FACE-VOICE ASSOCIATION TOWARDS MULTIMODAL-BASED AUTHENTICATION USING MODULATED SPIKE-TIME DEPENDENT LEARNING

Nooraini Yusoff[1], and Mohammed Fadhil Ibrahim[2]

[1] *School of Computing, UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia, nooraini@uum.edu.my*

[2] *Technical College of Management- Baghdad, Middle Technical University, 10047 Bab Al Muadham, Baghdad, Iraq, mfi@mtu.edu.iq*

ABSTRACT. We propose a reward based learning to associate face and voice stimuli. In particular, we implement learning in a spiking neural network paradigm using modulated spike-time dependent plasticity (STDP). The face and voice stimuli are paired with a temporal delay, and the network is trained to associate the paired face-voice with a target response. The learning rule is dependent on a reward policy in which the network is given a positive reward for a correct response to a face-voice stimulus pair, or the network receives a negative reward for an incorrect response. Despite a stochastic environment, the learning result of real images and sound indicates a good performance with 77.33% accuracy. The result demonstrates that a machine can be trained to associate a pair of biometric inputs to a target response.

Keywords: multimodal, associative learning, spiking neural network, spike-time dependent plasticity

INTRODUCTION

Biometric identification approach focuses on how to identify individuals based on their physiological or behavioral characteristics. It based on what the person is, and what the person does. Biometric traits include fingerprint, iris, gait, speech, palm geometry, face, retinal pattern, and signature. In contrast with the traditional approach (e.g., password or PIN code), biometrics are difficult to be stolen, borrowed, or forgotten, that makes this approach is more efficient and desirable in security performance (Jain & Nandakumar, 2012).

Generally, there are two approaches in biometric authentication namely unimodal and multimodal biometric systems (Long, 2012; S. Sahoo & Choubisa, 2012). Unimodal relies on single biometric trait to perform the authentication, while multimodal authentication system relies on two or more biometrics to perform the authentication. Thus, multimodal system provides higher immunity against spoofing attacks.

A variety of multimodal biometrics approaches have been recently studied by researchers. These approaches include facial features and fingerprints (Nayak & Narayan, 2012), face and iris (Kim, Shin, Lee, & Park, 2012), face and ear (Huang, Liu, Li, Yang, & Chen, 2013). Artificial neural network (ANN) is among commonly used recognition and classification techniques for biometrics (e.g. Borah, Sarma, & Talukdar, 2013; Hassan & Habeb, 2012; Hinton

et al., 2012). Furthermore, spiking neural network (SNN), the third-generation of ANN, is considered as a promising paradigm to generate new computational models (Kasabov, 2012). SNNs can model complex information processes because of their ability to integrate and represent different information dimensions (Kasabov, Dhoble, Nuntalid, & Indiveri, 2012), such as time, space, and frequency, as well as to deal with huge amounts of data in an adaptive and self-organised manner.

For this study, we train a network to learn association of biometric stimuli, i.e., face and voice, to a target response. Learning is implemented in a spiking neural network paradigm consisting of a number of neuronal groups that each group may represent stimuli or response. We follow the learning scheme as proposed by Yusoff, Grüning and Notley (2012), in which a reward based approach is applied using a modulated spike-time dependent plasticity (STDP). The task of learning is to associate a stimulus pair of (S_i,S_j) to a target response R_k. The stimulus pair consists of S_i, namely the face and S_j, its voice. The presentation of S_j follows after S_i in a sequence with an inter-stimulus interval (ISI). As a result of learning, R_k activates most spikes in response to its associated face-voice pair (S_i,S_j).

METHODS

In this study, we perform an association learning of face image and voice stimuli. We implement learning using a recurrent SNN in a reward-based learning paradigm. The learning rules are dependent on STDP and a reward function. The reinforcement signal from the reward function is used to modulate the STDP. For each face-voice stimulus pair, we assign a target response. Positive rewards are given for the network to associate a correct pair to its target response, otherwise a negative reward is given for an incorrect response. In our approach, a correct response refers to a group of neurons with highest number of active neurons in response to a pair of face and voice stimuli (inputs).

In our learning experiments, from a number of subjects, we captured their face images and recorded a speech sample of those subjects. The speech sample involved a speaking of the number "one". For image capturing, we used a camera with the specification as follows - Sony *a77*, lens Sony 16-50 2.8mm, 24mp, and for speech collection we used a Dell laptop compatible microphone.

For facial features, we have experimented two approaches namely principal component analysis (PCA)-based Eigenfaces and singular value decomposition (SVD). For speech features, we use wavelet packet decomposition (WPD). The experiments show that the PCA-based Eigenfaces feature extraction approach produces better results than SVD (not discussed in this paper).

Spiking Neural Network

For our learning simulation, we use a recurrent spiking neural network of 2000 neurons. The network model consists of excitatory neurons (80%) and inhibitory neurons (20%), the ratio of pyramidal cells (i.e. excitatory) to interneurons (i.e. inhibitory) in the cortical network, with sparse and random connectivity (no self-feedback). Each excitatory neuron is randomly connected to 100 neurons, and each inhibitory neuron is randomly connected to 100 excitatory neurons only, with synaptic transmission delays between 1 to 20 ms (Izhikevich, 2006).

Neurons are divided into subpopulations of stimulus groups (S), i.e. face and voice neurons with 100 neurons (to represent 100 randomly selected features from face or voice) encoding an input stimulus, response groups (R) with 250 neurons encoding a response, non-selective neurons (NS) and inhibitory pool (IH). A stimulus group is composed of a number of excitatory neurons that are selective to a stimulus. Meanwhile a response group consists of

both excitatory and inhibitory neurons. For lateral inhibition to competitor group(s), each excitatory neuron in the response group has random connections to neurons from its inhibitory pool. The inhibitory pool is connected randomly to its competitor's excitatory neurons. Therefore, triggering a response group would invoke its inhibitory neurons which consequently to prevent activation of its competitors. Neurons from the NS group are not selective to any stimulus but their activities also contribute to learning dynamics, and IH consists of inhibitory neurons.

Association Learning using Modulated STDP

In every learning simulation, for the first 100 ms, we initiate a network with random activity. For this purpose, we stimulate an arbitrary neuron with 20-pA current for every ms. This to simulate a network state with no enhanced activation of neuron groups. After the initialisation state, with the same random activity as the background, the network is given a set of paired face-voice-response mappings $(S_i,S_j) \rightarrow R_k$. For each learning trial, at time t_n we present the first stimulus (face), i.e S_i to the network by stimulating all neurons in S_i with a strong current of 20 pA. After ISI, we stimulate all neurons with the same amount of current to the second stimulus (voice), i.e. S_j to be associated to S_i. An optimal ISI is set to 15 ms, based on a preliminary experiment with a range of 10 – 50 ms. We have found that, for ISI < 15 ms, the network only strongly associates the first stimulus to its target response. Meanwhile, for ISI > 15 ms, the network "forgets" the first stimulus and only strongly associates the second stimulus to its target response. With ISI=15 ms, the response is obtained as a result of an association between the first and the second stimulus.

Each learning runs for 20 minutes simulated time with random presentation of stimulus pairs. Within a 20-ms time window from the onset of the voice stimulus, we count the number of activations in the response groups, i.e, R_k. The response group with the highest number of activations is considered to be the winner. To accelerate the learning, some bias current is supplied to the target winner. This is implemented via stimulation with a 20-pA current to arbitrary neurons (with probability of neurons to be selected is between p=0.25 to 0.5, weak to strong potentiation) in the target response group. The next face-voice pair is presented after a 100-ms delay from the offset of each response interval. Synapse reinforcement is implemented based on a reward policy. The reward policy determines the amount of synapse potentiation (i.e. strong or weak potentiation) or depression.

During the 20-ms response interval, for the first 10 ms, we reward the network based on the number of activations in the response inhibitory groups. This is to reinforce the synapses for connectivity between a stimulus and the target response inhibitory group for preventing the activation of response competitor groups. Then we reward the network for the number of activations in the response excitatory groups within the 20-ms response interval for synapse reinforcement from the stimulus group to the target response excitatory group. The bias current is supplied to both winners of the target response inhibitory and excitatory groups.

In our model, synaptic plasticity is implemented on excitatory synapses only for every 10-ms time step. The synaptic efficacy is dependent on a reinforcement signal (i.e. reward signal), r(t), derived from a reward policy in 2. The signal modulates the synaptic changes read from an STDP function (as in 1).

$$\Delta w_{stdp} = A_+ e^{-\Delta t/\tau +}, \Delta t \geq 0 \; ; \; A_- e^{\Delta t/\tau -}, \Delta t < 0. \qquad (1)$$

where $\Delta t = t_{post} - t_{pre}$, parameters τ_+ (τ_-) is the millisecond-scale time constant, and A_+ (A_-) represents the maximum of the change, Δw_{stdp}, when Δt is approaching 0 (Izhikevich, 2007).

The reinforcement signal, r(t), is obtained from a reward policy that is based on the number of neuron firings (F) of response groups within a response interval of 20 ms. The reward policy to derive r(t) is given by (2):

$$r(t) = \{r(t-1) + 0.5\ F_i \geq 2F_j;\ 1 - F_j/F_i,\ F_j < F_i < 2F_j;\ -0.1,\ F_i < F_j\} \quad (2)$$

where F_i and F_j are the number of firings of a target response group, and non-target group, respectively. The reinforcement signal rate is computed based on the type of reward; strong positive, weak positive and negative reward. The signal determines the amount of modulation to the summation of Δw_{stdp} . Therefore, the reward modulated STDP learning holds (Florian, 2007; Izhikevich, 2007):

$$\Delta w(t) = [\alpha + r(t)]\ z(t) \quad (3)$$

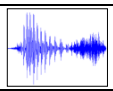
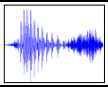
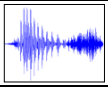
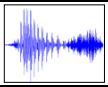
where α is the activity-independent increase of synaptic weight, $r(t)$ and $z(t)$ are the reinforcement signal (2) and eligibility trace, respectively. $z(t)$ represents the summation of Δw_{stdp} obtained from (1). Excitatory and inhibitory weights are initialised to 1.0 and -1.0, respectively. To avoid infinite saturation, weights are kept to be in the range between 0 to 4 mV.

LEARNING RESULTS

All learning simulations were implemented in C++ and testing or probe trials were performed in MATLAB. The proposed learning algorithm was applied to a recurrent excitatory-inhibitory spiking network composed of 1600 excitatory and 400 inhibitory neurons with synaptic interconnections, as described in the earlier section. With the same network and stimulus presentation settings, we implement a series of probe trials (i.e. testing) to recall learned stimulus pairs. The result shows the average percentage of performance over a number of trials, i.e. performance = (number of correct recall/number of trials)*100, for both learning and testing.

For this article, we report a series of experiment of 4 face-voice stimulus pairs (see Table 1). Fig. 1 shows the stimulation activities that have been recorded during the simulation of the training process. Each 'dot' represents an active neuron and an intense activity in a group indicates a significant stimulation of input or a response. From Fig. 1, as learning progresses, the association between a face-voice pair and its target response becomes stronger. As a result of learning, the network is able to associate all the face-voice pairs with their target response with accuracy of 79.11% and 77.33% for training and testing, respectively.

Table 5. Real Dataset for Training and Testing

Pair (F-V #)	Face image	Wave sound	Target response
1			B
2			A
3			A
4			B

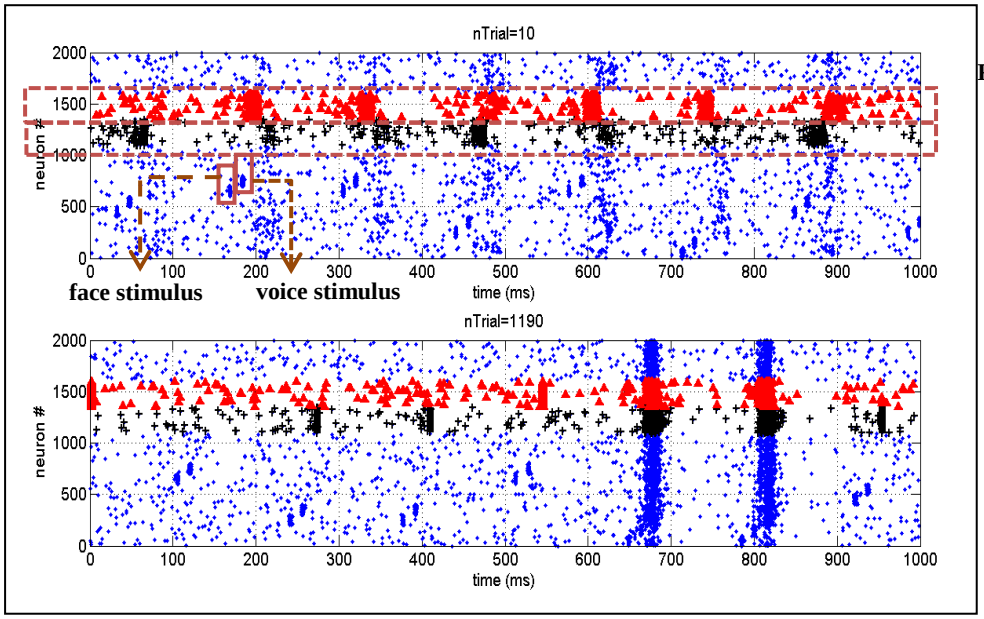


Figure 4. Spike Raster Plot for the Network Activity During the Simulation of Real Data Experiment

CONCLUSION AND DISCUSSION

We propose an associative learning scheme to associate face and voice stimuli to a target response. In our approach, stimuli and responses are represented by a pool of neurons population with random and recurrent connection. The network is trained to provide a target response within a 20-ms time interval upon a presentation of a face stimulus followed by its voice stimulus.

The reinforcement of learning is implemented via a reward policy, in which the under training network is given a positive reward for a correct response in associating the face and its voice, otherwise the network is negatively rewarded. From a series of experiments, it has been shown that, as learning progresses the reinforcement signal eventually influence the network to associate a pair of face-voice stimuli to its target response with a good accuracy.

For synaptic plasticity (i.e. weight updates) we use STDP modulated by the reinforcement signal derived from the reward policy. We follow the learning scheme as suggested in Yusoff, Gruning and Notley (2012), with a different neural network structure. Furthermore, we train a network with real data of biometric features. For preliminary study, we have also performed a number of experiments that using existing face samples from the Olivetti Research Laboratory (ORL) dataset, and speech samples from TIDigits. The results show that the proposed learning model can associate the face with its voice at optimum performance rates of 77.26% and 82.66% for training and testing, respectively. For the experiments reported in this paper, the percentage of training and testing are 79.11% and 77.33%, respectively. Hence, the findings indicate a feasible way to train a machine to associate biometric inputs towards developing a multimodal authentication system.

In several multimodal recognition studies using ANN with Backprogation (BP), e.g., Xu et al. (2013) and Maind & Dehankar (2014), learning is based on gradient-decent method. The

goal of learning is to minimise errors between an output and its target. In such supervised approach, learning requires a template of a target response. In our reward-based learning, no learning template is required, in which an enhanced activity in a response group indicates a strong association between a face-speech pair to its target response. Moreover, learning is simple and natural with minimal assumption on the network dynamic.

Even though, many studies have proven the success of ANN with BP in multimodal recognitions, learning an association between a delayed input pair to its target is complex and challenging. Learning requires an additional mechanism that sometimes may involve massive computation. Realising the needs to learning of complex data, we propose a spiking neural network. The new class of computational models uses time as a resource for coding information, computationally more powerful than the ANN with BP models.

ACKNOWLEDGMENTS

This research has been funded under the Research Acculturation Grant Scheme from Ministry of Higher Education (Malaysia), and supported by Middle Technical University (MTU) - Technical College of Management- Baghdad (Iraq).

REFERENCES

Borah, T. R., Sarma, K. K., & Talukdar, P. H. (2013). Fingerprint Recognition using Artificial Neural Network. *International Journal of Electronics Signals and Systems (IJESS), 3*(1), 98-101.

Florian, R.V. (2007). Reinforcement learning through modulation of spike-timing dependent synaptic plasticity. *Neural Comput., 6,* 1468-1502.

Hassan, Y. F., & Habeb, N. (2012). Hybrid system of PCA, rough sets and neural networks for dimensionality reduction and classification in human face recognition. *IJIIP: International Journal of Intelligent Information Processing, 3*(1), 16-24.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE, 29*(6), 82-97.

Huang, Z., Liu, Y., Li, C., Yang, M., & Chen, L. (2013). A Robust Face and Ear based Multimodal Biometric System using Sparse Representation. *Pattern Recognition*.

Izhikevich, E. M. (2003). Simple Model of Spiking Neurons. *IEEE Trans. Neural Networks, 14*(6), 1569-1572.

Izhikevich, E. M. (2006). Polychronization: Computation with Spikes. *Neural Computation, 18,* 245-282.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb Cortex, 17,* 2443-2452.

Jain, A. K., & Nandakumar, K. (2012). Biometric Authentication: System Security and User Privacy. *IEEE Computer, 45*(11), 87-92.

Kasabov, N. (2012). Evolving spiking neural networks and neurogenetic systems for spatio-and spectro-temporal data modelling and pattern recognition. *Advances in Computational Intelligence* (pp. 234-260): Springer.

Kasabov, N., Dhoble, K., Nuntalid, N., & Indiveri, G. (2012). Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition. *Neural Networks*.

Kim, Y. G., Shin, K. Y., Lee, E. C., & Park, K. R. (2012). Multimodal biometric system based on the recognition of face and both irises. *Int J Adv Robotic Sy*, *9*(65).

Long, T. B. (2012). *Hybrid Multi-Biometric Person Authentication System.* Paper presented at the World Congress on Engineering and Computer Science, San Francisco, USA.

Maind, S. B. & Dehankar, A. V. (2014). A Review on Hand and Speech based Interaction with Mobile Phone. *Journal of Computer Engineering (IOSR-JCE), 4*, 40-44.

Nayak, P. K., & Narayan, D. (2012). Multimodal Biometric Face and Fingerprint Recognition Using Neural Network. *International Journal of Engineering, 1*(10).

Sahoo, S., & Choubisa, T. (2012). Multimodal Biometric Person Authentication: A Review. *IETE Technical Review, 29*(1), 54.

Xu, C., Du, P., Feng, Z., Meng, Z., Cao, T., & Dong, C. (2013). Multi-Modal Emotion Recognition Fusing Video and Audio. *Appl. Math. Inf. Sci., 7*(2), 455-462.

Yusoff, N., Grüning, A. & Notley, S. (2012). Pair-associate Learning with Modulated Spike-Time Dependent Plasticity. *Artificial Neural Networks and Machine Learning – ICANN 2012, LCNS,* 7552/22012, 137-144, doi: 10.1007/978-3-642-33269-2_18.