaction Learning

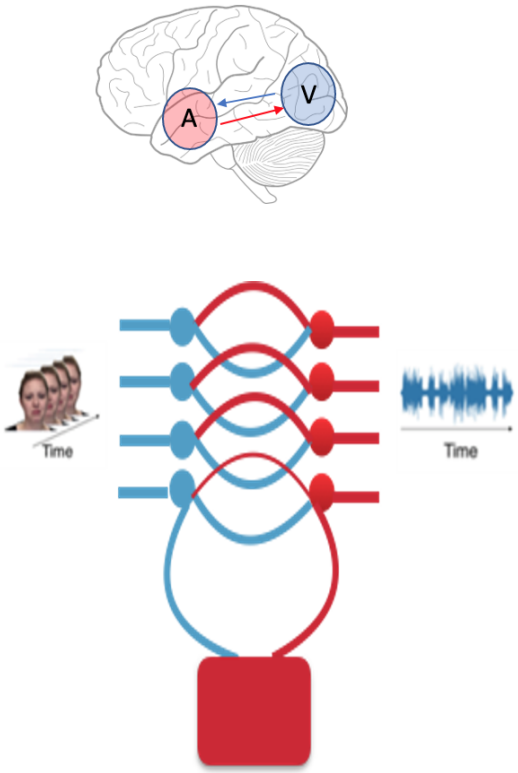


Figure 5.7: High level model for integration through neural synchrony

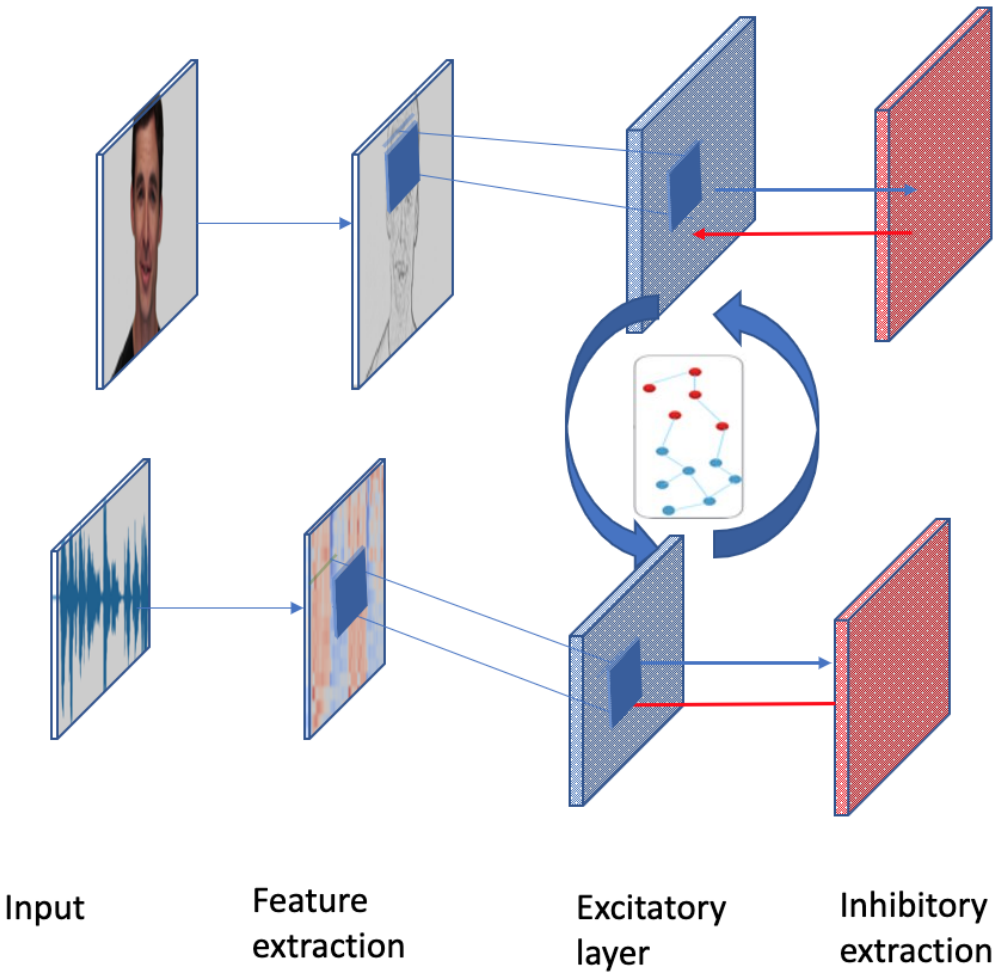


Figure 5.8: Workflow of the synchrony model. First features are extracted from both visual and audio data, and then fed to a SNN where multisensory integration is simulated. After training, neuron activities are recorded, based on which a graph is constructed.

Figure 5.8 describes the process of multisensory integration and interaction in a SNN for graph construction. Input from audio and visual modality are captured through two different neurons groups, representing each a modality. The input neurons groups are connected to the neurons groups in the excitatory layer, where each modality's neurons are connected to their corresponding excitatory groups. Recurrent connections are set at the excitatory layer where neuron groups receive connections from each other. This enables constant cross-talk between modalities.

Given a video input, we obtain these two connected neuron groups, which will form into a

subgraph. We compose each subgraph from videos in a complete graph, where neurons between subgraphs are connected if they share the same class label; *i.e.*, the same emotional state.

5.5.4.3 Multisensory Emotion Recognition via Graph Convolution Neural Network

We define emotion recognition as a subgraph classification problem; that is, assigning a class label to each subgraph. The architecture is based on semi-supervised GCN model [145], which is applied to node classification in GCN. It employs layer wise propagation rule based on first-order approximation using spectral convolutions. Spectral convolutions represent filters as graph signal processing based on spectral theory. Introducing the first-order approximation [145] allows a simplification of the model and faster training time. Their model is particularly useful for our neural synchrony model as it can better capture global complex patterns in graphs compared to spatial convolutions methods that capture more local areas of nodes. Training the whole graph instead of node batches helps maintain the neural synchrony structure. The reason is that the classification of emotions is conveyed by the neural synchrony pattern instead of individual nodes.

The model is adapted from [145] for the subgraph classification by introducing an additional general pooling layer [69]. This is applied in order to have a higher representation of the features learned at a node level. It results in features for each subgraph (video input). This is an essential step, reducing the size of the overall graph and propagating the learned features for each subgraph representing a video input.

Figure 5.9 shows the main architecture for graph convolution network for neural synchrony, which stacks up multiple convolution layers. We have used a deeper architecture compared to the one introduced in [145] by adding a hidden layer. Having a deeper network helps in aggregating and translating the complex relationship between nodes to subgraphs.

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:

$$H^{(l+1)} = f(H^{(l)},A), \tag{5.2}$$

$H^{(l)}$ 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 $x \in \mathbb{R}^N$ (which is a scalar value for every node) with a filter $g_\theta = diag(\theta)$ where $\theta \in \mathbb{R}^N$ is in the Fourier domain [145]. The spectral convolution can be translated by:

$$g_\theta * x = U_{g_\theta} U^T x \tag{5.3}$$

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 subgraphs each representing neural activities of a video input. The network consists of three layers followed by a pooling layer over graph [69] in order to combine features from all subgraphs and enable the classification of subgraph. The main learning model and propagation rule can be defined as follows:

$$Z = f(X,A) = \texttt{softmax}(\hat{A}\sigma(\hat{A}\sigma(\hat{A}W^{(0))})W^{(1)})W^{(2)}), \tag{5.4}$$

where weights are defined by weights 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 layer 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:

$$\mathcal{L} = -\sum_{d \in y_D}^{\cdot} \sum_{c=1}^{C} Y_{d,c} ln Z_{d,c} \tag{5.5}$$

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 [145].

5.6 Summary

This chapter details the main contribution of the thesis: designing computational models for bio-inspired multisensory integration. Three models are attempted following three different

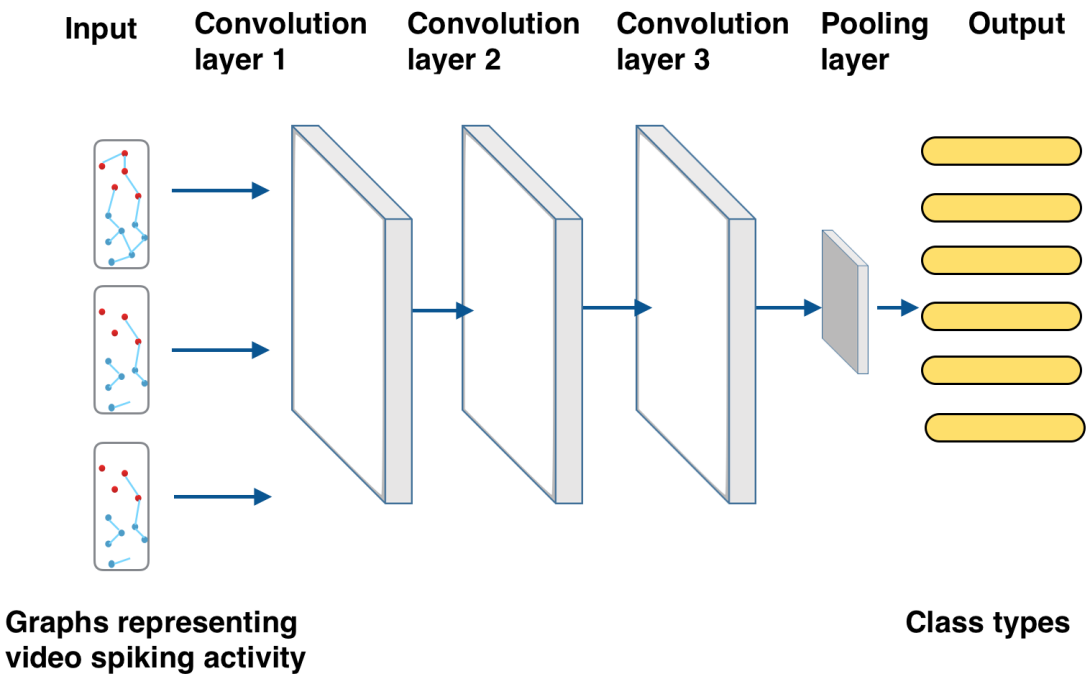


Figure 5.9: Architecture of graph convolutional network for neural synchrony

pathways for multisensory integration in the brain.

1) Multisensory integration by convergence represents the first proposed model for multisensory emotions recognition. Built on the basic SNN for a unisensory model, this first model is inspired by the multisensory integration in the higher-order regions such as Superior Colliculus. This region receives input from different sensory modalities and integration happens at a higher order.

2) The second model, early cross-modal enhancement, represents integration at the lower sensory region at early stages, where one modality precedes and enhances the other. This happens particularly for multisensory integration in speech processing, where visual information in facial features predicts the sound in the auditory area, thus enhancing the auditory signal.

3) The third proposed model, multisensory integration through neural synchrony, represents the latest finding on the integration of multisensory social signals of emotion in the brain through neural synchrony. Neural synchrony plays a role in driving information and multisensory learning.

Table 5.1 summarises the three proposed models. The three models represent different pathways of multisensory integration happening at different stages of multisensory integration.

Table 5.1: Bio-inspired multisensory integration models

Model	**Brain Pathway**	**Characteristic**	**Learning Method**
Integration by convergence	Superior Colliculus (SC)	Unisensory modalities converging into one multisensory area	Unsupervised
Early cross-modal enhancement	Auditory and Visual cortex	Visual modality connected directly to the auditory one	Unsupervised
Integration through synchrony	Auditory and Visual cortex	Constant cross-talk between modalities, (decentralised). Temporal coherence / Stimulus driven	Unsupervised and Semi-supervised

The three models aim to address the challenges faced by current fusion techniques. Convergence and enhancement models use unsupervised learning, whereas the integration through synchrony model uses a combination of unsupervised and semi-supervised learning for classification. The next chapters 7 and 8 will detail experimental evaluations for both unisensory models detailed in Chapter 4 and multisensory models detailed in this chapter.

Chapter 6

Evaluation Methodology and Experiment Setup

6.1 Introduction

This chapter describes experimental setup for evaluating the proposed models. It introduces different datasets being experimented, tools, features selection and extraction techniques, baseline techniques, and model configuration.

6.2 Evaluation Objectives

The main objectives of this thesis are to address the following four questions:

1. Are -inspired architectures effective for unisensory social signals of emotions recognition tasks?

   To answer this question, we design experiments to evaluate the effectiveness of bio-inspired architectures for unisensory models such as facial expression and speech emotion recognition tasks.

2. Does applying bio-inspired models in multisensory integration increase the effectiveness of multisensory recognition systems?

To answer this question, we design experiments to evaluate three bio-inspired models; convergence, enhancement and synchrony and compare their performance to state-of-the-art multisensory integration techniques.

3. Do the bio-inspired models 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 testing on another dataset?

   To answer this question, we design cross-dataset experiments; where we train all models on one dataset and test on another dataset. We then compare the accuracy of the proposed models against a collection of state-of-the-art techniques.

4. Do bio-inspired architectures present *robustness* to noise; that is, the accuracy of recognising emotions is not influenced by noise in sensory modality?

   To answer this question, we evaluate all the proposed model with various types of noise on both visual and audio data and assess the degradation level and accuracy of the bio-inspired models against state-of-the-art techniques.

6.3 Datasets

Evaluations of the models presented in this thesis are conducted using still images, audio and video data. The datasets selected are third party datasets that are widely used in state-of-the-art emotion recognition. The use of third party datasets enables a more meticulous evaluation of the proposed models and a fair comparison to state-of-the-art techniques using the same datasets. Table 6.1 summaries the datasets. Most of the datasets used, present posed emotions, where participants were instructed on how to elicit each emotion. This kind of data collection enables a cleaner data for training the models. All the datasets include six basic emotions (anger, happiness, fear, surprise, sadness, and disgust). The next sections introduce in details the specificity for each dataset chosen in this thesis.

Table 6.1: Multimodal emotions datasets

Dataset	Facial expression	Subjects	Video	Colour	Type
RAVDESS	happiness, sadness surprise, fear, anger,disgust and neutral	24	1026	color	Acted
eNTERFACE'05	happiness, sadness, surprise, fear, anger, disgust and neutral	42	1290	color	Posed
CK+	happiness, sadness, surprise, fear, anger,disgust and neutral	123	593	color	Posed
JAFFE	happiness, sadness, surprise, fear, anger,disgust and neutral	10	213	grey	Posed

6.3.1 The Extended Cohen-Kanade Dataset CK+

The Extended Cohen-Kanade dataset plus is one of the largest and most popular publicly available datasets for facial expression [170]. Subjects recorded in the dataset include 123 adults aged between 18 and 50 years with a higher proportion of females and different ethnic background [125]. The completed dataset consists of multiples frames per subject and per emotional type. The images are mostly grey colour with a size of 640 $\times$ 490.

The dataset consists of a total of 3297 samples of seven basic emotions (Neutral, Surprised, Sad, Disgusted, Fearful and Angry). For each emotion, there exist various intensities. For each subject and each image only, the most intensified emotion frames are retained. The onset of the frames is discarded as they represent the neural emotion class [170]. Figure 6.1 shows a sample from the dataset with a wide range of emotion class labels.

6.3.2 Japanese Female Facial Expressions (JAFFE) Dataset

JAFFE dataset consists of a total of 213 images with seven basic emotions (Angry, Disgusted, Fearful, Happy, Surprised, and Neutral). The dataset comprises ten female Japanese subjects. The facial expressions executed by the subjects are in a controlled environment and acted, where subject are asked to act different emotions. All images are in grey scale, with a size of 256$\times$ 256. Figure 6.2 shows samples for some of the emotions present in the dataset.

Figure 6.1: Sample of Cohen-Kanade dataset

Figure 6.2: Sample of JAFFE dataset

6.3.3 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Dataset

RAVDESS [166] is a multimodal database with videos composed of basic emotions through speech and songs. The dataset consists of 24 participants with a balanced gender number. All subjects are professional actors, reading sentences in emotional states such as happy, sad, angry, fearful, surprise and disgust. The dataset also consists of actors singing in different emotional states. The recordings are available through video, audio, and audio-video options. In this thesis, we use only spoken sentences subset with a total of 1026 video files all in colours with a size of 1280×720.

6.3.4 eNTERFACE'05 Dataset

The eNTERFace dataset [217] is an acted dataset of 42 subjects from 14 nationalities, with a proportion of 81% males and 19% females. This dataset includes subjects with glasses and facial hair. The audio is recorded at 48000Hz in 16-bit format. Each subject records the same six basic emotions. Subjects are recorded while reacting to particular emotions. Recordings results to videos with an average of 4 seconds long.

6.4 Tools

This section describes all software, libraries and tools used in implementing experiments for the evaluation of the proposed bio-inspired models in this thesis.

OpenCV

OpenCV [32] is an open source computer vision library. OpenCV includes multiple algorithms for general image processing, segmentation, camera calibration, stereo 3D vision and features extraction techniques. It is widely used in various applications of computer vision. It supports various operating systems such Windows, Mac OS or Android.

Librosa Librosa [185] is a python package for audio signal processing and analysis. It is also used for music processing and represents a basis for music creation. It is mainly useful for audio features extraction such as MFCCs, and Mel Spectrogram.

Keras Keras is an open source library in written in Python [51]. Keras is mainly used for the implementation of deep learning network with different architectures. It contains various building blocks such as layers, activation functions, optimisers or objects. It also contains a range of functions for image and text processing. Keras is particularly useful for convolution neural networks and supports layers such as batch normalisation or pooling.

BRIAN Simulator BRIAN [92] is a spiking neural network simulator python library. This tool is used to compute networks of neurons and has a range of neuron models. It is particularly useful for models of networks of spiking neurons. Networks implementations are set through differential equations, defining neuron groups and synapses connections.

Scikit-learn Scikit-learn is the most comprehensive machine learning library in Python [216]. It is part of SciPy (Scientific Python) group and is principally used for scientific computing

data analysis. The package comprises various areas of machine learning algorithms, such as supervised learning, unsupervised, data processing and transformation and models evaluation. Scikit-Learn is widely used for classification and regression in supervised learning, and clustering and dimensionality reduction in unsupervised learning.

6.5 Data Experimental Setup

For all experiments presented in this thesis, we chose repeated holdout method with 10 trials, then average results [228]. Trials results are presented in APPENDIX B. We randomly shuffle the data and split into 60% for training, 20% validation and 20% testing for CNN and GCN based models. Data is split into 80% training and 20% testing for SNN and SVM based models [305].

6.6 Baseline Models

This section describes baseline models representing some basic implementation as representative of state-of-the-art techniques in emotions recognition tasks. They are implemented and used to evaluate the proposed bio-inspired models.

6.6.1 FER Models

To evaluate SNN for facial expression recognition, we compare the performance of the most basic models that are representative of state-of-the-art techniques. We have considered a range of manual and automatic feature extraction techniques, including HOG, LBP and geometrical/ coordinates based features applied with an SVM classifier [205], [57]. We have also experimented the implementation of a basic CNN and training the last layer of a pre-trained network [141]. The experiments are conducted on two facial expressions datasets CK+ and JAFFE.

The selected automatic and manual features extraction models are described as follows:

Coordinates based In this first approach, facial landmarks coordinates are extracted for each image and added to a feature vector. Facial landmarks consist of detecting nose, eyebrows, eyes, mouth and jawline areas. We use the open source library OpencV and dlib library to extract

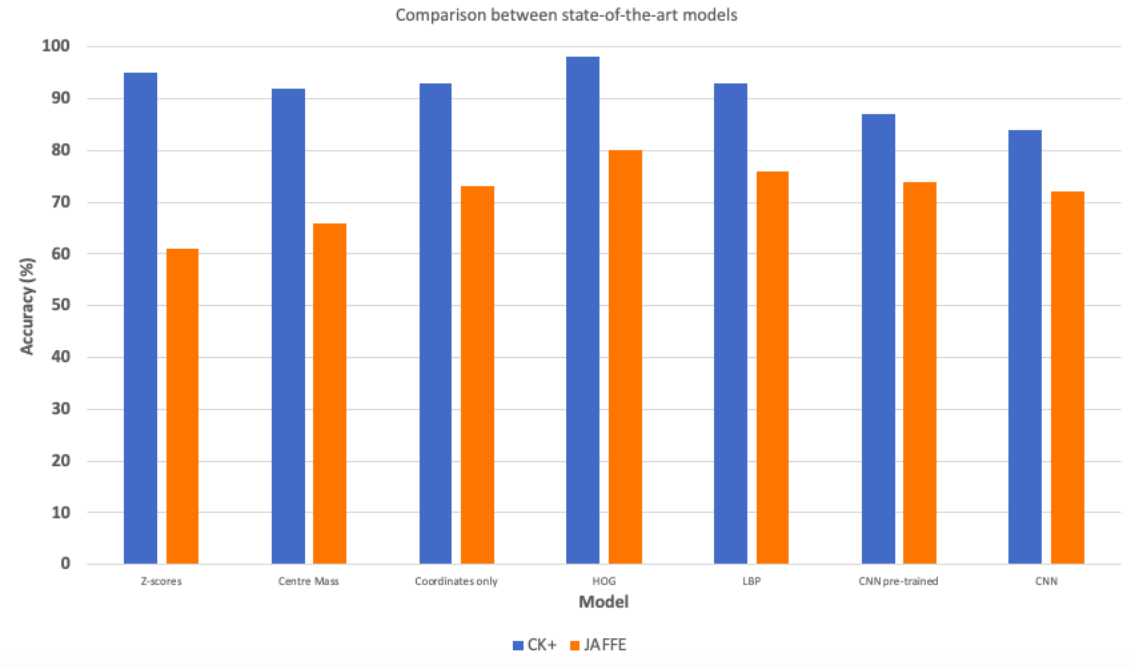


Figure 6.3: Comparison of overall results for automatic and manual FER models

68 facial landmark features. Then each coordinate is added the the main features vector. The coordinates features are fed to an SVM for training.

Distance to centroid The centre of mass or the centroid represents the mean location of distribution of the coordinates in an image. The centre of mass of a 2D image is calculated using the centroid at each axis:

$$X = \Sigma(x * m)/\Sigma(m) \tag{6.1}$$

$$Y = \Sigma(y * m)/\Sigma(m) \tag{6.2}$$

Where x and y are the coordinates and m is the intensity of each pixel.

We extract regions of interest in the image, that is the main facial landmarks points that represent the eyes, brows, jawline, mouth and nose and calculate the centre of mass using the coordinates of facial landmarks, as including all pixels of the image will lead to a shift of the centroid.

Statistical standardisation The statistical method consists of applying statistical standardisation to facial landmarks coordinates. Data standardisation is very common when having

data belonging to different categories in statistics. It consists of adjusting data according to a population mean and standard deviation. Applying data standardisation such as z-scores enables to prepare data when working with coordinates in order to accurately classify and compare coordinates from different images sets. We first use standard deviation method on each image facial landmarks. Then standard deviation values are used to obtain z-scores for the 68 facial landmarks coordinates. Z-scores are calculated by the flowing formula

$$Z = X - m/\sigma \tag{6.3}$$

Where X is the observed data, m is the mean of the data and is the standard deviation of the population.

Local Binary Pattern Local Binary Pattern (LBP) represents one of the most popular feature extraction techniques for facial expression recognition. The main algorithm consist of computing local representation of texture for an image. This is achieved by comparing each pixel and its neighbourhood by defining specific radius of neighbour pixels. LBP values are calculated for the centre pixel. The main formula for generating is as follows:

$$LBP_{p,R}(X_c, Y_c) = \sum_{p=0}^{p-1} s(g_p - g_c) \tag{6.4}$$

where

$$s(x) = \begin{cases} 1, x >= 0 \\ 0 \; otherwise \end{cases} \tag{6.5}$$

and g_c represents the central value,g_p are the equally spaced pixels value within a neighbourhood P.

Each image is first converted into grey scale. Then LBP is applied to each image in the dataset using the open source Skimage python library. The algorithm radius is defined to 16x16 blocks. We apply the LBP algorithm to pre-processed and cropped image with only facial area for each image. we obtain a feature vector of 16900 features.

Histogram of gradient orientation The HOG features extraction is performed using the the library Skimage HOG extracting function. After images are pre-processed by resizing and selecting the face region. The Histogram of Gradient Orientation algorithm is applied.We

Table 6.2: CNN baseline architecture

Input (256x256)
convolution 1 (5x5) (1,2)
maxpool (3x3) (2,1)
convolution 2 (4x4) (1,1)
maxpool (3x3) (1,2)
convolution 3 (5x5) (1,2)
maxpool (3x3) (2,1)
fully connected
output

obtained features vector of size 22500.

CNN small network

The network consists of three main convolution layers followed by maxpool layer each and a dropout layer to reduce over-fitting and a final fully connected layer. The three main layers consists of a Relu activation. An output layer is added with a Softmax activation and categorical cross entropy which suits best the multi-class problem. Each hidden layer includes a pooling layer which reduces or down-samples the resolution of the feature extracted from the previous layer. The architecture is similar to the baseline identified by [256]. We have used the same parameters for the network architecture which are shown in Table 6.2. For hyperparameters tuning we focus on learning rate, dropout, as we keep the model as basic as possible. We have experimented on parameters such as dropout (0.25 -0.5), learning rate (0.1-0.0001). For the final model we train it for 500 epochs and set the dropout value to .50 and learning rate 0.001.

Pre-trained CNN network We use a pre-trained network method to extract features. We use the pre-trained network VGG16 for image classification on the ImageNet dataset. The pre-trained network VGG16 is used to produces features vectors, with the last layer not being included. Then, we create a fully connected layer where the features are fed for classification. The first step in training consist of producing a vector of 512 features for each sample. Then the obtained features are fed into a one layer dense network with 256 nodes with a softmax activation.

The best performing techniques are selected for both automatic and manual features as show in figure 6.3, including HOG feature extraction for manual methods and a pre-trained Convolution Neural Network with transfer learning for automatic methods, as the representatives for manual and deep learning feature extraction techniques respectively. Scikit-image library

Table 6.3: SER CNN baseline architecture

Input (40x388)
convolution 1 (5x5) (1,2)
maxpool (3x3) (2,1)
convolution 2 (4x4) (1,1)
maxpool (3x3) (1,2)
convolution 3 (5x5) (1,2)
maxpool (3x3) (2,1)
fully connected
output

[289] is used to extract the HOG features from each image, resulting in 22500 features. The features are then fed to a linear SVM for classification. SVM has demonstrated as one of the most popular and straightforward classifiers for FER [177].

6.6.2 SER Models

In order to evaluate SNN model for SER we have implemented some classic methods for SER classification with SVM and CNN [270]. First MFCCs features are extracted from audio input, using Librosa [185], with a total number of feature of 40 and temporal feature length of 388. Choosing a higher number of features results in a higher level of spectral details. MFCCs as an input for a simple SVM classifier. Many kernels have been experimented such as linear, polynomial or radial basis function (RBF). The linear kernel have been retained as they produce the best overall accuracy. CNN represents an effective way of extracting features for SER [142]. Here we choose a baseline CNN architecture that consists of three sets of convolution layers, each followed by a max-pooling and batch normalisation. It also consist of a fully connected layer [16]. The networks are implemented using Keras framework in python [51] on a GPU.

The network consists of three main convolution layers followed by maxpool layer each and a dropout layer to reduce over-fitting and a final fully connected layer. The three main layers consists of a Relu activation. An output layer is added with a Softmax activation. Each hidden layers includes a pooling layers which reduces or down-samples the resolution of the feature extracted from the previous layer. We have used the same parameters for the network architecture in [256] which are shown in Table 6.3 :

6.6.3 Baseline Multisensory Model

To evaluate the proposed bio-inspired multisensory models, we design a baseline deep learning architecture for multisensory emotion recognition. Following state-of-the-art models introduced in Chapter 2 CNN have been identified as a good candidate for features extraction and fusion [37]. The model uses a basic architecture of fusion of features through CNN models. First, we use CNN for the extracting visual features. The network is designed with two convolution layers, followed by a single max pooling layer each. Similarly, the audio modality is designed using a CNN with two convolution layers followed by a max pooling layer. The concatenated features represent an input to a Multi-Perceptron (MLP) classifier with three fully connected layers using a simple cross-entropy loss. MLPs has been successfully used as a feature fusion technique [88]. Figure 6.4 shows the architecture of this baseline model. First, audio and visual CNNs are used to extract features from audio and visual input respectively. The obtained features are concatenated and passed to an MLP with fully connected layers. Training of the network happens in two stages. First, features are extracted from raw input in both modalities. These features then represent input for CNN in both modalities.

6.7 Bio-inspired Models Implementation Details

This section describes the implementation details, which starts from unisensory and moves to the configurations for multisensory models. Section 6.7.1 illustrates the process of feature extraction and input encoding for both visual (in Section 6.7.1.1) and audio input (in Section 6.7.1.2), and SNN configuration for unisensory emotion recognition (in Section 6.7.1.3). Section 6.8 describes specific SNN configurations for each of the three integration models: integration through convergence, named as *Convergence* (in Section 6.8.2), early cross-model enhancement, named as *Enhancement* (in Section 6.8.3), and neural synchrony, named as *Synchrony* (in Section 6.8.4).

6.7.1 Unisensory Configuration

Chapter 4 has detailed theoretical methods of the main methods used for feature extraction in both visual and audio data. This section describes the implementation and parameters for feature

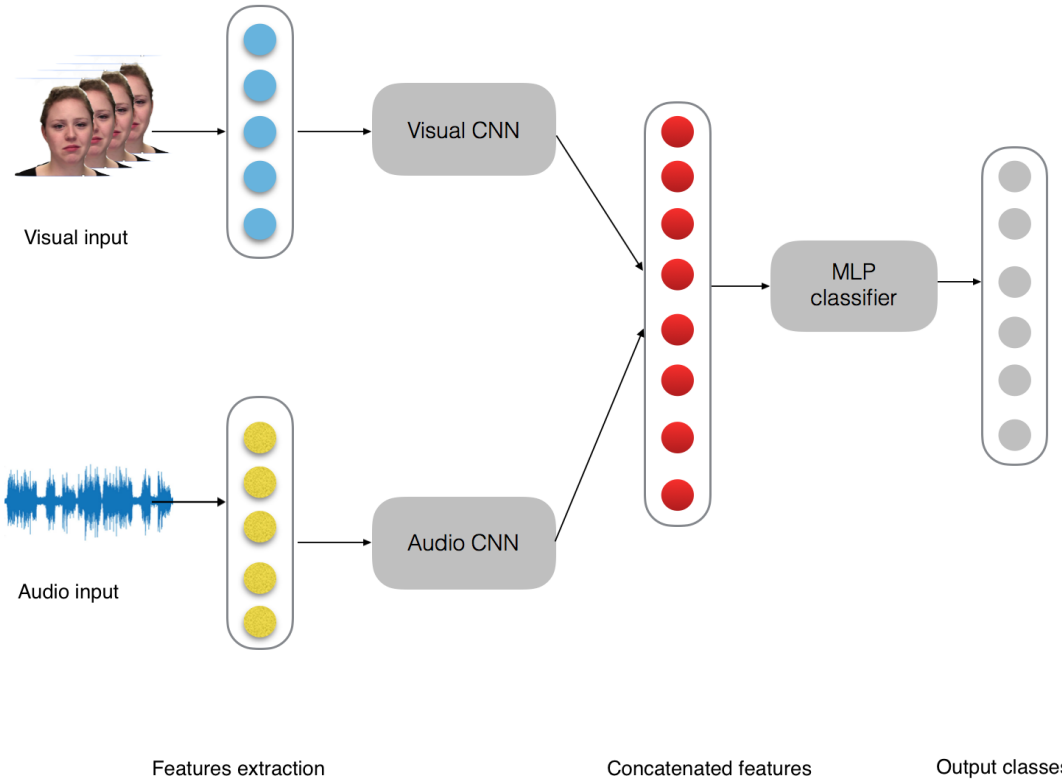


Figure 6.4: Multisensory baseline model

extraction techniques used in experiments conducted for the evaluation of bio-inspired methods.

6.7.1.1 Visual Feature Extraction and Spike Train Encoding

Input images are first uniformly resized to 100×100 pixels, then converted to grey scale using OpenCV Library. OpenCV is also used to detect facial area and crop the image to include only the facial area. Then we apply the Gaussian filter to smooth and remove noise on each facial image, and then apply the Laplacian filter to locate edges and corners of the image. Figure 6.5 presents the effect of LoG on a raw image input.

After input pre-processing and LoG features extraction, we obtain a feature vector with a size of 100 × 100. These features are encoded into Poisson spike train, where each element represents a neuron. Neuron rates are proportional to the intensity of the corresponding feature point.

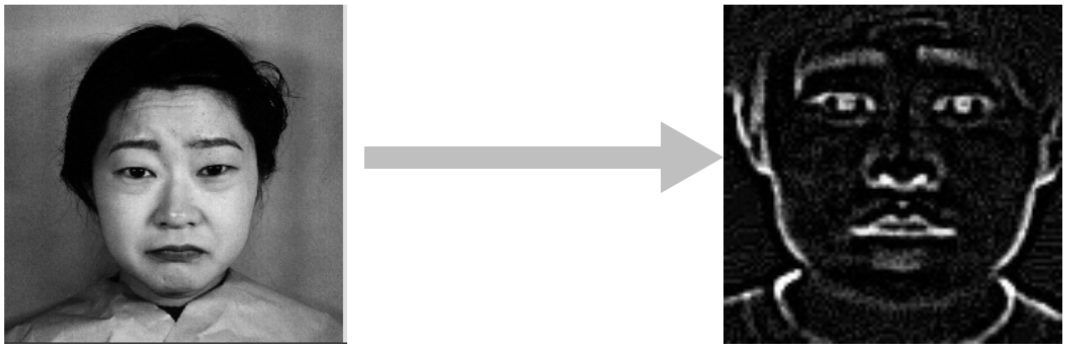


Figure 6.5: Laplacian applied on an image with Gaussian filter

6.7.1.2 Audio Feature Extraction and Encoding

Mel-scale Spectrogram features

For each audio sequence, Mel-scale Spectrogram are extracted using Fast Fourier Transform (FFT) [274]. First the magnitude spectrogram is calculated from the raw input signal. Then it is mapped onto the Mel scale with a power spectrum. In this thesis, the FFT window with length of 128 is chosen. This enables to transform the time domain signal into a frequency domain. The Mel-scale features are then computed using Librosa python library [185]. The maximum frequency used to the input is 8000 and the number of Mel bands is set to be 128. Although using a higher maximum frequency gives better precision, this choice gives a smaller input to the network input layer, which will be more computationally advantageous. Figure 6.6 represents result of Mel-scale spectrogram from the eNTERFACE'05 dataset sample with 'angry' emotion.

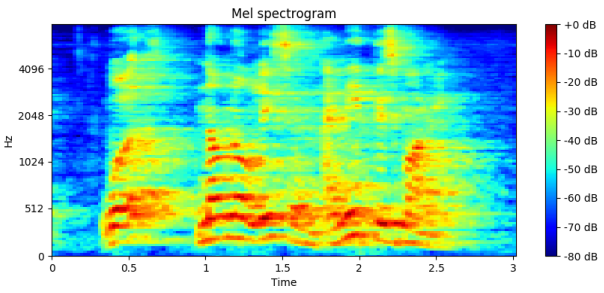


Figure 6.6: Mel-scale spectrogram sample for 'angry' emotion class

MFCCs features

MFCCs are 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 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. MFCCs are computed using the python library Librosa [185]. In summary, feature extraction goes though the following steps:

- Fourier transform is applied to audio signal
- After extracting the power of the spectrum, it is then mapped to Mel-scale
- For each Mel frequency, logs of power are calculated
- The final step involves computing Discrete Cosine Transform on Mel log powers. The log Mel spectrum is then converted back to temporal signal

The number of energies of filter banks is set at 40. All audio features are unified to have a temporal length of 388. Audio signals which result in smaller size are padded to match the chosen setting. A sample of the MFCCs features for 'angry' emotion is shown in Figure 6.7.

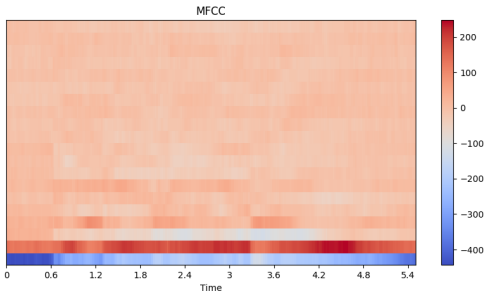


Figure 6.7: MFCCs features sample for 'angry' emotion label

Poisson Spike-train Poisson distribution is used to encode both Mel-scale Spectrogram and MFCCs into spike train. In the end, the number of input neurons is 128 $\times$ 130 for Mel-scale spectrogram and 40 $\times$ 388 for MFCC. Rates are obtained from MFCCs are shown in Figure 6.8.

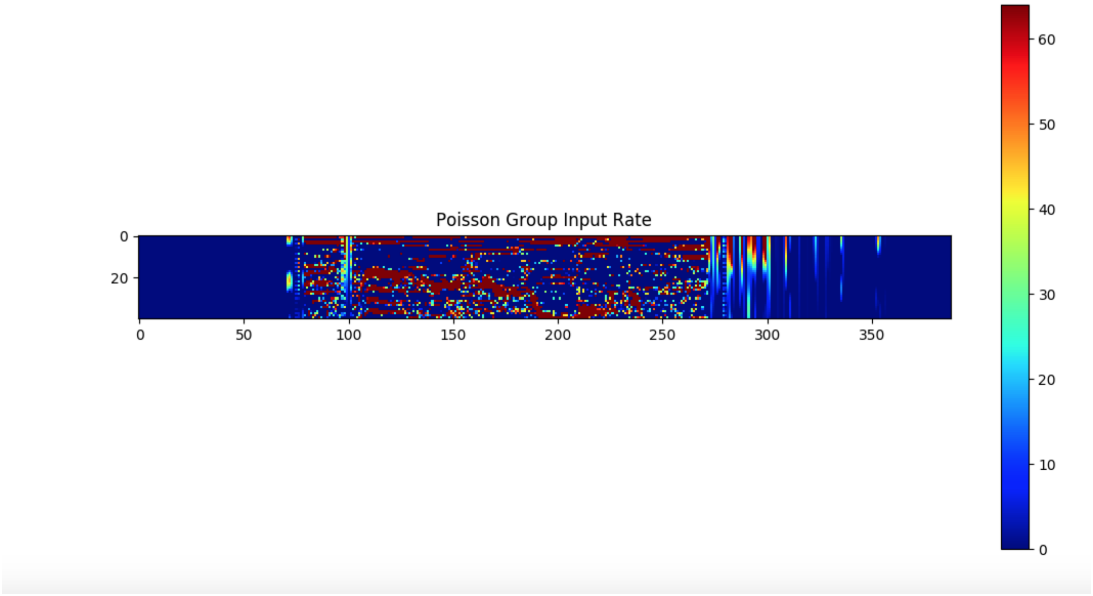


Figure 6.8: MFCCs spike-train generation

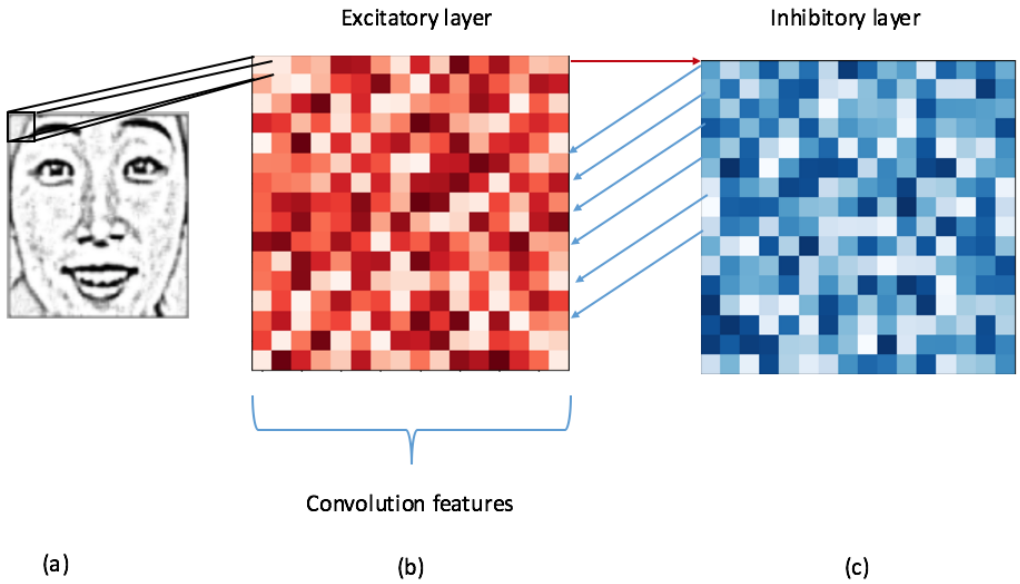


Figure 6.9: SNN workflow for FER: (a) LoG filters are applied to raw input, then the input is processed to create Poisson spikes train. (b) Excitatory convolutional layer where a number of features, stride and convolution window are chosen. (c) Inhibitory layer where each neuron inhibits all convolutional feature neurons apart from the one it receives input from.

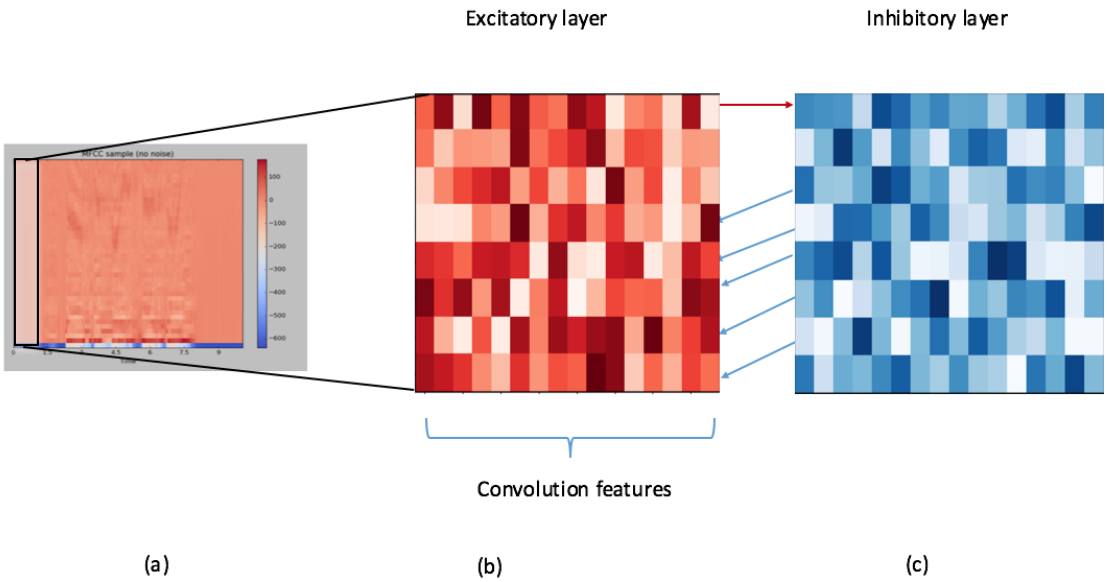


Figure 6.10: SNN workflow for SER: (a) MFCC features are extracted and Poisson spike train are created. (b) Excitatory convolution layer where a number of features, stride and convolution window are chosen and convolution moves through temporal axis. (c) Inhibitory layer where each neuron inhibits all convolution features apart from the one it receives input from.

6.7.1.3 SNN Configurations for FER and SER

Once features are extracted from raw inputs and transformed into spike-trains, they constitute the input to SNNs for training, as shown in Figure 6.9 and 6.10. We apply a convolution layer across input, as it proves beneficial for increasing the overall accuracy by defining various features [247]. That is, each input is divided into several features of the same size and a stride window that moves throughout the whole input. Each convolution window represents a feature, which constitutes an input to the excitatory layer. The number of neurons O in the convolutional layer are calculated through the formula:

$$O = \frac{(in_{size} - c_{size}) + 2P}{c_{stride}} + 1 \tag{6.6}$$

where in_{size} is the input image size in the input layer, c_{size} is the size of each feature in the convolutional layer, c_{stride} is the size of the stride in the convolutional layer, and P is the padding. O is the convolutional output size that represents the squared root of the number of neurons in the convolutional layer.

The third layer represents an inhibitory layer where feature neurons are inhibited apart

from the one that a neuron is connected to. The number of neurons in the inhibitory layer is proportional to the number of patches in the excitatory layer.

For **FER tasks**, the chosen network configuration consists of a convolutional layer containing 50 features, with a stride size of 15 and convolution size of 15. This configuration was retained as it performed the best. The input data are all resized to 100×100. Thus, the number of neurons in the input layers is set to 10000. At the convolutional layer, the number of neurons is calculated using the chosen number of strides and convolution size according to Equation 6.6.

For **SER tasks**, two different network architectures are experimented. The first approach consists of splitting the input into different frames where each frame represents an input to the network with a 1D convolution layer. The second experimented approach takes extracted features such as MFCCs and input them to the convolution layer running across the temporal axis. The input layer of the network architecture consists of a group of neurons representing audio spike-train with a dimension of 40×388. The input layer is then connected to an excitatory convolution layer which is laterally connected to an inhibitory layer. Each audio input is divided into convolutional windows where a stride window moves through the temporal axis of the audio input. The initial number of features is set to 60 with a convolution window size and stride size to 10.

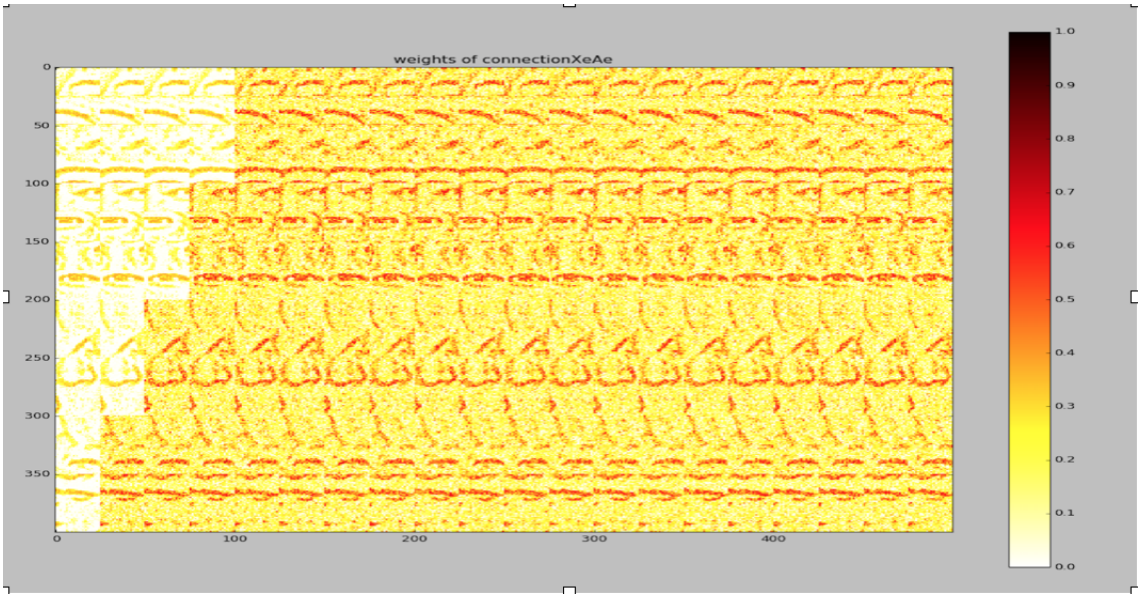


Figure 6.11: Weight learning for FER through convolution SNN with size 25, stride 25, and feature size 20

Network learning Weights are learned through STDP by either being increased when a post-synaptic neuron fires after a spike reaches a synapse, or decreased when the post-synaptic spike fires before a spike arrives at a synapse. Figure 6.11 shows an example of features learning through time with a configuration of 20 convolution features with a window size of 25 and stride size of 25. This configuration is too coarse to capture fine-grained features, therefore the actual configuration used experiments in this thesis is of a larger feature size 50 with the smaller convolution size 15 and the smaller stride size 15. When an input is presented for 350ms, spikes are recorded for both excitatory and inhibitory layers as shown in Figure 6.12, where a group of neurons spike for different features. The network weights are updated after each 100 input interval.

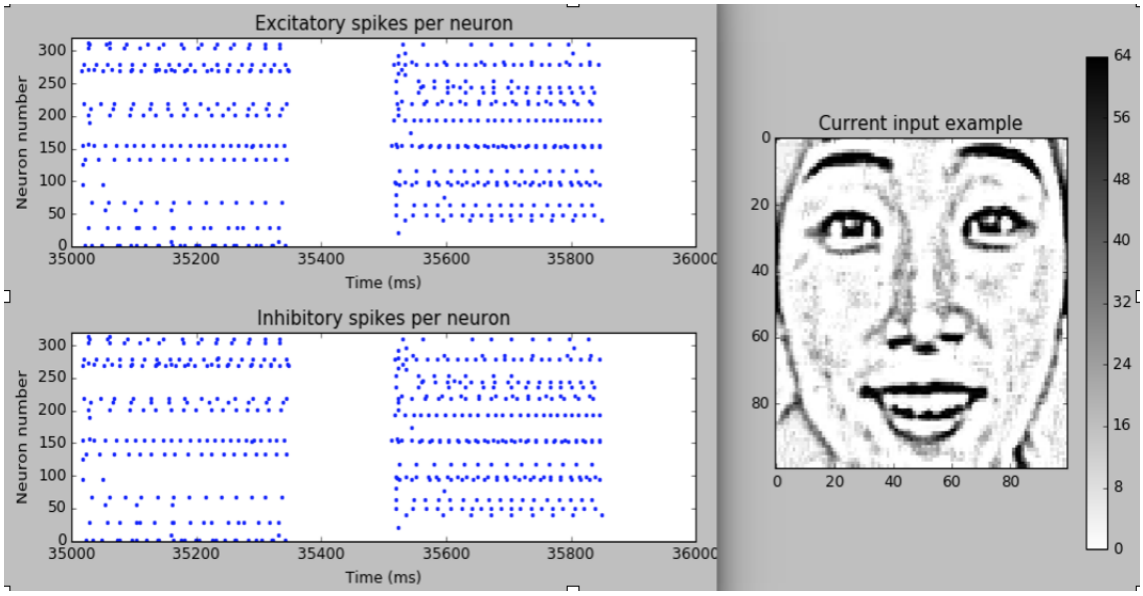


Figure 6.12: Spike activity in excitatory and inhibitory layer for FER through convolution SNN

Bio-inspired architectures for FER and SER are implemented using the SNN simulator BRIAN [92]. We use the same network parameters in terms of input firing rates, membrane threshold and resting phase duration as the work presented in [65]. These are detailed in Table 6.4. Classification and labelling works by assigning labels to most spiked neurons. After each training interval, neurons groups representing individual input features are assigned to a label class according to their spiking activity.

Table 6.4: SNN implementation parameters

Parameter	Description	Value
τ	Membrane time constant	20 ms
E_{rest}	Inhibitory rest potential	-60 mV
E_{rest}	Excitatory rest potential	-65 mV
E_i	Inhibitory equilibrium potential	-25 mV
E_e	Excitatory equilibrium potential	-25 mV
V_e	Excitatory threshold potential	-65 mV
V_i	Inhibitory threshold potential	-60 mV
$nu_e e_p re$	Pre-synaptic learning rate	0.0001
$nu_e e_p ost$	Post-synaptic learning rate	0.01

Table 6.5: Convolution parameters in unisensory SNN

Experiment	**Features**	**Convolution Window**	**Sride**
FER	50	15	15
SER	80	10	10

6.7.2 SNN Convolution Parameters

Table 6.5 summarises the convolution window parameters in facial expression and speech emotion models. Figure 6.13 presents how facial features are learnt over time in SNN. Each feature is learnt by a group of neurons, where weights are updated according to their spikes.

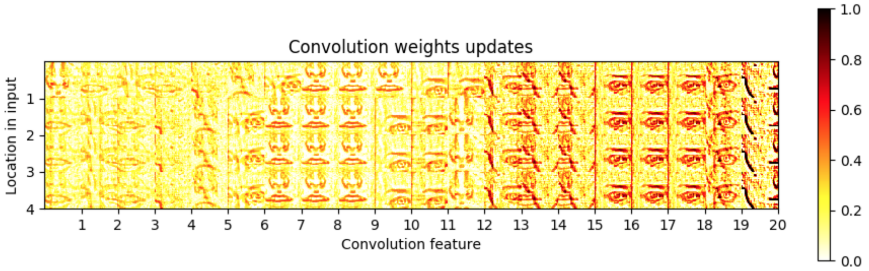


Figure 6.13: Facial feature learning over time in SNN

6.8 Multisensory Configuration

In this section, we detail the configurations for each of the multisensory integration models, which will build on the above configuration of unisensory modalities; that is, from raw input to features, to spike-trains, neurons models and their STDP unsupervised learning. Each integration model will take sensory modalities spike-trains as input.

6.8.1 Feature Extraction and Encoding

Feature extraction and encoding follows the same process as the one described in Section 6.7. Features from audio and visual modalities are extracted separately. Features are extracted from two datasets, eNTERFACE'05 and RAVDESS. Both dataset contain balanced number in each emotion class.

6.8.2 Convergence Setup

This section describes different design processes for evaluating the proposed models in this thesis and answering the main research questions.

Similarly to unisensory SNN implementation, the general network dynamics for the convergence are governed by the LIF model, and learning uses unsupervised STDP as detailed in Section 6.7.1.3. Although there are some similarities with unisensory models architectures, modelling the proposed multisensory integration by convergence goes through various complex steps, and the network comprises the following layers:

Input layer

Unlike unisensory SNNs, the input layer consists of two distinct neuron groups each representing a modality. After features extraction from each modality through MFCCs for audio and LoG for visual data, inputs are encoded into meaningful Poisson spike train input for SNN networks.

Excitatory layer

The excitatory layer comprises three distinct neuron groups. The first two groups define each modality. The third is a multisensory group where integration happens. Visual neuron group receives input from the visual input neuron group, and excitatory audio group receives input

from the audio input group. Multisensory excitatory neurons receive connections from both the visual and audio excitatory neurons. Similar to the unisensory model, a convolution layer is applied. Neurons are connected to all neurons in the excitatory layer apart from the one receiving information. 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, the visual and audio both have different configuration in terms of convolution windows, the number of features and the total excitatory neurons. We have experimented with various configurations and have chosen the best-performing ones using 10 for the window an stride size in the auditory modality and 10 for the visual. We set the number of features to 60 for the auditory modality and 60 for the visual modality.

Inhibitory layer

The inhibitory layer contains three distinct neuron groups. Two neuron groups are connected laterally to each excitatory group in the corresponding modality. Then this group is connected to the multisensory neuron group.

Connections

First, connections are set between input groups in each modality and their corresponding excitatory neuron group. The second set of connections links the excitatory neuron group of each modality to their corresponding inhibitory group. Neurons in the multisensory excitatory group are also linked to the corresponding inhibitory group. A third set of connections are set between excitatory neurons in each modality and the multisensory excitatory group. There is no direct link between neurons from unisensory modalities. The main learning happens in the multisensory convergence area, where it receives inputs from both modalities. During the learning phase each modality drives information into the multisensory area. The multisensory area also sends feedback for each unisensory area. When neurons in each group spike for the same features, their connections get stronger.

The number of neurons for each group in the input layer is equivalent to the dimension of the input. The input dimensions are set to 40×388 for audio and 100×100 for the visual inputs. Similar to unisensory models, a convolution layer is added in each modality and the number of features is set to 60 and the window size is set to 10 for both modalities.

Neurons in the excitatory layer receive input from the input layer through connections. The number of neurons in the excitatory layer corresponds to the total number of neurons in the auditory and visual groups. The multisensory neuron group contains connections receiving information from both the visual and auditory excitatory groups. SNN is trained by first receiving input from the visual modalities through the pre-processing video frames. It then received information from the auditory input. Patterns are learned through the excitatory layer using an unsupervised STDP learning. Weights are initialised randomly and are identical for all connections in the network. Neurons spiking for the same labels have stronger connections between audio and multisensory and visual and multisensory groups. Therefore patterns are learned mainly in the multisensory excitatory neuron group.

6.8.3 Enhancement Setup

Modelling multisensory through cross-modal enhancement follows a similar approach as modelling multisensory integration through convergence, the general network dynamics for the multisensory integration by convergence. Learning is governed by the LIF model and STDP. The setup of the enhancement architecture is described as follows.:

Input layer

The input layer consists of two distinct neuron groups each representing a modality. After features extraction from each modality through MFCC sfor audio and LoG for visual data, inputs are encoded into meaningful Poisson spike train input for SNN networks.

Excitatory layer

The excitatory layer contains two distinct neuron groups representing each modality. Visual neurons group receive input from the visual input neuron group. Audio excitatory group receives input from the audio input group. Similar to the convergence model, convolution window size and features number share the same parameters with 10 for the window and stride size and 60 features.

Inhibitory layer

The inhibitory layer contains three distinct neuron groups. Two neuron groups are connected laterally to each excitatory group for each modality. Then this group is connected to the multisensory neuron group.

Connections

Connections are set between different layers and neuron groups. First connections are set between inputs of each modality and their corresponding excitatory neuron group. The second set of connections links the excitatory layer of each modality to their corresponding inhibitory neurons group. To model cross-modal enhancement, connections are set from visual excitatory modality neurons to audio excitatory modality neurons. The connections enable learning from preceding visual information.

After processing the visual frames input, 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.

6.8.4 Synchrony Setup

This section describes experiments conducted to evaluate the third proposed model for multisensory integration through neural synchrony. This method consists of two main steps:

1) Training a multisensory neural network and recording spiking activities of audio and visual modalities.

2) Constructing neural synchrony graph, and implementing Graph Convolution Network (GCN) for the classification of emotion through neural synchrony patterns.

Network topology Neural synchrony network is implemented through SNN using LIF for neurons behaviour modelling and STDP learning, as detailed in Section 6.7. Although this model shares the main learning functions and network dynamics with both unisensory models, multisensory by convergence and cross-modal enhancement models, there are differences mainly in the network topology and connections set between various neuron groups.

Input layer

The input layer consists of two distinct neuron groups each representing a modality. After features extraction from each modality through MFCCs for audio and LoG for visual data, inputs are encoded into meaningful Poisson spike train input for SNN networks.

Excitatory layer

The excitatory layer receives connections from the input layer and comprises two distinct excitatory groups. Each excitatory group represents a modality :that is; audio and visual. Excitatory groups receive input directory from the input layer groups. In this model we apply a convolution window and stride size of 20 and number of features of 40. These parameters are different from the two previous models; convergence and enhancement due to computational cost.

Inhibitory layer

The inhibitory layer contains two distinct neuron groups. Two neuron groups are connected laterally to each excitatory group for each modality.

Connections

Connections between modalities are set at the excitatory level. Recurrent connections are set between audio and visual excitatory neuron groups. This enables cross-talk between modality between the whole learning process. Weights are learning using STDP learning.

6.8.4.1 Implementation and Network Configuration

We have designed the excitatory layer as a convolution layer for a better feature representation [179], [247]. After features extraction and transforming inputs into Poisson spike trains [65] for both audio and visual, we set the number of neurons for each group proportionally to the dimension of inputs; that is, 40×388 for audio and 100×100 for visual neuron groups.

We have set convolution parameters separately each modality with a convolutional window of 40 with an initial number of features of 20 for each modality. 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. Recurrent connections between modalities are applied at the excitatory layer. This will enable the cross-talk between audio and visual modalities and help simulate multisensory interaction where modalities influence each other during the learning process. After training the network on the whole dataset, spike timing and locations are recorded for the whole dataset.

6.8.4.2 Graph Neural Network For Neural Synchrony

The constructed neural synchrony graph on RAVDESS dataset consists of 814 sub-graphs and 130008 nodes in total. On the eNTERFACE'05 dataset we have obtained 1260 sub-graphs and 201600 nodes in total. After obtaining the basic structure for each graph, the input is prepared for the GCN. Three-layer GCN is trained with a semi-supervised learning and have initialised the weights randomly [145]. To determine the appropriate architecture for GCN, we have experimented with different numbers of layers.

Figure 6.15 shows the learning curve of neural synchrony graph on training and validation data for a three layer network. The loss for training and validation decreases to around 200 epochs and then stabilises. The figure also shows that there is a small gap between the training and validation loss compared to the original architecture with two layers from [145] in Figure 6.14. Figure 6.14 shows that the gap of the loss is bigger between validation and training meaning that when using only two layers, the model is underfitting and needs more training epochs. When applying GCN with only two layers to graph classification, they have a tendency to underfit, this has also been demonstrated in [304]. Having a third layer increases the learning capacity of the network as the training and validation loss converge quicker than with a 2 layers network.

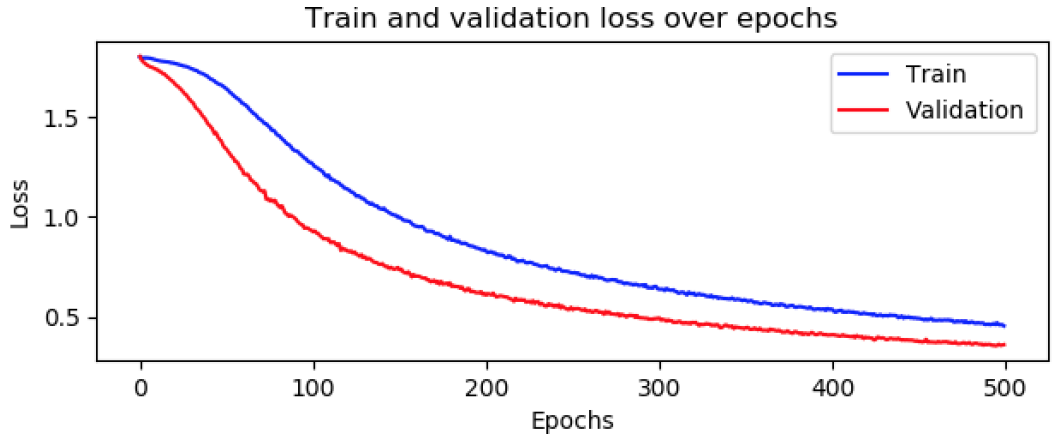


Figure 6.14: Loss on 2-layer GCN with RAVDESS dataset

Input is randomly shuffled and split by 60% for training 20% for validation and 20% for testing. 10 trials are conducted with results detailed in Appendix A. The model uses the following parameters:

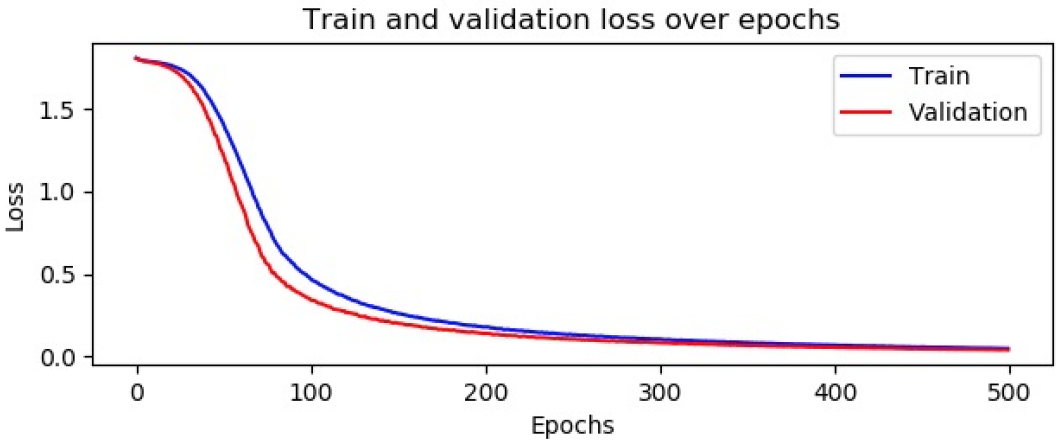


Figure 6.15: Loss on 3-layer GCN with RAVDESS dataset

- Adam optimisation
- Leaning rate of 0.0001. There was an experimentation of learning rate from 0.01, and the best performing rate is selected.
- We experimented with two and three layers with 64 units.
- The network initial training epochs are 200. Then increased to 500 epochs.
- Dropout rate is set to 0.5.

6.8.5 Multisensory SNN Convolution Parameters

This section summarises the convolution window parameters used in different experiments presented in this thesis. The choice of convolution window size and number of features affects the accuracy of models as detailed in 7. The higher the features and the smaller the convolution window leads to higher accuracy. Figure 6.16 shows an example of weight learning with a convolution window of 40 and the number of feature of 10. However, having higher values of features is computationally costly and leads to slow processing times.

Table 6.6: SNN implementation convolution parameters

Experiment	Features	Convolution Window	Stride
Convergence	60	10	10
enhancement	60	10	10
Synchrony	40	20	20

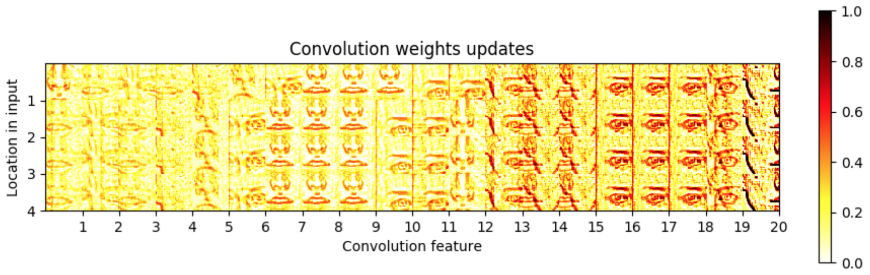


Figure 6.16: Facial feature learning over time with wider convolution windows

6.9 Summary

This chapter has described all experimental setups used in this thesis. It first outlines the datasets used for the evaluation of the proposed bio-inspired models. It then outlines the different type of software and tools used in this thesis. After that, a description of different video, audio and image processing and feature extraction techniques used. It also provides details on the configuration of Spiking Neural Network used in all experiments. Finally, this chapter details the implementation setup for all proposed models in this thesis for both unisensory and multisensory emotion recognition tasks. The next chapters 7 and 8 will present the evaluation results for unisensory and multisensory models respectively.

Chapter 7

Results and Discussion On Unisensory Emotion Recognition

7.1 Introduction

Chapter 4 has introduced novel models for applying bio-inspired architectures for emotions and social signals recognition. It describes the application of bio-inspired architectures using SNN for unisensory emotions recognition from facial expressions and non-verbal speech emotion recognition tasks. Using the design and architectures described in Chapter 4 and experimental setup described in Chapter 6, several experiments are conducted to evaluate the proposed models for unisensory emotion recognition. This chapter presents the evaluation results for unisensory models for facial expression and speech emotion recognition. It also presents evaluations of generalisation capacity and investigate robustness to visual and auditory noise. Part of the experiments in this chapters have been published in [179] and [180].

7.2 Facial Emotion Recognition (FER)

This section will present and discuss the results on facial emotion recognition in terms of accuracy (in Section 7.2.1), cross-corpus generalisation (in Section 7.2.2), and robustness to noise (in Section 7.2.3). To do so, we compare with a range of state-of-the-art techniques. We have considered a range of manual and automatic feature extraction techniques, including HOG,

LBP and geometrical/ coordinates based features applied with a SVM classifier [205], [57]. We have also experimented the implementation of a CNN and training the last layer of a pre-trained network [141]. The models are described in Chapter 6.

We present results on CK+ and JAFEE datasets. On each dataset, we apply repeated holdout with 10 trials by splitting data into 60% is used for training, 20% for validation and 20% for testing in CNN based models and 80% for training and 20% for testing in SNN and SVM models. Data is shuffled randomly with a balanced distribution within classes on both training and testing data. We obtain the accuracy for the 10 trials and average the accuracy. Appendix A presents the results for the full trials.

For the SNN we use the parameters described in 6.4. For the SVM we use the linear kernel. For the CNN we use the architecture and setting described in Chapter 6.6.1. The pre-trained CNN model starts by extracting features from VGG16. We have done some hyperparameters tuning to find the optimal parameters for the model. We have experimented on parameters such as dropout (0.25 -0.5), learning rate (0.1-0.0001). We choose the activation function softmax, and adam optimizer. We experience with epoch from 250 too 500. For the final model we train it for 500 epochs and set the dropout value to .25 and learning rate 0.001.

7.2.1 FER Accuracy

Figure 7.1 shows FER accuracy for SNN, HOG, and CNN on JAFFE and CK+ datasets. On CK+ dataset, SNN achieves an average accuracy of 95.0%, which outperforms CNN by 14% while lower than the HOG+SVM model by 4%.

On JAFFE dataset, SNN achieves an average recognition accuracy of 94.0% , similar to HOG+SVM and exceeds CNN by 23%. CNN model experiences the lowest performance which is mainly due to the small training size of JAFFE dataset compared to CK+ dataset; that is, not enough to train the network to generate effective feature representations without any data augmentation or pre-processing [167].

Tables 7.1, 7.2, 7.3, 7.4 present the confusion matrices of SNN and CNN on CK+ and JAFFE datasets. On JAFFE dataset, the highest accuracy is 100% for all emotion classes apart from the class 'sad' where the accuracy is 66.7%. The same pattern can be noticed for CNN accuracy on JAFFE dataset in Table 7.4 where the lowest accuracy is for the 'sad' class with 33.0% accuracy

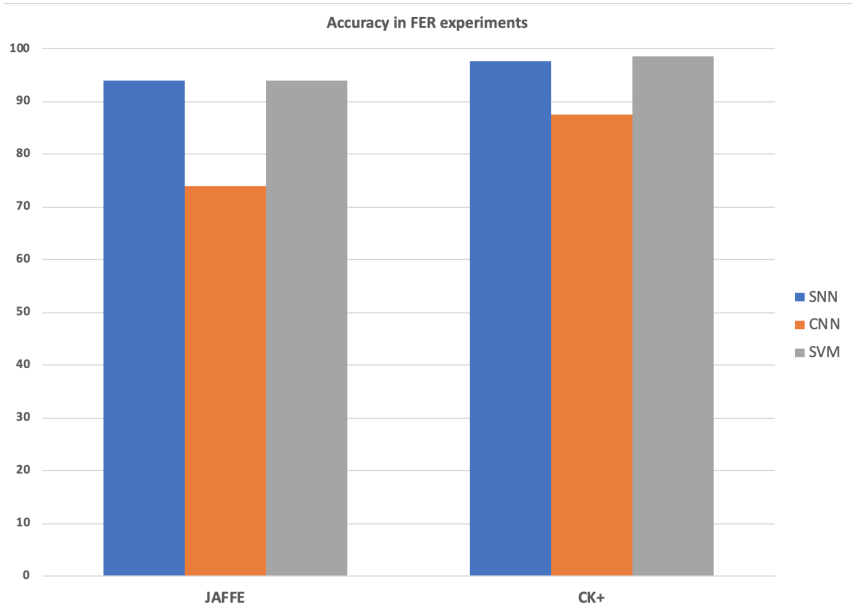


Figure 7.1: Comparision of FER accuracy on SNN, HOG+SVM and CNN with models on CK+ and JAFFE

and 66.7% are either classified as 'happy', 'fearful' or 'disgusted'.

On CK+ dataset, the highest accuracy for SNN is 97.3% on 'disgusted', and the lowest accuracy is 90.5% on 'fearful'. SNN exhibits a more balanced accuracy for individual emotions classes, where for both datasets the accuracy are over 60%. Whereas CNN exhibits very low accuracy in some emotions labels for both datasets such as 26.0% for 'angry' in CK+ and 'sad' with 33.3 %.

7.2.2 FER Cross-corpus Generalisation Results

We have performed cross-datset generalisation experiments by training models on one dataset and testing them using a different dataset with different distribution of data. Figure 7.2 presents the accuracy of SNN, HOG, and CNN on generalisation capacity with cross-dataset validation. In both cases, SNN has achieved consistently high accuracy: 85.0% – trained on CK+ and tested on JAFFE, and 92.0% – trained on JAFFE and tested on CK+, which significantly exceed the HOG+SVM and CNN techniques.

Tables 7.5, 7.6, and 7.7 present the confusion matrices of SNN, CNN and HOG+SVM on cross-dataset validation. SNN has the best performance in all classes compared to CNN and HOG+SVM. The highest class accuracy for both methods is 'surprise', where SNN achieved

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	96.3	0.0	0.0	0.0	3.7	0.0
Disgusted	0.0	97.3	0.0	0.0	0.0	2.7
Fearful	1.6	0.0	90.5	1.6	1.6	4.8
Happy	0.0	0.7	0.0	97.2	0.0	2.1
Sad	5.9	0.0	0.0	0.0	94.1	0.0
Surprised	3.6	0.0	0.7	1.4	3.6	90.6

Table 7.1: Confusion matrix SNN on the CK+ dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.0	0.0	0.0	0.0	0.0	0.0
Disgusted	0.0	100.0	0.0	0.0	0.0	0.0
Fearful	0.0	0.0	100.0	0.0	0.0	0.0
Happy	0.0	0.0	0.0	100.0	0.0	0.0
Sad	16.7	0.0	16.7	0.0	66.7	0.0
Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table 7.2: Confusion matrix for SNN on the JAFFE dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	26.0	8.0	4.0	0.0	62.0	0.0
Disgusted	4.0	94.0	0.0	1.0	1.0	0.0
Fearful	0.03	0.0	78.0	0.0	14.0	8.0
Happy	0.0	0.0	7.0	92.0	0.0	0.0
Sad	0.0	0.0	0.0	0.0	100.0	0.0
Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table 7.3: Confusion matrix for CNN on the CK+ dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.00	0.0	0.0	0.0	0.0	0.0
Disgusted	0.0	60.0	20.0	0.0	20.0	0.0
Fearful	0.0	0.0	83.3	0.0	0.0	16.7
Happy	0.0	16.7	0.0	83.3	0.0	0.0
Sad	0.0	16.7	16.7	33.3	33.3	0.0
Surprised	0.0	0.0	0.0	33.3	0.0	66.7

Table 7.4: Confusion Matrix for CNN on JAFFE dataset

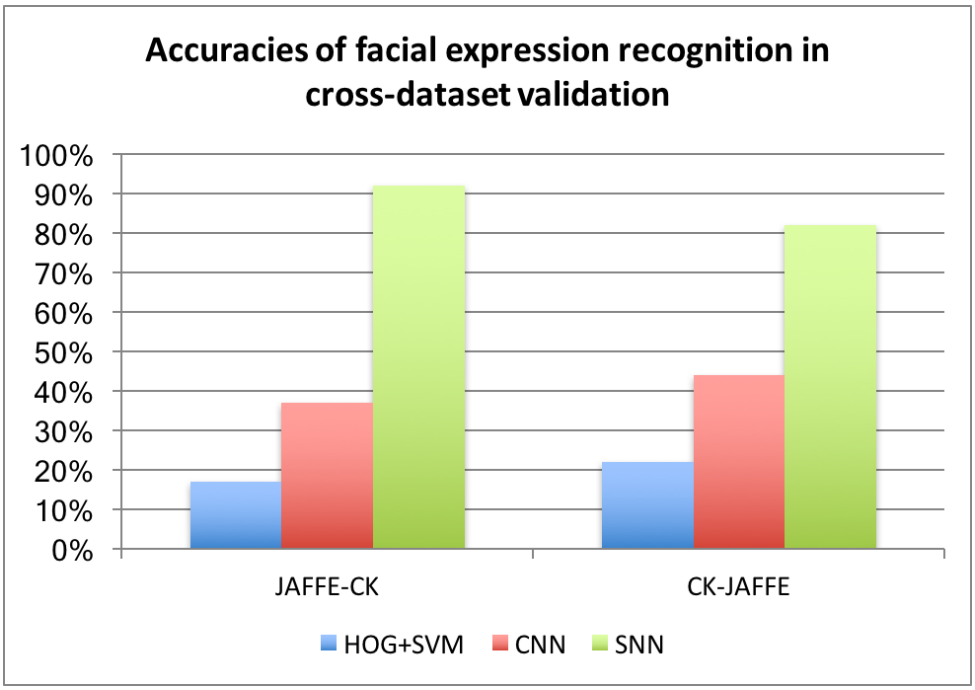


Figure 7.2: Comparison of FER accuracy on SNN, HOG+SVM and CNN with models on cross-dataset

100.0% and CNN 75.0%. Whereas the highest class accuracy using HOG features is 'fearful', and all classes are mainly classified as 'fearful'.

The supervised learning used in both CNN and SVM expects training and testing data to have the same distribution and are more biased by the dataset used for training. They also work better with larger datasets. Using JAFFE dataset with only ten subjects has a negative impact on the accuracy for CNN and SVM, due to limited variation in faces, facial expressions and cultural differences. JAFFE dataset has exclusively Japanese females subjects, whereas the CK+ dataset includes more diverse subjects. Similar findings have also been reported in [278, 167]. SNN accuracy does not seem to be affected much by this issue. The combination of applying Laplacian Of Gaussian (LoG) filter, unsupervised learning, and convolutional layer enables the model to generalise well without expecting the same distribution of the data, and the accuracy is dependent on the number of features/patches chosen. LoG filters help define contours and highlight key facial features.

		Predicted					
		Angry	Disgusted	Fearful	Happy	Sad	Surprised
Actual	Angry	91.7	8.3	0.0	0.0	0.0	0.0
	Disgusted	4.3	91.3	0.0	0.0	0.0	4.3
	Fearful	4.0	0.0	80.0	0.0	0.0	16.0
	Happy	4.2	0.0	4.2	91.7	0.0	0.0
	Sad	0.0	8.3	12.5	4.2	70.8	4.2
	Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table 7.5: Confusion matrix for generalisation SNN trained on CK+ and tested on JAFFE

		Predicted					
		Angry	Disgusted	Fearful	Happy	Sad	Surprised
Actual	Angry	4.2	29.2	12.5	8.3	45.8	0.0
	Disgusted	0.0	47.8	17.4	21.7	8.7	4.3
	Fearful	0.0	0.0	48.0	12.0	20.0	20.0
	Happy	0.0	0.0	41.7	54.2	0.0	4.2
	Sad	4.2	8.3	37.5	16.7	33.3	0.0
	Surprised	0.0	0.01	20.8	4.2	0.0	75.0

Table 7.6: Confusion matrix for generalisation CNN trained on CK+ and tested on JAFFE

		Predicted					
		Angry	Disgusted	Fearful	Happy	Sad	Surprised
Actual	Angry	1.0	44.0	44.0	11.0	0.0	0.0
	Disgusted	0.0	12.0	88.0	0.0	0.0	0.0
	Fearful	0.0	0.0	67.0	33.0	0.0	0.0
	Happy	0.0	0.0	89.0	11.0	0.0	0.0
	Sad	0.0	33.0	44.0	22.0	0.0	0.0
	Surprised	0.0	0.0	100.0	0.0	0.0	0.0

Table 7.7: Confusion matrix for generalisation HOG+SVM trained on CK+ and tested on JAFFE

7.2.3 FER Robustness To Noise Results

Various types of noise have been used in the literature to assess the sensitivity of models for image recognition tasks. There exist various ways of assessing model robustness to image degradation such as colours changing, noise such as salt and pepper or Gaussian noise [128]. Noise degradation is also used to assess the sensitivity of different CNN models (ALexNet, VGG, and GoogleNet) [128].

We have experimented with different intensity parameters of salt and pepper noise degradation ranging from 0 to .5. Salt and pepper noise represent intensity and sparse disturbances to an image where original pixels are randomly replaced with black and white pixels. from 0.5 to 1 noise intensity probability, we have noticed that the image is completely covered, thus not any

more useful to get more insight of the performance. Figure 7.3 shows the samples of salt and pepper noise degradation of input images.

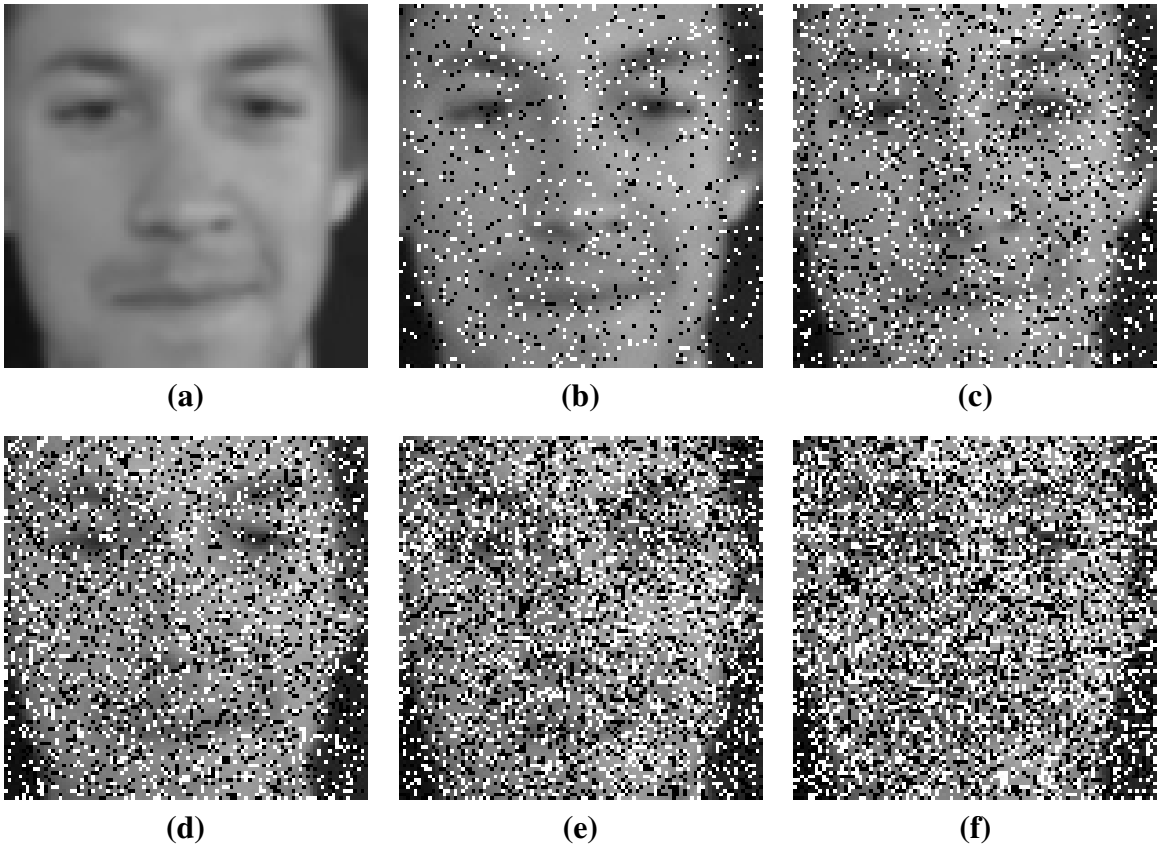


Figure 7.3: (a) Image no noise, (b) 0.1 noise probability, (c) 0.2 noise probability, (d) 0.3 noise probability, (e) 0.4 noise probability, and (f) 0.5 noise probability.

Results for FER noise degradation tasks are summarised in Figure 7.4. The initial results of the three models where no noise is applied are quite close. SVM model experiences the highest accuracy with 99.6% overall, followed by CNN and SNN with 97.63% and 97.43% respectively.

Starting from the lowest probability of noise degradation of 0.2% we notice a drop in the overall accuracy for all three models. The drop for the SNN model down to 92.39% is not as significant as the drop in CNN to 84.96% or the significant drop for the SVM model to 32.6%. The higher noise intensity results in a lower overall accuracy for all three models. However, SNN performs best for all noise intensities. The lowest accuracy for SNN is using the 0.5 probability distribution of noise with only 56.2%. However, the lowest accuracy for CNN and SVM is significantly weaker: 22.64% and 14.10% respectively. SVM is the most affected by the artificial noise degradation. The drop in accuracy pattern in all models does follow the results obtained

in [238] and [127], where the increase of noise affect feature identification. Although accuracy has dropped for SNN, it still maintains an accuracy over 65% up to the noise intensity of 0.4. whereas we notice a quicker drop for the other models starting from intensity .1 for CNN and .02 for SVM. Figure 7.4 presents the trend of accuracy decreasing with the increase of noise ratio.

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.0	0.0	0.0	0.0	0.0	0.0
Disgusted	0.0	100	0.0	0.0	0.0	0.0
Fearful	0.0	0.0	100.0	0.0	0.0	0.0
Happy	0.0	0.0	0.0	100.0	0.0	0.0
Sad	0.0	0.0	2.0	0.0	98.0	0.0
Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table 7.8: Confusion matrix for SVM for FER task with no noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.0	0.0	0.0	0.0	0.0	0.0
Disgusted	99.0	0.0	1.0	0.0	0.0	0.0
Fearful	92.0	0.0	8.0	0.0	0.0	0.0
Happy	99.0	0.0	1.0	0.0	0.0	0.0
Sad	98.0	0.0	2.0	0.0	0.0	0.0
Surprised	98.0	0.0	0.0	0.0	0.0	0.0

Table 7.9: Confusion matrix for SVM for FER task with 0.1 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	97.0	0.0	0.0	3.0	0.0	0.0
Disgusted	97.0	0.0	0.0	3.0	0.0	0.0
Fearful	100.0	0.0	0.0	0.0	0.0	0.0
Happy	99.0	0.0	0.0	1.0	0.0	0.0
Sad	100.0	0.0	0.0	0.0	0.0	0.0
Surprised	98.0	0.0	1.0	1.0	0.0	0.0

Table 7.10: Confusion matrix for SVM for FER task with 0.2 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.0	0.0	0.0	0.0	0.0	0.0
Disgusted	100.0	0.0	0.0	0.0	0.0	0.0
Fearful	98.0	0.0	0.0	0.0	2.0	0.0
Happy	100.0	0.0	0.0	0.0	0.0	0.0
Sad	98.0	0.0	0.0	0.0	2.0	0.0
Surprised	99.0	0.0	0.0	0.0	1.0	0.0

Table 7.11: Confusion matrix for SVM for FER task with 0.5 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.0	0.0	0.0	0.0	0.0	0.0
Disgusted	1.3	97.3	0.0	1.3	0.0	0.0
Fearful	0.0	0.0	100.0	0.0	0.0	0.0
Happy	0.0	0.0	0.7	99.3	0.0	0.0
Sad	0.0	0.0	0.0	0.0	98.0	0.0
Surprised	0.0	0.0	0.0	0.0	1.4	97.8

Table 7.12: Confusion matrix for CNN for FER task with no noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	16.3	3.7	30.0	27.5	15.0	7.5
Disgusted	0.0	36.0	14.7	41.3	0.0	8.0
Fearful	0.0	0.0	65.1	33.3	0.0	1.6
Happy	0.0	0.0	1.4	98.6	0.0	0.0
Sad	0.0	2.0	33.3	7.8	43.1	13.7
Surprised	0.0	0.0	13.0	7.2	2.2	77.5

Table 7.13: Confusion matrix for CNN for FER task with 0.1 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	1.3	0.0	25.0	53.7	3.7	16.3
Disgusted	0.0	5.3	5.3	78.7	1.3	9.3
Fearful	0.0	0.0	31.7	68.3	0.0	0.0
Happy	0.0	0.0	2.1	97.9	0.0	0.0
Sad	0.0	2.0	41.2	45.1	2.0	9.8
Surprised	0.0	0.0	20.3	46.4	0.0	33.3

Table 7.14: Confusion matrix for CNN for FER task with 0.2 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	0.0	0.0	16.3	82.5	0.0	1.3
Disgusted	0.0	0.0	32.0	65.3	0.0	2.7
Fearful	0.0	0.0	28.6	71.4	0.0	0.0
Happy	0.0	0.0	25.5	71.0	0.0	3.4
Sad	0.0	0.0	25.5	70.6	0.0	3.9
Surprised	0.0	0.0	22.9	73.2	0.0	2.9

Table 7.15: Confusion matrix for CNN for FER task with 0.5 noise

For individual emotions classes, Tables 7.8 7.9, 7.10 and 7.11 show confusion matrices for noise application on the SVM model. Tables 7.12, 7.13, 7.14 and 7.15 show confusion matrices for noise application on the CNN model. Tables 7.16, 7.17, 7.18 and 7.19 show confusion matrices for noise application on the SNN model. All results show that the accuracy of SNN remains stable within different classes. The accuracy drops significantly for the SVM commencing from the noise probability of 0.1 for all classes, where all tests data are classified as 'angry'. SVM is less resilient to noise degradation and exhibits less robustness to image degradation. SNN is the most robust to noise compared to SVM and CNN.

7.3 Speech Emotion Recognition (SER)

This section presents results and discussion on speech emotion recognition in terms of accuracy (in Section 7.3.1), generalisation (in Section 7.3.2), and robustness to noise (in Section 7.3.3). Data experimental setup is described in Chapter 6.5. We follow repeated holdout with 10 trials with data split in 80% and 20% for SNN and SVM and 60% training 20% validation and 20% testing for the CNN model. For the SNN we use the parameters described in 6.4. For the SVM we use the linear kernel. For the CNN we use the parameters and setting described in Chapter 6.6.2. We have experienced various hyperparameters in CNN have been experienced with various parameters in terms of learning rate, dropout rate of (0.25 -0.5). We also experimented with various values in learning rate from (0.01- 0.0001). We started with the number of epoch from 500 to 1500. The best performing parameters chosen are 0.0001 for learning rate , dropout of 0.25 and 1500 epoch in training.

7.3.1 SER Accuracy

Table 7.20 presents accuracy for SER using two types of features: Mel-scale spectrogram and MFCCs. MFCCs achieves a higher accuracy than raw audio signals or Mel-scale spectrogram on both eNTERFACE and RAVDESS datasets with 72.2% and 80.29% respectively.

Results show that MFCCs features are more effective audio features for processing speech in SNN, which is also in line with state-of-the-arts methods where MFCCs outperforms other types of audio features in SER tasks [257].

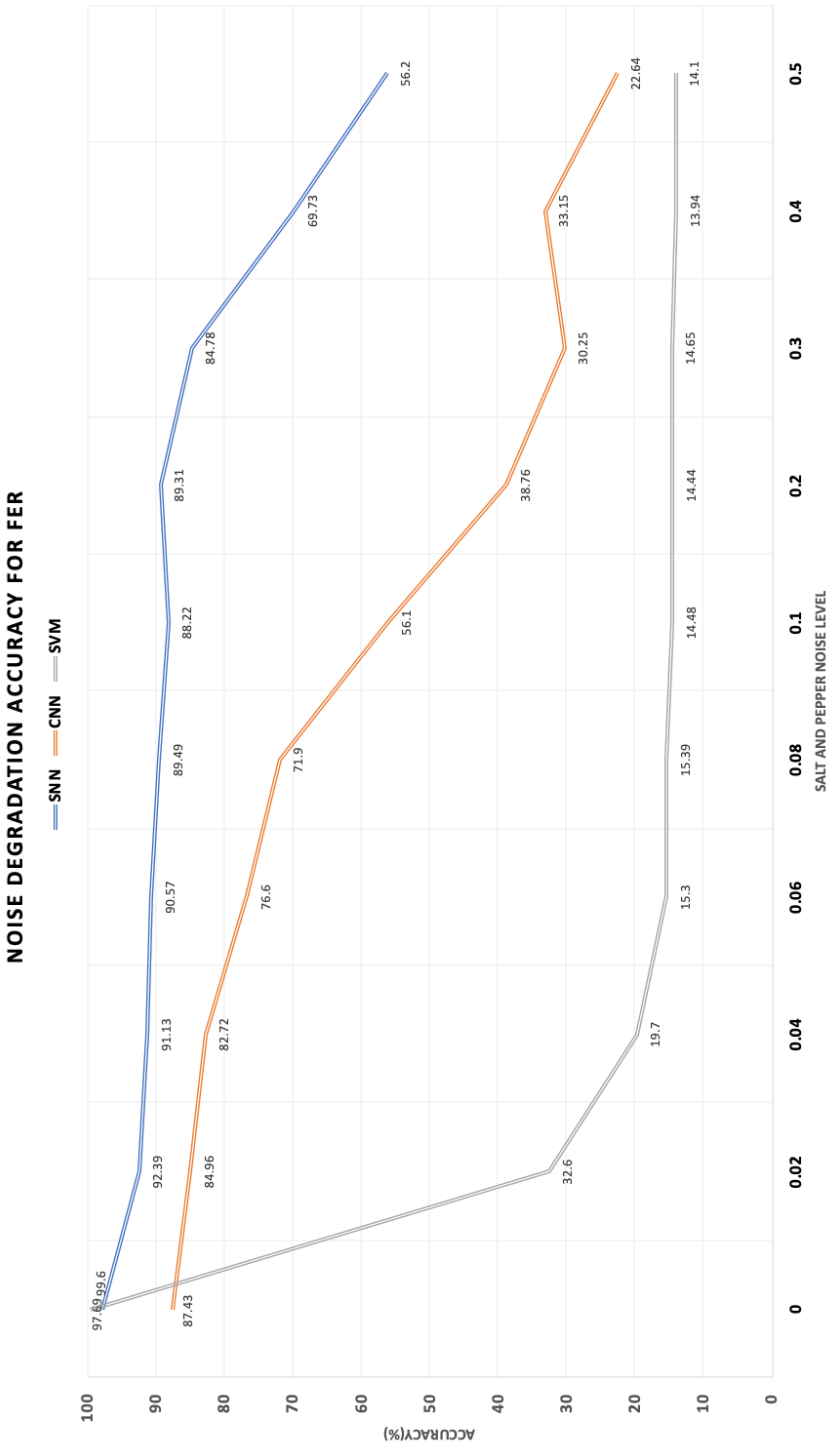


Figure 7.4: Models accuracy with different noise degradation intensity

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	88.7	0.0	1.3	0.0	10.0	0.0
Disgusted	0.0	90.7	1.3	6.7	0.0	1.3
Fearful	0.0	0.0	100.0	0.0	0.0	0.0
Happy	0.0	0.0	0.0	99.3	0.0	0.7
Sad	0.0	0.0	0.0	0.0	100.0	0.0
Surprised	0.0	0.0	0.7	2.2	0.0	97.1

Table 7.16: Confusion matrix for SNN for FER task with no noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	62.5	6.2	0.0	0.0	27.5	3.7
Disgusted	0.0	90.7	0.0	9.3	0.0	0.0
Fearful	0.0	0.0	92.1	1.6	6.3	0.0
Happy	0.0	0.0	0.0	0.0	98.0	2.8
Sad	0.0	0.0	0.0	0.0	98.0	2.0
Surprised	0.0	1.4	0.7	3.6	4.3	89.9

Table 7.17: Confusion matrix for SNN for FER task with 0.1 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	77.5	1.3	0.0	2.50	17.5	1.3
Disgusted	0.0	90.7	2.7	2.7	2.7	1.3
Fearful	6.3	0.0	87.3	0.0	6.3	0.0
Happy	2.1	4.8	3.4	80.0	4.1	5.5
Sad	0.0	0.0	0.0	0.0	100.0	0.0
Surprised	2.2	1.4	3.6	3.6	9.4	79.7

Table 7.18: Confusion matrix for SNN for FER task with 0.2 noise

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	77.5	6.2	7.5	1.3	7.5	0.0
Disgusted	4.0	88.0	4.0	0.0	2.7	1.3
Fearful	4.8	11.1	82.5	0.0	1.6	0.0
Happy	15.9	23.4	20.0	20.7	19.3	0.07
Sad	2.0	5.9	3.9	0.0	88.2	0.0
Surprised	13.0	23.2	12.3	2.2	26.8	22.5

Table 7.19: Confusion matrix for SNN for FER task with 0.5 noise

Table 7.20: Comparison of SER accuracy between Mel-scale spectrogram and MFCC coefficients

Feature	**eNTERFACE (%)**	**RAVDESS (%)**
Mel-scale Spectrogram	43.1	45.1
MFCCs	72.2	80.3

Table 7.21 compares SER accuracy between SNN, SVM and CNN implementations evaluated on RAVDESS dataset. SNN have performed better results than both classical methods using the same features, MFCCs. SNN outperforms CNN by 7% and SVM by 23%.

Table 7.21: Comparison of SER accuracies between SNN, SVM and CNN for RAVDESS dataset

Model	**Feature and classification**	**Learning type**	**Accuracy (%)**
SNN	**MFCCs, SNN**	**Unsupervised**	**80.3**
SVM	MFCCs, SVM	Supervised	60.5
CNN	MFCCs, CNN	Supervised	76.3

Table 7.22 compares SER accuracy between the SNN and the state-of-the-art techniques on eNTERFACE'05 dataset. Experimental setup used in previous work experiments are different from the ones used in this thesis. Noroozi et al [204] and Ozseven et al. [209] has both have used 10-fold cross validation. Fonnegra et al. [77] have use data split with 70% training and 30% testing. Fu et al. [78] have used person-dependent experimental setting with 50% samples for training, 25% for validation and 25% for testing. In this thesis we have used a repeated holdout with 10 trials with 80% training and 20% testing. Although having different experimental settings can have an impact on the overall accuracy [143], it is possible to compare the overall accuracy to existing work using the same datasets.

As an unsupervised learning technique, SNN has produced comparable results, and in some instances, it outperforms some state-of-the-art. The better performing techniques are Enhanced Sparse Local Discriminate Canonical Correlation Analysis (En-SLDCCA) approach proposed by Fu et al. [78], which uses multimodal feature learning representation. Fu et all [78] have produced the best results with 80.1 % with data augmentation, which is out of scope of this thesis. The proposed use of SNN with only one type of features MFCC is comparable to state-of-the-art without the use of any data augmentation techniques or other features. In addition the evaluation setup used in the presented models presents an advantage by using more data for training.

The accuracy of SNN for SER can be enhanced by choosing different parameters for number of features, window and stride size for the convolution window as shown in Figure 7.5).

Results in Figure 7.5 show that the overall accuracy increases when the convolutional size is smaller, and the number of features is higher. Increasing the number of features leads to an increase in the number of excitatory neurons; *i.e.,*, a better accuracy. The pattern is observed

Table 7.22: Comparison of SER accuracy between SNN and the state-of-the-art techniques on the eNTERFACE'05 dataset

Model	**Feature extraction**	**Experimental Setting**	**Accuracy (%)**
Noroozi et al [204]	RF/MFCC	10-fold cross-validation	47.1
Ozseven [209]	Acoustic analysis	10-fold	56.3
Ozseven [209]	Texture analysis of Spectrogram	10 fold cross-validation	60.9
SNN	**MFCCs**	**Repeated holdout**	72.2
Fu et all [78] En-SLDCCA	MFCC	Person-dependent	80.1
Fonnegra et al [77] Spread auto-encoder	MFCC	Data split 70/30	74.0

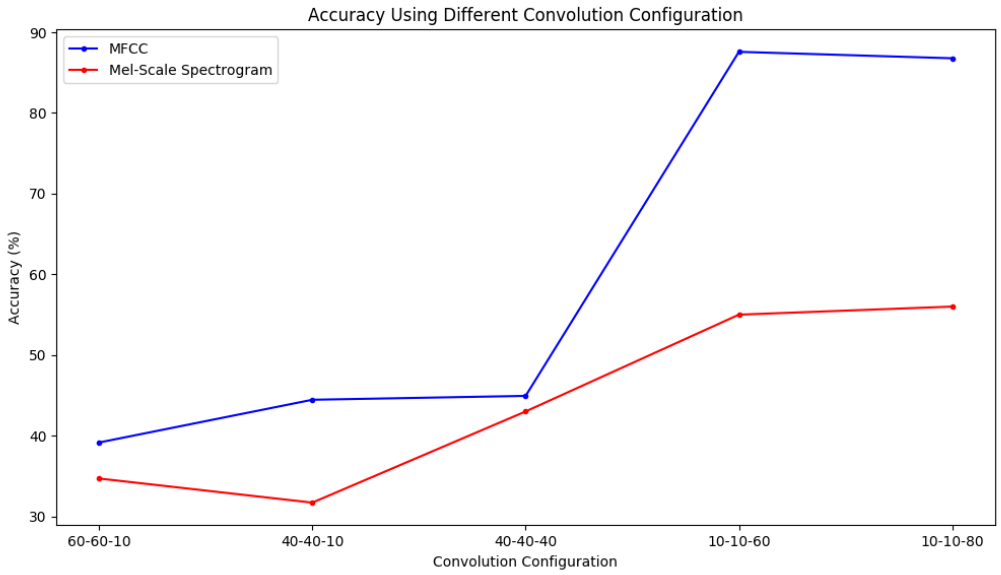


Figure 7.5: Effect of convolution window configuration on overall accuracy

using both MFCCs and Mel-scale Spectrogram features [65]. Having more features and more excitatory neuron leads to learning more features. However, having more excitatory neurons is more computationally costly.

7.3.2 SER Cross-corpus Generalisation Results

This section presents evaluation results for the investigation of generalisation capacity for cross-corpus of SNN on SER tasks. SNN accuracy is compared to baseline methods such as SVM and CNN. Models are trained using RAVDESS and tested using eNTERFACE'05. SNN exhibits the highest performance for both overall accuracy and generalisation, where it performs 69.47% compared to 31.28% and 20.80% for CNN and SVM respectively for RAVDESS dataset. Figure 7.6 shows generalisation results for models trained with RAVDESS and testing with eNTERFACE'05.

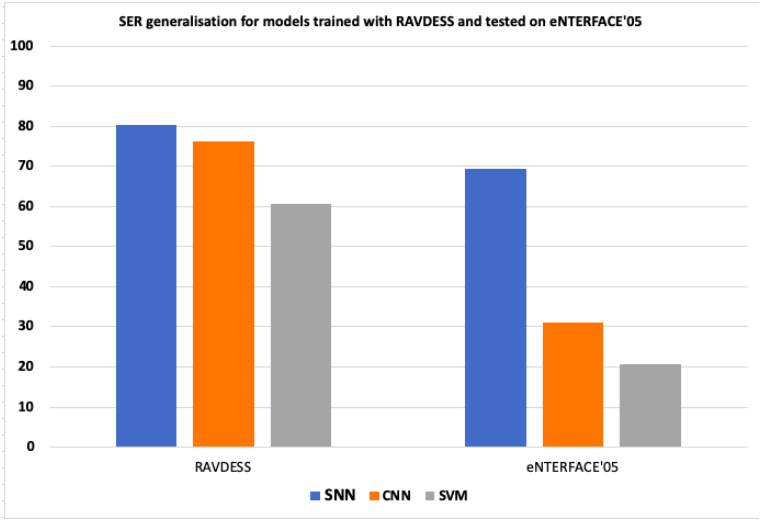


Figure 7.6: Generalisation results for SER tasks where model trained with RAVDESS

Figure 7.7 shows generalisation results for models trained with eNTERFACE'05 and tested on RAVDESS. Models trained with eNTERFACE'05 exhibits the same patterns for results as the models trained with RAVDESS, with SNN outperforming SVM and CNN baselines for generalisation using RAVDESS as a test dataset. That is, SNN achieves an overall accuracy of 70.80 % compared to 68.5% and 58.3% for SVM and CNN respectively. Exploiting the unsupervised learning using SNN and the feature learning using convolution layers, we obtain a more robust model that can learn features.

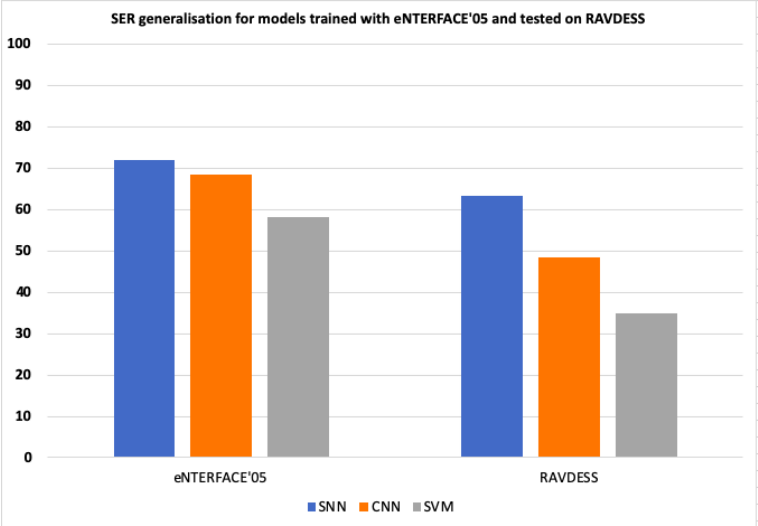


Figure 7.7: Generalisation results for SER tasks where model trained with eNTERFACE'05

7.3.3 SER Noise Robustness Results

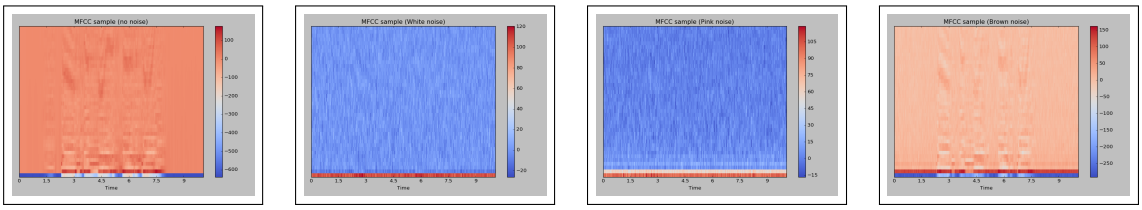
The work in this section investigate the sensitivity and robustness to noise for SER models. Noise is added artificially with three levels of noise with different power spectrum noises such as white, pink and brown noise. White noise is characterised by a flat frequency spectrum, where the noise has an equal power spectrum. Thus the white noise designates flat power. Pink and brown noise are represented by uneven power. These three levels of noise are used in speech recognition tasks to test the effect of noise in real-word error rate [52]. Figure 7.8 shows the effect of the application of different noises on the extracted MFCC features.

Table 7.23: Comparison of SER accuracy for noise degradation tasks for RAVDESS

Model	**No Perturbation (%)**	**Brown Noise (%)**	**White Noise (%)**	**Pink Noise (%)**
SNN	80.3	73.7	80.1	73.3
CNN	76.3	18.0	15.8	16.7
SVM	60.5	23.5	16.0	16.5

Table 7.24: Comparison of SER accuracy for noise degradation tasks for eNTERFACE'05

Model	**No Perturbation (%)**	**Brown Noise (%)**	**White Noise (%)**	**Pink Noise (%)**
SNN	77.2	70.3	68.5	73.2
CNN	68.5	34.4	28.1	32.2
SVM	58.3	32.1	13.5	19.4



(a) (b) (c) (d)

Figure 7.8: (a) MFCC feature with no noise, (b) White noise, (c) Pink noise, and (d) Brown noise.

The results of SER noise degradation experiments on RAVDESS are summarised in Table 7.23. The application of SNN to SER without noise perturbation results in much higher accuracy than CNN and SVM. Test results show an overall accuracy of 85.04% for SNN model, 76.31% for CNN and 60.52% for SVM. Applying noise leads to a significant drop on CNN and SVM accuracy. However, a much less significant drop is noticed in SNN with the lowest accuracy experienced with pink noise at 73.3%. However, the accuracy of CNN drops significantly to lower than 20%.

Table 7.24 shows noise degradation results on eNTERFACE'05. Similar to the results on RAVDESS, we have observe a degradation in accuracy for all audio noise with SNN performing best for the three audio noise effects. Noise affects the overall accuracy of all tested models. However, the less affected model for both tasks is SNN, as with unsupervised learning, it can overcome various degrees of noise degradation for both images and audio inputs. Results are consistently in line with the generalisation tasks results, where the best performing models is SNN.

7.4 Summary

This chapter has presented experiments for the evaluation of bio-inspired architectures for unisensory emotion recognition: visual and audio modalities. Both experiments have shown that SNN, using unsupervised learning technique, has achieved better or at least comparable accuracy to the state-of-the-art supervised learning techniques.

In addition, SNN exhibits better generalisation capability. In FER, facial features learned in the SNN models are less biased by training datasets diversity such as different facial dimensions,

diverse ways of expressing emotions through cultural differences or even different data capturing conditions. In SER, SNN abstracts away individual differences in gender and age from emotional characteristics inherent in MFCCs. SNN has also demonstrated robustness to noise degradation with different noise densities, compared to the state-of-the-art techniques such as SVM and CNN.

All these unisensory experiment results are promising and validate SNN as a viable option for emotion recognition. They also form the foundation for multisensory integration models, and each unisensory SNN will be used for extracting features; i.e., neuron groups on visual and audio signals. The next chapter will detail experimental evaluations of multisensory models proposed in this thesis. The chapter will describe results of multisensory models and evaluations compared to state-of-the-art models.

Chapter 8

Results and Discussion on Multisensory Emotion Recognition

8.1 Introduction

Chapter 5 has described the three multisensory integration models inspired by three different pathways of multisensory integration in the brain. The objective of this chapter is to evaluate the effectiveness of these models and answer the key research questions:

1) Does applying more bio-inspired architectures improve multisensory emotion recognition?

2) Do bio-inspired models present better generalisation capacity compared to state-of-the-art?

3) Are bio-inspired architectures robust to signal noise?

This chapter first evaluates results of multisensory integration by convergence (convergence) by implementing it using SNN. The second model to evaluate is early cross-modal enhancement (enhancement), where speech modality is enhanced through visual modality. Then, this chapter describes the experiments conducted for the evaluation of multisensory integration through neural synchrony (synchrony) model. Finally it details evaluation results for cross-corpus generalisation capacity and robustness to audio and visual noise.

Experimental evaluations in this chapter use the same parameters and design methodology introduced in Chapter 6. We also use the experimental setup explained in Chapter 6.5. The primary metrics for evaluating all models presented in this chapter is the overall accuracy. This choice of metrics is driven by the nature of the datasets used. Both RAVDESS and eNTERFACE'05 are

balanced and have the same number of data in each class. Parts of the evaluations results in this chapter have been published in [180] and [181].

8.2 Parameters And Hyperparameters Selection

Hyperparmeters selection represent a long process of tuning parameters resulting in an optimal model. In this thesis, most of the hyperparameters for the baseline models in terms of kernel size, padding and strides is derived from previous literature such as [256]. We mostly focus on tuning optimisers, learning rate and dropout rates. Model training is repeated using various parameters in order to find the optimal hyperparameters. Various parameters such as number of neurons in layers, activation functions, number of epochs or learning rate have been experimented. We adopt a grid search method using Keras.

8.2.1 Multisensory baseline model

For the Multisensory baseline model described in Chapter 6.6.3 we experiment with different configurations of the network. For the MLP fusion, we have made a hyper parameter search for the number of fully connected layers (2-3), learning rate (0.01-0.0001). We have also experimented with both SGD and Adam optimisers. The final parameters used for the model are are 3 fully connected layers, a learning rate of 0.001 and using SGD optimiser.

8.2.2 GCN model

We start by experimenting on the original parameters and architecture proposed by Kipf et al [145]. We then tune the parameters and architecture for neural synchrony data. We train the network initially for 200 than increased for 500 epochs due to loss not converging. We tune hyperparameters such as dropout (.25 - 0.5) , learning rate (0.1- 0.0001), hidden layer units (64-124) and weight-decay to 0.0005. The final values of the different values are shown in Table 8.1.

Table 8.1: GCN network parameters

Parameter	**Value**
Loss function	Adam
Learning rate	0.0001
number of layers	3
number of neurons in hidden layers	64
number of epochs	500
Dropout	0.5
weight-decay	0.0005

8.2.3 SNN models

The parameters used for SNN model are detailed in Chapter 6. We have used the parameters presented in the original work [65]

8.3 Accuracy of Multisensory Emotion Recognition Models

This section reports the results of experimental evaluation of the three proposed multisensory integration models; convergence, enhancement and synchrony. It first compares the proposed models' performance to state-of-the-art techniques. We choose state-of-the-art methods that have used the same datasets. This section then analyses and discusses the evaluations results.

Table 8.2 presents evaluation results for the three proposed models compared to the state-of-the-art on RAVDESS dataset. Table **??** presents the experimental setups of the experiments compared in 8.2. There exists various experimental setup in terms of model validation in the literature [143] such as n-fold cross validation, repeated cross-validation, bootstrap, holdout or repeated holdout . In this thesis we chose repeated holdout method to improve the reliability of the holdout estimate. We randomly split 80% and 20% for SNN based methods and 60% train, 20% validation and 20% testing for CNN based methods. For state-of-the-art methods Gibilisco et al [89] and Ghaleb et al [86] use a 10-fold cross-validation.

All reported accuracy are an average of the trials. The overall accuracy is 81.3%, 83.6% and 98.3% for the convergence, enhancement and synchrony models respectively. All these integration models have outperformed the state-of-the-art techniques. The best performing one is from Gibilisco et al. [89] with an overall accuracy of 80.2%. They have used facial landmarks

for extracting facial features, Fast Fourier Transform FFT for audio features. They have added natural language features. After extracting all modalities features they have used Random Forest for the classification of emotions on RAVDESS dataset. They have compared Random Forest classifier with simple MLP classifiers.

Among the three integration models, the synchrony model performs the best. Synchrony exploits both unsupervised learning to model neural synchrony and relationship between different neurons group and semi-supervised learning using GCN to classify multisensory emotions.

Table 8.2: Comparison of multisensory models to state-of-the-art for RAVDESS dataset

Model	**Visual Feature**	**Audio Feature**	**Experimental Setup**	**Fusion**	**Accuracy (%)**
Synchrony	LoG+SNN	MFCC+SNN	Repeated holdout 80/20	Synchrony with GCN	98.3
Enhancement	LoG	MFCC	Repeated holdout 80/20	Cross-modal enhancement	83.6
Convergence	LoG	MFCC	Repeated holdout 80/20	Convergence	81.3
CNN	Raw images	MFCC	Repeated holdout 80/20	Features features	81.0
Gibilisco et al. [89]	Facial Landmarks	FFT	10-fold Cross-validation	Random Forest	80.2
Ghaleb et al. [86]	Deep Metric Learning	Deep MEtric Learning	10-fold Cross-validation	AV-Gating Paradigm	67.7
Beard et al. 2018 [22]	Openface	COVAREP	N/A	LSTM+Global Conceptualised Attention	58.3
Dedeoglu et al [61]	CNN	CNN	N/A	late fusion DNN	87.3

Table 8.3 provides a comparison of the performance of the three proposed models and some state-of-the-art on the eNTERFACE'05 dataset. The proposed models, convergence, enhancement and synchrony experimental setup are described in Chapter 6. Previous work experiments have used different experimental setup. Similar to the RAVDESS dataset, the integration models

perform well achieving an overall accuracy of 96.8%, 86.3% and 80.0% for synchrony, enhancement and convergence respectively. The best performing state-of-the-art technique is from Zhang et al. [312] with the overall accuracy of 85.8% respectively. DBN fusion technique proposed by Zhang et al. [312] outperforms the convergence model by 5.5% with 85.8%. The DBN model enables features learning in a multisensory way and quickly translates the non-linear relationship between both modalities. However, the input data is segmented into several discrete temporal interval. Thus, there is the probability of missing information on continuous emotion information.

Noroozi et al. [204] outperforms the three proposed models with an accuracy of 98.7%. However, the visual features used consist of reduced frames in each video input and a set of facial geometrical features. Video inputs are summarised through key representative frames. After features extraction, they use three different classifiers for each feature type, one for facial features, the second for the representative frames and the third for MFCCs features. They then fuse the confidence score of each classifier. They report a very high accuracy, as they evaluate only on representative frames, which have summarised visual frame features and result in much fewer frames. This frames reduction could lead to missing critical information from the interaction of audio and visual modalities.

The proposed models: convergence, enhancement and synchrony make use of the whole visual and audio sequence in order to capture the whole dynamics. The proposed synchrony model extracts information from both modalities and learning happens simultaneously between audio and visual inputs. 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. GCN helps to model and learn synchrony patterns of neuron groups and enable multisensory emotion recognition across them. As a result of this strength, the synchrony model outperforms the DBN approach by 10.85%. Figure 8.1 presents the comparison on individual classes. The DBN approach achieves a lower accuracy of 80% on three classes: ‘sadness’, ‘fear’, and ‘surprise’, while neural synchrony has achieved a consistently high accuracy of $\geq$90% on all the classes.

In summary, compared to the other two presented models (convergence and enhancement), integration through neural synchrony outperforms convergence by 16.3% enhancement by 14% for RAVDESS dataset. For overall accuracy tasks, the synchrony model is the best performing

Table 8.3: Comparison of multisensory models to state-of-the-art for eNTERFACE'05 dataset

Model	Visual Feature	Audio Feature	Experimental setup	Fusion	Accuracy (%)
Synchrony	LoG+SNN	MFCC+SNN	Repeated holdout	Synchrony with GCN	**96.8**
Enhancement	LoG	MFCC	Repeated holdout	Cross-modal enhancement SNN	86.3
Convergence	LoG	MFCC	Repeated holdout	Convergence (SNN)	80.1
CNN	Raw images	MFCC	Repeated holdout	Features fusions (MLP)	79.0
Zhang et al. [312]	3DCNN		LOSO and LOGO	DBN	85.8
Fonnegra et al. [76]	CNN	RNN	5-fold	MLP	81.8
Di Nardo et al. [64]	CNN	CNN	10-fold cross-validation	3D Pyramid Neural Network	71.4
Wang et al. [294]	CNN	MFCC	data splitting	Decision level	83.0
Noroozi et al. [204]	AUC	MFCC	10-fold cross-validation	Random Forest Late	**98.7**
Ma et al [174]	CNN	LSTM	5-fold	autoencoder	85.4
Noor et al [203]	VCCR facial regions	MFCC+PLPC	Holdout 80/20	Early fusion KNN	96.6

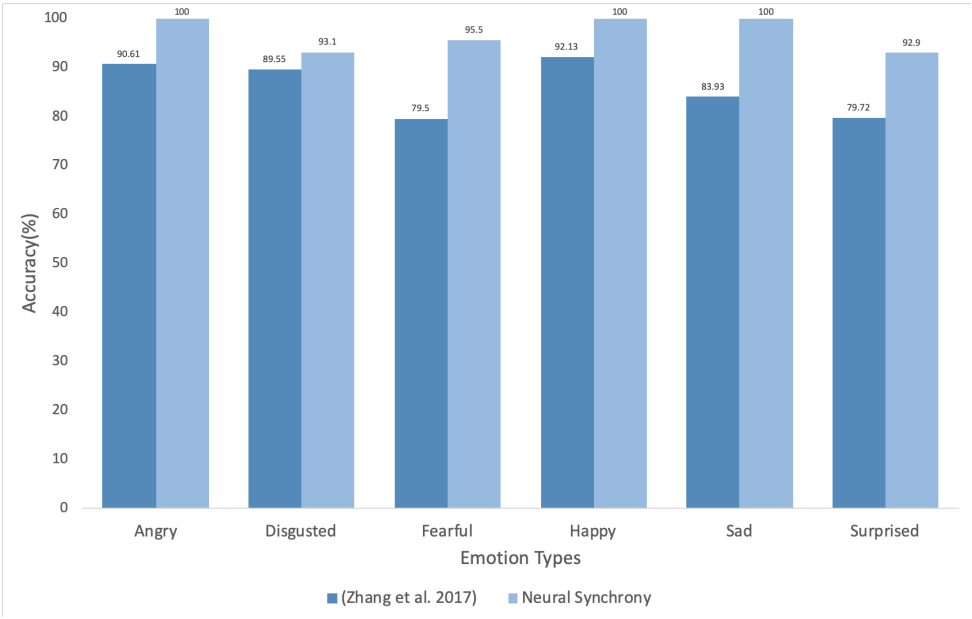


Figure 8.1: Comparison of accuracy by class type between state-of-the-art and neural synchrony on eNTERFACE'05

proposed model in this thesis. It also outperforms all presented state-of-the-art techniques.

Tables 8.6 and 8.7 present the confusion matrices of both datasets. We can observe a balanced accuracy for all classes with a lowest accuracy of 92.9% for surprise class on RAVDESS dataset and 92% for sad class on eNTERFACE'05 dataset.

8.3.1 Ablation Analysis

We have run ablation analysis on the unisensory and multisensory models. Because the convergence and enhancement models are unsupervised learning while the synchrony model is semi-supervised, we separate their results.

Table 8.8 shows that convergence and enhancement models both outperform unisensory models. Audio and visual-only models are implemented using the same convolution parameters as the convergence, and enhancement models with 60 features and a size of window and stride of 10. On RAVDESS dataset, the convergence outperforms unisensory model by 2.75% and 5.01% for audio and visual modalities respectively. Whereas, the enhancement model outperforms by 5.15% and 7.47% for audio and video, respectively. The same pattern can be noticed when using eNTEFFACE'05 dataset, where both multisensory models outperform audio and visual-only

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	90.0	5.0	0.0	0.0	5.0	0.0
Disgusted	6.1	84.8	0.0	0.0	9.1	0.0
Fearful	0.0	6.2	87.5	0.0	6.2	0.0
Happy	0.0	0.0	5.9	82.4	11.8	0.0
Sad	0.0	0.0	0.0	0.0	100.0	0.0
Surprised	15.2	9.1	12.1	3.0	24.2	36.4

Table 8.4: Confusion matrix of SNN audio only on RAVDESS datast

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	80.0	5.0	5.0	5.0	5.0	0.0
Disgusted	0.0	84.8	3.0	0.0	6.1	6.1
Fearful	0.0	0.0	93.8	0.0	6.2	0.0
Happy	5.9	5.9	0. 0	82.4	5.9	0.0
Sad	0.0	0.0	5.6	0.0	94.4	0.0
Surprised	3.0	9.1	12.1	3.0	6.1	66.7

Table 8.5: Confusion matrix of enhancement model on RAVDESS dataset

models. Convergence model outperforms the unisensory models by 9.8% and 4.8% for video and audio modalities. Enhancement model also outperforms unisensory models by 16% and 11% for visual and audio modalities, respectively.

Confusion matrices in Tables 8.4 and 8.5 show the difference between the recognition with SNN audio-only and cross-modal enhancement model. Although there is an enhancement of the overall accuracy, confusion matrices show different patterns depending on emotion classes. The accuracy on 'surprise' is at 36.4% using SNN with audio alone, and increases to 66.7% using the visual information enhancement. However 'sad' class accuracy decreases from 100%

		Predicted					
		Angry	Disgusted	Fearful	Happy	Sad	Surprised
Actual	Angry	100.0	0.0	0.0	0.0	0.0	0.0
	Disgusted	0.0	93.1	3.4	0.0	0.0	3.4
	Fearful	0.0	0.0	95.5	0.0	0.0	4.5
	Happy	0.0	0.0	0.0	100.0	0.0	0.0
	Sad	0.0	0.0	0.0	0.0	100.0	0.0
	Surprised	7.1	0.0	0.0	0.0	0.0	92.9

Table 8.6: Confusion matrix of synchrony model on RAVDESS datast

		Predicted					
		Angry	Disgusted	Fearful	Happy	Sad	Surprised
Actual	Angry	94.7	0.0	2.6	0.0	0.0	2.6
	Disgusted	0.0	95.5	2.3	0.0	2.3	0.0
	Fearful	0.0	0.0	100.0	0.0	0.0	0.0
	Happy	0.0	0.0	0. 0	100.0	0.0	0.0
	Sad	0.0	0.0	0.0	0.0	92.0	8.0
	Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table 8.7: Confusion matrix of synchrony model on eNTERFACE'05

accuracy from SNN with audio-only to 94.4% with visual cross-modal enhancement. On the other hand, 'happiness' and 'disgust' do not change their accuracy. The highest increase noticed in the 'surprise' emotion class is a result from a higher information in the visual modality and can translate the inverse-effectiveness [288] and early cross-modal enhancement of multisensory integration. The labels 'angry' and 'sad' experiencing a drop in accuracy in multisensory modal compared to the unisensory model, can be due to the visual modality bad effect, also translating the inverse-effectivness principle. When the auditory signal is high, it does not benefit from the multisensory signal if the visual signal is low.

Table 8.9 compares synchrony model to unisensory results with the same datasets. Unisensory models are trained and run for both RAVDESS and eNTERFACE'05 datasets with the same parameters and architecture as in 6.7.1. Using the same convolutional window and number of features as in multisensory integration through synchrony, and have run two separate SNNs for audio and visual data that is 40 for window size and 20 for number of features.

The accuracy gain of neural synchrony model is over 50% compared to unisensory models. This significant improvement in accuracy demonstrates the advantage of modelling and learning

connections between neuron groups in multisensory emotion recognition. Although unisensory models perform way better with smaller convolution window sizes and higher number of features, we have compared with the same number of features and window and stride sizes as integration through neural synchrony with number of features of 40 and window size of 20.

Table 8.8: Comparison of convergence and enhancement models to unisensory models for RAVDESS and eNTERFACE'05

Modality	**Feature extraction**	**eNTERFACE'05**	**RAVDESS (%)**
Video	LoG	70.3	76.1
Audio	MFCCs	75.3	78.5
Convergence	LoG,MFCCs, SNN	80.1	81.3
Enhancement	LoG,MFCCs, SNN	86.3	83.6

Table 8.9: Comparison of neural synchrony model to unimodal techniques

Modality	**Feature extraction**	**eNTERFACE'05**	**RAVDESS (%)**
Video	LoG	65.3	57.5
Audio	MFCCs	43.5	42.6
Synchrony	LoG,MFCCs, SNN	96.8	98.3

8.4 Cross-corpus Generalisation Results

This section evaluates cross-corpus generalisation capacity for the three proposed multisensory models: convergence, enhancement and synchrony. The proposed models' overall accuracy is compared to a state-of-the-art baseline; that is, setting up CNNs for extracting features on both auditory and visual modalities and concatenating these features for classification as detailed in Chapter 6. In the first set of experiments, models are trained using RAVDESS dataset and tested on eNTERFACE'05 dataset. These experiments aim to assess the ability to generalise learned features in the learning phase to new and never seen before data with a different distribution. First, all models are trained using RAVDESS datasets. The dataset is divided into 60% train, 20% validation and 20% testing. Results from the testing phase are recorded. In the second phase, all models are tested with eNTERFACE'05 dataset.

Table 8.14 shows results for multisensory models trained on RAVDESS dataset and tested with eNTERFACE'05. Generalisation capacity for each model is compared with the baseline

Actual \ Predicted	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	21.0	12.1	26.6
Fearful	8.6	10.1	28.1	13.5	8.6	31.1
Happy	11.2	9.0	25.6	23.7	8.7	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

Table 8.10: Confusion matrix for CNN baseline trained on RAVDESS and tested on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	69.8	0.0	4.6	7.0	11.6	7.0
Disgusted	9.5	42.9	19.0	7.1	16.7	4.8
Fearful	2.4	0.0	59.5	11.9	23.8	2.4
Happy	4.9	0.0	2.4	68.3	22.0	2.4
Sad	7.1	2.4	2.4	0.0	85.7	2.4
Surprised	2.5	2.4	7.3	22.0	14.6	51.2

Table 8.11: Confusion matrix for convergence trained on RAVDESS and tested on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	83.7	0.0	4.7	4.7	4.7	2.2
Disgusted	42.9	26.2	4.8	9.5	11.9	4.7
Fearful	33.3	0.0	40.5	4.8	21.4	0.0
Happy	22.0	0.0	2.4	58.5	14.6	2.5
Sad	26.2	0.0	2.4	7.1	64.3	0.0
Surprised	46.3	0.0	4.9	17.0	9.8	22.0

Table 8.12: Confusion matrix for enhancement trained on RAVDESS and tested on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	100.0	0.0	0.	0.0	0.0	0.0
Disgusted	0.0	100.0	0.0	0.0	0.0	0.0
Fearful	0.0	0.0	42.9	0.0	0.0	57.1
Happy	0.0	0.0	0.0	60.0	0.0	40.0
Sad	0.0	0.0	0.0	0.0	77.8	22.2
Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table 8.13: Confusion matrix for generalisation for synchrony trained on RAVDESS and tested on eNTERFACE'05

Table 8.14: Generalisation investigation on multisensory models trained on RAVDESS and tested on eNTERFACE'05

Modality	**RAVDESS(%)**	**eNTERFACE (%)**
CNN (Baseline)	81.0	22.6
Convergence	81.3	62.7
Ehancement	83.3	49.2
Synchrony	98.6	80.0

model. For overall evaluation, the synchrony model outperforms all other models with an overall accuracy of 98.6%. Convergence model performs worse among the three proposed model with 81.1%, nearly similar to the baseline network performance with 81.0% overall accuracy. Having the constant cross-talk in the enhancement and synchrony models enable a better performance compared to the classical convergence or the CNN baseline with features concatenation. Synchrony model exhibits the best performance due to the exploitation of the graph convolution network capacity.

To analyse generalisation capacity, models are tested with eNTERFACE'05. Synchrony model performs best with an overall accuracy of 80.0%. Integration through convergence and cross-modal enhancement performed 62.7% and 49.2% respectively. CNN, with only 22.6% exhibits the worst generalisation capacity. The baseline model is unable to generalise learnt features to a completely different dataset. Generalisation results present a similar pattern to the findings in unisensory modalities as detailed in Chapter 7.

Figures 8.10, 8.11, 8.12 and 8.13 show confusion matrices for the four models trained with RAVDESS and tested on eNTERFACE'05. The matrices show individual class performances. We notice that most balanced performances in individual emotion are exhibited in enhancement and synchrony models. The best-classified emotion for convergence model is 'sad' with 85.7%, and the worst one is 'disgusted' with 42.9%. The best-classified emotion synchrony are 'angry', 'disgusted' and 'surprised' with 100%. The best performing emotion for CNN is 'disgusted' with 33.9%. Synchrony model achieves the best classification results in individual emotion classes from the four models with 100% for three emotions. The 'angry' class has the most false positives in the enhancement models, whereas 'surprised' emotion has the most false positive in synchrony.

In the second set of experiments, all models are trained using eNTERFACE'05 and tested

using RAVDESS datasets. Table 8.15 shows results of generalisation evaluation on multisensory models trained on eNTERFACE'05 and tested on RAVDESS. Similar to results presented in Table 8.14, the results of the models trained with eNTERFACE'05 and tested on RAVDESS show that synchrony model performs well and has the highest generalisation accuracy of 77.8%. Training the models with eNTERFACE'05 and test them on RAVDESS leads to better results for the convergence model with 77.37% compared to 44.85% when the model is trained using RAVDESS and tested with eNTERFACE'05. We notice the same pattern for enhancement. However, results for synchrony dropped by 2.2%.

The synchrony model performs very well in generalisation tasks compared to the other presented models. Results are in line with the ones presented for the overall accuracy model evaluation. Exploiting the constant cross-talk, temporal synchrony and semantic similarity enables better feature learning.

The bio-inspired models presented in this thesis show a better generalisation capacity than the baseline features concatenation model using CNN.

Table 8.15: Generalisation investigation on multisensory models trained on eNTERFACE'05 and tested on RAVDESS

Modality	**eNTERFACE'05(%)**	**RAVDESS (%)**
CNN (Baseline)	79.0	20.0
Convergence	80.1	77.3
Enhancement	86.3	65.7
Synchrony	96.8	77.8

Figure 8.17, 8.18,8.19, and 8.16 show confusion matrices for generalisation evaluation for models trained with eNTERFACE'05 and tested with RAVDESS. For the CNN baseline model, the highest recognition rate is for 'surprised' class label with 33.2%, and the lowest classified class is 'disgusted' with only 13.3%. In contrast, the proposed models identified most class accurately with most of the classes with accuracy over 70%, with the expeption to 'surprised' for convergence and convrgence with 51.9%, and 19.5% and 'disgusted' in enhancement and synchrony with 42.9% and 48.5% respectively. Synchrony model is the best performing model for generalisation tasks with a balanced accuracy through the six class labels. All emotion classes are identified with over 75.0% accuracy except for 'disgusted' class label with only 48.5%. in cross-corpus generalisation tasks SNN models outperformed the baseline CNN consistently, this

is due to their ability to generalise learnt features. Whereas CNN perform badly if faced by different data distribution and data setting such as the background or the presence of facial hair or glasses as in eNTERFACE'05 dataset.

8.5 Noise Robustness Results

This section describes evaluation results on noise robustness for multisensory integration models proposed in this thesis. First, it describes results for the application of noise on visual data. Then outlines results for noise applied to audio data.

Results in this section are independent of the unisensory experiments results. SNN models have been implemented with different parameters in terms of convolution window size and number of features as detailed in Chapter 6. Less number of features and bigger convolution and size windows are used for computational speed efficiency.

This section evaluates the robustness of multisensory models described in the Chapter 5; that is, convergence, enhancement and synchrony. It compares the proposed model to the CNN baseline, similar to generalisation experiments.

8.5.1 Auditory Noise Evaluation

Table 8.20 shows results for the evaluation of the proposed models when applying three types of audio noise (brown, pink and white), similar to the evaluation of unisensory audio models for SER on RAVDESS dataset. Performance accuracy decrease for all models when applying audio noise to the test dataset. The baseline CNN model experienced the worst drop in accuracy when applying audio noise with 'white' noise experiencing the worst accuracy and dropping to 35.0%. The other models followed a different pattern and had their lowest accuracy when applying 'pink' noise. However, all the presented bio-inspired models perform well when presented with audio noise, experiencing a low drop of accuracy with accuracy over 60% for the three models. It suggests that the presented models are robust to noise and can compensate with the non-noisy modality.

The decrease of accuracy using noise in auditory data is in line with findings in [300] for all tested models. Results are in line with the unisensory experiments on SNN where applying pink

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	16.8	8.8	20.5	16.7	8.4	28.8
Disgusted	12.5	23.0	12.9	15.3	14.9	21.4
Fearful	15.7	16.5	26.6	10.5	16.1	14.6
Happy	13.7	10.4	28.9	26.5	8.5	11.8
Sad	10.2	22.3	15.7	21.2	11.3	19.3
Surprised	11.0	20.1	19.1	13.1	8.1	28.6

Table 8.16: Confusion matrix for generalisation in CNN baseline model trained with eNTERFACE'05 and tested on RAVDESS

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	71.4	19.0	0.00	0.0	4.8	4.8
Disgusted	3.0	83.9	7.0	0.0	6.1	0.0
Fearful	0.0	0.0	80.0	0.0	20.0	0.0
Happy	3.9	0.0	0.0	70.6	23.5	2.0
Sad	0.0	0.0	0.0	20.0	72.0	8.0
Surprised	0.0	26.2	0.0	0.0	21.9	51.9

Table 8.17: Confusion matrix for generalisation in convergence model trained with eNTERFACE'05 and tested on RAVDESS

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	88.4	0.0	4.7	0.0	4.7	2.2
Disgusted	33.3	42.9	2.4	2.4	19.0	0.0
Fearful	14.3	0.0	71.4	2.4	11.9	0.0
Happy	12.2	0.0	2.4	80.5	4.9	0.0
Sad	9.5	0.0	0.0	0.0	90.5	0.0
Surprised	34.1	0.0	9.8	2.4	34.2	19.5

Table 8.18: Confusion matrix for generalisation in enhancement model trained with eNTERFACE'05 and tested on RAVDESS

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	81.0	0.0	0.0	0.0	4.7	14.3
Disgusted	0.0	48.5	0.0	0.0	3.0	48.5
Fearful	0.0	0.0	81.2	0.0	0.0	18.8
Happy	0.0	0.0	0.0	76.5	5.9	17.6
Sad	0.0	0.0	0.0	0.0	83.3	16.7
Surprised	0.0	0.0	0.0	0.0	3.1	96.9

Table 8.19: Confusion matrix for generalisation model in synchrony trained with eNTERFACE'05 and tested on RAVDESS

Table 8.20: Audio noise evaluation on RAVDESS

Model	**No noise(%)**	**White (%)**	**Pink (%)**	**Brown (%)**
CNN	81.0	35.0	43.0	46.0
Convergence	81.3	76.6	72.5	78.1
Enhancement	83.6	77.3	72.3	76.6
Synchrony	98.3	67.0	63.4	65.2

Table 8.21: Audio noise evaluation on eNTERFACE'05

Model	**No noise**	**White (%)**	**Pink (%)**	**Brown (%)**
CNN	79.0	21.0	21.0	21.0
Convergence	80.1	71.3	70.3	71.3
Enhancement	86.3	77.3	71.7	72.9
Synchrony	96.8	70.2	50.9	77.9

noise experience the lowest accuracy.

Enhancement and convergence models experience the best accuracy using the three applied audio noise. Although the synchrony model exhibits the best overall accuracy for the model when tested without noise, its accuracy is lower when applying audio noise. Evaluating synchrony model on noise is challenging due to the nature of its implementation. The model is based on graph and an 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. The lower accuracy of synchrony compared to convergence and enhancement suggests that for the synchrony model, applying noise on one modality affects both modalities due to similar weights allocated given to both modalities.

To investigate individual emotion labels, 'brown' noise is chosen as a sample to analyse the effect of audio noise on individual emotions. Brown noise is selected as it represents the deepest audio noise applied. Brown noise represent a deepest version of the pink noise. The brown noise have all energy concentrated in low frequency [255].

Confusion matrices A.1, A.2, A.3 and A.4 show a sample of individual class accuracy with application of brown noise on RAVDESS dataset. The best performing emotion label in the convergence model is 'happy' with 88.2%. In contrast, the best performing label class in

synchrony and early enhancement models is ‘surprised’ with 96.9% and 100% for enhancement and synchrony respectively. ‘Surprise’ exhibits the highest false positive values for synchrony and enhancement models, whereas the class ‘sad’ has the highest false positive for the convergence model.

Table 8.21 shows the performance of the proposed models compared to the baseline when applying the three audio noise (brown, white and pink) on eNTERFACE’05. The lowest performing model is the baseline CNN model with 21.0% for all applied audio noise. The enhancement model performs the best under white noise with 77.3%. Although the three models performs comparably for the three noise levels, synchrony has the lowest accuracy for pink with 50.9%. It also exhibits the highest drop in accuracy.

Confusion matrices A.6, A.7, A.8 and A.5 show a sample for individual class accuracy for eNTERFACE’05 using ’pink’ noise. Similar to the ’brown’ noise, ’pink’ noise resulted in very low accuracy in all labels for CNN model. The synchrony model failed to classify data with emotion ’surprised’. Whereas the lowest performing emotion for enhancement is ’sad’ with only 48.0% positively classified samples in this emotion. The most balanced model in terms of individual emotion classification is the convergence where the lowest classified emotion is ‘angry’ with 57.1%.

8.5.2 Visual Noise Evaluation

Table 8.22 presents the results for the three proposed bio-inspired models, evaluating their robustness to visual noise. The experiments applied four degrees of noises probabilities from 0.1 to 0.8. The choice of noise levels follows the results of unisensory noise experiments in Chapter 7 where noise is applied from 0.1 to 0.8. Results do not change much from 0.3 to 0.5 therefore a higher level of noise is chosen to demonstrate and evaluate the robustness to visual noise.

Similarly to the experiments on audio noise, the baseline model CNN has the lowest accuracy for all degrees of noise with the lowest accuracy for the noise probability of 0.8, with only 22%. The three proposed models have a lower drop in accuracy with more stable accuracy with all noise levels. Enhancement model performs the best with only an insignificant drop for the highest noise level 0.8. Synchrony has the lowest accuracy when applying 0.8 with 66.5%.

The enhancement model is not as much affected by visual noise as much as the other models

with accuracy remaining stable for the five level of noise from 83.6% to 83.0%. This is due to 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.

Integration through neural synchrony has the best performance compared to all presented models. However the accuracy dropped to 66.46% when applying 0.8 noise probability. This accuracy is lower than the two other models with cross-modal enhancement model performing the best for 0.8 noise probability. Neural synchrony model is more sensitive to higher noise due to the fact that it is based on both modalities for and relies on temporal congruence between modalities.

For both enhancement and convergence models the accuracy dropped only when a high noise level is applied with a probability of 0.8%. The best performing model for the highest noise is the enhancement model. The results mean that using early cross-modal enhancement with noisy visual data is less affected by the inverse effectiveness principle of multisensory integration compared to the other models, which is in line with finding in neuroscience [288].

All the proposed models, convergence, enhancement and synchrony demonstrate a consistent higher accuracy with high visual noise level. The high accuracy of the proposed model results from the nature of feature extraction step and the overall model architecture. Using Laplacian of Gaussian (LoG) enhances the input an edges detection even in the presence of high noise. The convolution layers used in SNN enhances features learning.

The highest noise probability of 0.8% is chosen to analyse individual labels accuracy for RAVDESS dataset and confusion matrices are represented in Figure A.13,A.16, A.14, and A.15. Similar to audio noise, neural synchrony model exhibits the same pattern, where the highest false positive are for 'surprise' class. Whereas, 'sad' class is the highest false positive for early cross-modal and convergence models. Overall enhancement model exhibits the best accuracy for individual class labels.

Table 8.23 reports results for applying visual noise to the proposed models compared to the CNN baseline using eNTERFACE'05 dataset. We can notice the same pattern as for the RAVDESS results, with the baseline have the worst accuracy for all noise levels compared to

the proposed models. The most balanced and highest accuracy model is the enhancement model with accuracy ranging from 69.3% to 77.5% for noises from 0.1 to 0.8.

Confusion matrices in Figure 8.24, 8.25, 8.26 and 8.27 describe individual labels performance for the models tested on eNTERFACE'05 dataset. Following the same pattern as for RAVDESS dataset, the enhancement model show the best performances. Recognition rates for individual emotions is the most balanced compared to the other models. The best performing emotion is 'sad' with 78.6%, and the lowest performing is 'surprised' with 22.0%. For the convergence model, classes are also balanced with the highest classified emotion being 'sad' with 85.7% and the lowest emotion classification being 'surprised' with 51.2%, following the same pattern for the enhancement model. Similar to the results on RAVDESS, the model with less balanced label accuracy is the synchrony model with the most false positive for 'happy' emotion class.

Table 8.22: Visual salt and pepper noise evaluation of multisensory models on RAVDESS dataset

Model	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.8**
Synchrony	98.3	98.2	87.2	87.2	87.2	87.2	66.5
Enhancement	83.6	83.6	83.2	83.2	83.2	83.2	83.0
Convergence	81.3	81.0	81.0	81.3	81.3	81.3	77.4
CNN (Baseline)	81.0	65.0	63.0	35.0	35.0	35.0	22.0

Table 8.23: Visual salt and pepper noise evaluation of multisensory models on eNTERACE'05 dataset

Model	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.8**
Synchrony	96.8	63.1	63.0	62.1	62.0	62.0	60.0
Enhancement	86.3	77.5	70.5	70.0	70.0	70.0	69.3
Convergence	83.3	66.5	62.5	62.5	62.5	62.5	62.3
CNN (Baseline)	79.0	42.0	35.0	34.0	32.0	32.0	27.0

8.6 Summary

This chapter has presented evaluation experiments of the three proposed bio-inspired multisensory models: **convergence**, **enhancement** and **synchrony**. The three models are evaluated using third party datasets eNTERFACE'05 and RAVDESS. Results are compared to state-of-the-art methods on the same datasets.

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	14.3	8.0	18.6	18.0	12.1	18.6
Disgusted	7.5	28.4	13.5	24.1	18.2	20.8
Fearful	12.3	10.2	41.5	14.7	12.1	18.6
Happy	8.7	13.9	14.5	28.1	14.4	18.6
Sad	15.5	11.7	14.8	18.0	20.8	13.3
Surprised	19.4	15.7	9.7	14.7	8.7	26.1

Table 8.24: Confusion matrix CNN baseline with 0.8 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	69.8	0.0	4.7	7.0	11.6	7.0
Disgusted	9.5	42.9	19.0	7.1	16.7	4.8
Fearful	2.4	0.0	59.5	11.9	23.8	2.4
Happy	4.9	0.0	2.4	68.3	22.0	2.4
Sad	7.1	2.4	2.4	0.0	85.7	2.4
Surprised	2.4	2.4	7.3	22.0	14.6	51.2

Table 8.25: Confusion matrix convergence with 0.8 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	74.4	0.0	0.0	11.6	11.6	2.3
Disgusted	26.2	40.5	2.4	14.3	16.7	0.0
Fearful	14.3	0.0	50.0	4.8	31.0	0.0
Happy	9.8	0.0	0.0	68.3	22.0	0.0
Sad	14.3	0.0	2.4	4.8	78.6	0.0
Surprised	29.3	0.0	2.4	26.8	19.5	22.0

Table 8.26: Confusion matrix enhancement with 0.8 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	81.5	0.0	0.0	16.7	0.0	1.9
Disgusted	0.0	79.3	0.0	20.7	0.0	0.0
Fearful	0.0	3.9	66.7	27.5	0.0	2.0
Happy	0.0	0.0	0.0	90.0	10.0	0.0
Sad	3.7	0.0	0.0	92.6	3.7	0.0
Surprised	0.0	0.0	0.0	100.0	0.0	0.0

Table 8.27: Confusion matrix synchrony with 0.8 visual noise on eNTERFACE'05

Evaluations reveal that not only the proposed models' performances are comparable to state-of-the-art methods, they also exhibit an excellent generalisation capacity and robustness to noise. Each model is analysed in terms of performance and are compared to state-of-the-art and unisensory modalities.

The evaluation also demonstrates that modelling constant cross-modal interaction between different modalities using neural synchrony helps address challenges faced by current data fusion techniques such as late or features based fusion. Besides, cross-modal enhancement model architecture is more robust to visual and auditory noise to a simple CNN model. Having modalities constanting interacting and receiving feedback during training enables a more precise multisensory features representation, thus a better evaluation of the multisensory emotional content.

Chapter 9

Conclusions

The research work presented in this thesis represents an interdisciplinary project aiming at designing novel bio-inspired architectures and models for multisensory integration with applications in audio-visual social signals of emotions recognition. Multisensory integration of emotions is an essential area of research a represents a very challenging task due to the nature of emotions, being essentially multisensory. As discussed in Chapter 2, the majority of existing methods in the literature focus on applying fusion methods derived from engineering with the most popular methods being early, late or deep learning fusion techniques. This thesis is dedicated to exploring the use of bio-inspired models for emotion recognition and proposing novel computational models for multisensory integration based on different pathways of multisensory integration in the brain. The evaluation of the proposed models not only shows higher overall accuracy for both unisensory and multisensory emotion recognition tasks but also demonstrates excellent generalisation capacity and robustness to noise compared to state-of-the-art methods. Evaluation also shows that without the recourse to data augmentation, bio-inspired methods implemented through SNNs with unsupervised learning are comparable and even superior to state-of-the-art methods. The work presented in this thesis is divided into two main parts:

1) Explore the use of bio-inspired models in unisensory emotion recognition with applications for audio and visual modalities.

2) Design and implement three multisensory models based on three different pathways of multisensory integration in the brain.

9.1 Research Summary and Contributions

The research in this thesis reveals that by adopting bio-inspired models with unsupervised learning, we can achieve more accurate and precise multisensory integration. This thesis evaluates the proposed models on state-of-the-art datasets for both audio and visual data, representing continuous facials expressions and non-verbal speech features. Translating cross-talk between modalities facilitates the interpretation of multisensory integration, thus producing better accuracy, generalisation and robustness to noise.

9.1.1 Research Question 1 Contribution – Unisensory Emotion Recognition

"Are bio-inspired architecture effective for unisensory social signals of emotions recognition tasks? "

Chapters 4 and 7 aim at addressing the first research question. They describe the proposed models on the application of bio-inspired architectures in unisensory emotion recognition. Models are implemented and simulated using SNNs. Evaluations confirmed that the application of such models in facial expression recognition and speech emotion recognition tasks increases the overall accuracy. Furthermore, applying bio-inspired unsupervised learning provides a new insight on emotion recognition tasks. Experiments results also have presented promising accuracy which is comparable and often higher than most state-of-the-art techniques. With unsupervised learning, SNN models have achieved comparable accuracy to some of the most popular methods such as HOG features with SVM or CNN in facial expression recognition. SNN also proves useful for speech emotion recognition tasks. They can successfully translate the temporal dimension in audio signal processing.

This thesis has demonstrated that the exploration of different types of classifiers and more biologically inspired architectures can be beneficial for emotion recognition tasks. Providing an unsupervised STDP learning proves to be useful for features learning, with the reduced reliance on labelled training data and vast datasets or data augmentation.

9.1.2 Research Question 2 Contribution – Multisensory Emotion Recognition

"Does applying bio-inspired models in multisensory integration increase the efficiency of multisensory recognition systems?"

Chapters 5 and 8 aim at addressing and answering the second research question. The chapters propose novel computational models for multisensory integration with applications in multisensory audio-visual emotion recognition. It details three main models inspired by three pathways of multisensory integration in the brain for audio-visual data.

The three multisensory integration models are presented as follows:

Multisensory Integration Through Convergence

The first model in this thesis represents the most classical view on multisensory integration, which is integration by convergence in higher-order areas, using multisensory neurons. The model is implemented using SNN with three distinct neuronal groups, two representing each modality and a third group representing the multisensory area. Experimental evaluations on two third-party datasets confirmed that applying bio-inspired technologies with unsupervised learning is more effective than most state-of-the-art results from the most popular fusion techniques such as early or decision level.

Multisensory through Cross-modal Enhancement

The second proposed model is inspired by multisensory integration at early sensory areas in the brain. In this model one modality – auditory is preceded, and enhancement by another modality – visual. Using early cross-modal enhancement provides us with more accurate recognition compared state-of-the-art supervised learning techniques. Results show that using SNNs using early cross-modal enhancement either is equivalent or surpasses most state-of-the-art results for the same datasets in both implementations. Exploiting early cross-modal enhancement by using one modality to enhance and complement the other also proves more effective than the first proposed convergence model.

Multisensory Through Neural Synchrony

The third proposed model consist of using temporal and semantic coherence and synchrony between different sensory areas to drive multisensory integration. Exploiting SNN with STDP learning, temporal neural synchrony and the effectiveness of GCNs enables better feature

representation and multisensory interactions modelling. More specifically,

- SNN with STDP unsupervised learning enables feature learning and cross-talk between both modalities;
- Computing with neural synchrony with spike timing and stimuli enables integration of audio and visual data;
- GCN has demonstrated as a viable choice for modelling neuron activities and their interactions to facilitate learning complex patterns.

The third approach successfully translates the synchronous relation between audio and visual signals by using SNN representation of multisensory interaction. Using SNN to represent multisensory data can also alleviate the heterogeneity challenge of multisensory data. This is achieved by unifying all modalities features into a uniform input type – Poisson spike trains. In addition, representing data in graph addresses the fusion challenge by enabling data fusion while keeping the temporal and spatial relationship. Integration through neural synchrony can be particularly useful for robust in-the-wild emotion recognition, where there exists incongruous information between facial expression and vocal signals, or uncertainty in either of the modalities.

9.1.3 Research Question 3 Contribution – Generalisation

"Do bio-inspired models present better generalisation capacity compared to state-of-the-art? "

This thesis investigate generalisation capacity of the proposed bio-inspired models. Evaluation experiments are applied for both unisensory and multisensory models. First, unisensory SER and FER models are trained and evaluated using two distinct datasets with different data distribution. The bio-inspired models have achieved consistently better accuracy compared to the state-of-the-art techniques such as SVM with HOG features and CNN networks for FER tasks. They have also been compared to state-of-the art models for SER tasks such as SVM and CNN. The evaluation results show that bio-inspired models using SNNs have better generalisation capability for unisensory models for both FER and SER tasks. The models learning enables to learn features that can be generalised through never seen data.

The second part of experimental evaluations consist of investigating generalisation capacity for the proposed multisensory models. Convergence, enhancement and synchrony models are trained with one dataset and tested with a different one. Although performances drop when facing new data, the proposed models in this thesis have constantly achieved a stable accuracy that is constantly higher than the baseline model. Results confirm generalisation capacity of the three proposed models, with synchrony model performing the best for generalisation tasks. They also demonstrate that exploiting constant cross-talk between modalities is beneficial in multisensory integration with results demonstrated in enhancement and synchrony models.

9.1.4 Research Question 4 Contribution – Robustness to Noise

"Are bio-inspired architectures robust to signal noise?

This thesis also investigate the robustness to audio and visual noise in the proposed models. First, experiments evaluate the robustness to noise in unisensory models. SNN has demonstrated robustness to noise degradation with different noise densities, compared to state-of-the-art techniques such as SVM and CNN.

The second part of evaluations concerns the proposed multisensory models. All the proposed models, convergence, enhancement and synchrony, are evaluated and compared to state-of-the-art baseline model CNN under various audio and visual noise densities and types. The proposed models perform constantly better than the baseline with both audio and visual noise applications. Amongst the three proposed models, the enhancement model performs consistently better than the convergence and synchrony models for both visual and auditory noise. The enhancement model exploits early enhancement from visual data when applying audio noise. It is also less affected by visual noise than the other models. Evaluation results show that the proposed models can be successfully exploited in noisy environments.

9.2 Limitations

Despite the significant potential of applying bio-inspired architectures in both unisensory and multisensory, there exist some limitations in the implementations of such models. The main challenge is the computational cost, which are reflected in the following aspects.

1. **SNN parameters**

   Processing large datasets of audio-visual data presents some challenges in terms of computational efficiency. Parameters chosen for convolution window, stride and number of features are reduced due to computational cost. Having higher number of features and smaller window and stride size results in better accuracy results.

2. **Unisensory emotion recognition model**

   Unisensory models for both visual and auditory data is implemented using SNNs. Limitations have been encountered in processing audio data. Due to computational power limitations, audio data processing using temporal segments is limited.

3. **Multisensory models**

   The implementation of multisensory integration models proposed in this thesis face few limitations mainly linked in computational power. Early cross-modal enhancement model faces few limitations in the translation of temporal interaction between modalities. A more useful addition includes a temporal delay between different temporal segments. In the third proposed model, integration through neural synchrony, limitations have been encountered in the choice of spectral graph for graph learning. Although choosing spectral models with semi-supervised learning proved very beneficial in terms of overall accuracy, introducing new nodes and new subgraph requires to reload the whole graph. This operation is very computationally costly.

9.3 Future Work

The main motivation behind this research work is to explore novel methods for multisensory integration in general. The work presented in this thesis can be expanded in various forms as follows:

- **Apply multisensory integration models to different types of data** Application of the multisensory models presented in this thesis can be extended to include data from different modalities such as sensors data with general multisensory integration.

- **Include additional modalities such as body language** Another addition to the evaluation of the models with audio -visual data. Other modalities can be included such as body gesture or verbal speech information.

- **Extend the application to more complex and subtle emotions** The models included in this thesis may be suitable for more subtle emotions as they translate the constant cross-talk between modalities. The third model which consist of multisensory integration through neural synchrony is particularly useful for subtle emotion as it translate continuous emotions and cross-modal talks.

- **Explore the behaviour of the proposed models on cross-corpus experiments with different emotional state representation** The proposed models can be evaluated with different representation of emotional states such as using different emotion models. The models can be enhanced to include emotions intensity dimensions in addition to emotions categories.

- **Explore the behaviour of the proposed models on cross-subject experiments** The models performance and robustness can be evaluated cross-subject by using Leave One Subject Out (LOSO) technique. This can assess their performance for applications such as assistive technologies.

- **Explore the combination of the three models** The combination of the three models can translate the whole multisensory integration process in the brain. This operation is particularly useful for multisensory integration of various sensory modalities as opposed to bimodal integration.

- **Include attention mechanism** Attention module can be added to the main multisensory integration model to regulate the attention between modalities. It can also be used to add context for multisensory emotion recognition.

- **Extend the models to achieve recognition of different emotions representations** This can be achieved by introducing the use of hybrid emotion models such as Plutchik.

Finally, the research described in this thesis can help create systems such as sensory substitutions that can be particularly beneficial for applications in emotion recognition impairment. Applications can include assistive technologies for autism, dementia or schizophrenia.

Appendix A

Confusion Matrices For Multisensory Experiments

A.1 Audio Noise Experiments

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	34.0	11.2	12.6	8.8	14.0	19.5
Disgusted	4.4	39.1	14.5	13.7	12.5	15.7
Fearful	4.9	13.1	40.8	14.2	8.2	18.7
Happy	4.7	7.1	22.3	39.8	12.3	13.7
Sad	8.8	10.9	9.9	21.9	34.7	13.9
Surprised	6.4	6.7	11.3	11.3	16.3	48.1

Table A.1: Confusion matrix for CNN baseline with brown audio noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	85.7	0.0	0.0	0.0	9.5	4.8
Disgusted	0.0	81.8	0.0	0.0	12.1	6.1
Fearful	0.0	6.3	81.2	0.0	0.0	12.5
Happy	0.0	0.0	0.0	88.2	11.8	0.0
Sad	0.0	11.1	0.0	0.0	83.3	5.6
Surprised	0.0	3.1	3.2	0.0	12.5	81.2

Table A.2: Confusion matrix for convergence with brown audio noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	81.0	0.0	0.0	0.0	4.8	14.2
Disgusted	0.0	48.5	0.0	0.0	3.0	48.5
Fearful	0.0	0.0	81.2	0.0	0.0	18.8
Happy	0.0	0.0	0.0	76.5	5.9	17.6
Sad	0.0	0.0	0.0	0.0	83.3	16.7
Surprised	0.0	0.0	0.0	0.0	3.1	96.9

Table A.3: Confusion matrix for enhancement with brown audio noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	86.4	0.0	0.0	0.0	0.0	13.6
Disgusted	0.0	74.1	3.7	0.0	0.0	22.2
Fearful	0.0	0.0	85.0	0.0	0.0	15.0
Happy	0.0	0.0	0.0	55.6	0.0	44.4
Sad	0.0	0.0	0.0	0.0	0.0	100
Surprised	0.0	0.0	0.0	0.0	0.0	100

Table A.4: Confusion matrix for synchrony with brown audio noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	20.5	3.7	6.0	8.4	26.0	35.3
Disgusted	15.3	9.7	14.5	19.4	14.1	27.0
Fearful	9.7	11.6	13.1	13.9	14.6	37.1
Happy	12.8	9.0	16.1	21.3	16.1	24.6
Sad	13.5	6.6	19.0	16.4	20.8	23.7
Surprised	10.6	7.4	15.2	15.9	12.7	38.2

Table A.5: Confusion matrix for CNN baseline with pink audio noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	57.1	23.8	0.0	0.0	9.5	9.6
Disgusted	0.0	97.0	0.0	0.0	3.0	0.0
Fearful	0.0	0.0	87.5	0.0	12.5	0.0
Happy	0.0	5.9	5.8	64.7	11.8	11.8
Sad	0.0	5.6	0.0	0.0	94.4	0.0
Surprised	0.0	12.5	0.0	0.0	25.0	62.5

Table A.6: Confusion matrix for convergence with pink audio noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	72.1	2.3	0.0	14.0	9.3	2.3
Disgusted	9.5	76.2	0.0	11.9	2.4	0.0
Fearful	7.1	0.0	52.4	14.3	26.2	0.0
Happy	2.4	0.0	0.0	92.7	4.9	0.0
Sad	0.0	0.0	0.0	7.1	90.5	2.4
Surprised	14.6	0.0	0.0	19.5	19.5	46.4

Table A.7: Confusion matrix for enhancement with pink audio noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	88.4	0.0	0.0	4.6	7.0	0.0
Disgusted	0.0	87.8	0.0	7.3	4.9	0.0
Fearful	0.0	18.2	27.3	31.8	22.7	0.0
Happy	0.0	34.2	0.0	26.8	39.0	0.0
Sad	0.0	24.0	0.0	32.0	44.0	0.0
Surprised	0.0	50.0	0.0	50.0	0.0	0.0

Table A.8: Confusion matrix for synchrony with pink audio noise on eNTERFACE'05

A.2 Visual noise experiments

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	39.3	4.3	16.1	15.8	7.8	16.7
Disgusted	7.5	38.9	11.3	10.0	14.4	17.9
Fearful	8.3	6.5	50.4	12.2	7.1	15.5
Happy	6.0	13.9	12.9	42.6	13.6	11.0
Sad	11.9	6.8	12.3	16.0	38.3	14.7
Surprised	15.1	13.7	11.9	10.4	5.7	43.2

Table A.9: Confusion matrix for CNN baseline with 0.1 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	76.7	0.0	7.0	2.3	11.6	2.3
Disgusted	9.5	57.1	7.1	2.4	23.8	0.0
Fearful	14.3	0.0	76.2	0.0	9.5	0.0
Happy	4.9	0.0	7.3	75.6	12.2	0.0
Sad	4.8	0.0	2.4	0.0	92.9	0.0
Surprised	26.8	2.4	17.1	0.0	34.1	19.5

Table A.10: Confusion matrix for convergence with 0.1 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	76.2	19.0	0.0	0.0	0.0	4.8
Disgusted	3.0	75.8	0.0	0.0	21.2	0.0
Fearful	0.0	0.0	100.0	0.0	0.0	0.0
Happy	0.0	0.0	0.0	76.5	17.6	5.9
Sad	0.0	0.0	0.0	0.0	100.0	0.0
Surprised	3.1	0.0	0.0	3.1	34.4	59.4

Table A.11: Confusion matrix for enhancement with 0.1 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	52.9	0.0	0.0	0.0	0.0	47.1
Disgusted	0.0	80.6	0.0	0.0	0.0	19.4
Fearful	0.0	0.0	87.5	0.0	0.0	12.5
Happy	0.0	0.0	0.0	57.8	0.0	42.2
Sad	0.0	0.0	0.0	5.4	0.0	94.6
Surprised	0.0	0.0	0.0	0.0	0.0	100.0

Table A.12: Confusion matrix for synchrony with 0.1 visual noise on eNTERFACE'05

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	10.2	17.9	17.7	18.5	17.7	18.0
Disgusted	10.2	22.8	18.3	17.1	15.0	16.6
Fearful	8.8	19.4	30.3	15.1	11.3	15.5
Happy	8.4	9.1	26.7	29.0	11.9	14.9
Sad	7.9	13.1	16.1	23.1	20.4	19.4
Surprised	4.0	12.3	10.9	15.6	19.0	38.2

Table A.13: Confusion matrix for CNN baseline with 0.8 visual noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	81.0	4.8	0.0	4.7	9.5	0.0
Disgusted	0.0	75.8	3.0	0.0	21.2	0.0
Fearful	0.0	0.0	100.0	0.0	0.0	0.0
Happy	0.0	0.0	5.9	82.4	11.7	0.0
Sad	0.0	0.0	5.6	0.0	94.4	0.0
Surprised	12.5	3.1	9.4	0.0	21.9	53.1

Table A.14: Confusion matrix for convergence with 0.8 visual noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	88.4	0.0	4.7	0.0	4.6	2.3
Disgusted	33.3	42.9	2.4	2.4	19.0	0.0
Fearful	14.3	0.0	71.4	2.4	11.9	0.0
Happy	12.2	0.0	2.4	80.5	4.9	0.0
Sad	9.5	0.0	0.0	0.0	90.5	0.0
Surprised	34.1	0.0	9.8	2.4	34.1	19.6

Table A.15: Confusion matrix for enhancement with 0.8 visual noise on RAVDESS dataset

Actual \ Predicted	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	81.0	0.0	0.0	0.0	4.7	14.3
Disgusted	0.0	48.5	0.0	0.0	3.0	48.5
Fearful	0.0	0.0	81.2	0.0	0.0	18.8
Happy	0.0	0.0	0.0	76.5	5.9	17.6
Sad	0.0	0.0	0.0	0.0	83.3	16.7
Surprised	0.0	0.0	0.0	0.0	3.1	96.9

Table A.16: Confusion matrix for synchrony with 0.8 visual noise on RAVDESS dataset

Appendix B

Experiments Repeated Holdout Trials Results

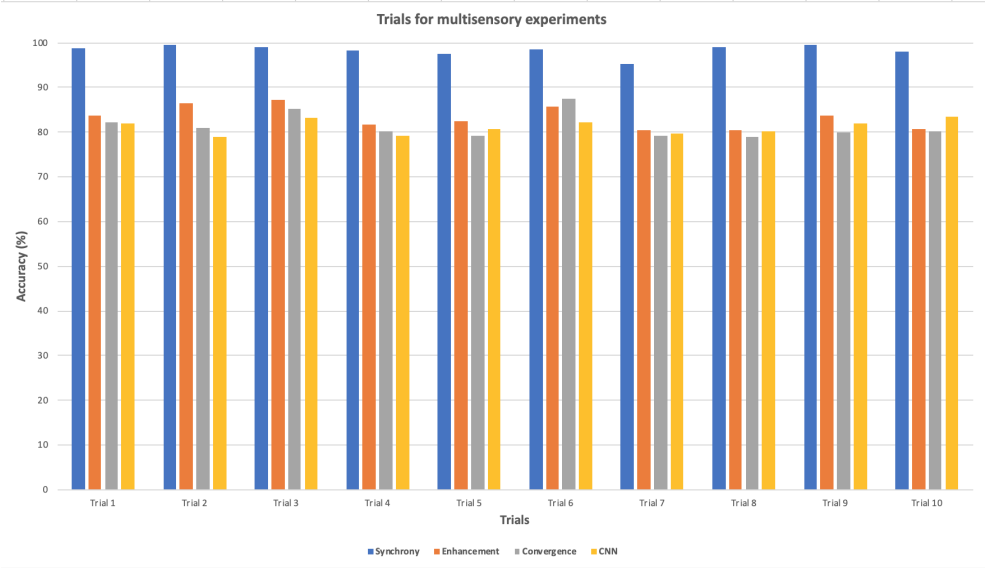


Figure B.1: Multisensory experiments repeated holdout trials results on RAVDESS dataset

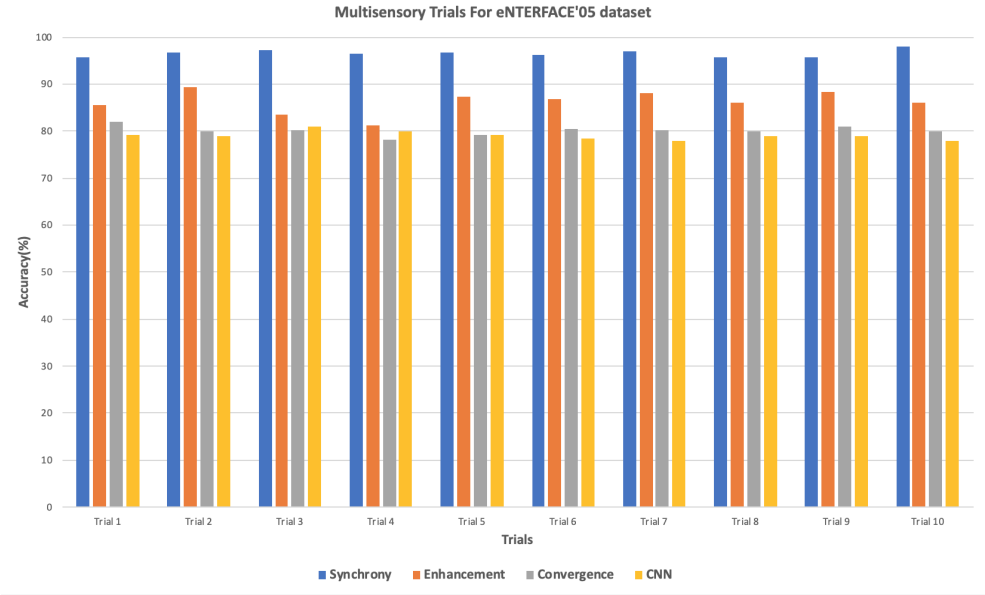


Figure B.2: Multisensory experiments repeated holdout trials results on eNTERFACE'05 dataset

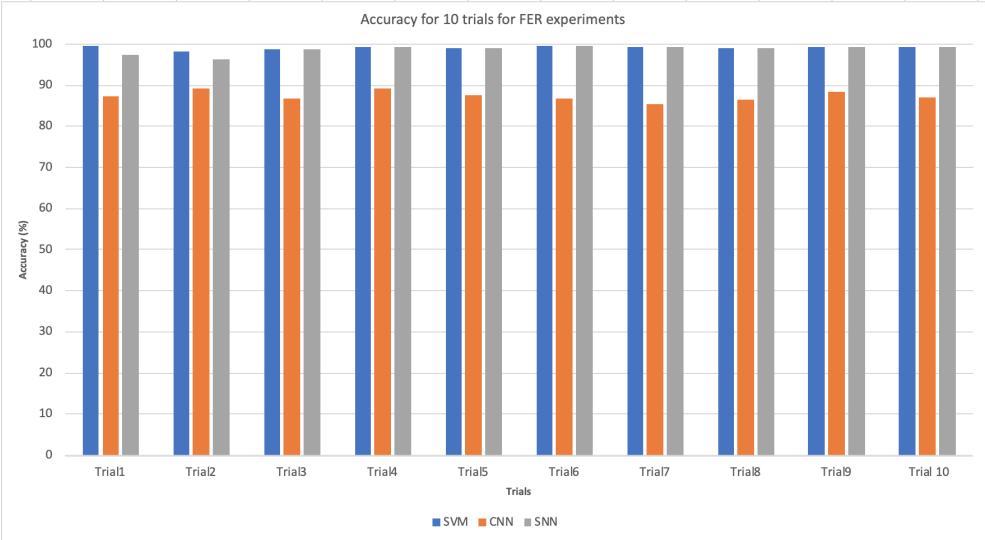


Figure B.3: FER experiments repeated holdout trials results on CK+ dataset

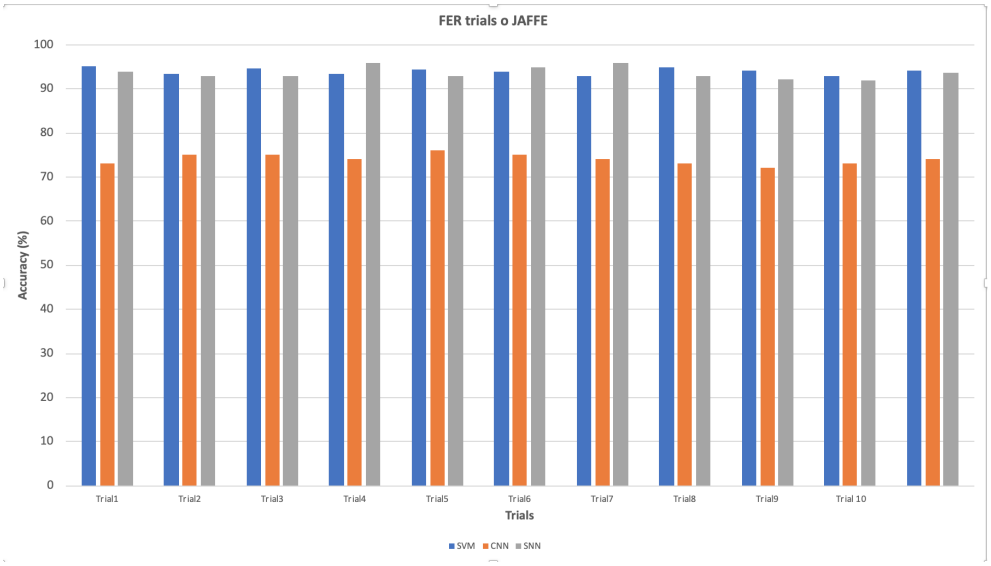


Figure B.4: FER experiments repeated holdout trials results on JAFFE dataset

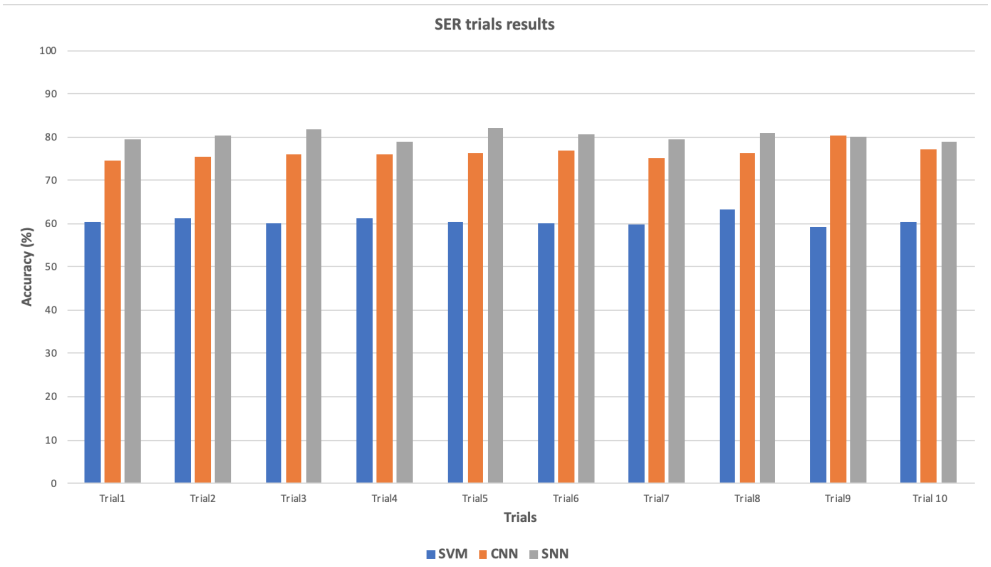


Figure B.5: SER experiments repeated holdout trials results on RAVDESS dataset

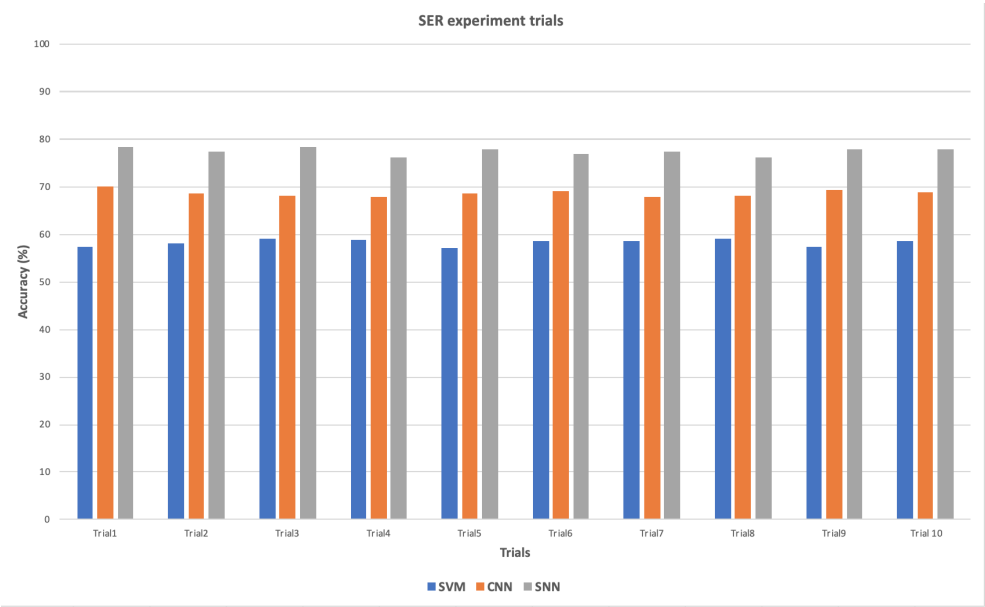


Figure B.6: SER experiments repeated holdout trials results on eNTERFACE'05 dataset

References

[1] D. A and P. J. Pattern recognition using spiking neural netwoks with temporal encoding and learning. In *IEEE 9th International Conference on Intelligent Systems and Control (ISCO)*, pages 1–5, Jan 2015. https://doi.org/10.1109/ISCO.2015.7282233.

[2] L. F. Abbott and P. Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

[3] M. Abeles. *Local cortical circuits: An electrophysiological study*, volume 6. Springer Science & Business Media, 2012.

[4] T. Ahonen, A. Hadid, and M. Pietikäinen. *Face Recognition with Local Binary Patterns*, pages 469–481. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. https://doi.org/10.1007/978-3-540-24670-1_36.

[5] M. M. R. Al-Yasari and N. A. S. Al-Jamali. Modified training algorithm for spiking neural network and its application in wireless sensor network. *Energy*, 5(10), 2018.

[6] O. Al Zoubi, A. Mayeli, M. Awad, and H. Refai. Hierarchical fusion evolving spiking neural network for adaptive learning. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 86–91. IEEE, 2018.

[7] 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*, pages 292–301. ACM, 2018.

[8] H. Alshamsi, V. Kepuska, H. Alshamsi, and H. Meng. Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. In

2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pages 730–738. IEEE, 2018.

[9] 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*, 20 Feb 2017. https://doi.org/10.1007/s11263-017-0997-7.

[10] 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*, 126(2-4):440–456, 2018.

[11] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 01 Feb" 2015. https://doi.org/10.1007/s10462-012-9368-5.

[12] B. Antje, G. M, J. Matthias, Wieser, and W. A. Georg. Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. *Front Psychol*, 2014.

[13] L. H. Arnal, B. Morillon, C. A. Kell, and A.-L. Giraud. Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, 29(43):13445–13453, 2009.

[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*, 97(3):640–655, 2018.

[15] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010. https://doi.org/10.1007/s00530-010-0182-0.

[16] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

[17] 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, 32(5):415–430, 2016. https://doi.org/10.1080/10447318.2016.1159799.

[18] M. Balconi and A. Carrera. Cross-modal integration of emotional face and voice in congruous and incongruous pairs: the p2 erp effect. *Journal of Cognitive Psychology*, 23(1):132–139, 2011.

[19] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[20] A. Barutchu, C. Spence, and G. W. Humphreys. Multisensory enhancement elicited by unconscious visual stimuli. *Experimental Brain Research*, 236(2):409–417, 01 Feb 2018. https://doi.org/10.1007/s00221-017-5140-z.

[21] A. Barutchu, C. Spence, and G. W. Humphreys. Multisensory enhancement elicited by unconscious visual stimuli. *Experimental brain research*, 236(2):409–417, 2018.

[22] R. e. a. Beard. Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 251–259, 2018.

[23] M. S. Beauchamp. Using multisensory integration to understand the human auditory cortex. In *Multisensory Processes*, pages 161–176. Springer, 2019.

[24] 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*, 17(2):11–21, 2018.

[25] L. E. Bernstein, E. T. Auer Jr, M. Wagner, and C. W. Ponton. Spatiotemporal dynamics of audiovisual speech processing. *Neuroimage*, 39(1):423–435, 2008.

[26] 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*, pages 215–228. Springer, 2019.

[27] Z. Bing, C. Meschede, K. Huang, G. Chen, F. Rohrbein, M. Akl, and A. Knoll. End to end learning of spiking neural network based on R-STDP for a lane keeping vehicle. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[28] J. D. Birdwell, M. E. Dean, M. G. Drouhard, and C. D. Schuman. Method and apparatus for constructing a neuroscience-inspired artificial neural network with visualization of neural pathways, Jan. 18 2018. US Patent App. 15/689,925.

[29] S. M. Bohte, H. La Poutré, and J. N. Kok. Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer rbf networks. *IEEE Transactions on neural networks*, 13(2):426–435, 2002.

[30] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 2019.

[31] J. Bower and D. Beeman. The book of genesis. 1998. *Telos*, 1998.

[32] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[33] A. B. Brandwein, J. J. Foxe, J. S. Butler, H.-P. Frey, J. C. Bates, L. H. Shulman, and S. Molholm. Neurophysiological indices of atypical auditory processing and multisensory integration are associated with symptom severity in autism. *Journal of autism and developmental disorders*, 45(1):230–244, 2015.

[34] R. Brette. Computing with neural synchrony. *PLoS computational biology*, 8(6):e1002561, 2012.

[35] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[36] C. A. Buscicchio, P. Górecki, and L. Caponetti. Speech emotion recognition using spiking neural networks. In F. Esposito, Z. W. Raś, D. Malerba, and G. Semeraro, editors, *Foundations of Intelligent Systems*, pages 38–46, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[37] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, S. Han, P. Liu, M. Chen, and Y. Tong. Feature-level and model-level audiovisual fusion for emotion recognition in the wild. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 443–448. IEEE, 2019.

[38] Y. Cao, Y. Chen, and D. Khosla. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113(1):54–66, 01 May 2015. https://doi.org/10.1007/s11263-014-0788-3.

[39] N. Caporale and Y. Dan. Spike timing–dependent plasticity: a hebbian learning rule. *Annu. Rev. Neurosci.*, 31:25–46, 2008.

[40] C. Cappe, E. M. Rouiller, and P. Barone. Cortical and thalamic pathways for multisensory and sensorimotor interplay. In *The neural bases of multisensory processes*. CRC Press/Taylor & Francis, 2012.

[41] K. D. Carlson, M. Richert, N. Dutt, and J. L. Krichmar. Biologically plausible models of homeostasis and STDP: stability and learning in spiking neural networks. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013.

[42] C. Chandrasekaran. Computational principles and models of multisensory integration. *Current opinion in neurobiology*, 43:25–34, 2017.

[43] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Audio visual emotion recognition with temporal alignment and perception attention. *arXiv preprint arXiv:1603.08321*, 2016.

[44] 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*, pages 2752–2756. IEEE, 2016.

[45] J. Chatterjee, V. Mukesh, H.-H. Hsu, G. Vyas, and Z. Liu. Speech emotion recognition using cross-correlation and acoustic features. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 243–249. IEEE, 2018.

[46] Y. Chavhan, M. Dhore, and P. Yesaware. Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, 1(20):6–9, 2010.

[47] K. Chen and L. Wang. *Trends in neural computation*, volume 35. Springer, 2006.

[48] Z. Chen, D. Huang, Y. Wang, and L. Chen. Fast and light manifold cnn based 3d facial expression recognition across pose variations. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 229–238. ACM, 2018.

[49] F. Cheng, J. Yu, and H. Xiong. Facial expression recognition in jaffe dataset based on gaussian process classification. *Trans. Neur. Netw.*, 21(10):1685–1690, Oct. 2010. https://doi.org/10.1109/TNN.2010.2064176.

[50] T. V. B. E. S. R. Chevallier C, Kohls G. The social motivation theory of autism. *Trends in Cognitive Sciences*, pages 231–239, 2012. https://doi.org/10.1016/j.tics.2012.02.007.

[51] F. Chollet. keras. https://github.com/fchollet/keras, 2015.

[52] 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 *International Conference on Speech and Computer*, pages 354–361. Springer, 2016.

[53] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000.

[54] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller. An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 478–484. ACM, 2017.

[55] 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*, 4:6, 2010.

[56] H. Dai. The low prediction accuracy problem in learning. In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 367–371, Nov 1994. https://doi.org/10.1109/ANZIIS.1994.396990.

[57] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005. https://doi.org/10.1109/CVPR.2005.177.

[58] C. Darwin. *The Expression of Emotion in Man and Animals*. CreateSpace Independent Publishing Platform, 2017. Online at https://books.google.co.uk/books?id=DcMqngAACAAJ.

[59] J. Davies-Thompson, G. V. Elli, M. Rezk, S. Benetti, M. J. Van Ackeren, and O. Collignon. Hierarchical brain network for face and voice integration of emotion expression. *BioRxiv*, page 197426, 2018.

[60] B. De Gelder and J. Vroomen. The perception of emotions by ear and by eye. *Cognition & Emotion*, 14(3):289–311, 2000.

[61] M. Dedeoglu, J. Zhang, and R. Liang. Emotion classification based on audiovisual information fusion using deep learning. In *International Conference on Data Mining Workshops (ICDMW)*, pages 131–134. IEEE, 2019.

[62] V. Demin and D. Nekhaev. Recurrent spiking neural network learning based on a competitive maximization of neuronal activity. *Frontiers in neuroinformatics*, 12, 2018.

[63] 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*, pages 179.1–179.12, 2015. https://doi.org/10.5244/C.29.179.

[64] E. Di Nardo, A. Petrosino, and I. Ullah. EmoP3D: A brain like pyramidal deep neural network for emotion recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[65] P. Diehl and M. Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, 9, 2015.

[66] K. Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current opinion in neurobiology*, 10(6):732–739, 2000.

[67] L. Duan, H. Ge, Z. Yang, and J. Chen. *Multimodal Fusion Using Kernel-Based ELM for Video Emotion Recognition*, pages 371–381. Springer International Publishing, Cham, 2016. https://doi.org/10.1007/978-3-319-28397-5_29.

[68] A. Durmuşoğlu and Y. Kahraman. Facial expression recognition using geometric features. In *2016 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5, May 2016. https://doi.org/10.1109/IWSSIP.2016.7502700.

[69] D. K. Duvenaud and et al. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[70] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[71] S. Elaiwat, M. Bennamoun, F. Boussaid, and A. El-Sallam. 3-d face recognition using curvelet local features. *IEEE Signal Processing Letters*, 21(2):172–175, Feb 2014. https://doi.org/10.1109/LSP.2013.2295119.

[72] M. A. Farries and A. L. Fairhall. Reinforcement learning with modulated spike timing–dependent synaptic plasticity. *Journal of neurophysiology*, 98(6):3648–3665, 2007.

[73] H. M. Fayek, M. Lech, and L. Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60 – 68, 2017. Advances in Cognitive Engineering Using Neural Networks, https://doi.org/https://doi.org/10.1016/j.neunet.2017.02.013.

[74] J. I. Feldman, K. Dunham, M. Cassidy, M. T. Wallace, Y. Liu, and T. G. Woynaroski. Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 2018.

[75] C. Felipe, M. Luis J, and N. Pedro. A novel multimodal emotion recognition approach for affective human robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[76] R. D. Fonnegra and G. M. Díaz. Deep learning based video spatio-temporal modeling for emotion recognition. In *International Conference on Human-Computer Interaction*, pages 397–408. Springer, 2018.

[77] R. D. Fonnegra and G. M. Díaz. Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model. In M. Kurosu, editor, *Human-Computer Interaction. Theories, Methods, and Human Issues*, pages 385–396, Cham, 2018. Springer International Publishing.

[78] J. Fu, Q. Mao, J. Tu, and Y. Zhan. Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. *Multimedia Systems*, 22 Mar 2017. https://doi.org/10.1007/s00530-017-0547-8.

[79] S.-Y. Fu, G.-S. Yang, and Z.-G. Hou. Spiking neural networks based cortex like mechanism: A case study for facial expression recognition. In *The 2011 International Joint Conference on Neural Networks*, pages 1637–1642. IEEE, 2011.

[80] C. Gallicchio, A. Micheli, and L. Pedrelli. Deep reservoir computing: a critical experimental analysis. *Neurocomputing*, 268:87–99, 2017.

[81] H. Gao, Z. Wang, and S. Ji. Large-scale learnable graph convolutional networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1416–1424. ACM, 2018.

[82] 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*, 12:244, 2018.

[83] A. V. Gavrilov and K. O. Panchenko. Methods of learning for spiking neural networks. a survey. In *2016 13th International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, volume 2, pages 455–460. IEEE, 2016.

[84] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 7538–7550, 2018.

[85] W. Gerstner and W. M. Kistler. Mathematical formulations of hebbian learning. *Biological cybernetics*, 87(5-6):404–415, 2002.

[86] E. Ghaleb, M. Popa, and S. Asteriadis. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *8th International Conference on Affective Computing & Intelligent Interaction (ACII 2019), Cambridge, United Kingdom*, 2019.

[87] D. Ghimire and J. Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013. https://doi.org/10.3390/s130607714.

[88] D. Ghosal, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya. Deep ensemble model with the fusion of character, word and lexicon level information for emotion and sentiment prediction. In *International Conference on Neural Information Processing*, pages 162–174. Springer, 2018.

[89] H. Gibilisco, M. Laubenberger, V. Spiridonov, J. Belga, J. O. Hallstrom, and P. R. Peluso. A multi-modal approach to sensing human emotion. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2499–2502. IEEE, 2018.

[90] S. Z. Gilani, A. Mian, F. Shafait, and I. Reid. Dense 3d face correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. https://doi.org/10.1109/TPAMI.2017.2725279.

[91] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, and F. Schwenker. *Kalman Filter Based Classifier Fusion for Affective State Recognition*, pages 85–94. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. https://doi.org/10.1007/978-3-642-38067-9_8.

[92] D. Goodman and R. Brette. Brian: a simulator for spiking neural networks in python. *Frontiers in Neuroinformatics*, 2:5, 2008. https://doi.org/10.3389/neuro.11.005.2008.

[93] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[94] P. Goyal and E. Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018.

[95] A. Grüning and S. Bohte. Spiking neural networks: Principles and challenges. In *Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning–ESANN*, 2014.

[96] D. Gupta, P. Bansal, and K. Choudhary. The state of the art of feature extraction techniques in speech recognition. In S. S. Agrawal, A. Devi, R. Wason, and P. Bansal, editors, *Speech and Language Processing for Human-Machine Communications*, pages 195–207, Singapore, 2018. Springer Singapore.

[97] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[98] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[99] R. J. Harris, A. W. Young, and T. J. Andrews. Brain regions involved in processing facial identity and expression are differentially selective for surface and edge information. *NeuroImage*, 97:217 – 223, 2014. https://doi.org/https://doi.org/10.1016/j.neuroimage.2014.04.032.

[100] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. Human neural systems for face recognition and social communication. *Biological psychiatry*, 51(1):59–67, 2002.

[101] H. Hazan, D. Saunders, D. T. Sanghavi, H. Siegelmann, and R. Kozma. Unsupervised learning with self-organizing spiking neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[102] H. Hazan, D. J. Saunders, H. Khan, D. T. Sanghavi, H. T. Siegelmann, and R. Kozma. Bindsnet: A machine learning-oriented spiking neural networks library in python. *Frontiers in neuroinformatics*, 12:89, 2018.

[103] D. O. Hebb. *The organization of behavior: a neuropsychological theory*. Science Editions, 1962.

[104] D. Heeger. Poisson model of spike generation. *Handout, University of Standford*, 5:1–13, 2000.

[105] M. Henaff. *Deep Networks for Forward Prediction and Planning*. PhD thesis, New York University, 2018.

[106] U. Hess and S. Hareli. The influence of context on emotion recognition in humans. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 03, pages 1–6, May 2015. https://doi.org/10.1109/FG.2015.7284842.

[107] M. L. Hines and N. T. Carnevale. Discrete event simulation in the neuron environment. *Neurocomputing*, 58:1117–1122, 2004.

[108] M. L. Hines and N. T. Carnevale. Translating network models to parallel hardware in neuron. *Journal of neuroscience methods*, 169(2):425, 2008.

[109] G. E. Hinton, T. J. Sejnowski, and T. A. Poggio. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[110] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, 1952.

[111] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 352–358. IEEE, 2017.

[112] F. S. Hsu, W. Y. Lin, and T. W. Tsai. Automatic facial expression recognition for affective computing based on bag of distances. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4, Oct 2013. https://doi.org/10.1109/APSIPA.2013.6694238.

[113] Y. Huang, J. Yang, P. Liao, and J. Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience*, 2017, 2017.

[114] Y. Humeau and D. Choquet. The next generation of approaches to investigate the link between synaptic plasticity and learning. *Nature neuroscience*, pages 1–8, 2019.

[115] T. Iakymchuk, A. Rosado-Muñoz, J. F. Guerrero-Martínez, M. Bataller-Mompeán, and J. V. Francés-Víllora. Simplified spiking neural network architecture and STDP learning algorithm applied to image classification. *EURASIP Journal on Image and Video Processing*, 2015(1):4, 19 Feb 2015. https://doi.org/10.1186/s13640-015-0059-4.

[116] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016.

[117] J. Irwin, T. Avery, L. Brancazio, J. Turcios, K. Ryherd, and N. Landi. Electrophysiological indices of audiovisual speech perception: beyond the mcgurk effect and speech in noise. *Multisensory Research*, 31(1-2):39–56, 2018.

[118] E. M. Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

[119] S. Jessen and S. A. Kotz. On the role of crossmodal prediction in audiovisual emotion perception. *Frontiers in Human Neuroscience*, 7:369, 2013.

[120] S. Jessen, J. Obleser, and S. A. Kotz. and voices interact in early emotion perception. *PLOS ONE*, 7(4):1–9, 04 2012. https://doi.org/10.1371/journal.pone.0036070.

[121] S. Jessen, J. Obleser, and S. A. Kotz. How bodies and voices interact in early emotion perception. *PLoS One*, 7(4):e36070, 2012.

[122] M. B. R. F. T. Joel, L. Davis, et al. *Synaptic plasticity: molecular, cellular, and functional aspects*. MIT Press, 1993.

[123] J. T. Jose, J. Amudha, and G. Sanjay. A survey on spiking neural networks in image processing. In E.-S. M. El-Alfy, S. M. Thampi, H. Takagi, S. Piramuthu, and T. Hanne, editors, *Advances in Intelligent Informatics*, pages 107–115, Cham, 2015. Springer International Publishing.

[124] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, pages 671–678. SCITEPRESS-Science and Technology Publications, Lda, 2014.

[125] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53, 2000. https://doi.org/10.1109/AFGR.2000.840611.

[126] E. R. Kandel, J. H. Schwartz, T. M. Jessell, D. of Biochemistry, M. B. T. Jessell, S. Siegelbaum, and A. Hudspeth. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[127] S. Karahan, M. Kilinc Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel. How image degradations affect deep cnn-based face recognition? In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, Sep. 2016. https://doi.org/10.1109/BIOSIG.2016.7736924.

[128] 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 *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2016.

[129] N. Kasabov and E. Capecci. Spiking neural network methodology for modelling, classification and understanding of eeg spatio-temporal data measuring cognitive processes. *Information Sciences*, 294:565 – 575, 2015. Innovative Applications of Artificial Neural Networks in Engineering, https://doi.org/https://doi.org/10.1016/j.ins.2014.06.028.

[130] N. K. Kasabov. Neucube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. *Neural Networks*, 52:62 – 76, 2014. https://doi.org/https://doi.org/10.1016/j.neunet.2014.01.006.

[131] A. K. Katsaggelos, S. Bahaadini, and R. Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, Sept 2015. https://doi.org/10.1109/JPROC.2015.2459017.

[132] C. Kayser and N. K. Logothetis. Do early sensory cortices integrate cross-modal information? *Brain structure and function*, 212(2):121–132, 2007.

[133] J. Keil and D. Senkowski. Neural oscillations orchestrate multisensory processing. *The Neuroscientist*, 24(6):609–626, 2018.

[134] J. Keil and D. Senkowski. Neural network dynamics and audiovisual integration. In *Multisensory Processes*, pages 201–220. Springer, 2019.

[135] G. Keren, A. E.-D. Mousa, O. Pietquin, S. Zafeiriou, and B. Schuller. Deep learning for multisensorial and multimodal interaction. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pages 99–128. 2018.

[136] K. Kessler and R. G. "brain oscillations and connectivity in autism spectrum disorders (asd): new approaches to methodology, measurement and modelling". *Neuroscience and Biobehavioral Reviews*, 71(Supplement C):601 – 620, 2016.

[137] L. Kessous, G. Castellano, and G. Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1-2):33–48, 2010.

[138] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier. Stdp-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, 99:56–67, 2018.

[139] P. Khorrami, T. L. Paine, and T. S. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 19–27, 2015.

[140] N. Kilian-Hütten, E. Formisano, and J. Vroomen. *Multisensory Integration in Speech Processing: Neural Mechanisms of Cross-Modal Aftereffects*, pages 105–127. Springer US, Boston, MA, 2017. https://doi.org/10.1007/978-1-4939-7325-5_6.

[141] B. K. Kim, S. Y. Dong, J. Roh, G. Kim, and S. Y. Lee. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–1508, June 2016. https://doi.org/10.1109/CVPRW.2016.187.

[142] J. Kim, K. P. Truong, G. Englebienne, and V. Evers. Learning spectro-temporal features with 3d cnns for speech emotion recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 383–388. IEEE, 2017.

[143] J.-H. Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745, 2009.

[144] R. S. Kim, A. R. Seitz, and L. Shams. Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS One*, 3(1):e1532, 2008.

[145] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[146] K. V. K. Kishore and P. K. Satish. Emotion recognition in speech using mfcc and wavelet features. In *2013 3rd IEEE International Advance Computing Conference (IACC)*, pages 842–847, Feb 2013. https://doi.org/10.1109/IAdCC.2013.6514336.

[147] E. I. Knudsen. Supervised learning in the brain. *Journal of Neuroscience*, 14(7):3985–3997, 1994.

[148] S. D. Koehler and S. E. Shore. Stimulus-timing dependent multisensory plasticity in the guinea pig dorsal cochlear nucleus. *PloS one*, 8(3):e59828, 2013.

[149] J. Kokinous, S. A. Kotz, A. Tavano, and E. Schröger. The role of emotion in dynamic audiovisual integration of faces and voices. *Social Cognitive and Affective Neuroscience*, 10(5):713–720, 2014.

[150] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: A review. *Int. J. Speech Technol.*, 15(2):99–117, June 2012. https://doi.org/10.1007/s10772-011-9125-1.

[151] B. J. Kröger and T. Bekolay. Speech acquisition. In *Neural Modeling of Speech Processing and Speech Learning*, pages 71–84. Springer, 2019.

[152] R. S. Larsen, D. Rao, P. B. Manis, and B. D. Philpot. STDP in the developing sensory neocortex. *Frontiers in synaptic neuroscience*, 2:9, 2010.

[153] C. Lau, F. A. Manno, C. M. Dong, K. C. Chan, and E. X. Wu. Auditory-visual convergence at the superior colliculus in rat using functional mri. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5531–5536. IEEE, 2018.

[154] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. https://doi.org/10.1038/nature14539.

[155] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. https://doi.org/10.1109/5.726791.

[156] A. K. Lee and M. T. Wallace. Visual influence on auditory perception. In *Multisensory Processes*, pages 1–8. Springer, 2019.

[157] C. Lee, G. Srinivasan, P. Panda, and K. Roy. Deep spiking convolutional neural network trained with unsupervised spike-timing-dependent plasticity. *IEEE Transactions on Cognitive and Developmental Systems*, 11(3):384–394, 2018.

[158] J. Lee and I. Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth annual conference of the international speech communication association*, 2015.

[159] R. Legenstein, D. Pecevski, and W. Maass. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS computational biology*, 4(10):e1000180, 2008.

[160] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[161] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[162] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran. An event driven fusion approach for enjoyment recognition in real-time. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 377–386, Orlando, Florida, USA, 2014. https://doi.org/10.1145/2647868.2654924.

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

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

[165] Y. Liu, Y. Cao, Y. Li, M. Liu, R. Song, Y. Wang, Z. Xu, and X. Ma. Facial expression recognition with pca and lbp features extracting from active facial patches. In *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 368–373, June 2016. https://doi.org/10.1109/RCAR.2016.7784056.

[166] 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*, 13(5):1–35, 05 2018. https://doi.org/10.1371/journal.pone.0196391.

[167] 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*, 61(Supplement C):610 – 628, 2017. https://doi.org/https://doi.org/10.1016/j.patcog.2016.07.026.

[168] R. Lotfidereshgi and P. Gournay. Biologically inspired speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5135–5139, March 2017. https://doi.org/10.1109/ICASSP.2017.7953135.

[169] D. M. Lovinger. Communication networks in the brain: neurons, receptors, neurotransmitters, and alcohol. *Alcohol Research & Health*, 2008.

[170] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanad dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[171] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, June 2010. https://doi.org/10.1109/CVPRW.2010.5543262.

[172] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[173] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, Dec 1999. https://doi.org/10.1109/34.817413.

[174] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang. An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1144–1149. IEEE, 2019.

[175] 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*, 46:184–192, 2019.

[176] W. Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659 – 1671, 1997. https://doi.org/https://doi.org/10.1016/S0893-6080(97)00011-7.

[177] A. Majumder, L. Behera, and V. K. Subramanian. Automatic facial expression recognition system using deep network-based data fusion. *IEEE Transactions on Cybernetics*, PP(99):1–12, 2016. https://doi.org/10.1109/TCYB.2016.2625419.

[178] E. Mansouri-Benssassi. A decentralised multimodal integration of social signals: A bio-inspired approach. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 633–637, Glasgow, UK, 2017. https://doi.org/10.1145/3136755.3137032.

[179] 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*, pages 426–437. Springer, 2018.

[180] E. Mansouri-Benssassi and J. Ye. Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[181] E. Mansouri Benssassi and J. Ye. Synch-graph: multisensory emotion recognition through neural synchrony via graph convolutional networks. In *Procceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*. AAAI Press, 2020.

[182] N. T. Markov, J. Vezoli, P. Chameau, A. Falchier, R. Quilodran, C. Huissoud, C. Lamy, P. Misery, P. Giroud, S. Ullman, et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259, 2014.

[183] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275(5297):213–215, 1997.

[184] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London Series B*, 23:187–217, 1980.

[185] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[186] M. Meredith. On the neuronal basis for multisensory convergence: a brief overview. *Cognitive Brain Research*, 14(1):31 – 40, 2002. Multisensory Proceedings, https://doi.org/https://doi.org/10.1016/S0926-6410(02)00059-9.

[187] M. Methu and K. Scherer. A psycho-ethological approach to social signal processing. *Cognitive Processing*, 2012.

[188] R. L. Miller, B. E. Stein, and B. A. Rowland. Multisensory integration uses a real-time unisensory–multisensory transform. *Journal of Neuroscience*, 37(20):5183–5194, 2017.

[189] B. Mishra, S. L. Fernandes, K. Abhishek, A. Alva, C. Shetty, C. V. Ajila, D. Shetty, H. Rao, and P. Shetty. Facial expression recognition using feature based techniques and model based techniques: A survey. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, pages 589–594, Feb 2015. https://doi.org/10.1109/ECS.2015.7124976.

[190] V. Mitra and H. Franco. Time-frequency convolutional networks for robust speech recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 317–323, Dec 2015. https://doi.org/10.1109/ASRU.2015.7404811.

[191] 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*, 14(1):115 – 128, 2002. Multisensory Proceedings, https://doi.org/https://doi.org/10.1016/S0926-6410(02)00066-6.

[192] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, March 2016. https://doi.org/10.1109/WACV.2016.7477450.

[193] J. Morrow, C. Mosher, and K. Gothard. Multisensory neurons in the primate amygdala. *Journal of Neuroscience*, 39(19):3663–3675, 2019.

[194] A. Mouraud and D. Puzenat. Simulation of large spiking neural networks on distributed architectures, the "damned" simulator. In *International Conference on Engineering Applications of Neural Networks*, pages 359–370. Springer, 2009.

[195] M. Mozafari, M. Ganjtabesh, A. Nowzari-Dalini, S. J. Thorpe, and T. Masquelier. Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks. *Pattern Recognition*, 94:87–95, 2019.

[196] M. Mukeshimana, X. Ban, N. Karani, and R. Liu. Multimodal emotion recognition for human-computer interaction: A survey. *System*, 9:10, 2017.

[197] M. M. Murray and M. T. Wallace. *The neural bases of multisensory processes*. CRC Press, 2011.

[198] A. Nandi, H. Schättler, J. T. Ritt, and S. Ching. Fundamental limits of forced asynchronous spiking with integrate and fire dynamics. *The Journal of Mathematical Neuroscience*, 7(1):11, 2017.

[199] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes. Deep spatio-temporal features for multimodal emotion recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1223. IEEE, 2017.

[200] F. Nian, X. Chen, S. Yang, and G. Lv. Facial attribute recognition with feature decoupling and graph convolutional networks. *IEEE Access*, 2019.

[201] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan. Improvement on speech emotion recognition based on deep convolutional neural networks. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, ICCAI 2018, pages 13–18, Chengdu, China, 2018. https://doi.org/10.1145/3194452.3194460.

[202] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288, 2016.

[203] S. Noor, E. A. Dhrubo, A. T. Minhaz, C. Shahnaz, and S. A. Fattah. Audio visual emotion recognition using cross correlation and wavelet packet domain features. In *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 233–236. IEEE, 2017.

[204] F. Noroozi, M. Marjanovic, A. NjeguÅ¡, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, PP:1–1, 06 2017. https://doi.org/10.1109/TAFFC.2017.2713783.

[205] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996. https://doi.org/https://doi.org/10.1016/0031-3203(95)00067-4.

[206] K. Okada, J. H. Venezia, W. Matchin, K. Saberi, and G. Hickok. An fmri study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PloS one*, 8(6):e68959, 2013.

[207] P. Opiełka, J. T. Starczewski, M. Wróbel, K. Nieszporek, and A. Marchlewska. Application of spiking neural networks to fashion classification. In *International Conference on Artificial Intelligence and Soft Computing*, pages 172–180. Springer, 2019.

[208] 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.

[209] T. Özseven. Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. *Applied Acoustics*, 142:70 – 77, 2018. https://doi.org/https://doi.org/10.1016/j.apacoust.2018.08.003.

[210] J. P. O'Doherty, S. W. Lee, and D. McNamee. The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1:94–100, 2015.

[211] C. P, D. C. M, L. M, and D. C. Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4, 2015. https://doi.org/http://doi.org/10.1186/s40064-015-1427-3.

[212] F. Pan, L. Zhang, Y. Ou, and X. Zhang. The audio-visual integration effect on music emotion: Behavioral and physiological evidence. *PloS one*, 14(5):e0217040, 2019.

[213] M. Papakostas, E. Spyrou, T. Giannakopoulos, G. Siantikos, D. Sgouropoulos, P. Mylonas, and F. Makedon. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5(2), 2017. https://doi.org/10.3390/computation5020026.

[214] P. Patel, A. Chaudhari, R. Kale, and M. Pund. Emotion recognition from speech with gaussian mixture models & via boosted gmm. *International Journal of Research In Science & Engineering*, 3, 2017.

[215] H. Paugam-Moisy. Spiking neuron networks a survey. Technical report, IDIAP, 2006.

[216] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[217] I. Pitas, I. Kotsia, O. Martin, and B. Macq. The enterface?05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)(ICDEW)*, volume 00, page 8, 04 2006. https://doi.org/10.1109/ICDEW.2006.145.

[218] L. Piwek, F. Pollick, and K. Petrini. Audiovisual integration of emotional signals from others' social interactions. *Frontiers in Psychology*, 6:611, 2015. https://doi.org/10.3389/fpsyg.2015.00611.

[219] B. Plakke and L. M. Romanski. Audiovisual integration in the primate prefrontal cortex. In *Multisensory Processes*, pages 135–159. Springer, 2019.

[220] R. Plutchik and H. Kellerman. *Theories of emotion*, volume 1. Academic Press, 2013.

[221] F. Ponulak and A. Kasinski. Introduction to spiking neural networks: Information processing, learning and applications. *Acta neurobiologiae experimentalis*, 71(4):409–433, 2011.

[222] F. Ponulak and A. Kasinski. Introduction to spiking neural networks: Information processing, learning and applications. *Acta neurobiologiae experimentalis*, 71(4):409–433, 2011.

[223] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98 – 125, 2017. https://doi.org/https://doi.org/10.1016/j.inffus.2017.02.003.

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

[225] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[226] C. Pramerdorfer and M. Kampel. Facial expression recognition using convolutional neural networks: State of the art. *CoRR*, abs/1612.02903, 2016. Online at http://arxiv.org/abs/1612.02903.

[227] H. Ranganathan, S. Chakraborty, and S. Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016. https://doi.org/10.1109/WACV.2016.7477679.

[228] S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[229] 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*, 2018.

[230] 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*, pages 1–11, 2018. https://doi.org/10.1109/TETCI.2018.2872014.

[231] S. Ratnasingam and T. M. McGinnity. A comparison of encoding schemes for haptic object recognition using a biologically plausible spiking neural network. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3446–3453. IEEE, 2011.

[232] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. Avec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM, 2015.

[233] E. Rolls. *Emotions Explained*. Oxford university press, 2005.

[234] E. T. Rolls. Precis of the brain and emotion. *Behavioral and brain sciences*, 23(2):177–191, 2000.

[235] J. Rong, Y. P. Chen, M. Chowdhury, and G. Li. Acoustic features extraction for emotion recognition. In *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*, pages 419–424, July 2007. https://doi.org/10.1109/ICIS.2007.48.

[236] L. A. Ross, D. Saint-Amour, V. M. Leavitt, S. Molholm, D. C. Javitt, and J. J. Foxe. Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophrenia research*, 97(1-3):173–183, 2007.

[237] B. A. Rowland and B. E. Stein. Multisensory integration produces an initial response enhancement. *Frontiers in integrative neuroscience*, 1:4, 2007.

[238] P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018.

[239] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[240] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28, 2018.

[241] B. Saleh, A. M. Elgammal, and J. Feldman. The role of typicality in object classification: Improving the generalization capacity of convolutional neural networks. *CoRR*, abs/1602.02865, 2016.

[242] P. Sanders, B. Thompson, P. Corballis, and G. Searchfield. On the timing of signals in multisensory integration and crossmodal interactions: a scoping review. *Multisensory research*, 32(6):533–573, 2019.

[243] E. Sariyanidi, H. Gunes, and A. Cavallaro. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, 26(4):1965–1978, April 2017. https://doi.org/10.1109/TIP.2017.2662237.

[244] W. Sato, T. Kochiyama, S. Uono, S. Yoshikawa, and M. Toichi. Direction of amygdala–neocortex interaction during dynamic facial expression processing. *Cerebral Cortex*, 27(3):1878–1890, 2017. https://doi.org/10.1093/cercor/bhw036.

[245] A. Satt, S. Rozenberg, and R. Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. *Proc. Interspeech 2017*, pages 1089–1093, 2017.

[246] A. Satt, S. Rozenberg, and R. Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. In *INTERSPEECH*, pages 1089–1093, 2017.

[247] D. J. Saunders, H. T. Siegelmann, R. Kozma, et al. STDP learning of image patches with convolutional spiking neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[248] 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*, pages 73–97. CRC Press, sep 2013.

[249] S. Schliebs and N. Kasabov. Evolving spiking neural network—a survey. *Evolving Systems*, 4(2):87–98, 2013.

[250] E. Schreuder, J. van Erp, A. Toet, and V. L. Kallen. Emotional responses to multisensory environmental stimuli: a conceptual framework and literature review. *SAGE Open*, 6(1):2158244016630591, 2016.

[251] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 1, pages I–401, July 2003. https://doi.org/10.1109/ICME.2003.1220939.

[252] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[253] D. Senkowski, T. R. Schneider, J. J. Foxe, and A. K. Engel. Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in neurosciences*, 31(8):401–409, 2008.

[254] A. Shaked and G. L. Clore. Breaking the world to make it whole again: Attribution in the construction of emotion. *Emotion Review*, 9(1):27–35, 2017. https://doi.org/10.1177/1754073916658250.

[255] H. Sheikhzadeh and H. R. Abutalebi. An improved wavelet-based speech enhancement system. In *Seventh European conference on speech communication and technology*, 2001.

[256] M. Shin, M. Kim, and D.-S. Kwon. Baseline cnn structure analysis for facial expression recognition. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 724–729. IEEE, 2016.

[257] A. Sonawane, M. U. Inamdar, and K. B. Bhangale. Sound based human emotion recognition using mfcc amp;amp; multiple svm. In *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, pages 1–4, Aug 2017. https://doi.org/10.1109/ICOMICON.2017.8279046.

[258] S. Song, K. D. Miller, and L. F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, 3(9):919, 2000.

[259] T. Song, W. Zheng, P. Song, and Z. Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 2018.

[260] I. Sporea and A. Grüning. Classification of distorted patterns by feed-forward spiking neural networks. In *International Conference on Artificial Neural Networks*, pages 264–271. Springer, 2012.

[261] M. Spüler, S. Nagel, and W. Rosenstiel. A spiking neuronal model learning a motor control task by reinforcement learning and structural synaptic plasticity. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015.

[262] G. Srinivasan, S. Roy, V. Raghunathan, and K. Roy. Spike timing dependent plasticity based enhanced self-learning for efficient pattern recognition in spiking neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1847–1854. IEEE, 2017.

[263] B. E. Stein. *The new handbook of multisensory processing*. Mit Press, 2012.

[264] B. E. Stein and M. A. Meredith. *The merging of the senses.* The MIT Press, 1993.

[265] B. E. Stein, T. R. Stanford, and B. A. Rowland. Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience*, 15:520–535, 2014.

[266] G. S. Stent. A physiological mechanism for hebb's postulate of learning. *Proceedings of the National Academy of Sciences*, 70(4):997–1001, 1973.

[267] R. A. Stevenson and M. T. Wallace. The multisensory temporal binding window: Perceptual fusion, training, and autism. *i-Perception*, 2(8):760–760, 2011.

[268] K. Strelnikov, J. Foxton, M. Marx, and P. Barone. Brain prediction of auditory emphasis by facial expressions during audiovisual continuous speech. *Brain topography*, 28(3):494–505, 2015.

[269] 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 *Proceedings of*

the 2015 ACM on International Conference on Multimodal Interaction, pages 497–502. ACM, 2015.

[270] M. Swain, A. Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[271] A. E. Symons. *Examining the Role of Temporal Prediction in Multisensory Emotion Perception*. PhD thesis, The University of Manchester (United Kingdom), 2018.

[272] 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*, 10:239, 2016. https://doi.org/10.3389/fnhum.2016.00239.

[273] J. C. Tapson, G. K. Cohen, S. Afshar, K. M. Stiefel, Y. Buskila, T. J. Hamilton, and A. van Schaik. Synthesis of neural networks for spatio-temporal spike pattern recognition and processing. *Frontiers in neuroscience*, 7:153, 2013.

[274] K. Tarunika, R. B. Pradeeba, and P. Aruna. Applying machine learning techniques for speech emotion recognition. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5, July 2018. https://doi.org/10.1109/ICCCNT.2018.8494104.

[275] I. J. Tashev, Z.-Q. Wang, and K. Godin. Speech emotion recognition based on gaussian mixture models and deep neural networks. In *Information Theory and Applications Workshop (ITA)*, pages 1–4, Feb 2017. https://doi.org/10.1109/ITA.2017.8023477.

[276] A. Tavanaei and A. S. Maida. Multi-layer unsupervised learning in a spiking convolutional neural network. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2023–2030, May 2017.

[277] D. Tavarez, X. Sarasola, A. Alonso, J. Sanchez, L. Serrano, E. Navas, and I. Hernáez. Exploring fusion methods and feature space for the classification of paralinguistic information. *Proc. Interspeech, Stockholm, Sweden*, pages 3517–3521, 2017.

[278] A. Teixeira Lopes, E. de Aguiar, A. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. In *Pattern Recognition*, volume 61, 07 2016.

[279] Y. Tie and L. Guan. A deformable 3-d facial expression model for dynamic human emotional state recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(1):142–157, Jan 2013. https://doi.org/10.1109/TCSVT.2012.2203210.

[280] J. L. Tracy and D. Randles. Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, 3(4):397–405, 2011. https://doi.org/10.1177/1754073911410747.

[281] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, March 2016. https://doi.org/10.1109/ICASSP.2016.7472669.

[282] H.-H. Tseng, M. G. Bossong, G. Modinos, K.-M. Chen, P. McGuire, and P. Allen. A systematic review of multisensory cognitive–affective integration in schizophrenia. *Neuroscience & Biobehavioral Reviews*, 55:444–452, 2015.

[283] E. Tsilionis and A. Vatakis. Multisensory binding: is the contribution of synchrony and semantic congruency obligatory? *Current Opinion in Behavioral Sciences*, 8:7–13, 2016.

[284] P. Uhlhaas, G. Pipa, B. Lima, L. Melloni, S. Neuenschwander, D. Nikolić, and W. Singer. Neural synchrony in cortical networks: history, concept and current status. *Frontiers in Integrative Neuroscience*, 3:17, 2009. https://doi.org/10.3389/neuro.07.017.2009.

[285] P. J. Uhlhaas and W. Singer. Neural synchrony in brain disorders: relevance for cognitive dysfunctions and pathophysiology. *neuron*, 52(1):155–168, 2006.

[286] M. Ursino, C. Cuppini, and E. Magosso. Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Networks*, 60:141–165, 2014.

[287] M. Ursino, C. Cuppini, and E. Magosso. Sensory fusion: A neurocomputational approach. In *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, pages 1–6, Sept 2016. https://doi.org/10.1109/RTSI.2016.7740583.

[288] L. P. H. van de Rijt, A. Roye, E. A. Mylanus, A. J. van Opstal, and M. M. Van Wanrooij. The principle of inverse effectiveness in audiovisual speech perception. *Frontiers in human neuroscience*, 13:335, 2019.

[289] S. van der Walt, J. L. Schenberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. a. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, June 2014. https://doi.org/10.7717/peerj.453.

[290] J. H. Venezia, S. M. Thurman, V. M. Richards, and G. Hickok. Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. *NeuroImage*, 186:647–666, 2019.

[291] D. Verstraeten, B. Schrauwen, M. d'Haene, and D. Stroobandt. An experimental unification of reservoir computing methods. *Neural networks*, 20(3):391–403, 2007.

[292] J. Wagner and E. André. Real-time sensing of affect and social signals in a multimodal framework: a practical approach. *The Handbook of Multimodal-Multisensor Interfaces*, pages 227–261, 2018.

[293] R. Walecki, V. Pavlovic, B. Schuller, M. Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017.

[294] X. Wang, X. Chen, and C. Cao. Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, page 115831, 2020.

[295] M. Wegrzyn, M. Riehle, K. Labudda, F. Woermann, F. Baumgartner, S. Pollmann, C. G. Bien, and J. Kissler. Investigating the brain basis of facial expression perception using

multi-voxel pattern analysis. *Cortex*, 69:131 – 140, 2015. https://doi.org/https://doi.org/10.1016/j.cortex.2015.05.003.

[296] A. Wendemuth, B. Vlasenko, I. Siegert, R. Böck, F. Schwenker, and G. Palm. *Emotion Recognition from Speech*, pages 409–428. Springer International Publishing, Cham, 2017. https://doi.org/10.1007/978-3-319-43665-4_20.

[297] 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*, 28(3):46–53, May 2013. https://doi.org/10.1109/MIS.2013.34.

[298] D. M. Wolpert, R. C. Miall, and M. Kawato. Internal models in the cerebellum. *Trends in cognitive sciences*, 2(9):338–347, 1998.

[299] J. Wu, Y. Chua, and H. Li. A biologically plausible speech recognition framework based on spiking neural networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2018. https://doi.org/10.1109/IJCNN.2018.8489535.

[300] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan. A spiking neural network framework for robust sound classification. *Frontiers in neuroscience*, 12, 2018.

[301] Q. Wu, T. M. McGinnity, L. Maguire, G. D. Valderrama-Gonzalez, and P. Dempster. Colour image segmentation based on a spiking neural network model inspired by the visual system. In D.-S. Huang, Z. Zhao, V. Bevilacqua, and J. C. Figueroa, editors, *Advanced Intelligent Computing Theories and Applications*, pages 49–57, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[302] Y. Wu, L. Deng, G. Li, J. Zhu, Y. Xie, and L. Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1311–1318, 2019.

[303] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[304] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[305] Y. Xu and R. Goodacre. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3):249–262, 2018.

[306] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W. Heinzelman, and M. Sturge-Apple. Enhanced multiclass svm with thresholding fusion for speech-based emotion classification. *International Journal of Speech Technology*, 20(1):27–41, 01 Mar 2017. https://doi.org/10.1007/s10772-016-9364-2.

[307] Q. Yu, K. C. Tan, and H. Tang. Pattern recognition computation in a spiking neural network with temporal encoding and learning. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, June 2012. https://doi.org/10.1109/IJCNN.2012.6252427.

[308] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643–649, 2018.

[309] M. Zhang, Y. Liang, and H. Ma. Context-aware affective graph reasoning for emotion recognition. In *ICME '19*, pages 151–156. IEEE, 2019.

[310] S. Zhang, L. Li, and Z. Zhao. Audio-visual emotion recognition based on facial expression and affective speech. In F. L. Wang, J. Lei, R. W. H. Lau, and J. Zhang, editors, *Multimedia and Signal Processing*, pages 46–52, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[311] S. Zhang, S. Zhang, T. Huang, and W. Gao. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 281–284. ACM, 2016.

[312] 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*, 28(10):3030–3043, 2017.