Speech Emotion Recognition With Early Visual Cross-modal Enhancement Using Spiking Neural Networks

Esma Mansouri-Benssassi and Juan Ye
School of Computer Science, University of St Andrews, UK
{emb24, juan.ye}@st-andrews.ac.uk

***Abstract*—Speech emotion recognition (SER) is an important part of affective computing and signal processing research areas. A number of approaches, especially deep learning techniques, have achieved promising results on SER. However, there are still challenges in translating temporal and dynamic changes in emotions through speech. Spiking Neural Networks (SNN) have demonstrated as a promising approach in machine learning and pattern recognition tasks such as handwriting and facial expression recognition. In this paper, we investigate the use of SNNs for SER tasks and more importantly we propose a new cross-modal enhancement approach. This method is inspired by the auditory information processing in the brain where auditory information is preceded, enhanced and predicted by a visual processing in multisensory audio-visual processing. We have conducted experiments on two datasets to compare our approach with the state-of-the-art SER techniques in both uni-modal and multi-modal aspects. The results have demonstrated that SNNs can be an ideal candidate for modeling temporal relationships in speech features and our cross-modal approach can significantly improve the accuracy of SER.**

***Index Terms*—Speech Emotion Recognition, Spiking Neural Networks, Unsupervised learning, Multisensory integration**

I. Introduction

Speech represents a crucial part in eliciting and understanding emotions, and both linguistic and para-linguistc parameters are useful for translating emotional states. Speech Emotion recognition (SER) has long been a popular task, which can be categorised into uni- and multi-modal approaches [1], [2]. Uni-modal approaches focus on extracting and learning auditory features contributing to emotion recognition [3], while multi-modal approaches integrate modalities from visual, audio, and other sources [4]–[6].

With the advantages of using multiple signal sources, multi-modal approaches have achieved better performance in SER, but the challenge still exists, especially where most multi-modal approaches have not investigated cross-modal learning and enhancement. In human brain, we process emotions in a cross-modal manner where one modality can predict, enhance or complete the other [7].

Turning towards more biologically inspired models would give an insight on how cross-modal learning and enhancement can be used to achieve better SER. Multisensory integration and learning in the brain does not occur via a simple fusion but follows specific principles and is based on a constant dialog between different modalities at an early level of information processing [8].

In this paper we investigate the feasibility of using bio-inspired models based on speech processing in the brain through both uni-modal and cross-modal aspects for the SER tasks. We explore the use of Spiking Neural Networks (SNN) to support unsupervised learning in SER and demonstrate comparable accuracy to the state-of-the-art techniques [9]. More importantly, we propose a novel cross-modal approach that enhances SER with visual information and achieves better performance compared to most commonly used data fusion approaches in multimodal emotion recognition [10]. This will lead to more simplified and computationally advantageous models. Also this enhancement model, rather than fusion, 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 not affect the overall recognition accuracy.

This paper is organised as follows. Section II gives an overview of the state-of-the-art techniques in Speech Emotion Recognition (SER) tasks. Section IV introduces a theoretical background of SNN and describes a novel approach based on early multisensory integration and enhancement of audio signals using visual signals. Section V reports experiments and results on evaluating the performance of SNN and our cross-modal approach on two datasets. Section VII draws conclusions and points to future directions in using more biologically inspired models for the SER task.

II. Related Work

This section will briefly introduce the state of the art in feature extraction, emotion recognition, and multimodal fusion in the application of speech emotion recognition.

A. Emotion Speech Features

One of the main interests in SER is feature extraction – learning best auditory features translating variations in affective and emotional states in speech. The features can be both linguistic and para-linguistic. In the following we list some of the most popular features for the SER task.

1) Mel Frequency Ceptral Coefficients MFCCs: MFCCs are calculated from the mel-frequency ceptrum (MFC) – a

linear cosine transform of a log power spectrum, which represents a short-term power spectrum of an audio signal. This method mimics the human auditory processing of sound [11].

2) Spectral Centroid: This feature represents the brightness of an audio signal, which is represented by the center mass of the magnitude of the spectrum. It indicates the rapid changes in the signal [12]. Cummins et al. have used this feature with a convolutional neural network for SER [13].

3) Energy: Energy represents 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.

B. Classifiers for SER

There are two types of learning in SER: static and dynamic. Static learning aims to recognise emotion through the whole utterance on auditory features [14]–[17], while dynamic learning partitions an auditory signal into frames and focuses on learning temporal relationships between frames in emotion recognition [18].

1) Deep Learning: In recent years, deep learning techniques have outperformed the classic machine learning techniques in SER [16], [19]. Most of deep learning work using hand-crafted features take the feature input as a whole regardless of the dynamic relation within time. These hand-crafted features are considered to represent the audio signal with a global level acoustic feature, where once extracted they tend to loose the dynamic relation in the temporal dimension.

Niu et al. [20] propose the application of Deep Retinal Convolution Neural Network (DRCNN), which consists of the following two steps: (1) data augmentation step where the principal of retina and convex lens imaging is used on the set of spectrogram features for each input, and (2) learning step that applies a deep convolution neural network on the extract spectrogram features. They have obtained an overall accuracy of 99% using the augmented data.

Satt et al. [21] have investigated two types of network on spectrogram features for the SER tasks. They first train a CNN on the extracted spectrogram data, using different network topologies. They then train a CNN in addition to a LSTM layer. Adding a LSTM layer has proved beneficial to the improvement of the overall accuracy; that is, it reaches 68% compared to 62% with the CNN layers alone. Other deep learning techniques have also been investigated; for example, Lee et al. have used a RNN to learn feature representations of audio signals [22].

2) Bio-inspired Approaches: Bio-inspired approaches have not been sufficiently explored in the literature of SER. Buscicchio et al. have made an early attempt of using a SNN for SER [23], where the work focuses on the the linguistic part of the speech by decomposing each sentence into different parts for each vowel occurrence. For each part MFCC features are extracted and encoded into spike trains using average rate coding. The network is trained using the reinforcement learning algorithm.

Lotfidereshgi et al. have used raw speech signal as an input and used Liquid State Machines (LSM) for classification [24]. LSM is a type of reservoir computing [25], where *reservoir* represents a SNN. The speech input goes through several preprocessing steps, where linear prediction analysis is applied to the audio signal. The overall classification task has achieved the accuracy of 82.35%, which is comparable to the state of the art for the same datasets.

3) Multisensory Integration: Research toward multisensory emotion recognition has focused on using various state-of-the-art data fusion techniques such as early, decision or model fusion for enhancing recognition accuracy.

State-of-the-art approaches offer a wide range of capabilities. However, they focus on choosing the right features rather than investigating novel approaches in terms of classifiers. Extracting emotional states solely from para-linguistic speech data can prove challenging, as such information cannot be fully translated into emotional states [26].

Naturally human processes emotion in a multisensory manner where the prediction from preceding visual information plays a significant role in facilitating the recognition of emotional states through speech [27]. In addition, external context such as the background and body gesture play a role in emotion recognition [28]. This is not translated in state-of-the-art techniques in SER either in audio alone or in a multimodal aspect. Although there has been extensive research in multimodal emotion recognition, data fusion techniques applied do not focus on the interaction between different modalities [29]. Therefore research needs to turn towards more realistic cross modal interaction instead of treating the tasks solely as a data fusion problem. In this paper, we propose a bio-inspired approach that aims to translate interactions between different modalities.

III. Background on Spiking Neural Networks

Humans perceive emotions through face expression and speech differently from other pattern recognition tasks. The process involves various brain regions constantly interacting to make meaningful precepts from the perceived linguistic, non linguistic and changes in facial expression [30]. Information in the brain is transmitted between neurons using action potentials via synapses. When a membrane potential reaches a certain threshold a spike is generated [31]. The computation of SNNs is based on the timing of spikes rather than their shape, where spikes that fire together get a stronger connection.

SNNs have been extensively used for translating neuro-computational processes in the brain and successfully applied to machine vision tasks and lately for speech signals [32]. In these tasks, SNNs have demonstrated as a promising candidate for modeling temporal data such as audio signals [33]. Here we focus on employing SNNs to translate temporal information of an audio signal for emotion recognition.

SNNs support different types of learning, including

- Unsupervised: unsupervised learning in spiking neural network follows the Hebbs' where the coupling between neurons is strengthened when neurons fire together. The

Hebbian plasticity is influenced by the timing of the pre-synatptic and post-synaptic spike. This type of learning is referred to Spike Timing Dependent Plasticity (STDP). STDP learning has been used in various clustering and pattern recognition tasks [34].

- Supervised: supervised learning has been implemented in a Hebbian way for biologically inspired models by adding a supervision signal to reinforce the firing at target times [35].
- Reinforcement: reinforcement learning enables learning directly from the environment where a SNN includes a rewarding signal spike [36].

IV. Proposed Work

In this section we will present an early cross-modal approach where visual information is used to enhance speech-based emotion recognition. Figure 1 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 [37], [38] to enhance audio signals using visual stimuli. In the following, we will detail the design on these two components.

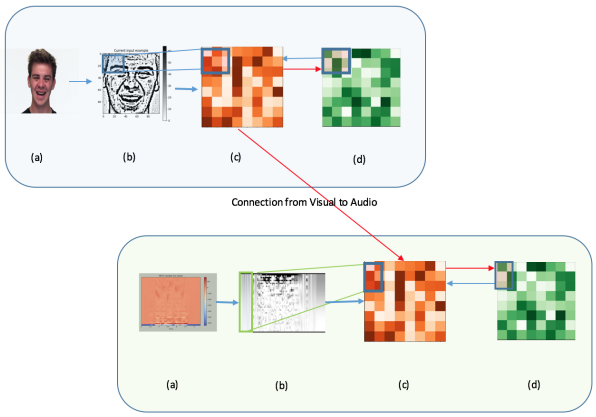


Fig. 1: The work-flow of our early cross-modal enhancement: (a) pre-processing both visual and audio input, (b) encoding input into spike trains, (c) excitatory layer, (d) inhibitory layer

A. SNN for Unimodal Learning

1) Neuron models: Here we present the application of SNN with unsupervised STDP learning for SER tasks. We have adapted the work from [34], where Leaky-Integrate-and-Fire (LIF) is used to model neurons dynamics and learning is achieved using STDP. As mentioned in Section III, neurons communicate through a series of spikes by firing spikes, and thus neurons at the excitatory layer can learn unique features that distinguish different emotional states. The membrane voltages of the neurons are translated by the following function:

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

V is the membrane voltage and E_{rest} represents the 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 certain time interval (5ms). This 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 a better network balance and stability, we have also employed *homoeostasis* – an adaptive membrane threshold V_{thresh} mechanism [34]. 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, homoeostatsis prevents some neurons firing for all presented inputs and as well as avoids few neurons from dominating emotional patterns [39]. We also employ lateral inhibition encouraging competition between neurons.

Modelling synapses is achieved by changes in conductance. Their 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:

$$\tau_{g_e} \frac{dg_e}{dt} = -g_e \quad (2)$$

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

2) STDP Learning: Learning is performed in an unsupervised manner from the input layer to the excitatory layer through unsupervised STDP learning [34]; that is, learning distinctive features for each emotional class label in an unsupervised manner. The STDP algorithm has been successfully used in pattern recognition and image classification tasks [40]. It represents a spike based type of Hebbian learning, where the connection between neurons is strengthened when they fire together. The plasticity is influenced by the timing of the pre-synatptic and post-synaptic spikes. The synaptic weight updates when a post-synaptic spike reaches a synapse, which is characterised by the following equation:

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

η is the learning rate. w_{max} is the maximum weight and x_{tar} is the target value of the pre-synaptic trace when the post-synaptic spike fires. This is used to enable the disconnection of neurons that seldom lead to firing, when the post-synaptic neuron is rarely active. μ 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 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 considered to be quite advantageous comparing to back-propagation as weights do not need to be learned through backward and forward pass [41].

B. SNN Adaptation for SER

The first step in using SNN is to encode the input into meaningful format for the SNN, especially for audio input. We encode extracted audio features into a population of Poisson spike trains [34]. Each extracted input represents a firing rate proportionate to its intensity and each feature value over time is transformed into firing rate between 0 and 63.75Hz. The input data is run through the network for 350ms [34]. After that the network enters a resting phase for 150ms, in order to get back to its initial equilibrium before receiving the next input.

Built on the basic SNN architecture [34], we have used a convolutional layer over the input; that is, the input layer is connected to a convolution excitatory layer coupled with an inhibitory layer. Each input is divided into convolutional features where a stride window moves through the input along the temporal axis of the audio features. Adding a convolution layer has demonstrated to be useful in improving the overall accuracy on unimodal learning from general image classification [42].

C. Cross-Modal Enhancement

Here we introduce an early cross-modal enhancement method which describes one of multisensory integration principles. Recent research findings have demonstrated the existence of early cross-modal interaction and integration between different brain areas in audio-visual processing at an early level [37], [38]. Multisensory areas in the brain such as the superior colliculus (SC) use STDP learning at a neural level for the interaction between different unisensory modalities [43]. It has been demonstrated that unisensory areas have a constant interaction at early sensory levels [43] during multisensory integration. This idea of early sensory enhancement represents one possibility of cross-modal prediction and interaction especially for audio and visual pathway in emotion processing [7], [44].

In our approach, cross-modal enhancement is achieved by designing two distinct neural groups, with different early pathways corresponding to the auditory and visual modality [43], as shown in Figure 1. We use spiking neural networks to translate this cross-modal enhancement where spiking patterns in the visual modality affect the auditory part. This translates early multisensory integration in the brain; that is, influencing auditory processing with visual neurons spikes. The auditory excitatory layer receives input from both the auditory input layer and the visual excitatory layer. Following the same pattern in the brain where visual information precedes by few milliseconds the auditory processing. This is different from early multimodal data fusion introduced in Section II, which simply concatenate features extracted from each modality while ignoring interactions between them. It is also different from the recent cross-model learning [26] where a cross-modal transfer from the visual to auditory data is applied. Our approach is more biologically plausible, where the auditory part does not use prediction from the visual but learns from the spiking patterns. This represents a multisensory learning, which helps propagate spikes from the visual group to the auditory group [45].

Here we summarise the main workflow in Figure 1. We generate Poisson spike trains from both image and audio input data, and feed the spike trains to both visual and audio SNNs to learn distinctive features of image and speech for each type of emotions. We use the same SNN architecture presented in Section IV-A to learn the visual modality. The input layer for each modality is then recurrently connected to an excitatory layer that is in turn connected to an inhibitory layer in a one-to-one aspect providing a lateral inhibition. Where neurons in the inhibitory layer are connected to all neurons from all features in the excitatory layer apart from the ones it receives input from. Similarly, we have added pre-processing and a convolution layer on the input layer for better feature learning [42].

We then connect the spiking activity of visual neurons at the excitatory layer to audio neurons at the excitatory layer in the audio SNN. These connections will activate additional neurons in the audio SNN. 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. The neurons spiking for the audio modality play a multisensory role, accepting input from the visual modality. Connections between the visual excitatory neuron groups and the auditory neuron groups help the 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. The evaluation of the unsupervised STDP learning is achieved in two stages. During the training, weights are updated after each training interval and spiked 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.

V. Experiments

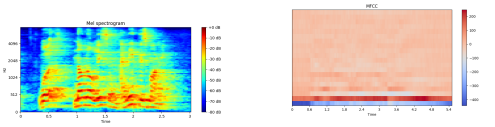
The main purpose of the experiments is to assess (1) the suitability of using SNN for speech emotion recognition with audio data alone and (2) the effect of early cross-modal enhancement of audio using visual modality. To do so, we have designed the following experiment methodology.

A. Datasets

We use some common, third-party datasets for our experiments. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [46] is a multimodal database composed of the basic emotions through speech and songs. The dataset consists of 24 participants with a balanced gender number. The subjects are professional actors, reading sentences in emotional states such as happy, sad, angry, fearful, surprised and disgust. The recordings are available through video, audio, and audio-video options. The eNTERFace dataset [47] is an acted dataset of 42 subjects from 14 nationalities. With a proportion of 81% males and 19% females, all being English speakers. This dataset includes subjects with glasses and beard. The audio is recorded as 48000 Hz in 16-bit format. Each subject records the same basic emotions as in the RAVDESS dataset. We split both datasets into 80% for training and 20% for testing. The number of runs in the training phase corresponds to the number of data in the training set.

B. Feature Extraction

We extract audio features that achieve best performance in SER in the community of speech recognition; *i.e.*, Mel-scale spectrogram, and MFCCs, as shown in Figure 2.



(a) Mel-Scale spectrogram sample (b) MFCCs features sample

Fig. 2: Mel-Scale spectrogram and MFCC features

For each audio sequence, Mel-scale spectrogram is extracted using Fast Fourier Transform (FFT) [48]. First the magnitude spectrogram is calculated from the raw input signal. Then it is mapped onto the Mel scale with a power spectrum. We choose the FFT window with a length of 128, which enables to transform the time domain signal into a frequency domain. The Mel-scale features are then computed using librosa python library [49].

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 2a presents result of Mel-scale spectrogram from the eNTERFACE dataset sample with 'angry' emotion.

MFCCs are extracted from the Mel-Scale spectrogram by applying logs of power which are calculated for each Mel frequency. Then Discrete Cosine Transform are 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 [49]. 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 results in smaller size are padded to match the chosen setting. A sample of the MFCCs features for 'angry' emotion is shown in Figure 2b.

C. SNN Architecture for SER

All experiments are implemented using the Spiking neural Network simulator BRIAN [50]. We have used the same parameters as [34] in terms of input firing rates, membrane threshold and resting phase duration. The input layer of the network architecture consists of two groups of neurons each representing a modality. The number of neurons for each input neuron group is proportional to the size of the input; that is, the size of the audio features and video frame features. We use 40*388 and 100*100 input neurons for the auditory and visual input respectively. The input layer is then connected to a convolution excitatory layer which is connected to an inhibitory layer with a lateral inhibition, where neurons are connected to all neurons in the excitatory layer apart from the one receiving information from. 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 convolutional window and the number of features and the total excitatory neurons. We have experimented with various configuration and have chosen the best performing ones which are using 10 as the window size and the stride size for the auditory and 10 for the visual. The number of features is set to 60 for the auditory modality and 60 for the visual modality.

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.

D. Multimodal Integration Baseline

In order to compare our proposed approach of early cross-modal enhancement, we have created a basic multimodal integration approach with audio and visual modalities. We achieve an early fusion by creating distinct multimodal neuron groups which gets input from both audio and visual excitatory layers.

VI. Results and Evaluation

This section will present experiments results, discussion and shed insights on the use of SNN and cross-modal interaction in speech emotion recognition. First we describe the effect of convolution configuration on the overall results. Then we investigate the results for unimodal SER tasks. Finally we discuss the results for early multisensory integration for the enhancement of SER tasks.

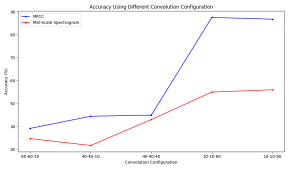


Fig. 3: Effect of convolutional window configuration on overall accuracy

1) Effect of The Number of Features And The Size of Convolutional window: The network is experimented with various convolutional window sizes and number of features. Results in Figure 3 show that the overall accuracy increases when the convolutional size is smaller and the number of features are large. Increasing the number of features leads to an increase in the number of excitatory neurons; *i.e.,*, a better accuracy. The pattern is observed using both MFCCs and Mel-scale spectrogram features [34], [39]. Having more features and more excitatory neuron leads to learning more features. In the end, we choose a convolution window of 10 and 60 number of features.

A. SER With Audio Data Alone

Table I presents the accuracy of SER using two types of features: Mel-scale Spectrogram and MFCC. MFCCs achieve a higher accuracy than raw audio signals or Mel-scale spectrogram for both eNTERFACE and RAVDESS datasets with 72.2% and 80.29% respectively. This shows that MFCCs are an effective type for audio features for processing speech data in SNN. This is also in line with state-of-the-arts methods where MFCCs outperforms other types of audio features in emotion recognition [51].

TABLE I: Comparison of SER accuracy with audio data only between different audio features

Feature extraction	eNTERFACE(%)	RAVDESS (%)
Mel-scale spectrogram	42.1	45.1
MFCCs	72.2	80.3

Table II compares the SER accuracy between the SNN and the state-of-the-art techniques. As an unsupervised learning technique, the SNN has produced comparable results, and in certain cases it outperforms some of the state-of-the-art. Fu et al. [9] introduce an Enhanced Sparse Local Discriminate Canonical Correlation Analysis approach (En-SLDCCA) using multi-modal feature learning representation. This presents an interesting future work for SNN, as we currently only use one feature representation – MFCC, and it would be beneficial to build multi-modal network to learn on each type of audio features. Fonnegra et al [52] have produced the best accuracy of 91.4% using a convolutional auto-encoder and a data augmentation technique. Data augmentation is out of scope of this paper, but it certainly presents an interesting direction to explore in the future. The proposed use of SNN with only one type of features, MFCC, is comparable to the state-of-the-art techniques without the use of any data augmentation or other features.

TABLE II: Comparison of SER accuracy between SNN and the state-of-the-art techniques on the eNterface dataset

Recognition method	Model	Learning	Accuracy (%)
Noroozi et al. [53]	RF/MFCC	supervised	47.1
Turgut [54]	Acoustic analysis	supervised	56.3
Turgut [54]	Texture analysis of spectrogram	supervised	60.9
Fu et al. [9]	En-SLDCCA	supervised	80.1
Fonnegra et al. [52]	Convolution auto-encoder	supervised	91.4
SNN	**SNN/MFCC**	**unsupervised**	**72.2**

TABLE III: Comparison of SER accuracy on cross-model enhancement between different audio features

Feature extraction	eENTRFACE(%)	RAVDESS (%)
Mel-scale Spectrogram	63.7	54.4
MFCCs	86.3	83.6

1) Cross-Modal Enhancement: Table III presents the SER accuracy for cross-modal enhancement. The accuracy has consistently improved from unimodal learning in Table I on both Mel-scals spectrogram and MFCC for both RAVDESS and eNTERFACE datasets. There is an improvement compared to SER with audio data only. Confusion matrices in Figure 4 and 5 show the difference between the recognition with SNN audio only and the cross-modal enhancement. Although there is an enhancement of the overall accuracy, the confusion matrices show different patterns depending on the emotion classes. '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% 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 can result from a higher information in the visual modality and can translate the inverse-effectiveness of multisensory integration.

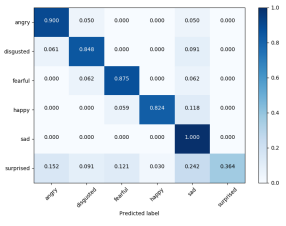


Fig. 4: Confusion matrix for the SNN with audio only on the RAVDES dataset

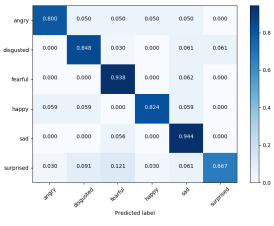


Fig. 5: Confusion matrix for the cross-modal enhancement on the RAVDES dataset

The best overall accuracy of cross-modal enhancement is comparable to the state-of-the-art approaches on the same

TABLE IV: Comparison of SER accuracy between cross-modal enhancement and the state-of-the-art techniques

Recognition method	**Model**	**Learning**	**Fusion**	**Accuracy (%)**
Zhangi et al. [10]	SVM	supervised	A+V	66.5
Noroozi et al. [53]	RF/PCA	supervised	A+V	99.9
SNN with cross-modal enhance-ment	**SNN/MFCC**	**unsupervised**	**A enhanced by V**	**86.3**

eNTERFACE dataset, as presented in Table IV. Our method is outperformed by Noroozi et al. [53]. Their method is a supervised learning approach consisting of late fusion of various classifier on various visual and audio features. They fuse the confidence score of each classifier. Although they report a very high accuracy, Their system can be challenging in a real time environment as they summarise visual frame features using fewer key frames. This could lead to missing key information from real-time data. Our proposed method makes use of the whole visual and audio sequence in order to capture the whole dynamics.

We have also compared our proposed early cross-modal enhancement to a basic multisensory integration architecture. Early cross-modal integration outperforms the basic multisensory integration implementation where it achieves the accuracy of 83.6% for RAVDESS dataset compared to 81.3% for the basic multimodal integration. The latter achieves nearly similar accuracy to unimodal implementation in Table V.

TABLE V: Comparison of SER accuracy between the proposed early cross-modal enhancement to the basic multimodal integration

Model	**Fusion**	**Accuracy (%)**
Multimodal integration	A+V early fusion	81.3
SNN with cross-modal enhancement	**A enhanced by V**	**83.6**

VII. Conclusion and Future Work

In this this paper we have demonstrated that the exploration of different types of classifiers and more biologically inspired architectures can be beneficial for SER tasks. Providing an unsupervised STDP learning proves to be effective for SER, with the reduced reliance on the labelled training data and very large datasets. Using cross-modal enhancement provides us with more accurate recognition compared to a uni-sensory SNN, early fusion or state-of-the-art supervised learning techniques. Using SNN is more computationally advantageous due to their effectiveness on small datasets [55], the support spatio-temporal data [56] and the ability to generalise features across different datasets [55]. Turning towards more biologically inspired architectures can be useful for computing a cognitive model and more accurate learning especially in mutimodal affective computing. In the future, we will extend the experiments to test on incongruent visual and auditory data to assess the robustness of our model.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro M5000 GPU used for this research.

References

[1] 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)*, July 2007, pp. 419–424.

[2] J. Deng, X. Xu, Z. Zhang, S. Frhholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, April 2017.

[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98 – 125, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517300738

[4] C. Vinola and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 14, no. 2, pp. 24–44, 2015. [Online]. Available: http://elcvia.cvc.uab.es/article/view/v14-n2-vinola-vimaladevi

[5] C. Felipe, M. Luis J, and N. Pedro, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[6] 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*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 377–386. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654924

[7] S. Jessen and S. Kotz, "On the role of crossmodal prediction in audiovisual emotion perception," *Frontiers in Human Neuroscience*, vol. 7, p. 369, 2013. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2013.00369

[8] N. Kilian-Hütten, E. Formisano, and J. Vroomen, *Multisensory Integration in Speech Processing: Neural Mechanisms of Cross-Modal Aftereffects*. Boston, MA: Springer US, 2017, pp. 105–127.

[9] 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*, Mar 2017.

[10] S. Zhang, L. Li, and Z. Zhao, "Audio-visual emotion recognition based on facial expression and affective speech," in *Multimedia and Signal Processing*, F. L. Wang, J. Lei, R. W. H. Lau, and J. Zhang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 46–52.

[11] D. Gupta, P. Bansal, and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition," in *Speech and Language Processing for Human-Machine Communications*, S. S. Agrawal, A. Devi, R. Wason, and P. Bansal, Eds. Singapore: Springer Singapore, 2018, pp. 195–207.

[12] 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*, pp. 3517–3521, 2017.

[13] 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*. ACM, 2017, pp. 478–484.

[14] Y. Chavhan, M. L. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine."

[15] 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*, vol. 20, no. 1, pp. 27–41, Mar 2017. [Online]. Available: https://doi.org/10.1007/s10772-016-9364-2

[16] 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*, vol. 5, no. 2, 2017.

[17] I. J. Tashev, Z.-Q. Wang, and K. Godin, "Speech emotion recognition based on gaussian mixture models and deep neural networks," in *2017 Information Theory and Applications Workshop (ITA)*, Feb 2017, pp. 1–4.

[18] 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*, vol. 43, no. 2, pp. 155–177, Feb 2015.

[19] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] 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*, ser. ICCAI 2018. New York, NY, USA: ACM, 2018, pp. 13–18. [Online]. Available: http://doi.acm.org/10.1145/3194452.3194460

[21] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," *Proc. Interspeech 2017*, pp. 1089–1093, 2017.

[22] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 2015.

[23] C. A. Buscicchio, P. Górecki, and L. Caponetti, "Speech emotion recognition using spiking neural networks," in *Foundations of Intelligent Systems*, F. Esposito, Z. W. Raś, D. Malerba, and G. Semeraro, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 38–46.

[24] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5135–5139.

[25] C. Gallicchio, A. Micheli, and L. Pedrelli, "Deep reservoir computing: a critical experimental analysis," *Neurocomputing*, vol. 268, pp. 87–99, 2017.

[26] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," *arXiv preprint arXiv:1808.05561*, 2018.

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

[28] 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)*, vol. 03, May 2015, pp. 1–6.

[29] 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)*, March 2016, pp. 1–9.

[30] 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*, vol. 69, pp. 131 – 140, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010945215001604

[31] J. T. Jose, J. Amudha, and G. Sanjay, "A survey on spiking neural networks in image processing," in *Advances in Intelligent Informatics*, E.-S. M. El-Alfy, S. M. Thampi, H. Takagi, S. Piramuthu, and T. Hanne, Eds. Cham: Springer International Publishing, 2015, pp. 107–115.

[32] J. P. Dominguez-Morales, Q. Liu, R. James, D. Gutierrez-Galan, A. Jimenez-Fernandez, S. Davidson, and S. Furber, "Deep spiking neural network model for time-variant signals classification: a real-time speech recognition approach," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[33] J. Wu, Y. Chua, and H. Li, "A biologically plausible speech recognition framework based on spiking neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[34] P. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 9, p. 99, 2015. [Online]. Available: https://www.frontiersin.org/article/10.3389/fncom.2015.00099

[35] H. Mostafa, "Supervised learning based on temporal coding in spiking neural networks," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 7, pp. 3227–3235, 2018.

[36] H. S. Seung, "Learning in spiking neural networks by reinforcement of stochastic synaptic transmission," *Neuron*, vol. 40, no. 6, pp. 1063–1073, 2003.

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

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

[39] N. Rathi and K. Roy, "Stdp-based unsupervised multimodal learning with cross-modal processing in spiking neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2018.

[40] Y. Cao, Y. Chen, and D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition", journal="international journal of computer vision," vol. 113, no. 1, pp. 54–66, May 2015. [Online]. Available: https://doi.org/10.1007/s11263-014-0788-3

[41] 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)*. IEEE, 2018, pp. 1–6.

[42] D. J. Saunders, H. T. Siegelmann, R. Kozma, and M. Ruszink, "Stdp learning of image patches with convolutional spiking neural networks," 2018.

[43] B. E. Stein, T. R. Stanford, and B. A. Rowland, "Development of multisensory integration from the perspective of the individual neuron," *Nature Reviews Neuroscience*, vol. 15, pp. 520–535, 2014.

[44] S. Molholm, W. Ritter, M. M. Murray, D. C. Javitt, C. E. Schroeder, and J. J. Foxe, "Multisensory auditoryvisual interactions during early sensory processing in humans: a high-density electrical mapping study," *Cognitive Brain Research*, vol. 14, no. 1, pp. 115 – 128, 2002, multisensory Proceedings. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926641002000666

[45] K. Strelnikov, J. Foxton, M. Marx, and P. Barone, "Brain prediction of auditory emphasis by facial expressions during audiovisual continuous speech," *Brain topography*, vol. 28, no. 3, pp. 494–505, 2015.

[46] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[47] 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)*, vol. 00, 04 2006, p. 8.

[48] 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)*, July 2018, pp. 1–5.

[49] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa : Audio and music signal analysis in python," 2015.

[50] D. Goodman and R. Brette, "Brian: a simulator for spiking neural networks in python," *Frontiers in Neuroinformatics*, vol. 2, p. 5, 2008. [Online]. Available: https://www.frontiersin.org/article/10.3389/neuro.11.005.2008

[51] 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)*, Aug 2017, pp. 1–4.

[52] R. D. Fonnegra and G. M. Díaz, "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model," in *Human-Computer Interaction. Theories, Methods, and Human Issues*, M. Kurosu, Ed. Cham: Springer International Publishing, 2018, pp. 385–396.

[53] F. Noroozi, M. Marjanovic, A. Njegu, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 06 2017.

[54] T. zseven, "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition," *Applied Acoustics*, vol. 142, pp. 70 – 77, 2018.

[55] E. Mansouri-Benssassi and J. Ye, "Bio-inspired spiking neural networks for facial expression recognition: Generalisation investigation," in *International Conference on Theory and Practice of Natural Computing*. Springer, 2018, pp. 426–437.

[56] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition," *Neural Networks*, vol. 41, pp. 188–201, 2013.