Noise Reduction using Neural Lateral Inhibition for Speech Enhancement

Yannan Xing, Weijie Ke, Gaetano Di Caterina, John Soraghan

***Abstract*—Recurrent spiking neurons with lateral inhibition connection play a vital role in human's brain functional abilities. In this paper, we propose a novel noise reduction method that is based on neuron rate coding and bio-inspired spiking neural network architecture. The excitatory-inhibitory topology in the network acts as the temporal characteristic synchrony and coincidence detector that removes uncorrelated noisy spikes. A LIF source encoder is introduced along with the network. The network uses generated binary Short-Time Fourier Transform (STFT) masks according to the rate of processed spike train, which is used to reconstruct the denoised speech signal. The technique is evaluated on noisy speech samples with 5 types of real-world additive noise with different noise strength.**

***Index Terms*—Spiking Neural Network, Speech Enhancement, Noise Reduction, Lateral Inhibition**

I. Introduction

Speech enhancement methods are central to many real-world application designs e.g. hearing aids. Hearing aid technologies are usually designed with the aim to provide reinforced speech signal to hearing impaired people to assist their auditory system processing information under varying noisy environments. Hearing aid technologies are usually portable devices with limited size, weight, and power (SWaP). Such desirable SWaP profile increases the level of challenge of developing speech enhancement algorithms in terms of the characteristic that can satisfy both noise suppression quality and energy efficiency.

In the past decades, numerous speech enhancement techniques have been investigated by researchers. Spectral subtraction [1] subtracts an estimated noise spectrum from the noisy signal to produce the denoised spectrum. Ephraim and Malah introduced the minimum mean-square error (MMSE) [2] that reduces the residual noise level without significantly affecting the original speech components. The optimally modified log spectral amplitude estimator (OMLSA) [3] and improved minima controlled recursive averaging (IMCRA) [4] offer high performance in speech enhancement tasks.

Early work using shallow neural networks [5], [6] estimated Signal-to-Noise-ratio (SNR) based on the spectrogram which is then subsequently used to reduce the noise in each frequency band. In [7] and [8] speech enhancement was considered as a classification problem to predict an ideal mask in the time frequency domain to estimate the presence of speech components. Modern mobile technologies tend to include multiple algorithms using one or more power demanding DSP or FPGA cores to obtain maximal real-world performance. A main challenge of future efficient mobile devices development lies in the trade-off between computing capability and power consumption. Although DNN based models can achieve effective noise reduction, they usually require large datasets to represent various types of noise and multiple hidden layers with a significant number of free parameters. The computational cost and power-hungry nature of DNN based speech enhancement technique makes it difficult for them to be applied on SWaP limited devices.

Compared to standard artificial neural networks, spiking neural networks (SNNs) can achieve significant power efficient computing by employing simplified bio-inspired neuron model as the fundamental processing unit and the event-based spike train as information carrier. Recently, SNNs have been successfully applied in the design of intelligent systems such as object detection[9], speech recognition[10] and speech enhancement[11][12]. Furthermore, many developed neuromorphic computing platforms have demonstrated tremendous potential in real-world power limited applications. The IBM TrueNorth[13] system consists of 5.4 billion transistors with only 70mW power density consumption, which accounts for 1/10000 of traditional processing units. The SpiNNaker[14] platform developed by Manchester provided ASIC solutions to hardware implementation to SNNs. It utilized multiple ARM cores and FPGAs to configure the hardware combined with PyNN[15] software API, which achieved completely scalable SNN hardware architecture with a large scale neuron capacity. The emergence of these hardware technologies demonstrates a strong suitability of applying power efficient neuromorphic computing into real world mobile units.

In this paper, we utilize the SNN's power efficient bio-inspired computing to propose a spectrogram-based rate coding method that can contribute to the efficient lateral inhibition SNN based speech enhancement. The connectivity of lateral inhibitory SNN is applied in a layerwise local to global competitive fashion. The proposed architecture does not need to be trained to react for specific noise type but only uses forward propagation with naturally event-based information processing.

The remainder of the paper is organized as follows. Section

Manuscript received October 28, 2019.
Yannan Xing is with the deep learning and neuromorphic lab, Centre for Image and Signal Processing, Electronics and Electrical Engineering Department University of Strathclyde, Glasgow, UK (e-mail: yannan.xing@strath.ac.uk).

II will introduce the concepts of neural synchronization with lateral inhibition in SNNs and indicate how this can be useful for speech enhancement. Section III proposes the method of transforming the speech to spike domain. Section IV provides the structural information of the spiking neural network. The experiment and evaluation process will be described in Section V. Section VI provides the SNR improvement results on 5 noise types. Conclusions are provided in Section VII.

II. NEURAL SYNCHRONIZATION

The use of lateral inhibition as neural synchrony and coincidence detector was investigated by Abbott[16]. A simple SNN is illustrated in Fig.1. It comprises 3 spiking neurons that each produce a spike train output. In Fig.1(a) neuron A and B are Leaky-integrated and fire (LIF) [17] neurons that interacts with each other via latera inhibitory connections. A simplified differential equation that describes the membrane potential dynamics of a LIF neuron model can be expressed as:

$$\frac{dv}{dt} = \frac{R_{mem} I(t) - v}{\tau_{mem}} \quad (1)$$

where v is the membrane potential, R_{mem} denotes the membrane resistance, τ_{mem} refers to the membrane constant and $I(t)$ stands for the synaptic input current. The LIF neuron reacts to input stimuli that raises a certain amount of membrane potential. Once the membrane potential is greater than a pre-defined membrane threshold, the neuron will emit constant amplitude spikes at a certain frequency which is dependent on the magnitude of membrane potential. The inhibition in the example is modelled as one decreasing its membrane potential due to the other neurons' firing activity. As illustrated in Fig.1(a), Neuron B is inhibited from firing if Neuron A is firing and vice versa. Output neuron C simply receives the output spikes from A and B to generate output spike trains. Fig 1(b) shows the input stimuli (synaptic input current) to neurons A and B are different over a certain time period t. One input excitation makes the neuron firing at constant rate of 25Hz (A in Fig 1(c)) while the other input makes the neuron's firing rate linear changing from 28Hz to 22Hz (B in Fig1 (c)) As shown in the highlighted red rectangular in C in Fig 1(c), the output neuron C will fire maximally in a short period when A and B's firing rates are approximately equal. When they have different firing rates, the two neurons tend to inhibit each other in turn leading to sparse events, until their firing rates reach the range of

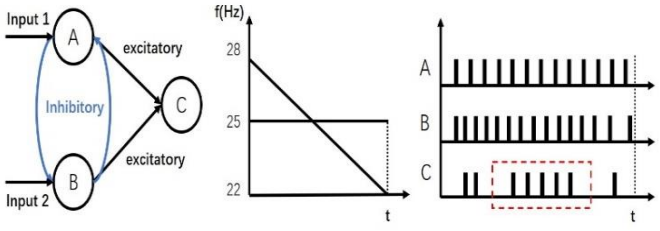


Fig.1. Illustration of two LIF neuron SNN. (a) two LIF neurons (A and B) with lateral inhibitory connections are receiving different input stimuli, the output of A and B are fed in to neuron C. (b) Input neurons (A and B) firing frequency change over a certain time t. Where A's firing frequency linear change from 28Hz to 22Hz, B's firing frequency remain constant at 25Hz. (c) The output spike train observed from A, B and C. The red dashed box highlights the neural synchronous behavior due to lateral inhibition

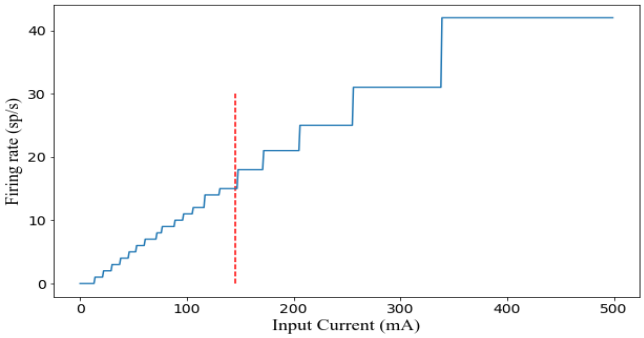


Fig. 2. The firing rate dynamics experiment of a LIF neuron with simulation time step of 100us, the input current linear changes from 0 to 500mA where the neuron firing rate becomes unstable after 115mA.

synchronization (nearly the same). Cornelius et al. [11] demonstrated that multiple fully connected lateral inhibitions are able to exploit the inhibitory process between neurons, to remove uncorrelated spikes (frequency difference). This mechanism can be extremely useful when the speech components are able to be coded into spike trains with similar frequencies.

III. SPEECH SOURCE CODING

We use the Short-Time-Fourier-Transform (STFT) to form the complex spectrogram. The STFT is formed using a Hamming window length of 1024 sample to provide high frequency resolution with 80% overlap which results in 514 frequency channels ranging 0 to 8kHz. The absolute value of output complex matrix from the STFT is log scaled and normalized to the input current to input neurons. The number of input neurons is set to be the same as the number of frequency channels according to the spectrogram.

The use of the lateral inhibitory connections preserves the spike trains that have approximately the same frequency. Each LIF neuron responds to noise and speech components to generate fixed frequency spike train during every single time resolution bin. The LIF neuron is expected to have a higher firing rate in a STFT time resolution bin to the speech components. In contrast, the noise components should be converted to low frequency spike trains which are easier to be distinguished by lateral inhibition. The firing frequency of a LIF neuron usually is proportional to its input, but this is not obvious when simulating it at a very small time step. Fig.2 shows the firing rate dynamics of a LIF neuron with input current linear changing from 0 to 500mA with the simulation time step of 100us. The firing rate profile displayed in Fig. 2 ensures the LIF neurons are able to response differently to a certain range of synaptic input currents during a single time bin of STFT temporal resolution (i.e. the time difference between two adjacent value in a same frequency band). The neuron firing frequency becomes unstable when the input current is over approximately 115mA. Thus, the range of input current from 0mA to 115mA is scaled to provide a balanced input current normalization. In our case, the input neuron firing rate ranges 0-15Hz which means there will be maximally 15 spikes that can be observed in a single time resolution bin of STFT. The full spectrogram is input to 514 LIF neurons by updating the input current of each input neuron based on STFT time resolution. Fig.4 shows the

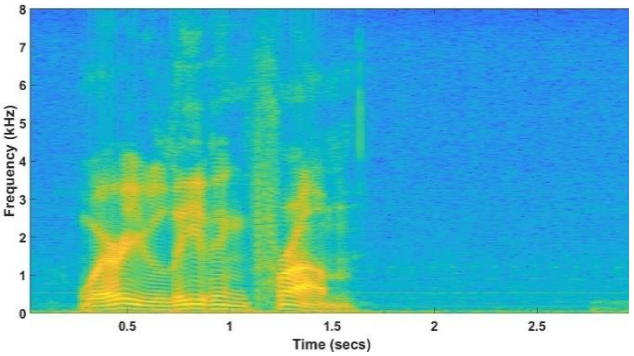


Fig.3. Log Spectrogram of clean speech signal

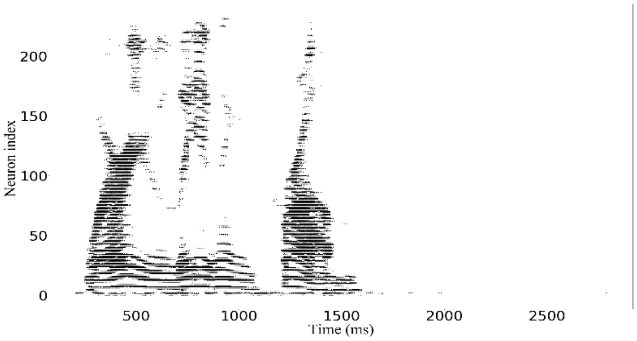


Fig.4. Spike coding result of clean speech samples

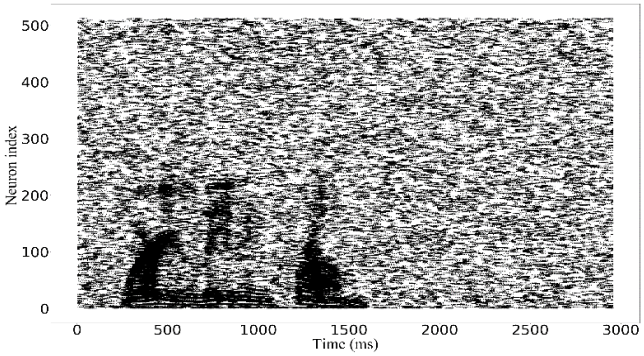


Fig. 5. Spike coding result of noisy speech signal (with white noise SNR = 1)

results of spike coding method applied on the clean speech sample. For comparison purposes, the spectrogram of the same speech sample is shown in Fig.3. It can be seen that the method can represent the temporal-frequency pattern of the speech signal. Fig.5 demonstrates the coding results for white noise corrupted speech sample with SNR=1dB. The resulting raster plot shows the speech components are densely packed (high frequency spike trains). The goal of the SNN is to remove the sparse distributed spikes resulting from noise and preserve the speech elements.

IV. SNN WITH LATERAL INHIBITION

In [12], lateral inhibitory SNN with neighborhood connectivity [18] has been successful demonstrated on Gaussian while noise corrupted speech. Unlike the approaches described in [12] and [18] which uses global inhibitory actions we consider a local to global inhibitory connection strategy. The lateral inhibitory connections are built with different inhibitory radius for each layer. The inhibitory radius defines how neurons in a layer that are close to one another are connected laterally. The basic idea is described in Fig.6. The dynamic inhibitory radius, results in a local to global neural temporal competition while the lateral

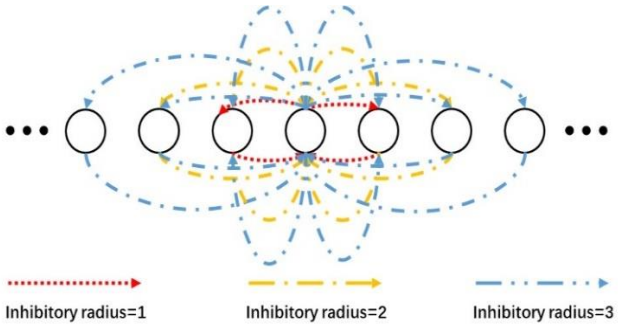


Fig.6. Lateral inhabitation in terms of connection radius which defines how close the adjacent neurons can be lateral connected

inhibition simultaneously removes spikes that are sparsely distributed in time from small set of frequency channel to all frequency bands. To adapt the different inhibitory connection radius for each layer, the inhibitory strength is modelled to exponentially decay in terms of synapse length (distance between two neurons) rather than a constant inhibitory strength for all inhibitory connections:

$$W_n = A_n \cdot \exp\left(-\frac{|i-j|}{D_n}\right),\ i \neq j \qquad (2)$$

where n is the layer index, W represents the weights (strength) of the inhibitory synapse, A represents the maximum inhibitory strength, i and j denote the neuron index of two lateral connected neurons, and D is decay constant. This is because it was discovered through simulation that strong inhibitory for large inhibitory radius (global competition) will result in information loss. On the contrary weak inhibitory for small inhibitory radius (local competition) can have little effect on removing unsynchronized spikes.

V. EXPERIMENT

Test clean speech corpus was obtained from the VoxForge open public dataset[18], where speech sample is sampled at 16kHz. We selected 5 types of real-world environmental noise corrupted speech from the DEMAND noise database [19], including: living, office, river, kitchen and white noise. A range of SNR were chosen for test performance of proposed model (-5, 0.1, 1, 5 and 10)dB.

To simulate the SNN we use Python and the BRIAN simulator[20]. We built a 3-layer SNN with each layer contains 514 LIF neurons that are laterally connected with different inhibitory radii. The connection between each layer is one to one using excitatory synapses. The inhibitory radius from layer one to three are set as: 10, 50 and 250 respectively. The inhibitory synapses parameters are set as {A_1=7, A_2=5, A_3=1, D_1=5, D_2=30, D_3=250}. The output from SNN is spike times of 514 neurons which represents 514 frequency channels of STFT. The quantity of spikes can accurately represent the log intensity of correspond time frequency element. Thus, the processed spectrogram can be obtained by summing the number of spikes and linear decoding for each time resolution cell in the spectrogram. Due to the lack of phase in the log spectrogram we use the

processed spectrogram as a binary mask for the original complex spectrogram. The mask is constructed by comparing the number of spikes in a spectrogram cell to a certain threshold, which determines the ON(1)/OFF(0) status of correspond location. The mask is then element by element

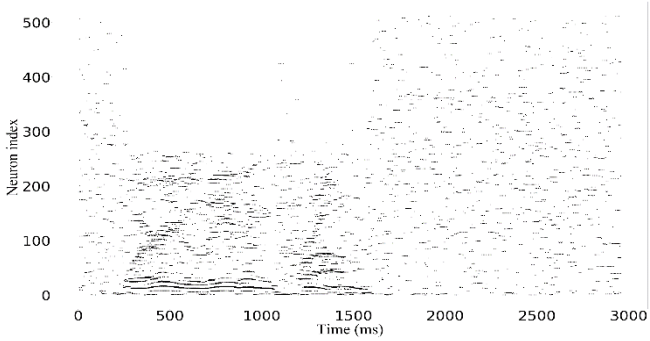


Fig. 6 The SNN spike output, noisy spikes are significantly reduced compared to Fig.5 (white noise SNR=1)

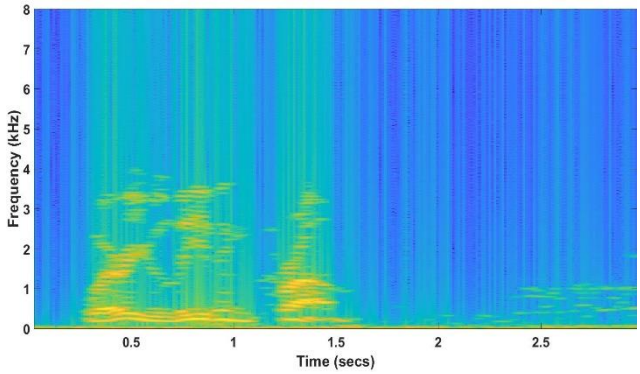


Fig. 7 Reconstructed spectrogram (white noise SNR =1) Note sparse spikes are removed during the thresh hold process

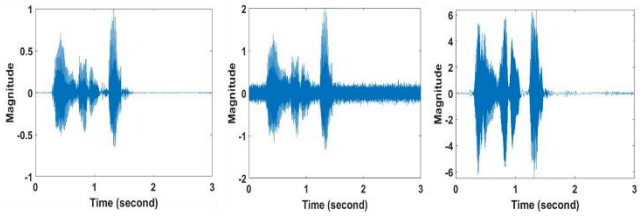


Fig. 8. The time domain signal representation. (a) Time domain clean speech. (b) Time domain noisy speech with white noise SNR=1 (c) Time domain reconstructed denoised signal

multiplied to the both real and imaginary part of original noisy spectrogram. This preserves the original phase information

from the complex numbers of spectrogram which can be used to perform ISTFT to reconstruct time domain signal.

Fig.6 shows an example of raster plot output from SNN. The spikes that are not densely packed in Fig. 6 are further reduced by the thresholding during decoding process (from raster plot to spectrogram). Fig. 7 demonstrate the reconstructed spectrogram using the binary mask. Compare to the Fig.6, most sparsely distributed noise is supressed. Fig 8 further shows an example of time domain signal comparison that (a) is the original clean speech, (b) is the noisy speech corrupted by white noise with SNR = 1dB and (c) is the noise reduced speech signal.

VI. RESULT AND DISCUSSION

Five types of noise sources were used to evaluate the effectiveness of the SNN based method. During the setup of simulations, informal listening was carried out to subjectively determine how successive the method performed. To determine the numerical speech improvement, we pass the clean signal through SNN and used the output signal as the reference target speech signal. The power of residual noise signal is determined by estimating the signal power during the time when no speech is presented. (This assumed that the noise signal is stationary over the presence of speech). The reason we do not using the subtraction to obtain the residual noise is because the non-linear and unsynchronized information processing property of SNN.

Table 1 presents the results on 5 types of noise with different noise level. The lateral inhibitory based SNN can have an average improvement of SNR of 10.915dB among 5 types noises. However, the improvement is noise type dependent, for example it improved only approximately 19dB for white noise but only 8dB for living noise. We strongly believe that the parameter of proposed SNN should be dynamically tuned in terms of the noise level and type. In the future work, we will investigate an automated way of tuning the parameters.

Table 1. Experiment Result

Type of Noise	Original SNR	Measured SNR	Enhanced SNR
White	-5	-5.033	13.62
	0.1	0.0662	19.48
	1	0.9662	23.08
	5	4.9662	24.48
	10	9.9662	24.62
Living	-5	-5.0436	3.22
	0.1	0.0109	8.29
	1	1.0111	9.61
	5	5.0254	13.95
	10	10.0229	15.022
Office	-5	-4.9554	2.515
	0.1	0.1049	8.016
	1	1.0234	9.07
	5	4.995	13.95
	10	10.00	15.022
River	-5	-4.8716	3.26
	0.1	0.0109	8.28
	1	1.149	9.75
	5	5.0254	13.96
	10	10.0229	18.72
Field	-5	-4.133	4.62
	0.1	0.103	9.66
	1	0.9961	10.68
	5	4.991	15.48
	10	9.996	20.8
Kitchen	-5	-4.8275	7.08
	0.1	0.1059	11.91
	1	1.0022	13.48
	5	5.0094	18.60
	10	9.9077	23.83
Average Improvement			10.915

The work presented has demonstrated successive noise reduction on real world noise using multilayer lateral inhibitory spiking neural networks. The lateral inhibitory strengthens the correlations in the time-frequency domain and naturally suppress the noise which are usually sparsely distributed. Unlike standard artificial neural networks, the lateral inhibitory based SNN does not need to train with datasets. However, during our experiment, we noticed that the performance of lateral inhibition is highly dependent on the presence of speech. In Fig.6 and Fig. 7, the noise cannot be efficiently removed by lateral inhibition without the competition from speech. This is nothing to do with the SNN structure but is due to the natural property of inhibitions. A possible solution to this is to separate the speech element from the noise using effective speech detection algorithms. In next stage of our work, we will further improve the performance of lateral inhibitory SNN by applying speech detection algorithms to detect the presence of speech.

VII. Conclusion

In this paper, we have presented a novel spectrogram coding method and a lateral inhibitory SNN structure that naturally suppresses uncorrelated noise in time-frequency domain. It demonstrated an average of 10.915dB SNR improvement on 5 types noise. In addition to the using speech detection algorithms and tuning the network, in future work we would like to extend this research to the hardware implementation. Furthermore, with the emergence of SNN unsupervised learning rule i.e. spiking-time-dependent-plasticity, it is envisaged that the creation of a large scale SNN with ability of denoising and recognition.

References

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust.*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] Q. Zhang and M. Wang, "Speech enhancement for nonstationary noise environments," in *International Conference on Communication Technology Proceedings, ICCT*, 2018, vol. 2017–October, pp. 1663–1667.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[5] Fei Xie and D. Van Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," 2002, p. II/53-II/56.

[6] J. Tchroz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 184–192, 2003.

[7] Yuxuan Wang and DeLiang Wang, "Towards Scaling Up Classification-Based Speech Separation," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[8] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 7092–7096.

[9] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier, "STDP-based spiking deep convolutional neural networks for object recognition," *Neural Networks*, vol. 99, pp. 56–67, 2018.

[10] M. Dong, X. Huang, and B. Xu, "Unsupervised speech recognition through spike-timing-dependent plasticity in a convolutional spiking neural network," *PLoS One*, vol. 13, no. 11, 2018.

[11] C. Glackin, L. Maguire, L. McDaid, and J. Wade, "Lateral inhibitory networks: Synchrony, edge enhancement, and noise reduction," in *Proceedings of the International Joint Conference on Neural Networks*, 2011, pp. 1003–1009.

[12] J. Wall, C. Glackin, N. Cannings, G. Chollet, and N. Dugan, "Recurrent lateral inhibitory spiking networks for speech enhancement," in *Proceedings of the International Joint Conference on Neural Networks*, 2016, vol. 2016–October, pp. 1023–1028.

[13] F. Akopyan *et al.*, "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, 2015.

[14] A. D. Brown *et al.*, "SpiNNaker: A 1-W 18-Core System-on-Chip for Massively-Parallel Neural Network Simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, 2013.

[15] A. P. Davison, "PyNN: a common interface for neuronal network simulators," *Front. Neuroinform.*, vol. 2, 2009.

[16] L. F. Abbott, "The timing game," *Nature Neuroscience*, vol. 4, no. 2. pp. 115–116, 2001.

[17] "The frequency of nerve action potentials generated by applied currents," *Proc. R. Soc. London. Ser. B. Biol. Sci.*, vol. 167, no. 1006, pp. 64–86, 2006.

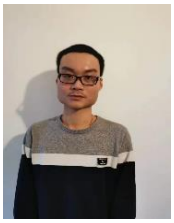
[18] "VoxForge Open Source speech dataset." [Online]. Available: http://www.voxforge.org/home.

[19] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [sound]," Edinburgh, 2017.

[20] D. Goodman, "Brian: a simulator for spiking neural networks in Python," *Front. Neuroinform.*, vol. 2, 2008.

Yannan Xing received his B.Sc. (First Class honors) degree in Electronics and Electrical Engineering from University of Strathclyde, UK in 2015. He is now a full-time Ph.D student at Deep learning and Neuromorphic Computing lab, Centre for Signal and Image Processing, University of Strathclyde.

His current research interests are neuromorphic computing, machine and deep learning, speech processing and event-based data processing.



Weijie Ke received his B.Sc. degree in University of Strathclyde in 2017. He's now a full time PHD student in Centre for Signal & Image Processing(CeSIP), Electronic & electrical engineering department. His research interest now is deep learning and neuromorphic technology for prosthesis.

Gaetano Di Caterina received the B.Eng. degree in computer engineering from the University of Naples in 2005, and the M.Eng. degree in computer and electronic system and the Ph.D. degree from the University of Strathclyde in 2009 and 2013,respectively, with a focus on image and video processing in the context of CCTV system, with an interest on embedded solutions. He is currently a Research Fellow with the University of Strathclyde, where he is involved within the Signal Processing Algorithms and Applications Group, CeSIP. He has authored and presented his work at several international conferences and in academic journals.

John Soraghan received the B.Eng. (Hons.) and M.Eng.Sc. degrees from University College Dublin, Dublin, Ireland, in 1978and 1983, respectively, and the Ph.D. degree from the University of Southampton, Southampton, U.K., in 1989, all in electronic engineering. His doctoral research focused on synthetic aperture radar processing on the distributed array processor. After graduating, he worked with the Electricity Supply Board in Ireland and with Westinghouse Electric Corporation in the United States. In 1986,he joined the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, U.K., as a Lecturer and became a Senior Lecturer in 1990, a Reader in 2000, and a Professor in signal processing in September2003, within the Institute for Communications and Signal Processing (ICSP). in December 2005, he became the Head of the ICSP. He currently holds the Texas Instruments Chair in Signal Processing with the University of Strathclyde. He was a Manager of the Scottish Transputer Centre from 1988 to 1991 and a Manager of the DTI Parallel Signal Processing Centre from 1991 to 1995.His main research interests include signal processing theories, algorithms, and architectures with applications to remote sensing, telecommunications, biomedicine, and condition monitoring.