# Sound Recognition System Using Spiking and MLP Neural Networks

Elena Cerezuela-Escudero, Angel Jimenez-Fernandez, Rafael Paz-Vicente,
Juan P. Dominguez-Morales, Manuel J. Dominguez-Morales,
and Alejandro Linares-Barranco

Robotic and Technology of Computers Lab,
Department of Architecture and Technology of Computers,
University of Seville, Seville, Spain
ecerezuela@atc.us.es

**Abstract.** In this paper, we explore the capabilities of a sound classification system that combines a Neuromorphic Auditory System for feature extraction and an artificial neural network for classification. Two models of neural network have been used: Multilayer Perceptron Neural Network and Spiking Neural Network. To compare their accuracies, both networks have been developed and trained to recognize pure tones in presence of white noise. The spiking neural network has been implemented in a FPGA device. The neuromorphic auditory system that is used in this work produces a form of representation that is analogous to the spike outputs of the biological cochlea. Both systems are able to distinguish the different sounds even in the presence of white noise. The recognition system based in a spiking neural networks has better accuracy, above 91 %, even when the sound has white noise with the same power.

**Keywords:** Neuromorphic auditory hardware · Address-Event representation · Spiking neural networks · Sound recognition · Spike signal processing

## 1 Introduction

By the information provided from the hearing system, the human being can identify virtually any kind of sound (sound recognition) and where it comes from (sound localization) [1]. If this ability could be reproduced by artificial devices, many applications would emerge, from support devices for people with hearing loss to security devices.

Sound recognition is commonly treated as a two stages problem: filtering and classification [2–6]. Filtering is the stage where the signal is processed to extract acoustic features, so only relevant information will pass to the classification stage, where the sound will be identified. There are some factors that make sound recognition a hard task: the presence of electric noise in the signal, the environment's noise level and reverberation, the fact that the signal is a complex time series data and the wide dynamic range of sound. Biological cochlea has a huge dynamic range, is adapted to a wide variety of listening environments and it has high noise immunity [7]. In order to take advantage of these characteristics, in this work we use in the first recognition stage a Neuromorphic

Auditory System (NAS) that decomposes an audio signal into different frequency bands, which produces spikes, in the same way a biological cochlea processes and sends the audio information coded in spikes to the brain.

Artificial Neural Network is a generic classification method that can deal with several kinds of information and has found great success in the area of pattern recognition. However, standard artificial neuron models require input signals to be transformed into static vectors by windowing processes, as, for example, the Time Delay Neural Network [8]. Another approach for processing temporal data is the use of Spiking Neural Network (SNN) [9]. The neurons within this kind of network deal with input signals on the form of pulse (also called spike) trains, using a potential as a reference for generating pulses on its output. Spiking models can directly deal with temporal data and can be efficiently implemented in hardware, due to its simple structure. In this work, we present two classification systems based on two kinds of neural network: Multilayer Perceptron Neural Network (MLPNN) and SNN.

It is very common the use of techniques based in Fourier Transforms for filtering stage. In [2] the Fast Fourier Transform and the Harmonic Product Spectrum are proposed for the filtering stage and an MLPNN for the classification stage. The system achieved 97.5 % recognition accuracy for 12 musical notes using 20 neurons in the first hidden layer and 10 neurons in the second one. The sound classification model proposed in [3] extracts the pitch of the signal using the Harmonic Product Spectrum. Based on the pitch estimation, features are created and used in a probabilistic model. The accuracy of the model is 99.95 % for 3 classes of sounds.

Although techniques based in Fourier Transformations can have remarkably successful, their underpinnings are somewhat removed from the spiking, highly parallelized nature of the mammalian auditory perception systems. The work presented here is an attempt to work within a more biologically realistic framework, both for the formation of sound descriptors, and for the task of sound classification itself.

There are previous works that presents bio-inspired models of cochlea and neural coding scheme. Reference [10] presents a phenomenological model of the cochlea consists of a bank of nonlinear time-varying parallel filters and an active distributed feedback and reference [11] simulates a model of auditory nerve and cochlear nucleus neurons. Both models have several realistic properties. Reference [5] presents a sound recognition system using a bank of band-pass filters and pulsed generator implemented in software for extracting sound frequency characteristics and a hardware implementation of pulsed neural network to classify. The accuracy of the system is 98.7 % for 6 classes of sounds. In [6] the cochlea response is simulated with a gammatone filterbank and classification task is performed using a time-domain reservoir neural network known as the echo state network [12]. The accuracy of the system is 45 % for 5 classes. The system proposed in [13] uses an MLPNN to classify sounds between 5 vowel phonemes with percentage of success of 93.99 %. The characteristics extraction stage is not bio-inspired because it is based on electromyogram signals.

## 2 Neuromorphic Auditory System

Neuromorphic systems, because of their high level of parallelism, interconnectivity, and scalability, carry out complex processing in real time, with a good relation between quality, speed and resource consumption [7]. The signals in these systems are composed of short pulses in time, called spikes or events. The information can be coded in the polarity and spike frequency, often following a Pulse Frequency Modulation (PFM) scheme, or in the inter-spike-interval time (ISI) [14], or in the time-from-reset, where the most important (with the highest priority) events are sent first [15]. Address-Event Representation (AER), proposed by Mead lab in 1991 [16], faced the difficult problem of connecting silicon neurons along chips that implement different neuronal layers using a common asynchronous digital bus multiplexed in time, the AER bus. This representation gives a digital unique code (address) to each neuron, which is transmitted using a simple four-phase handshake protocol [17].

In the filtering stage of the audio recognition problem, we use a neuromorphic device which decomposes an audio signal into different frequency bands of spiking information, in the same way a biological cochlea sends the audio information to the brain. The biological cochlea performs the transduction between the pressure signal representing the acoustic input and the neural signals that carry information to the brain. Due to the physical characteristics of a part of cochlea, the basilar membrane, cochlea divides an input signal into its frequency components. Thousands of hair cells on the membrane generate action potentials, or spikes, that travel along nerve fibers to higher-order auditory brain areas [7]. The first silicon cochlea was proposed by Lyon and Mead [18]. In their design, the membrane basilar was modeled by a cascade of 480 second-order filter sections. There are several VLSI implementations of the cochlea based on Lyon's design (for example, [19–21]). Digital models of the cochlea process audio signals using classical Digital Signal Processing techniques [22–24].

The NAS is innovate respect previous systems because it processes information directly encoded as spikes with a Pulse Frequency Modulation (PFM), performing Spike Signal Processing techniques [25, 26], and using AER interfaces. The architecture of the NAS is shown in Fig. 1. The system's input is the digitalized audio streams, which represent the audio signals of a monaural system. A Synthetic Spike Generator [27] converts this digital audio source into a spike stream. Then, the cascade band pass filter bank splits the spike streams in 64 (64 is the number of channels of the NAS) frequency bands using 64 different spiking outputs that are combined by an AER monitor block into an AER output bus [28], which encodes each spike according to AER and transmits this information to the classification systems. All the elements required for designing the NAS components (Synthetic Spike Generators, cascade filter bank and the AER monitor) have been implemented in VHDL and designed as small spike-based building blocks [26]. Table 1 shows the NAS characteristics. The NAS has been used before in [29] to measure the speed of DC motor and in [30] that proposes a convolutional spiking neural network for audio sound classification. Although, for this work, the gain of the band pass filters have been modified looking for improving the recognition system accuracy.

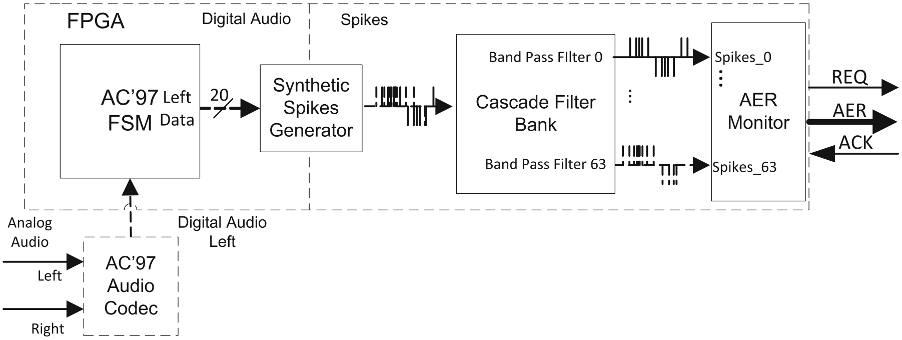

**Fig. 1.** NAS Architecture

**Table 1.** NAS characteristics

| Number bands | Frequency range | Dynamic range | Max. Event rate | Clock frequency |
|---|---|---|---|---|
| 64 | 9.6 Hz–14.06 kHz | 75 dB | 2.19 Mevents/s | 27 MHz |

# 3 Classification Systems

## 3.1 Multilayer Perceptron Neural Network

The topological structure of the MLPNN used consists of a two-layer feed-forward network, with a sigmoid transfer function in the hidden and output layer. The training algorithm used is the back-propagation. During this research, the optimal number of hidden units was found by running different performance tests, where a new MLPNN was created, trained and tested using a varying number of neurons in the hidden layer. This kind of neural network requires static vectors as input. The number of the network inputs is similar than the number of characteristics. The spiking signal has been transformed by windowing process and organized by characteristics like this: the spiking information of each NAS band has been integrated during 20 ms, generating a 64-element vector. We select 20 ms because the shortest audible sound ranges from 10 to 40 ms [31].

## 3.2 Spiking Neural Network

The SNN has been implemented by a two-layer neural network. The input layer consists of Integrate and Fire neurons [9]. The optimal number of input units was found by running different performance tests. The output layer has as many Winner-Take-All neurons as classes to classify. The SNN input are the 64 spiking streams from the NAS. This classification system was implemented in a FPGA, and the Integrate and Fire neuron hardware architecture is shown if Fig. 2, where $W$ are the neuron weights and $\theta$

is the neuron threshold. The SNN training is performed by a SNN simulation implemented in software.

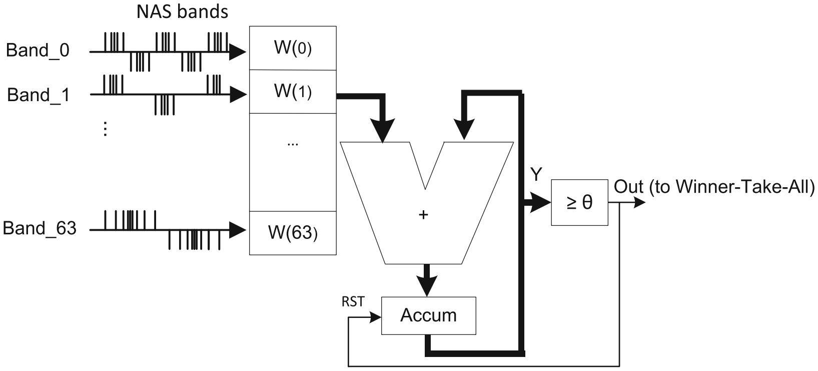

**Fig. 2.** Hardware architecture of Integrate and Fire Neuron of the SNN

## 4 Experimental Results

We have evaluated the capabilities of our sound recognition system using pure tone sounds in the presence of white noise. The fundamental frequencies of sounds are shown in Table 2. White noise was added to check the noise tolerance of the recognition system. The test set consisted of 50 20-millisecond samples of each tone. White noise was added to each sound with a SNR sweep from 46.05 to −21.97 dB (30 different values of SNR). Therefore, in total there are 1.500 samples of each tone.

**Table 2.** Fundamental frequency of sounds to recognize

| Freq | 130,81 | 174,61 | 261,62 | 349,22 | 523,25 | 698,45 | 1046,5 | 1396,91 |
|---|---|---|---|---|---|---|---|---|

The MLPNN was created, trained and tested using a varying number of neurons in the hidden layer. 70 % randomly-selected samples were used to training the MLPNN, including noisy samples. The results shows in Fig. 3 are obtained using 10 neurons in the hidden layer and 8 neuron in the output layer. The system achieves 98.95 % recognition accuracy for tones without white noise. Figure 3 (left) shows accuracy for each pure tone in the presence of different white noise powers as a color-map, being the X-axis the frequencies between 130.813 and 1.89 kHz, the Y-axis the SNR between 46.05 to −21.97 dB, and the color represents the percentage of successes.

The SNN, with 8 neurons in the input layer and 8 neuron in the output layer, training with 70 % noiseless samples, obtains the accuracy shown in Fig. 3 (right). The system achieves a mean success rate of 100 % for tones without white noise. Figure 3 shows that the hit rate decreases with the increase of white noise, and that there is a frequency less robust to white noise (698,45 Hz). Most of the pure tones have a hit rate over 90 % with a SNR over −8 dB.

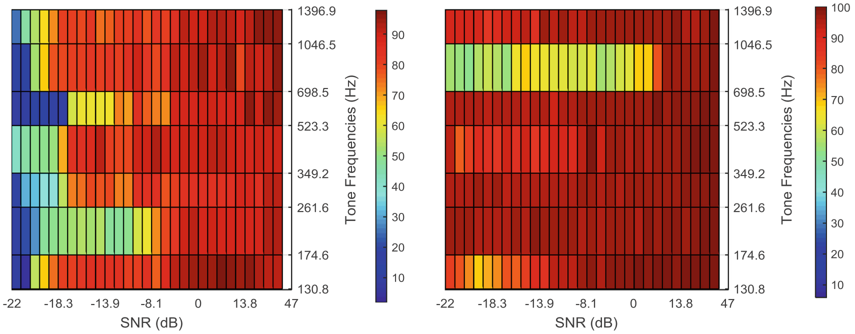

**Fig. 3.** Hit rate of sound recognition system using MLPNN (left) and using SNN (right)

The results shown in Fig. 3 (right), in general, are better than the results obtains with MLPNN. Furthermore, SNN uses 16 hardware implementation neurons and MLPNN uses 18 neurons.

## 5 Conclusions

In this study, we show that recognizing pure tones in presence of white noise can success using MLP neural network and SNN. The SNN achieves better accuracy with less neurons than MLPNN. In addition, SNN has been implemented in hardware. Audio information acquisition is carried out by a novel neuromorphic auditory system, which provided streams of spikes representing audio frequency components. As a future work, it would be valuable to evaluate the performance of the SNN with the models of the auditory system proposed in [10, 11], as well as performance of more complex sound recognition, like vowels [13].

In the audio context, traditional digital systems have to process several samples in a buffer, because sound makes sense along time, where Fast Fourier Transform (FFT) calculation prior to specific processing. However, NAS provide audio directly and continuously decomposed into its frequency components as a spike stream. This allows real time audio processing (without the need for buffering), using neuromorphic processing layers as SNN do.

The SNN-based system and MLPNN-based system have a percentage of success above 91 % even when the sound has white noise with the same power. When the SNR is −18.3 dB, the SNN-based system accuracy is kept on 85.3 %, but the MLPNN-based system accuracy is only 12.5 %. The SNN-based system achieves a mean success rate of 100 % for tones without white noise.

Most recognition systems exposed in the introduction cannot be implemented in dedicated hardware because of its high computational cost, however, we present a recognition system efficiently implemented in hardware, due to its simple structure. Furthermore, SNN architecture proposed is highly parallelizable. Regarding the classification stage, all the works presented in the introduction have more computational cost than SNN. For example, the method proposed in [2] achieved 97.5 % accuracy for 12 sounds using a MLPNN with 30 neurons. The bio-inspired recognition system proposed in [5] has a percentage of succeed of 98.7 % for 6 kinds of sounds, less than our recognition system and it is not fully implementable in hardware.

The system presented in this paper is being applied to animal behavior recognition, for example for horse behaviors, through a SNN-based sound recognition system associate to the animal movements.

**Acknowledgements.** This work is supported by the Spanish government grant BIOSENSE (TEC2012-37868-C04-02) and by the excellence project from Andalusian Council MINERVA (P12-TIC-1300), both with support from the European Regional Development Fund.

## References

1. Pickles, J.O.: An Introduction to the Physiology of Hearing. Emerald, London (2012)
2. Guerrero-turrubiates, J.J., Gonzalez-reyna, S.E., Ledesma-orozco, S.E., Avina-cervantes, J.G.: Pitch estimation for musical note recognition using artificial neural networks. In: International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 53–58 (2014)
3. Nielsen, A.B., Hansen, L.K., Kjems, U.: Pitch based sound classification. In: 2006 Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), vol. 3, pp. 788–791 (2006)
4. Pishdadian, F., Nelson, J.K.: On the transcription of monophonic melodies in an instance-based pitch classification scenario. In: Proceedings of 2013 IEEE Digital Signal Processing and Signal Processing Education Meeting, DSP/SPE 2013, pp. 222–227 (2013)
5. Iwasa, K., Kugler, M., Kuroyanagi, S., Iwata, A.: A sound localization and recognition system using pulsed neural networks on FPGA. In: International Joint Conference on Neural Networks, IJCNN 2007, pp. 902–907. IEEE, August 2007
6. Newton, M.J., Smith, L.S.: Biologically-inspired neural coding of sound onset for a musical sound classification task. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1386–1393 (2011)
7. Liu, S.C.: Event-Based Neuromorphic Systems. Wiley (2015)
8. Waibel, A., Hanazawa, T., Hinton, G., Shiano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. IEEE Trans. Acousti. Speech Sig. Process. **37**(3), 328–339 (1989)
9. Gerstner, W., Kistler, W.M.: Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press, Cambridge (2002)
10. Robert, A., Eriksson, J.L.: A composite model of the auditory periphery for simulating responses to complex sounds. J. Acoust. Soc. Am. **106**(4), 1852–1864 (1999)
11. Eriksson, J.L., Robert, A.: The representation of pure tones and noise in a model of cochlear nucleus neurons. J. Acoust. Soc. Am. **106**(4), 1865–1879 (1999)
12. Jaeger, H.: Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach. GMD Report 159, German National Research Center for Information Technology (2002)
13. Mendes, J.A., Robson, R.R., Labidi S., Barros A.K.: Subvocal Speech recognition based on EMG signal using independent component analysis and neural network MLP. In: Congress on Image and Signal Processing, CISP 2008, vol. 1, pp. 221–224 (2008)

14. Indiveri, G., Chicca, E., Douglas, R.: A VLSI array of low-power spiking neurons and bistables synapses with spike-timig dependant plasticity. IEEE Trans. Neural Netw. **17**(1), 211–221 (2006)
15. Thorpe, S.J., Brilhault, A., Perez-Carrasco, J.A.: Suggestions for a biologically inspired spiking retina using order-based coding. In: 2010 IEEE International Symposium on Circuits and Systems Nano-Bio Circuit Fabrics and Systems, ISCAS 2010, pp. 265–268 (2010)
16. Mahowald, M.: VLSI analogs of neuronal visual processing: a synthesis of form and function, Ph.D. dissertation, California Institute of Technology, Pasadena (1992)
17. Boahen, K.A.: Communicating Neuronal Ensembles between Neuromorphic Chips. Neuromorphic Systems. Kluwer Academic Publishers, Boston (1998)
18. Lyon, R.F., Mead, C.: An analog electronic cochlea. IEEE Trans. Acoust. Speech Sig. Process. **36**, 1119–1134 (1988)
19. Wen, B.: Boahen, K.A silicon cochlea with active coupling. IEEE Trans. Biomed. Circ. Syst. **3**, 444–455 (2009)
20. Hamilton, T.J., Jin, C., van Schaik, A., Tapson, J.: An active 2-d silicon cochlea. IEEE Trans. Biomed. Circ. Syst. **2**, 30–43 (2008)
21. Liu, S-C., Van Schaik, A., Minch, B.A., Delbruck, T.: Event-based 64-channel binaural silicon cochlea with Q enhancement mechanisms. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), 30 May−2 June 2010, pp. 2027–2030 (2010)
22. Leong, M.P., Jin, C., Leong, P.: An FPGA-based electronic cochlea. EURASIP J. Appl. Sig. Process. **2003**(7), 629–638 (2003)
23. Dundur, R., Latte, M.V., Kulkarni, S.Y., Venkatesha, M.K.: Digital filter for cochlear implant implemented on a field-programmable gate array. Int. J. Electr. Comput. Energ. Electron. Commun. Eng. **2**(7), 468–472 (2008)
24. Thakur, C.S., Hamilton, T.J., Tapson, J., van Schaik, A., Lyon, R.F.: FPGA Implementation of the CAR model of the Cochlea. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1853−1856 (2014)
25. Domínguez-Morales, M., Jimenez-Fernandez, A., Cerezuela-Escudero, E., Paz-Vicente, R., Linares-Barranco, A., Jimenez, G.: On the designing of spikes band-pass filters for FPGA. In: Honkela, T. (ed.) ICANN 2011, Part II. LNCS, vol. 6792, pp. 389–396. Springer, Heidelberg (2011)
26. Jimenez-Fernandez, A., Linares-Barranco, A., Paz-Vicente, R., Jiménez, G., Civit, A.: Building blocks for spike-based signal processing. In: International Joint Conference on Neural Networks, IJCNN, pp. 1–8 (2010)
27. Gomez-Rodriguez, F., Paz, R., Miro, L., Linares-Barranco, A., Jimenez, G., Civit, A.: Two hardware implementations of the exhaustive synthetic AER generation method. In: Cabestany, J., Prieto, A.G., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 534–540. Springer, Heidelberg (2005)
28. Cerezuela-Escudero, E., Dominguez-Morales, M.J., Jiménez-Fernández, A., Paz-Vicente, R., Linares-Barranco, A., Jiménez-Moreno, G.: Spikes monitors for FPGAs, an experimental comparative study. In: Rojas, I., Joya, G., Gabestany, J. (eds.) IWANN 2013, Part I. LNCS, vol. 7902, pp. 179–188. Springer, Heidelberg (2013)
29. Rios-Navarro, A., Jimenez-Fernandez, A., Cerezuela-Escudero, E., Rivas, M., Jimenez, G., Linares-Barranco, A.: Live demostration: real-time motor rotation frequency detection by spike-based visual and auditory sensory fusion on AER and FPGA. In: Wermter, S., Weber, C., Duch, W., Honkela, T., Koprinkova-Hristova, P., Magg, S., Palm, G., Villa, A.E.P. (eds.) ICANN 2014. LNCS, vol. 8681, pp. 847–848. Springer, Switzerland (2014)

30. Cerezuela-Escudero, E., Jimenez-Fernandez, A., Paz-Vicente, R., Dominguez-Morales, M., Linares-Barranco, A., Jimenez-Moreno, G.: Musical notes classification with Neuromorphic Auditory System using FPGA and a Convolutional Spiking Network. In: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2015)
31. Lass, N.: Contemporary Issues in Experimental Phonetics. Elsevier (2012)