Neurocomputing 465 (2021) 103–113

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

HuRAI: A brain-inspired computational model for human-robot auditory interface

Jibin Wu [a,1], Qi Liu [a,1], Malu Zhang [a,*], Zihan Pan [a], Haizhou Li [a], Kay Chen Tan [b]

[a] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 29 May 2020
Revised 11 July 2021
Accepted 31 August 2021
Available online 3 September 2021
Communicated by Zidong Wang

Keywords:
Human-computer interface
Deep spiking neural networks
Neuromorphic computing
Voice activity detection
Speaker localization
Voice command recognition

ABSTRACT

The deep learning era endows immense opportunities for ubiquitous robotic applications by leveraging big data generated from widespread sensors and ever-growing computing capability. While the growing demands for natural human-robot interaction (HRI) as well as concerns for energy efficiency, real-time performance, and data security motive novel solutions. In this paper, we present a brain-inspired spiking neural network (SNN) based Human-Robot Auditory Interface, namely HuRAI. The HuRAI integrates the voice activity detection, speaker localization and voice command recognition systems into a unified framework that can be implemented on the emerging low-power neuromorphic computing (NC) devices. Our experimental results demonstrate superior modeling capabilities of SNNs, achieving accurate and rapid prediction for each task. Moreover, the energy efficiency analysis reveals a compelling prospect, with up to three orders of magnitude energy savings, over the equivalent artificial neural networks that running on the state-of-the-art Nvidia graphics processing unit (GPU). Therefore, integrating the algorithmic power of large-scale SNN models and the energy efficiency of NC devices offers an attractive solution for real-time, low-power robotic applications.

1. Introduction

The recent advances in artificial intelligence (AI) have created immense opportunities for the development of robotics, ranging from traditional industrial robots to social robots, assistive robots, and cognitive robots, in applications such as localization [1,2], navigation [3–5] and human-robot interaction [6–8].

Deep neural networks (DNNs), as one of the most profound AI approaches, have spurred on-going interests in robotic applications due to their superior classification and regression capabilities. In [9], DNN-based approaches have been investigated for learning top-level multimodal interaction logic by imitating from a number of human–human interaction examples, which demonstrate high efficiency in reproducing human-like behaviors. In [10], to meet the specific requirement of disabled users, a human-robot interaction (HRI) system incorporating a motor-imagery-based brain teleoperation control is developed to provide assistance and support. In [11], the cascaded convolutional neural network and long short-term memory (LSTM) network are proposed to explicitly model the motion dynamics across image frames, whereby it significantly improves the object recognition accuracy.

Robotic vision, as the major perception peripheral in the aforementioned robotic systems, is sensitive to illumination variation and complex occlusions, which limits its applications under complex visual conditions. Other sensory modalities, for instance, the auditory, tactile and olfactory, also play a critical role in the robotic perception. The speech communication, as the most natural way of human–human interaction, has inspired decades of research on human language technologies. Recently, a novel robot audition system HARK[2] [12] is developed to ease the difficulty of human-robot interaction, by integrating these emerging human language technologies. Moreover, the increased capabilities of data-driven robotic systems also come along with challenges for efficiently analyzing and processing generated big data, during the procedures of data acquisition, transmission, and storage. With an ever-growing demand for DNN-based robotic systems, these challenges are expected to be exacerbated.

The neuroscience research offers a bountiful source of inspiration for building human-like robotic systems. Notably, the brain-

[*] Corresponding author.
E-mail address: maluzhang@nus.edu.sg (M. Zhang).
[1] Jibin Wu and Qi Liu contributed equally in this work.

[2] “Hark” is an old English word meaning “listen”.

https://doi.org/10.1016/j.neucom.2021.08.115
0925-2312/© 2021 Elsevier B.V. All rights reserved.

inspired spiking neural networks (SNNs), which is considered the third generation of neural network models, have shown great potential for high performance, energy-efficient computing [13]. Unlike traditional rate-based artificial neural networks (ANNs), the biologically realistic SNN models explicitly incorporate the concept of time into the computation. They encode and represent information by the precisely timed spikes, therefore, making SNN a promising candidate for processing temporally rich signals such as speech [14,15] and action [16,17]. By incorporating the time into the computation, it has been demonstrated that SNN models are potentially more efficient than ANNs [18,19] for data processing.

Following the computational principle of brains, the SNNs adopt the event-driven computation paradigm, which is carried out through discrete spike trains (i.e., all-or-nothing impulses). By removing the redundancy during the data generation process, SNNs are highly energy-efficient computational models for data-driven applications [20–23]. Recently, a growing number of neuromorphic computing chips are introduced to support ultra-low-power and real-time operation of SNNs, instances include TrueNorth [13], Loihi [18], and Tianjic [24]. By leveraging the massively parallel computing units and asynchronous spike-based computation, they have shown significant improvements in real-time performance and power efficiency in many data-driven applications. Therefore, integrating the algorithmic power of deep SNN models and the energy efficiency of emerging neuromorphic computing architectures represents an intriguing solution for real-time and low-power data-driven applications.

In this work, we present an SNN-based Human-Robot Auditory Interface, that is called HuRAI, by integrating the voice activity detection (VAD), speaker localization (SL), voice command recognition (VCR) systems into a unified framework. The VAD system detects the arrival of voice, while SL and VCR systems interpret ‘where’ and ‘what’ information, so as to activate the appropriate robotic services. As HuRAI is developed based on energy-efficient SNN computational models, it serves as an effective solution for energy-aware mobile robots. To the best of our knowledge, this is the first work that successfully applied SNNs for the aforementioned tasks with competitive accuracies to their ANN counterparts.

The rest of the paper is organized as follows: In Section 2, we give an overview of the spiking neural networks and explain how to effectively train these models. In Section 3, we explain each individual system of the proposed human-robot auditory interface in detail. Then, we present the experimental results on the learning capability, energy efficiency, and robustness of the proposed systems in Section 4. Finally, we conclude the paper in Section 5.

2. Spiking neural network

The brain-inspired spiking neural networks represent an energy-efficient solution for sensory signal processing, as it mimics the biological process of the human brain that evolved by nature. Following the same connectionism principle, SNNs share the same network structure, either feedforward or recurrent, with conventional ANNs. Different from the data representation of the ANNs, the information is represented and exchanged via stereotypical action potentials or spikes in the SNNs. The firing rate and temporal structure [25,26] of the spike train are both considered as important information carriers in the biological neural systems.

In this work, we use the current-based leaky integrate-and-fire (LIF) model for all the spiking neurons, which has been used intensively in computational neuroscience studies. The subthreshold membrane dynamics of the LIF neuron can be described by the following first-order differential equation:

$$\tau_{\mathrm{m}}\frac{dU_i^l(t)}{dt} = -\left[U_i^l(t) - U_{rest}\right] + R_m I_i^l(t) \tag{1}$$

where $U_i^l(t)$ is the membrane potential of neuron i in layer l and U_{rest} is the resting potential. τ_{m} is the membrane time constant that controls the decay rate of the membrane potential. R_m is the membrane resistance and $I_i^l(t)$ is the resulting synaptic current from incoming spikes. According to Eq. (1), the spiking neuron effectively acts as a leaky integrator of synaptic currents. By simplifying the complex chemical transduction process that happened at the synapse, the value of the synaptic current can be determined as follows

$$I_i^l(t) = \underbrace{\sum\nolimits_j W_{ij}^l S_j^{l-1}(t)}_{feedforward} + \underbrace{\sum\nolimits_k V_{ik}^l S_k^l(t)}_{recurrent} \tag{2}$$

where W_{ij} and V_{ik} refer to the strength of feedforward and recurrent connection that converge to neuron i, respectively. $S_j^{l-1}(t)$ denotes the incoming spike train from the presynaptic neuron j of layer $l-1$, and $S_j^{l-1}(t) \in \{0,1\}$. As shown in Fig. 2, the neuron generates an output spike from the soma whenever the membrane potential surpasses the firing threshold ϑ, and transmits the spike to the subsequent neurons along the axon. This spike generation process can be described by a Heaviside step function Θ as follows

$$S_i^l(t) = \Theta\left(U_i^l(t) - \vartheta\right) \tag{3}$$

Without loss of generality, we assume a unitary membrane resistance R_m and firing threshold ϑ in this work. In addition, we set the resting potential U_{rest} to 0. With a small time step Δt, typically 0.1 ms or 1 ms, the continuous function of Eq. (1) can be well approximated by the following discrete-time formulation

$$U_i^l[t] = \alpha U_i^l[t-1] + I_i^l[t] - \vartheta S_i^l[t-1] \tag{4}$$

where α denotes the membrane potential decay rate that takes a value of $exp(-\Delta t/\tau_{\mathrm{m}})$. The membrane potential is reset to the resting potential immediately after the spike generation by the last term in Eq. (4).

The stateful nature of spiking neurons, arising from the intrinsic subthreshold membrane potential dynamic, can be effectively modeled as recurrent ANNs by fixing the recurrent connections at the constant decay rate α [27]. By drawing equivalence to the recurrent ANNs, the network parameters can thus be optimized by unrolling the network through all time steps and applying the BP through time (BPTT) algorithm [28]. The non-differentiable nature of the spike generation function, as revealed by Eq. (3), poses a challenge to directly apply the BPTT algorithm as the derivatives at the time instant of spike generation is ill-defined. While this problem can be effectively overcome by replacing it with a continuously differentiable function during the gradient backpropagation, whereby a surrogate derivative can be derived and applied. In practice, a number of continuously differentiable functions are experimentally proved to be effective [29–33]. Here, we use a triangular function that monotonically increases towards the firing threshold as per Eq. (5). As such, it provides an unbiased gradient approximation for the non-differentiable spike generation function.

$$\frac{dS_i^l[t]}{dU_i^l[t]} \equiv \max\left(0, 1 - \left|\frac{U_i^l[t] - \vartheta}{\vartheta}\right|\right) \tag{5}$$

SNNs naturally deal with spike trains, therefore, a special mechanism is required to encode the analog-valued feature vectors into spike trains. In this work, we consider the feature vectors determined from the raw audio signals as the synaptic currents, and directly apply them into Eq. (4). As such, the first layer of spiking neurons performs neural encoding, and from this layer onward the information is represented as spike trains.

3. HuRAI: Human-robotic auditory interface

In this section, we first give an overview of the proposed human-robot auditory interface HuRAI. We then describe each subsystem in detail.

3.1. HuRAI system architecture

As summarised in [12], an ideal robot audition interface should provide a set of subsystems for signal processing that can be easily chosen and combined. In reality, the different working environments, for instance at home and in factories, typically have diversified acoustic conditions and requirements, which makes joint optimization and evaluation a challenging task. Therefore, we design the proposed HuRAI in the way that individual subsystems can be developed separately with limited interference from each other. As such, the proposed auditory interface not only has the flexibility to be optimized with data from different environments but can also function more robustly under the abnormality of a particular subsystem.

As shown in Fig. 1, the raw audio signals captured by the microphone array are first analyzed by the voice activity detection subsystem. The subsequent systems are turned on only when the voice is detected. With the VAD system, we significantly reduce the computational load by controlling access to speaker localization and voice command recognition services. Given the movement speed and the speaking rate of the speaker is widely different, the speaker localization and voice command recognition systems are designed to operate in parallel with different sampling rates, so as to reduce the total computational cost. Finally, the output speaker location and recognized command information from the associated subsystems are further processed by the onboard controller to determine a desired control signal for the robot.

3.2. Voice activity detection

Voice activity detection is an essential pre-processing step in modern speech processing systems, which distinguishes human voices from background noises. For instance, in the context of speech and speaker recognition, VAD reduces the computational load by discarding the non-voice segments. Meanwhile, VAD extracts the noise segments, which can be used for noise modeling and speech enhancement.

A typical VAD system can be constructed by cascading feature extraction and classification processes. Many existing solutions are available that differ in terms of the feature representation and classifier used. Traditionally, statistical models are utilized to model the data distribution of voiced- and non-voiced segments, with a combination of features [34] including the zero-crossing rate, discrete Fourier transform coefficients, cepstral coefficients, etc. Recently, machine learning models are studied to automatically extract a discriminative feature representation for the VAD task [35,36].

As shown in Fig. 3, we first segment the audio signals into overlapping frames. Then, we apply logarithmic triangular Mel-scaled filters to the resulted power spectrum from the Short-time Fourier transform (STFT) analysis and calculate the power at each Mel frequency band (FBANK). The recurrent SNN takes the FBANK features as input, and it is trained to make VAD decisions for each individual frame at run-time. By sharing the FBANK feature with the VCR subsystem that will be introduced in Section 3.4, the overall computational costs are significantly reduced. Moreover, with a recurrent neural network architecture, the latency of the proposed VAD subsystem is significantly reduced compared to other machine learning models (e.g., MLP, CNN) that require buffering and accumulating multiple frames into a spectrum images [36]. Due to its always-on nature, the VAD system requires an energy-efficient solution. As will be demonstrated in our experimental study, by exploiting the temporal and spatial sparsity with an event-driven SNN model, the proposed VAD model offers up to three orders of magnitude energy saving compared to their ANN-based counterparts.

3.3. Speaker localization

Speaker localization plays a vital role in the HRI, where the robot requires a good understanding of ‘where’ the sound is from, so as to respond to the speaker appropriately. Moreover, the location information is a useful cue for beamforming in speech enhancement.

Traditionally, given the physical layout of the acoustic environment, the speaker location can be determined analytically using signal processing techniques. While the performance may degrade severely if those assumptions are violated in the real environment. Recently, without the knowledge of the physical layout, learning-based approaches are studied to directly map the localization cues

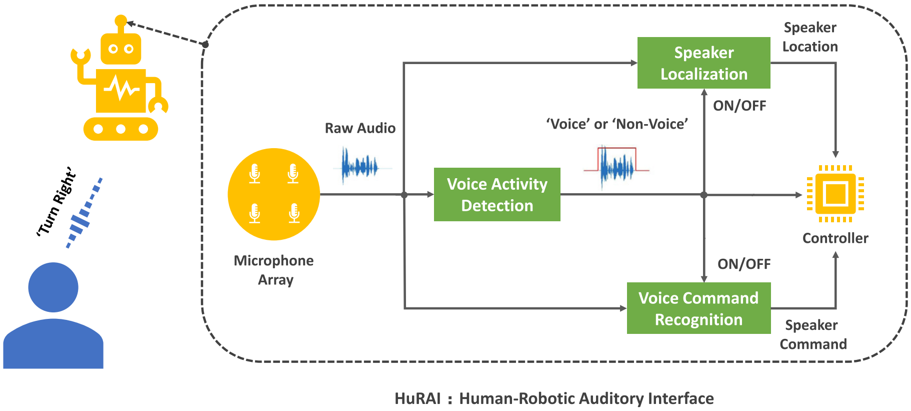


Fig. 1. Illustration of the proposed human-robot auditory interface (HuRAI). The raw audio signals collected from the microphone array will first be analyzed by the voice activity detection system to check whether the human voice is available, and switch on the speaker localization and voice command recognition systems only when the human voice is detected. The speaker location and command determined from the associated systems will be consumed by the embedded controller to generate an appropriate control signal to the robot.

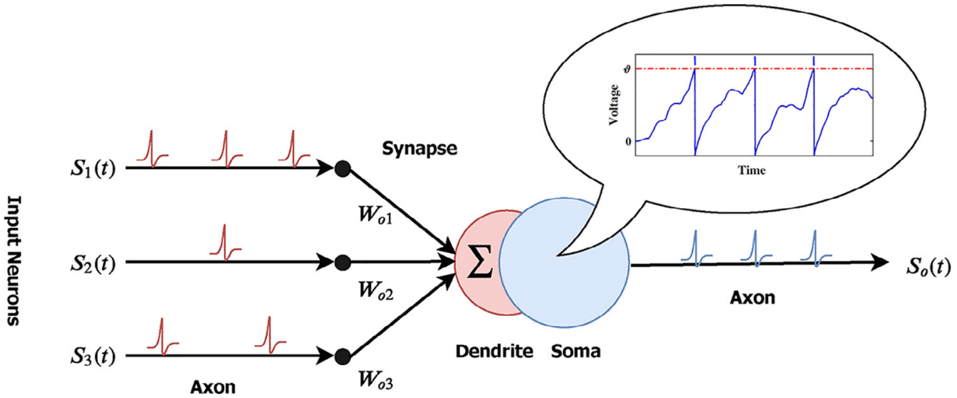


Fig. 2. Illustration of the spiking neuron model.

to the direction-of-arrival (DOA). In practice, with a sufficient amount of domain-specific training data, the ANN-based systems perform exceedingly well as demonstrated in a number of studies [37,38]. Motivated by the compelling performance of the learning-based approaches, we apply the deep SNN to exploit the localization cues derived from the microphone array. Specifically, we apply the generalized cross-correlation with phase transform (GCC-PHAT) to all microphone pairs, and calculate the GCC-PHAT coefficients with delay τ ranging from -25 to 25 samples. The formula for GCC-PHAT coefficient calculation is given as follows

$$\gamma_{\text{ij}}(\tau) = \sum_{\omega} R\left(\frac{X_i(\omega)X_j^*(\omega)}{\left|X_i(\omega)X_j^*(\omega)\right|}e^{j\omega\tau}\right) \tag{6}$$

where $X_i(\omega)$ and $X_j(\omega)$ refer to the inputs of microphone channel i and j, which are calculated from the STFT analysis at the discrete angular frequency ω. The time delay of arrival (TDOA) of the microphone pair ij can be associated with the τ that achieves a maximum GCC-PHAT coefficient. While the signals received are usually corrupted by noise and reverberation, with varying degrees of distortion across different microphone channels, in the real environment. Therefore, instead of using only the peak coefficient, we preserve all the GCC-PHAT coefficients and use them as the input feature to the SNN. As shown in Fig. 3, the multi-layer fully-connected SNN has trained to maps the input GCC-PHAT coefficients to a particular azimuth direction.

3.4. Voice command recognition

To ensure a natural human-robot interaction, the robot is also required to understand the verbal content of the speech signals, so as to respond to the voice command or return the queried information to the speaker. Following the same history of automatic

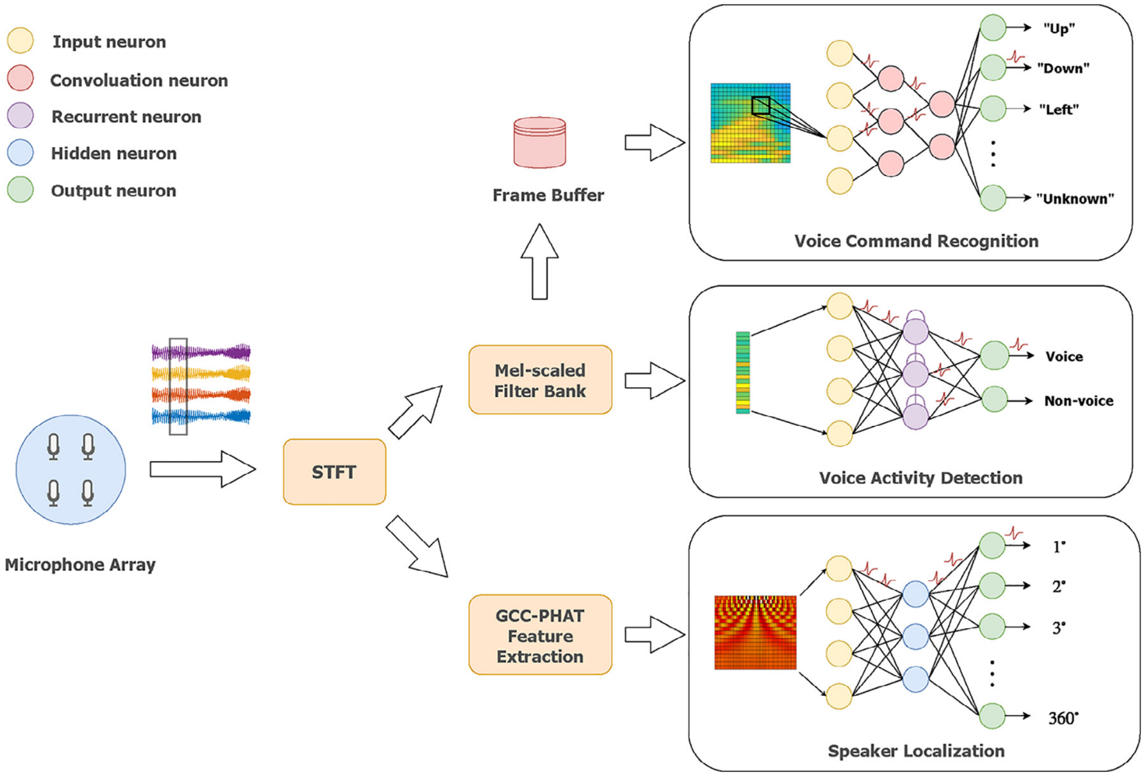


Fig. 3. The SNN implementation of three subsystems in the proposed human-robot auditory interface, HuRAI.

speech recognition (ASR) research, the voice command recognition techniques evolved from the traditional template matching [39], which compares the unknown speech to the pre-recorded speech samples to find the best match, to modern statistical-based machine learning models such as the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) model [40]. Recently, by leveraging a large amount of training data, DNNs have demonstrated superb modeling capability for rich acoustic states and their dynamic transitions, resulting in the ASR systems approaching the performance of human listeners [41].

The speech signals exhibit distinctive spectro-temporal dynamics and different words can be visually discriminated from the spectrogram. Inspired by this observation, convolutional neural networks (CNN) that are initially proposed for visual pattern recognition have been investigated in speech recognition [42] and speaker recognition [43] tasks. The translation invariance property offered by the CNN improves the robustness to small frequency shifts that are common in speech patterns. Meanwhile, CNN employs a weight-sharing scheme that reduces the number of parameters over that of a fully connected DNN.

Hence, we apply a spiking-based CNN for the voice command recognition task. As shown in Fig. 3, instead of processing speech signals in a frame-by-frame fashion, we store frame-wise FBANK feature vectors in a buffer and concatenate them into a fixed temporal dimension before inputting them to the SCNN. Together with speaker localization results, the recognized voice command will be consumed by the robot for action planning and social interaction.

4. Experimental evaluation

In this section, we first introduce the experimental setups to evaluate each subsystem. Then, we present the experimental results for each subsystem in detail. Finally, we discuss the attractive properties of rapid inference, energy-efficient computation, real-time operation, and high noise robustness of the proposed SNN-based framework.

4.1. Experimental setups

We perform all the experiments with Pytorch [44], which provides accelerated and memory-efficient training with GPUs. Following the discrete-time formulation introduced in Section 2, we implement the LIF neuron with a continuous surrogate derivative derived from the triangular function and leverage the autograd capabilities offered by Pytorch for efficient SNN training. The experimental datasets, model structures, implementation details, and evaluation metrics of each subsystem are provided in details in the following subsections. For each subsystem, instead of optimizing the network structure to achieve the optimal performance on the selected evaluation dataset, we focus on demonstrating the superior modeling capability of SNNs to their non-spiking ANN counterparts as well as to shed light on their high computational efficiency.

4.1.1. Voice activity detection

We evaluate the proposed VAD system on the QUT-NOISE-TIMIT corpus [35]. This corpus consists of 600 h of noisy speech sequences by mixing clean speech recordings from the TIMIT dataset with noise recordings from the cafe, home, street, car, and reverberant environments. The noisy speech sequences are constructed at 6 different signal-to-noise ratios (SNRs) ranging from -10 dB to 15 dB with a sample length of 60 or 120 s, so as to provide a comprehensive evaluation for different acoustic conditions. The speech sequences are further segmented into frames with a duration of 160 ms and 50% overlap. We comply with the evaluation protocol specified in the corpus and divide the corpus into groups A and B with noise recorded at different locations. We perform training and testing using group A and B in alternation (i.e., train on A (or B) and test on B (or A)) under 3 different noise levels: Low Noise (SNR = 15 and 10 dB), Medium Noise (SNR = 5 and 0 dB) and High Noise (SNR = -5 and -10 dB).

We extract 40-dimensional FBANK features and use them as the input to the recurrent SNN. The recurrent SNN model is constructed with 1,024 hidden spiking neurons and two linear readout neurons that correspond to 'Voice' and 'Non-Voice', respectively. The cross-entropy loss function is used with framewise labels extracted from the corpus. We trained the model using Stochastic Gradient Descent (SGD) optimizer for 100 epochs with a batch size of 16. The initial learning rate is set to 0.01 and decays to one-tenth after every 40 epochs. Following the evaluation protocol specified in the corpus, we report the half-total error rate (HTER), which takes the average of the miss rate (MR) and false alarm rate (FAR).

4.1.2. Speaker localization

We use the RSL2019 corpus [45] to evaluate the proposed speaker localization system. This corpus consists of 25,920 speech utterances spoken by well-mixed male and female speakers. The corpus is recorded inside a professional recording studio using a four-channel microphone array. The speech utterances are played through a loudspeaker with the DOA varying from 0 to 360 degrees at an interval of 5 degrees, and we use both the subsets recorded at the distance of 1.0 and 1.5 meters. We sample 170 ms frames with a stride of 85 ms from the speech utterances, resulting in a total of 133,103 training and 24,792 testing samples.

The fully-connected SNN model used in this work includes 3 hidden layers with 1,000 spiking neurons each and 360 linear output neurons correspond to 360 different azimuth directions with 1-degree intervals. The output label is encoded into a 360-dimension vector using a normalized Gaussian-like function with a standard deviation of 8, wherein the entries correspond to the posterior probability of each azimuth direction. We trained the model using the SGD optimizer for 30 epochs, with a learning rate of 0.1 and a batch size of 256. The mean square error loss function is used during training, and we report the mean absolute error (MAE) by taking the highest peak of the output as the predicted azimuth direction.

4.1.3. Voice Command Recognition

The voice command recognition system is evaluated on the Speech Commands corpus [46], which is designed for the keyword spotting task. All utterances are post-processed to have a uniform length of 1 s. Following the same data preparation procedure of a previous study [47], we select 10 words out of 31 words in the corpus, that are most relevant to robotic commands, namely 'Yes', 'No', 'Up', 'Down', 'Left', 'Right', 'On', 'Off', 'Stop', 'Go'. To ensure the system responds appropriately to the out-of-vocabulary words, we create an 'Unknow' class by randomly selected 20% samples outside the 10 words, excluding the 'Background_noise' class which is handled directly by the VAD system. We further randomly split the samples into train and test set with a ratio of 80% to 20%, resulting in a total of 25,493 train and 6,388 test utterances, respectively. The utterances are segmented into frames of 25 ms length with 50% overlap. The FBANK features are extracted and concatenated into a spectrogram before input to the spiking CNN for classification. We apply data augmentation to the resulted spectrogram by randomly masking blocks of data in both frequency and temporal dimension [48] during training.

The spiking CNN model used in this work is inspired by the AlexNet [49] with a network structure of 24c3-24c3-48c3-48c3-96c3-256–11, where the numbers before and after 'c' refer to the number of convolution kernels at a particular layer and the kernel size respectively. The kernels are applied at a stride of 2 for layer 1, 2, 3 and 5, so as to reduce the dimension of the resulted feature maps. The model is trained using the cross-entropy loss function and SGD optimizer for 100 epochs, with a batch size of 64. The initial learning rate is set to 0.01 and decays to one-tenth after every 30 epochs.

4.2. Voice activity detection results

We benchmark the results of the implemented VAD system against seven baseline systems, including the standard-based European Telecommunications Standards Institute (ETSI) [50] and G729B [51], the statistical-based Long Term Spectral Divergence (LTSD) [52] and Sohn [53], the machine learning-based GMM [35] and CNN [36]. Recently, an SNN-based VAD system has been introduced, which is called Bin e. [54]. This system integrates a novel bin coding mechanism to encode the spectral information and classify each frame by a single layer of spiking neurons.

As the results are shown in Fig. 4, our recurrent SNN-based system outperforms all the baseline systems under Low Noise (Fig. 4)) and Medium Noise (Fig. 4(b)) scenarios, with an HTER of only 2.72% and 6.67%, respectively. Under the High Noise(Fig. 4(c)) scenario, our system still surpasses most of the baseline systems with an HTER of 15.0%, except for the CNN baseline that has a slightly lower FAR. This may due to the translation invariance property offered by CNN, whereby exhibits better noise robustness. It worth noting that our VAD system outperforms the only existing SNN-based system by a large margin across all different noise scenarios, which may credit to the better modeling capacity with a recurrent network structure.

To allow a better evaluation of our system, we have provided a test example under the Low Noise scenario as shown in Fig. 5. It is obvious from Fig. 5(c) that the activation of hidden spiking neurons is self-organized into two distinct groups that are synchronized to either the 'voice' or the 'non-voice' segments. Moreover, as shown in Figs. 5(d) and 5(e), the prediction of our model largely aligns with the ground truth, while misses out only at a few abrupt transitions.

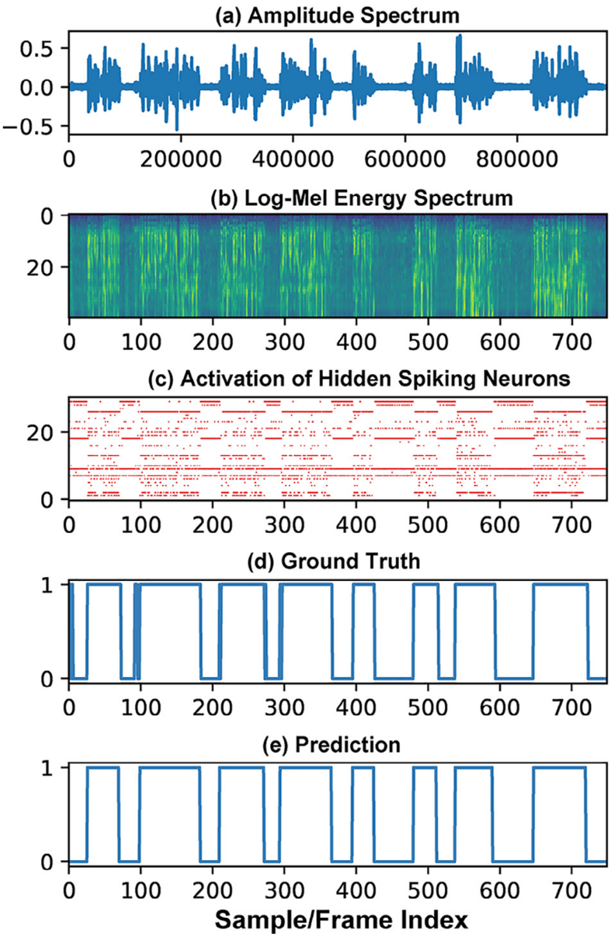


Fig. 5. Illustration of the voice activity detection system results. The activity of the hidden spiking neurons is self-organized and synchronized to either the 'voice' or the 'non-voice' segments. The prediction of our VAD system largely aligns with the ground truth and misses out on only a few abrupt transitions.

4.3. Speaker localization results

As reported in Fig. 6(b), at the distance of 1.5 meters, the proposed speaker localization system achieves an MAE of 1.19 degrees on the test set with an encoding time window of only one. It outperforms the baseline MUltiple SIgnal Classification (MUSIC) algorithm, which reported an MAE of 2.35 degrees [45]. At the distance of 1.0 meter, the proposed system achieves an MAE of 3.86 degrees that is also competitive to the traditional MUSIC algorithm that reported an MAE of 3.79 degrees [45]. The performance drop from 1.5 meters to 1.0 meter can be explained by the relatively large physical size of the loudspeaker to the sound source, which adversely affects the measurement precision [45]. These results highlight the superior performance of a learning-based approach

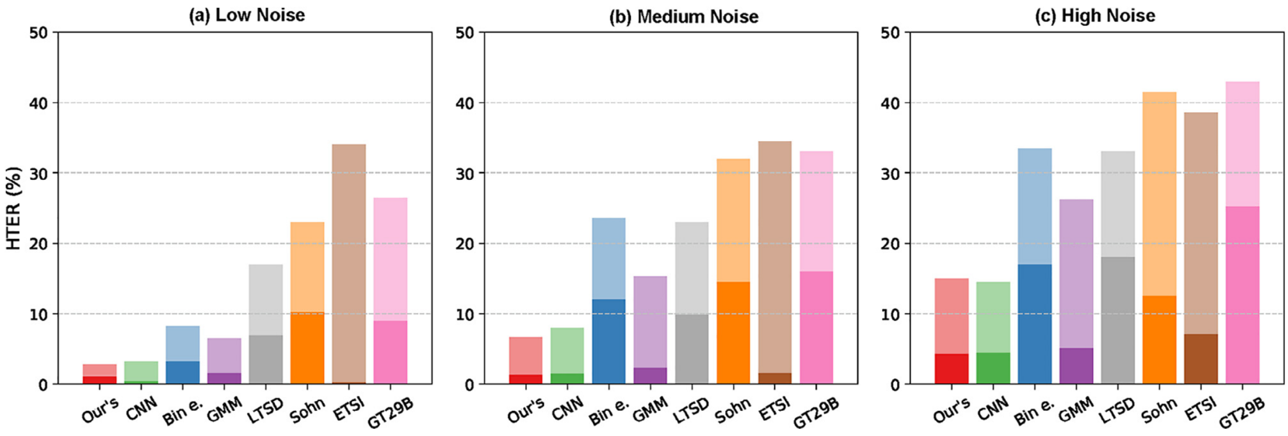


Fig. 4. HTER performance of the proposed SNN-based VAD system and other baseline systems over three noise levels.

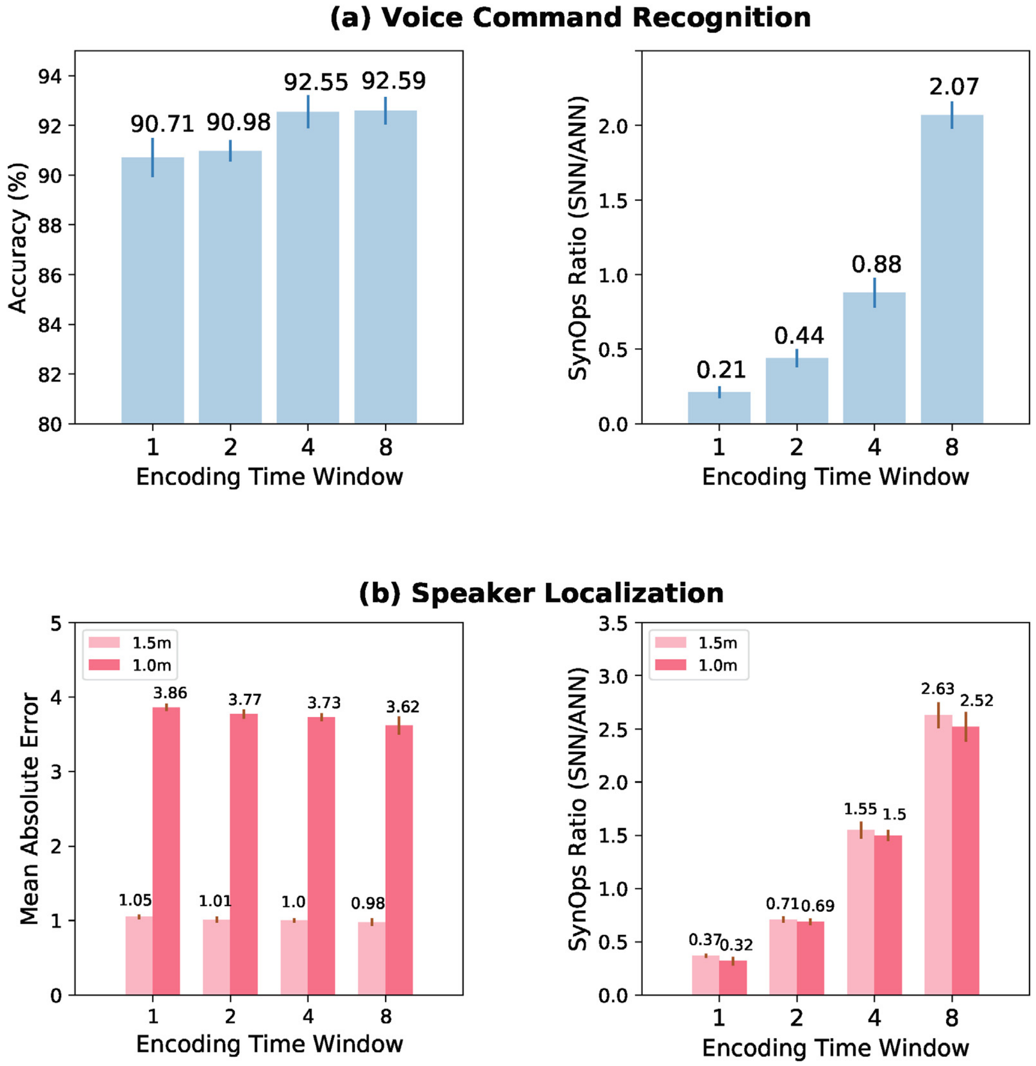


Fig. 6. Summarize the performance of (a) voice command recognition and (b) speaker localization systems. The results are obtained from 3 independent runs.

compared to conventional signal processing techniques on the speaker localization task. Moreover, we achieve these results with only one simulation time step, suggesting accurate and rapid localization can be achieved simultaneously. We observe a better performance with a longer encoding time window, the results converge when the encoding window reaches a size of four. We notice that the synaptic operations (SynOps) ratio, an important benchmark for energy efficiency of SNN-based systems, increases linearly with the encoding window size, which will translate into the same growth rate for energy consumption. Therefore, in practice, a suitable encoding window size should be decided, as a trade-off between accuracy and energy efficiency.

4.4. Voice command recognition results

In Fig. 6(a), we report the results for the voice command recognition system. With an encoding window size of only one, we observe that competitive classification accuracy of 90.71% is achieved on the test set. The classification accuracy can be further improved to 92.59% when a larger time window size of eight is provided. While the energy consumption grows linearly with time window size, which follows a similar trend as the SNN-based speaker localization system. Using the same experimental setup, a recent study has achieved a test accuracy of 93.59% with a non-spiking CNN implementation. This result is on par with our SNN-based system, suggesting the superior modeling capability can be achieved with spiking neural networks.

4.5. Energy efficiency analysis

Beyond superior modeling capabilities, HuRAI holds great potentials to improve energy efficiency when implemented on the emerging low-power neuromorphic computing chips. To shed light on this attractive prospect, we follow the convention of the neuromorphic computing community by calculating the average SynOps on a random batch of testing samples and comparing them to the equivalent ANN implementations on the state-of-the-art Nvidia GPU.

For ANNs, the Multiply-and-Accumulate (MAC) operations are typically used. Given a particular network architecture, the total SynOps is a constant number that can be calculated as follows

$$SynOps(ANN) = \sum_{l=1} f_{in}^{l} N_l \tag{7}$$

where f_{in}^{l} is the average number of fan-in connections to neurons in layer l and N_l is the total number of neurons in that layer.

Table 1
System performance with different extent of network activity regularization. The results for voice command recognition and speaker localization systems are obtained with a time window of 4.

System	λ	SynOps Ratio	SynOps Savings	Acc. (%)	Acc. Loss (%)
Voice Command Recognition	0	0.88	1x	92.03	0
	0.01	0.79	1.11x	92.49	0.46
	0.1	0.60	1.47x	92.55	0.52
	1.0	0.29	3.03x	89.76	−2.27
	λ	SynOps Ratio	SynOps Savings	MAE (%)	MAE Loss (%)
Speaker Localization	0	1.57	1x	1.14	0
	0.0001	1.50	1.05x	1.14	0
	0.001	1.20	1.31x	1.15	−0.01
	0.01	0.35	4.49x	1.25	−0.11
	λ	SynOps Ratio	SynOps Savings	HTER (%)	HTER Loss (%)
Voice Activity Detection	0	0.28	1x	2.65	0
	0.01	0.194	1.44x	2.85	−0.20
	0.1	0.048	5.83x	2.99	−0.34
	1.0	0.013	21.54x	4.09	−1.44

While cheaper Accumulate (AC) operations are used in SNNs, and total SynOps are positively correlated with the number of fan-out connections to the subsequently layer, average firing rate of spiking neurons and total inference time as expressed in the following equation.

$$SynOps(SNN) = \sum_{t=1}\sum_{l=1}\sum_{i=1} S_i^l[t] f^l_{out,i} \tag{8}$$

where $S_i^l[t]$ indicates whether a spike is generated by neuron i of layer l at time instant t. In order to reduce the total power consumption of SNN models, we add a penalty term L_{fr} related to the average neuronal firing rate alongside the task-specific loss term L_{task} to the loss function as follows

$$L = L_{task} + \lambda L_{fr} \tag{9}$$

where λ is a hyperparameter that controls the extent of firing rate regularization.

As reported in Table 1, without any firing rate regularization, the SynOps ratio of SNN over ANN [SynOps(SNN)/SynOps(ANN)] is 0.88, 1.57 and 0.28 times for VCR, SL and VAD systems, respectively. Taking benefits from the massively parallel non-von Neumann computing architecture with co-located memory and computing units, the neuromorphic computing chips TrueNorth [13] and Tianjic [24] achieved a throughput of 400 and 649 giga synaptic operations per second (GSOPS) per watt respectively, which is 9.1 and 14.6 times more energy-efficient than the state-of-the-art Nvidia Titan RTX architecture with a throughput of 44.4 GSOPS/W at single precision [55]. Therefore, 10.34 (16.59), 5.80 (9.30), and 32.5 (52.14) times energy savings can be achieved for the VCR, SL, and VAD systems when implemented on the TrueNorth (Tianjic) chip.

Given the always-on nature of the VAD system and information is sparsely transmitted over time, the energy consumptions of the human-robot auditory interface are dominated by the VAD system. Notably, by adding the firing rate regularization with $\lambda = 1$, an additional 21.54 times saving can be achieved for the VAD system with the HTER degraded by only 1.44% under the Low Noise scenario, resulting in an overall energy savings of 700.05 (1123.10) times with the aforementioned neuromorphic chips. It worth mentioning that the TrueNorth and Tianjic chips are research prototypes, and a substantial amount of architectural and implementation level improvements can be integrated to further boost the performance gain.

4.6. Real-time performance and noise robustness analysis

4.6.1. Real-time performance

The proposed HuRAI framework has three subsystems that have sampling periods of 80 ms (VAD), 85 ms (SL), and 1 s (VCR). As described in Fig. 1, the main performance bottleneck is the VAD and SL path, which requires the overall computation (i.e., VAD + SL) to be accomplished within 85 ms. As a recent experiment reported, the inference of a 10-layer feedforward SNN, with 2560 neurons at each layer, can be finished within 10 ms on the Intel

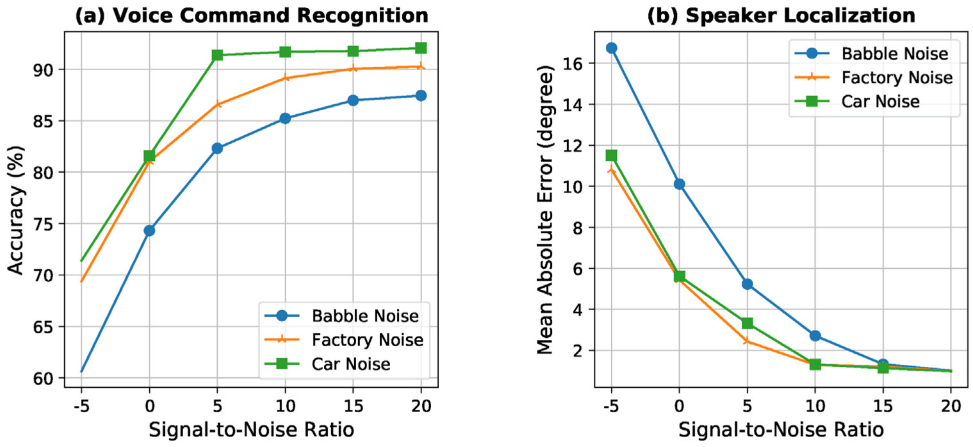


Fig. 7. Summarize the performance of (a) voice command recognition and (b) speaker localization systems under different environmental noises.

Loihi neuromorphic computing chip [56]. Given the network structure of the proposed VAD and SL subsystems are less complicated than that in the aforementioned work, their total computation time can be bounded within 20 ms. As such, it suggests the real-time performance can be guaranteed for the proposed HuRAI framework.

4.6.2. Noise robustness

An ideal human-robot auditory interface should function robustly under different working environments. Although neural network models perform reasonably well under the noise-free condition, their performance usually degrades rapidly in face of environmental noises [15,57,58]. To address this challenge, we explored the multi-condition training strategy, whereby both the clean and noisy audio samples are utilized for the network training. Specifically, we apply additive noise from the NOISEX-92 [59] dataset to the original clean samples at signal-to-noise (SNR) ratios varying from -5 to 20 dB. To cover a more diversified testing environment, we have randomly segmented the noise samples from three testing conditions (Babble Noise, Factory Noise, Car Noise).

As the results shown in Fig. 7, the VCR and SL subsystems perform robustly under the Low noise level (15, 20 dB), with only slight degradation from the results obtained under the clean condition. Notably, the performance remains competitive under the Medium noise level (5, 10 dB). Although the performance degrades rapidly under more adverse acoustic conditions, it remains reasonably well for real-time applications. Particularly, at the SNR of -5 dB, the mean absolute error of the SL subsystem can remain within 18 degrees. Among the three testing environments evaluated, the system performs worst under the Babble Noise, which can be explained by the high spectrum overlap with the clean audio samples.

5. Conclusion

The remarkable progress in deep learning has revolutionized the human-robot interface. The growing demands for natural interaction through the human-robot auditory interface have raised great concerns on energy efficiency, real-time performance, on-device computing capability, and data security, etc. In this paper, we present a novel brain-inspired SNN-based human-robot auditory interface to address all these concerns. In this framework, we introduce a VAD system at the front-end to discriminate the ‘voice’ segments from background noises. It significantly reduces the computational load by controlling access to other onboard services. The VCR and SL systems work separately to extract both the ‘what’ and ‘where’ information of the speaker from the voiced audio segments, which allow appropriate decisions to be made for action planning and control.

The experimental results demonstrate that accurate, real-time, and robust prediction can be achieved for each subsystem, suggesting a great modeling capability of SNNs. Moreover, our energy efficiency analysis reveals an attractive energy benefit of the proposed framework, with up to three orders of magnitude savings compared to the equivalent ANNs implemented on the state-of-the-art GPUs. Therefore, by integrating the algorithmic power of large-scale SNN models with low-power neuromorphic computing devices, we are expecting to enable the voice interface for ubiquitous robotic applications running locally. In view that the neuromorphic computing architectures and ultra-low power non-volatile memory devices are undergoing rapid developments, we are expecting to receive a sizeable performance boost in the near future.

CRediT authorship contribution statement

Jibin Wu: Conceptualization, Methodology, Writing - original draft. **Qi Liu:** Conceptualization, Methodology, Writing - original draft. **Malu Zhang:** Conceptualization, Writing - review & editing. **Zihan Pan:** Writing - original draft. **Haizhou Li:** Supervision, Writing - review & editing. **Kay Chen Tan:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research work was supported in part by A*STAR under its RIE2020 Advanced Manufacturing and Engineering Human (AME) Programmatic Grant (Grant No. A1687b0033, Project Title: Spiking Neural Networks (SNN)), the IAF, A*STAR, SOITEC, NXP and National University of Singapore under FD-fAbrICS: Joint Lab for FD-SOI Always-on Intelligent & Connected Systems (Award I2001E0053). The work of J. Wu was also partially supported by the Zhejiang Lab (No.2019KC0AB02). The work of M. Zhang was aslo partially supported by the National Natural Science Foundation of China under Grant No. 62106038, China Postdoctoral Science Foundation under Grant No. 2020M680148, Zhejiang Lab’s International Talent Found for Young Professionals. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the supporting institutions and companies.

References

[1] J.C. Murray, H.R. Erwin, S. Wermter, Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks, Neural Networks 22 (2) (2009) 173–189.

[2] E.A. Antonelo, B. Schrauwen, D. Stroobandt, Event detection and localization for small mobile robots using reservoir computing, Neural Networks 21 (6) (2008) 862–871.

[3] J. Wang, V.A. Shim, R. Yan, H. Tang, F. Sun, Automatic object searching and behavior learning for mobile robots in unstructured environment by deep belief networks, IEEE Trans. Cognitive Dev. Syst. 11 (3) (2018) 395–404.

[4] W. Huang, H. Tang, B. Tian, Vision enhanced neuro-cognitive structure for robotic spatial cognition, Neurocomputing 129 (2014) 49–58.

[5] H. Tang, W. Huang, A. Narayanamoorthy, R. Yan, Cognitive memory and mapping in a brain-like system for robotic navigation, Neural Networks 87 (2017) 27–37.

[6] A.H. Qureshi, Y. Nakamura, Y. Yoshikawa, H. Ishiguro, Intrinsically motivated reinforcement learning for human–robot interaction in the real-world, Neural Networks 107 (2018) 23–33.

[7] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, et al., The icub humanoid robot: An open-systems platform for research in cognitive development, Neural Networks 23 (8–9) (2010) 1125–1134.

[8] F. Stramandinoli, D. Marocco, A. Cangelosi, The grounding of higher order concepts in action and language: a cognitive robotics model, Neural Networks 32 (2012) 165–173.

[9] P. Liu, D.F. Glas, T. Kanda, H. Ishiguro, Two demonstrators are better than one-a social robot that learns to imitate people with different interaction styles, IEEE Trans. Cognitive Dev. Systems 11 (3) (2019) 319–333, https://doi.org/10.1109/TCDS.2017.2787062.

[10] J. Li, Z. Li, Y. Feng, Y. Liu, G. Shi, Development of a human-robot hybrid intelligent system based on brain teleoperation and deep learning slam, IEEE Trans. Autom. Sci. Eng. 16 (4) (2019) 1664–1674, https://doi.org/10.1109/TASE.2019.2911667.

[11] I. Bogun, A. Angelova, N. Jaitly, Object recognition from short videos for robotic perception, ArXiv abs/1509.01602.

[12] H.G. Okuno, K. Nakadai, Robot audition: Its rise and perspectives, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5610–5614, https://doi.org/10.1109/ICASSP.2015.7179045.

[13] P.A. Merolla, J.V. Arthur, R. Alvarez-Icaza, A.S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, et al., A million spiking-neuron

integrated circuit with a scalable communication network and interface, Science 345 (6197) (2014) 668–673.

[14] J. Wu, E. Yi̇lmaz, M. Zhang, H. Li, K.C. Tan, Deep spiking neural networks for large vocabulary automatic speech recognition, Front. Neurosci. 14 (2020) 199.

[15] J. Wu, Y. Chua, M. Zhang, H. Li, K.C. Tan, A spiking neural network framework for robust sound classification, Front. Neurosci. 12 (2018) 836.

[16] S.-C. Liu, B. Rueckauer, E. Ceolini, A. Huber, T. Delbruck, Event-driven sensing for efficient perception: Vision and audition algorithms, IEEE Signal Process. Mag. 36 (6) (2019) 29–37.

[17] Q. Xu, J. Peng, J. Shen, H. Tang, G. Pan, Deep covdensesnn: A hierarchical event-driven dynamic framework with spiking neurons in noisy environment, Neural Networks 121 (2020) 512–519.

[18] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S.H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, et al., Loihi: A neuromorphic manycore processor with on-chip learning, IEEE Micro 38 (1) (2018) 82–99.

[19] J. Wu, C. Xu, D. Zhou, H. Li, K.C. Tan, Progressive tandem learning for pattern recognition with deep spiking neural networks, arXiv preprint arXiv:2007.01204..

[20] Q. Yu, H. Tang, K.C. Tan, H. Li, Rapid feedforward computation by temporal encoding and learning with spiking neurons, IEEE Trans. Neural Networks Learning Syst. 24 (10) (2013) 1539–1552.

[21] Y. Zheng, S. Li, R. Yan, H. Tang, K.C. Tan, Sparse temporal encoding of visual features for robust object recognition by spiking neurons, IEEE Trans. Neural Networks Learning Syst. 29 (12) (2018) 5823–5833.

[22] M. Zhang, J. Wu, Y. Chua, X. Luo, Z. Pan, D. Liu, H. Li, Mpd-al: an efficient membrane potential driven aggregate-label learning algorithm for spiking neurons, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 1327–1334..

[23] M. Zhang, X. Luo, Y. Chen, J. Wu, A. Belatreche, Z. Pan, H. Qu, H. Li, An efficient threshold-driven aggregate-label learning algorithm for multimodal information processing, IEEE J. Selected Topics Signal Processing 14 (3) (2020) 592–602, https://doi.org/10.1109/JSTSP.2020.2983547.

[24] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, et al., Towards artificial general intelligence with hybrid tianjic chip architecture, Nature 572 (7767) (2019) 106–111.

[25] M. Zhang, J. Wang, Z. Zhang, A. Belatreche, J. Wu, Y. Chua, H. Qu, H. Li, Spike-timing-dependent back propagation in deep spiking neural networks, arXiv preprint arXiv:2003.11837..

[26] Z. Pan, M. Zhang, J. Wu, H. Li, Multi-tones' phase coding (mtpc) of interaural time difference by spiking neural network, arXiv preprint arXiv:2007.03274..

[27] E.O. Neftci, H. Mostafa, F. Zenke, Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks, IEEE Signal Process. Mag. 36 (6) (2019) 51–63.

[28] P.J. Werbos, Backpropagation through time: what it does and how to do it, Proc. IEEE 78 (10) (1990) 1550–1560.

[29] Y. Wu, L. Deng, G. Li, J. Zhu, L. Shi, Direct training for spiking neural networks: Faster, larger, better, arXiv preprint arXiv:1809.05793..

[30] J. Wu, Y. Chua, M. Zhang, Q. Yang, G. Li, H. Li, Deep spiking neural network with spike count based learning rule, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–6.

[31] S.B. Shrestha, G. Orchard, Slayer: Spike layer error reassignment in time, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates Inc, 2018, pp. 1412–1421.

[32] F. Zenke, S. Ganguli, Superspike: Supervised learning in multilayer spiking neural networks, Neural Comput. 30 (6) (2018) 1514–1541.

[33] J. Wu, Y. Chua, M. Zhang, G. Li, H. Li, K.C. Tan, A Tandem Learning Rule for Efficient and Rapid Inference on Deep Spiking Neural Networks, arXiv e-prints (2019) arXiv:1907.01167..

[34] H. Ghaemmaghami, B.J. Baker, R.J. Vogt, S. Sridharan, Noise robust voice activity detection using features extracted from the time-domain autocorrelation function, Proceedings of Interspeech (2010).

[35] D.B. Dean, S. Sridharan, R.J. Vogt, M.W. Mason, The qut-noise-timit corpus for the evaluation of voice activity detection algorithms, Proceedings of Interspeech (2010).

[36] D.A. Silva, J.A. Stuchi, R.P.V. Violato, L.G.D. Cuozzo, Exploring convolutional neural networks for voice activity detection, in: Cognitive Technologies, Springer, 2017, pp. 37–47.

[37] X. Xiao, S. Zhao, X. Zhong, D.L. Jones, E.S. Chng, H. Li, A learning-based approach to direction of arrival estimation in noisy and reverberant environments, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 2814–2818.

[38] W. He, P. Motlicek, J.-M. Odobez, Deep neural networks for multiple speaker detection and localization, in: in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 74–79.

[39] J. Wilpon, L. Rabiner, A. Bergh, Speaker-independent isolated word recognition using a 129-word airline vocabulary, J. Acoust. Society Am. 72 (2) (1982) 390–396.

[40] M. Gales, S. Young, et al., The application of hidden markov models in speech recognition, Foundations Trends Signal Processing 1 (3) (2008) 195–304.

[41] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig, Achieving human parity in conversational speech recognition, arXiv preprint arXiv:1610.05256..

[42] O. Abdel-Hamid, L. Deng, D. Yu, Exploring convolutional neural network structures and optimization techniques for speech recognition., in: Interspeech, Vol. 2013, 2013, pp. 1173–5..

[43] M. McLaren, Y. Lei, N. Scheffer, L. Ferrer, Application of convolutional neural networks to speaker recognition in noisy conditions, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inform. Processing Systems (2019) 8024–8035.

[45] R. Sheelvanth, B. Sharma, M. Madhavi, R. Das, S. Prasanna, H. Li, Rsl 2019: A realistic speech localization corpus, in: Oriental COCOSDA, 2019..

[46] P. Warden, Speech commands: A dataset for limited-vocabulary speech recognition, arXiv preprint arXiv:1804.03209..

[47] T.M. Inc., Speech command recognition using deep learning (2019). https://www.mathworks.com/help/deeplearning/examples/deep-learning-speech-recognition.html..

[48] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, Specaugment: A simple data augmentation method for automatic speech recognition, arXiv preprint arXiv:1904.08779..

[49] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105..

[50] E.S. Doc, Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms, ETSI ES 202 (050) (2002) v1.

[51] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, J.-P. Petit, Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications, IEEE Communications Magazine 35 (9) (1997) 64–73.

[52] J. Ramìrez, J.C. Segura, C. Benìtez, A. De La Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information, Speech Commun. 42 (3–4) (2004) 271–287.

[53] J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Processing Letters 6 (1) (1999) 1–3.

[54] G. Dellaferrera, F. Martinelli, M. Cernak, A bin encoding training of a spiking neural network-based voice activity detection, arXiv preprint arXiv:1910.12459..

[55] N. Corporation, Nvidia Titan RTX architecture (2020). https://www.nvidia.com/en-us/deep-learning-ai/products/titan-rtx/..

[56] P. Blouw, X. Choo, E. Hunsberger, C. Eliasmith, Benchmarking keyword spotting efficiency on neuromorphic hardware, in: Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop, 2019, pp. 1–8.

[57] Q. Liu, J. Wu, Parameter tuning-free missing-feature reconstruction for robust sound recognition, IEEE J. Selected Topics Signal Processing 15 (1) (2021) 78–89, https://doi.org/10.1109/JSTSP.2020.3038054.

[58] J. Wu, Z. Pan, M. Zhang, R.K. Das, Y. Chua, H. Li, Robust sound recognition: A neuromorphic approach., in: INTERSPEECH, 2019, pp. 3667–3668..

[59] A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. 12 (3) (1993) 247–251.

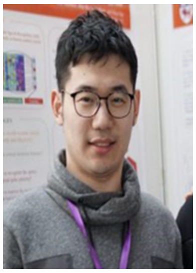
Jibin Wu received the B.E. and Ph.D degree in Electrical Engineering from National University of Singapore in 2016 and 2020, respectively. He is currently a Research Fellow of HLT lab in Department of Electrical and Computer Engineering, National University of Singapore. His research interests include Spiking Neural Network, Neuromorphic Computing, Auditory Modelling, and Automatic Speech Recognition.

Qi Liu received received his B. Eng. degree in Measuring & Control Technology and Instrumentations, and M. Eng. degree in Control Science and Engineering from Harbin Engineering University, Harbin, China, in 2013 and 2016, respectively, and the Ph.D. degree in Electrical Engineering from City University of Hong Kong, Hong Kong, China, in 2019. His research interests broadly lie in image restoration, machine learning, optimization methods and applications.

From 2018 to 2019, he was a Visiting Scholar at Department of Electrical and Computer Engineering, University of California, Davis, CA, USA. During this period, he was invited to give talks in University of

California, San Diego and California Institute of Technology, respectively. Currently, he works as a Research Fellow in Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He was the recipient of the Best Paper Award at the 2019 IEEE International Conference on Signal, Information and Data Processing.

Malu Zhang received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2019. He is now a Research Fellow of the HLT lab in the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include spiking neural networks, neural spike encoding, and the neuromorphic applications of speech recognition and sound source localization. Dr. Zhang is a guest associate editor for Frontiers in Neuroscience (Neuromorphic Engineering), and also reviews for several international journals, such as IEEE TPAMI, IEEE TNNLS, IEEE TCYB.



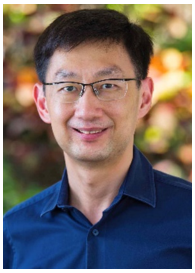
Zihan Pan received his B.Eng and M.Eng degrees in communication engineering from Tianjin University and Nanyang Technological University, in 2013 and 2015 respectively. He obtained his PhD degree in National University of Singapore, 2020. His research interests include spiking neural network, neural spike encoding, with the neuromorphic application of speech recognition and sound source localization.

Haizhou Li received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include automatic speech recognition, speaker and language recognition, natural language processing, and neuromorphic computing.

Prior to joining NUS, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor at CRIN in France (1994–1995), Research Manager at the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), Vice President in InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003–2016).

Dr Li has served as the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015–2018), a Member of the Editorial Board of Computer Speech and Language since 2012, a Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019.

Dr Li is a Fellow of the IEEE, and a Fellow of the ISCA. He was a recipient of the National Infocomm Award 2002, and the President's Technology Award 2013 in Singapore. He was named Nokia Visiting Professor in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019 by the University of Bremen, Germany.



Kay Chen Tan received the B.Eng. degree (First Class Hons.) in electronics and electrical engineering and the Ph.D. degree in evolutionary computation and control systems from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively. He is currently a Chair Professor (Computational Intelligence) of the Department of Computing, the Hong Kong Polytechnic University. He has published over 300 refereed articles and seven books.

Prof. Tan is currently the Vice-President (Publications) of IEEE Computational Intelligence Society, USA. He has served as the Editor-in-Chief for IEEE Computational Intelligence Magazine from 2010 to 2013 and the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION from 2015 to 2020, and currently serves as the Editorial Board Member for more than ten journals. He is also an IEEE Distinguished Lecturer Program (DLP) Speaker and the Chief Co-Editor of Springer Book Series on Machine Learning: Foundations, Methodologies, and Applications.