Audio classification systems using deep neural networks and an event-driven auditory sensor

Ceolini, Enea ; Kiselev, Ilya ; Liu, Shih-Chii

Abstract: We describe ongoing research in developing audio classification systems that use a spiking silicon cochlea as the front end. Event-driven features extracted from the spikes are fed to deep networks for the intended task. We describe a classification task on naturalistic audio sounds using a low-power silicon cochlea that outputs asynchronous events through a send-on-delta encoding of its sharply-tuned cochlea channels. Because of the event-driven nature of the processing, silences in these naturalistic sounds lead to corresponding absence of cochlea spikes and savings in computes. Results show 48% savings in computes with a small loss in accuracy using cochlea events.

DOI: https://doi.org/10.1109/sensors43011.2019.8956592

Accepted for publication at 2019 IEEE Sensors Conference

Audio classification systems using deep neural networks and an event-driven auditory sensor

Enea Ceolini, Ilya Kiselev, and Shih-Chii Liu
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland
Email: [enea.ceolini,kiselev,shih]@ini.uzh.ch

***Abstract*—We describe ongoing research in developing audio classification systems that use a spiking silicon cochlea as the front end. Event-driven features extracted from the spikes are fed to deep networks for the intended task. We describe a classification task on naturalistic audio sounds using a low-power silicon cochlea that outputs asynchronous events through a send-on-delta encoding of its sharply-tuned cochlea channels. Because of the event-driven nature of the processing, silences in these naturalistic sounds lead to corresponding absence of cochlea spikes and savings in computes. Results show 48% savings in computes with a small loss in accuracy using cochlea events.**

***Index Terms*—event-driven audio, edge computing, spiking cochlea, deep learning, sound classification, low-power cochlea**

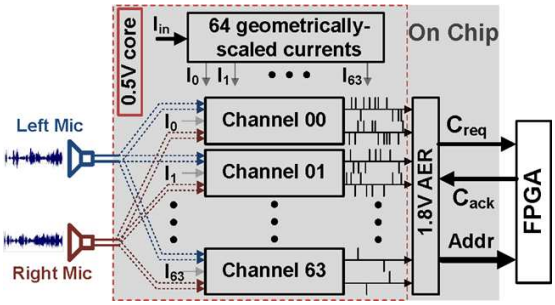

Fig. 1: Architecture of the DASLP. Adapted from [1].

I. Introduction

Spiking ASIC cochlea chip designs [1]–[3] have gradually matured over the years. These sensors implement circuits that model partial functionality of the biological cochlea. Applications of this sensor modality are still relatively unexplored, especially in comparison to the spiking visual Dynamic Vision Sensor. Early work showed that cochlea spikes can be used to measure an azimuthal audio source location with similar localization accuracy as the cross-correlation of binaural microphone samples [4]. The localization comparison in [5], showed that the number of computational operations from this data-driven method can be more than an order of magnitude lower than using cross-correlation of binaural samples.

A few studies have looked at the use of the spiking cochlea output on classification tasks such as speaker verification and digit recognition [6], [7]. These early investigations use SVM classifiers on features created using constant time bins and constant bin spike samples. These studies demonstrate that the accuracy using the spike features is close to the accuracy using conventional features such as log-filter banks or Mel Frequency Cepstral Coefficient (MFCC) features.

Because deep networks have been very successful for many machine learning tasks, they are also being tested together with the spikes for classification tasks [8]. In this case, constant time features or exponential features computed using a 5 ms time window for a digit recognition task show similar accuracies as the conventional log-filterbank or MFCC features usually computed using a 10-20 ms time window [9]. The cochlea has also been used as a front-end in a pipeline that incorporates the localization model, keyword spotting and source separation in the case of a speech mixture [10].

The latency of a data-driven system using an asynchronous data-driven sensor such as the spiking cochlea can be lower if processing is initiated only when cochlea spikes are generated when sounds are present in an environment. The results of these applications especially with the postprocessing of deep networks can help to identify key building blocks that allow one to construct a more general audio ASIC chip useful for applications such as IoT. Prototyping systems on an FPGA can be a useful intermediate step [11] especially with the increasing chip fabrication costs. In the remainder of this paper, we show an example of a sound classification task for naturalistic audio environments. The cochlea used in this study, the dataset, the input feature extraction and the architectures are described in Section II and the results in Section III.

II. Methods

A. DASLP silicon cochlea

The Dynamic Audio Sensor Low Power (**DASLP**) spiking cochlea used in this work is the latest low-power (LP) ASIC binaural design with 64 frequency channels per side and asynchronous spiking outputs [1]. This cochlea design uses a parallel bank of 64 filters ranging from best characteristic frequencies of 20 Hz to 20 kHz. The best frequency of each filter is generated by the 64 geometrically-scaled current block in Fig. 1. The fourth-order bandpass filter (BPF) design in each channel consists of two cascaded power-efficient second-order source-follower-based BPFs, followed by an asynchronous delta modulator (ADM) with on-chip asynchronous arbitration circuits for transmitting events off chip.

This work was partially supported by the Swiss National Science Foundation, HEAR-EAR, 200021_172553.

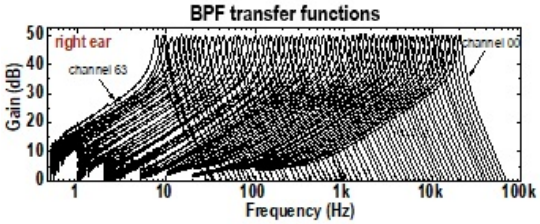

Fig. 2: Transfer function across the 64 channels. Adapted from [1].

The filters model the functionality of the basilar membrane of the biological cochlea. Each filter channel produces both ON and OFF spikes unlike other DAS cochleas which do not produce the dual polarity spikes. The dual polarity spikes are produced by the send-on-delta scheme used for generating the asynchronous spikes. These spikes are transmitted off-chip through the asynchronous event representation (AER) protocol. The AER block arbitrates among all the active channels. The asynchronous handshaking signals C_{ack} and C_{req} are used to transmit the chosen channel address on $Addr$ to the external device. The analog block operates on a power supply of 0.5V and consumes only 55 μW. This design has good matching properties of the quality factor, Q, of the filter across the different channels, and $Q > 10$ can be achieved across the entire array [1] as shown in Fig. 2.

The DASLP board that holds the chip is similar to the DAS board [2]. It is USB powered and interfaces to our Java-based jAER software [12] for setting the chip biases, recording, and processing sensor output. On-board ADCs also sample the output of the microphones to allow direct comparison of cochlea versus sampled audio algorithms. Data packets comprising both the output spike data and microphone samples are transmitted to a computer via USB.

B. Dataset

The naturalistic sound classification task is performed on a subset of AudioSet [13], a dataset consisting of about 2 million audio samples extracted from YouTube videos. Each sample is 10 seconds long with a total of more than 5000 hours of audio recording. We used a subset of 5 classes, from a total of 527 sound categories, for the classification task, namely *Wind*, *Siren*, *Gun*, *Dog*, *Car*. The classes of this dataset are not mutually exclusive, i.e., a sample can belong to multiple classes. For this reason the task consists of 5 binary classifications, one for each considered class. In order to evaluate the performance we use mean average precision (mAP) as done in [14].

C. Input features

We evaluate classification accuracy and computational cost for two different front ends: using analog-to-digital converter (ADC) samples from the microphones and using the DASLP spikes. In both cases, binned analog features are used.

The features extracted from the ADC samples are the log-filterbank features as described by [14]. That is, the Short-

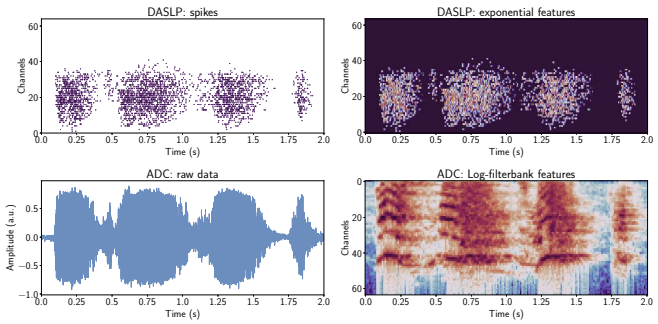

Fig. 3: Sample from AudioSet of the class 'Dog'. Top left panel shows the DASLP spike responses. Top right panel shows the exponential features, the bottom left panel shows the raw sampled audio waveform, and the bottom right panel shows the log-filterbank features.

Time Fourier Transform (STFT) is computed using a frame length of 25 ms and a frame step of 10 ms. A set of 64 Mel-frequency filters is then applied to the STFT. For each sample of length 10 s, we obtain a matrix of size 1000×64. The features generated from the DASLP spikes consist of analog exponential time bin features as described in [9] using a 10 ms frame length, 10 ms frame step and an exponential decay τ=5 ms. These features have been used in a digit recognition task and are simple enough to implement on an ASIC chip as a digital counter [15] or a real-time embedded system with an FPGA [11]. For each sample of length 10 s, we get a matrix of size 1000×64 where each 1 s frame corresponds to a matrix of size 100×64.

Because the input feature dimension is the same for both the spikes and ADC samples, we can process them with the same network configuration (see Section II-D). This makes the comparison fair, because the number of model parameters is the same for both feature types. Figure 3 shows the audio waveform of one sample from the dataset, the DASLP spikes, and the features extracted from the ADC samples and the DASLP spikes. The patterns of activity show the same temporal evolution for both feature types. Note that there are temporal periods in exponential feature plots where there are no spikes in comparison to the dense log-filterbank plot.

D. Network architecture and training parameters

We compare different deep learning architectures in terms of classification accuracy and number of operations required. This knowledge is useful when considering scenarios where computational requirements need to be low (e.g. detection of environmental sound classes in an Internet of Things (IoT) setting).

We investigate 3 architectures, namely a multilayer perceptron (MLP), a convolutional neural network (CNN), and a recurrent neural network (RNN) with a CNN front end. Each of the 3 architectures has two stages: a first stage that applies the same network layers to each 1 s frame in the 10 s sample and a second stage that takes the 10 concatenated processed

TABLE I: Classification accuracies of different models on subset of environmental sounds using DASLP exponential features or log-filterbank features.

Model	Params	MOp	mAP	
			Exp features	Log-filterbank
MLP	3.6 M	34	66%	68%
CNN	0.8 M	38	70%	78%
RNN	1.0 M	39	73%	81%

frames as input and produces the predicted output for each of the classes. The first architecture, i.e, the MLP, consists of 3 layers $(512/256/128)$ in the first stage and 2 layers $(128/5)$ in the second stage where 5 is the number of classes considered.

The second architecture (the CNN) is similar to the architecture presented in [14] and has a first stage with 4 convolutional layers of filter sizes $11/7/5/3$ and 3 channels followed by a 2-layered MLP $(256/128)$, and a second stage that consists of a 2-layered MLP $(128/5)$. The third architecture (the RNN) has the same first stage as the CNN, and a second stage that consists of a recurrent layer with 128 long-short term memory (LSTM) units, 5 attention heads [16] (one for each class) and a 2-layered MLP $(64/5)$. The attention mechanism produces an embedding (used by the last MLP to classify) which is a weighted sum over the frames. For all architectures, dropout $(p = 0.5)$ is used between every layer except for the last one. Training was done over 100 epochs using Adam optimizer [17] and learning rate $1e^{-4}$.

III. Results

The results are shown in Table I. The MLP has the worst performance and yields similar results for both feature types. A better performing architecture is the CNN. As pointed out in [14], the CNN is more successful than the MLP in extracting meaningful features from the 1 s frames in the first stage, making it easier for the second stage to classify the full 10 s frame. With the CNN architecture, we start to see that the log-filterbank features yield better results than the DASLP features. The same trend, but with better results, can be seen when using the RNN. The RNN is the architecture that performs the best. This is due to the fact that the RNN puts together the CNN feature extraction capabilities along with the RNN power of dealing with sequences. In particular, the attention mechanism of the RNN helps to filter out empty frames or frames that are not useful for the final classification.

When comparing the architectures in terms of computational cost, Table I shows that the MLP is again the worst choice. Even though the number of operations is increased in the CNN, there is a significant decrease in the number of stored parameters and the classification accuracy increased by $>$ 10%. In the RNN case, there is a small increase in both the number of operations and parameters compared to the CNN but the classification accuracy increased significantly.

Table II reports the RNN accuracy per class for the two features. From the table, we can see that the overall higher classification accuracy from the log-filterbank features is

TABLE II: Breakdown of the classification accuracy per class for the two input features types. The RNN is used here.

Feat	Class					Total
	Wind	*Siren*	*Gun*	*Car*	*Dog*	
Exp features	67%	85%	65%	70%	75%	73%
Log-filterbank	70%	87%	74%	95%	75%	81%

mostly due to two classes, namely the *Car* and *Gun* classes. By inspection of the samples in these classes, we noticed that a lot of samples in the *Gun* class have gunshot sounds in the background and with very low volume. For this reason, these samples do not trigger cochlea spikes and therefore provide no features for classifying the sample. The same is true for the *Car* samples. Conversely, for the *Siren* and *Dog* classes, results using the DAS features are on par with the results from the ADC features. This is because the volume amplitude of the samples led to the generation of enough spikes for creating useful spike features.

Data-driven sensory processing means that processing is not needed if no audio spikes are generated by the sensor. In this dataset, there are periods of silence for the 10 s file. We first compute the percentage of 10 ms frames where no spikes are detected. The average is around 48%. If the spike features feed into a hardware accelerator such as NullHop [18] where the computation is skipped over zero pixels in a vector, 48% of the computes will be saved. Moreover, if we compute the number of full 1 s frames that are empty, this number goes down to 18%. Therefore, even without a specialized architecture, these frames can be dropped completely by the model leading to savings on the overall number of computes.

IV. Discussion

Low-power spiking audio sensors can drive potential always-on low-power smart audio technology for IoT ([1], [2]). These sensors including the spiking retinas drive the field of data-driven processing where computation is only carried out when the spikes are present. Data-driven sensor processing systems demonstrate properties of low-latency low-energy tradeoff. Systems using these sensors can be used to do initial preprocessing for tasks that do not require much complexity, e.g., detecting speech in a scene. This can then be used to drive a more complex processing pipeline which requires more computes and power. The deep network architectures investigated here can be implemented on low-power hardware accelerators [19], [20] or the network circuits can be combined with the spiking cochlea filter channel circuits for an ASIC design that solves a particular task, e.g., the ASIC for a VAD task dissipates only 1μW [21]. Using information from multiple spiking modalities to better solve a task is one future direction, e.g., by combining visual and audio spikes for solving a task [22].

V. Acknowledgments

We acknowledge Tobi Delbruck and the INI Sensors Group.

REFERENCES

[1] M. Yang, C. Chien, T. Delbruck, and S.-C. Liu, "A 0.5V 55μW 64$\times$2 channel binaural silicon cochlea for event-driven stereo-audio sensing," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2554–2569, 2016.

[2] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with 2$\times$64$\times$4 channel output," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 453–464, Aug 2014.

[3] V. Chan, S. C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, Jan 2007.

[4] H. Finger and S.-C. Liu, "Estimating the location of a sound source with a spike-timing localization algorithm," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2011, pp. 2461–2464.

[5] S.-C. Liu, A. van Schaik, B. Minch, and T. Delbrück, "Asynchronous binaural spatial audition sensor with $2 \times 64 \times 4$ channel output," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453–464, 2014.

[6] A. Zai, S. Bhargava, N. Mesgarani, and S.-C. Liu, "Reconstruction of audio waveforms from spike trains of artificial cochlea models," *Frontiers of Neuromorphic Engineering: Special Issue on Benchmarks and Challenges in Neuromorphic Engineering*, 2015.

[7] M. Abdollahi and S.-C. Liu, "Speaker-independent isolated digit recognition using an aer silicon cochlea," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BIOCAS)*, Nov 2011, pp. 269–272.

[8] D. Neil and S.-C. Liu, "Effective sensor fusion with event-based sensors and deep network architectures," in *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, 2016, pp. 2282–2285.

[9] J. Anumula, D. Neil, T. Delbruck, and S.-C. Liu, "Feature representations for neuromorphic audio spike streams," *Frontiers in Neuroscience*, vol. 12, p. 23, 2018.

[10] E. Ceolini, J. Anumula, S. Braun, and S.-C. Liu, "Event-driven pipeline for low-latency low-compute keyword spotting and speaker verification system," in *Proc. IEEE ICASSP 2019*, 2019, pp. 7953–7957.

[11] C. Gao, S. Braun, I. Kiselev, J. Anumula, T. Delbruck, and S. Liu, "Real-time speech recognition for IoT purpose using a delta recurrent neural network accelerator," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2019, pp. 1–5.

[12] "jAER project," http://jaerproject.org.

[13] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[14] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 131–135.

[15] M. Yang, C. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1μW voice activity detector using analog feature extraction and digital deep neural network," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 346–348.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] A. Aimar, H. Mostafa, E. Calabrese, A. Rios-Navarro, R. Tapiador-Morales, I. Lungu, M. B. Milde, F. Corradi, A. Linares-Barranco, S. Liu, and T. Delbruck, "Nullhop: A flexible convolutional neural network accelerator based on sparse representations of feature maps," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 644–656, March 2019.

[19] V. Sze, Y. Chen, T. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec 2017.

[20] W. Y. Tsai, D. Barch, A. Cassidy, M. Debole, A. Andreopoulos, B. Jackson, M. Flickner, J. Arthur, D. Modha, J. Sampson, and V. Narayanan, "Always-on speech recognition using TrueNorth, a reconfigurable, neurosynaptic processor," *IEEE Transactions on Computers*, vol. PP, no. 99, pp. 1–1, 2016.

[21] M. Yang, C. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1μW voice activity detector using analog feature extraction and digital deep neural network," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 346–348.

[22] X. Li, D. Neil, T. Delbruck, and S. Liu, "Lip reading deep network exploiting multi-modal spiking visual and auditory sensors," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2019, pp. 1–5.