

A Novel Approach for Hardware Based Sound Classification

Mauricio Kugler, Victor Alberto Parciannelo Benso,
Susumu Kuroyanagi, and Akira Iwata

Department of Computer Science and Engineering
Nagoya Institute of Technology
Showa-ku, Gokiso-cho, 466-8555, Nagoya, Japan

mauricio@kugler.com, benso@mars.elcom.nitech.ac.jp,
{bw,iwata}@nitech.ac.jp

Abstract. Several applications would emerge from the development of efficient and robust sound classification systems able to identify the nature of non-speech sound sources. This paper proposes a novel approach that combines a simple feature generation procedure, a supervised learning process and fewer parameters in order to obtain an efficient sound classification system solution in hardware. The system is based on the signal processing modules of a previously proposed sound processing system, which convert the input signal in spike trains. The feature generation method creates simple binary features vectors, used as the training data of a standard LVQ neural network. An output temporal layer uses the time information of the sound signals in order to eliminate the misclassifications of the classifier. The result is a robust, hardware friendly model for sound classification, presenting high accuracy for the eight sound source signals used on the experiments, while requiring small FPGA logic and memory resources.

1 Introduction

By the information provided from the hearing system, the human being can identify 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 safety devices.

In contrast to sound localization, systems capable of identifying the nature of non-speech sound sources were not deeply explored. Some authors study the application of speech-recognition techniques [2], while others attempt to divide all possibly mixed sound sources and apply independent techniques for each kind of signal [3]. Sakaguchi, Kuroyanagi and Iwata [4] proposed a sound classification system based on the human auditory model, using spiking neural networks for identifying six different sound sources.

Due to its high computational cost, sound classification systems are often implemented in hardware for real-time applications. A preliminary hardware

implementation of the model proposed in [4] was presented in [5], while a full system implementation, including the signal processing modules, was proposed in [6]. The spiking neurons based model proposed in [4, 5], although presenting acceptable results, requires the adjustment of several critical parameters, while requiring a large FPGA area, in spite of claims of implementation efficiency of spiking neural networks in digital hardware.

This paper proposes a new approach for sound classification and its correspondent hardware implementation. While still based on spikes, a new feature generation method enables high accuracy with an efficient implementation in hardware. The proposed method also presents few non-critical parameters on the learning process.

The organization of the paper goes as follows: a short description of the signal processing and pulse generation modules is presented in Section 2, and Section 3 introduces the proposed model, which hardware implementation is presented in Section 4. Section 5 presents experimental results, and Section 6 concludes the paper with analysis of the results and suggests possible future extensions.

2 Signal Preprocessing and Spikes Generation

The main structure of the sound classification system is shown in Figure 1(a), with its first block is detailed in Figure 1(b). The use of spikes is required due to the use of the proposed method in a larger system, described previously in [6], which also includes sound localization and orientation detection, in which sharing of signal processing modules is mandatory.

The sound signal is sampled at 48kHz, converted to single-precision floating-point representation and sent to the filter bank module, which divides it in N frequency channels. After, the signals' envelopes are extracted and their amplitude used for controlling the period of the spikes generators. All spike trains $p_n(t)$ ($n = 1 \dots N$) become the input data of the sound classification module.

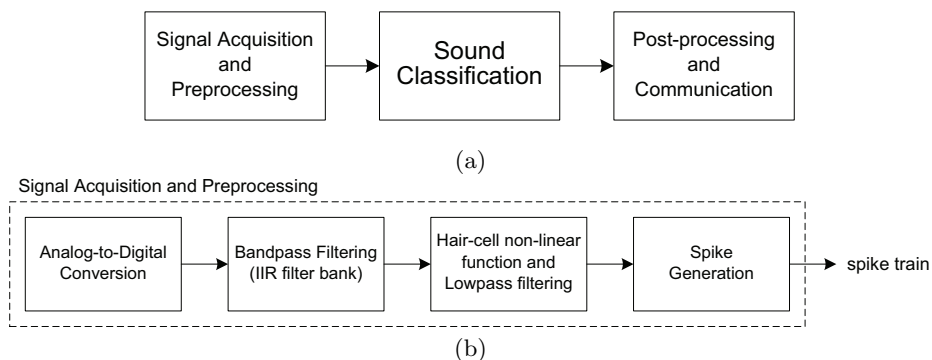


Fig. 1. Sound classification system (a) main structure and (b) signal processing

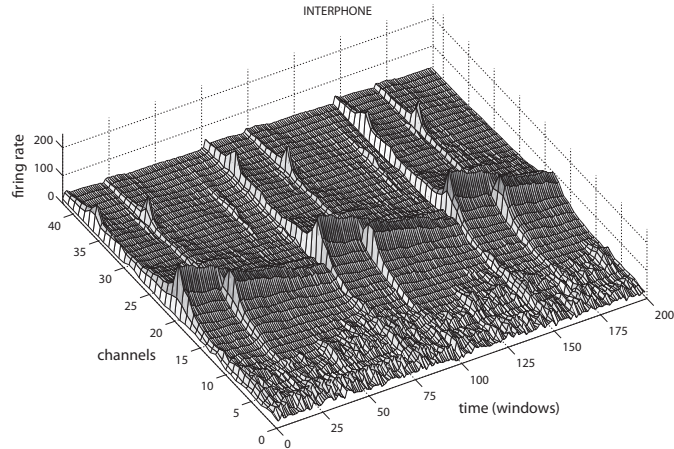


Fig. 2. Spike map for a segment of the *interphone* sound signals

3 Proposed Model

The previous sound classification model proposed in [4, 5] presents several parameters that must be correctly tuned in order to obtain a good accuracy. The tuning process is not straightforward and requires several retrials. Moreover, as the model is based on a non-supervised learning model, a successful learning cannot be always guaranteed and the training process is very time-consuming.

The approach proposed in this paper combines a simple feature generation procedure, a supervised learning process and fewer parameters in order to obtain an efficient sound classification system solution. This solution presents high accuracy and uses small resources (e.g. FPGA area and memory). The following sections present each of the modules in details.

3.1 Feature Generation

As the firing rate of the spike trains corresponds to the amplitude of the signal, a straightforward approach for detecting the energy $x_n(t)$ of each n^{th} frequency channel is to count the number of spikes in a time window of length W :

$$x_n(t) = \sum_{i=0}^{W-1} p_n(t-i) \quad (1)$$

where $p_n(t)$ is the spike train value on time t , $n = 1 \dots N$. The vector \mathbf{x} hence represents the sound pattern in time t . Figure 2 shows the plot of the number of spikes per time window for the sound signal *interphone*, used on the experiments in Section 5.

If this vector is naively used as the input sample \mathbf{z} for the classifier, different signal amplitudes would result in different patterns, what is not desirable. A

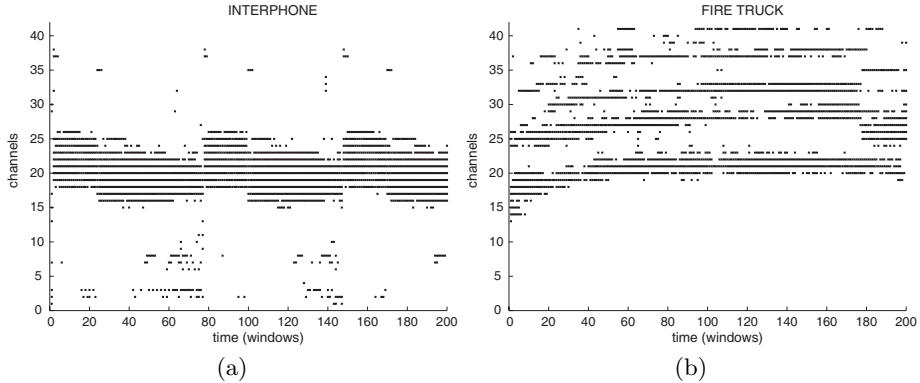


Fig. 3. Binary features for the (a) *interphone* and (b) *fire truck* sound signals

different approach is to consider the highest firing rate channels' indexes as the feature vector for the classifier:

$$z_m = \arg \max_n^m p_n \quad (2)$$

where $\arg \max^i$ represents the index of the i^{th} highest element in a vector, $m = 1 \dots M$ and M is the number of features (number of channels to be considered). The drawback of this approach is that, due to the use of the channels' indexes, small frequency shifts result in large vectorial distances. Therefore, standard distance measurements cannot be applied for comparing the patterns.

Even though the order of the highest firing rates may contain information about the pattern, a very robust set of features can be obtained by ignoring this information. Thus, the new feature vector becomes a binary vector defined as:

$$z_n = \begin{cases} 1 & \text{if } p_n \geq \max_i^M p_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where \max^i represents the i^{th} highest element in a vector. Figure 3 shows the binary features of the *interphone* and *fire truck* sound signals.

3.2 Classification

The standard Learning Vector Quantization (LVQ) neural network [7] was used as the classifier. The learning rate α was reduced linearly along the training epochs by a β factor. The clusters centers were initialized using the Max-Min Distance clustering algorithm [8].

As the patterns were reduced to simple binary vectors, they can be compared by Hamming distance:

$$d(\mathbf{z}, \omega) = \sum_{i=1}^N |z_i - \omega_i| \quad (4)$$

where the elements of sample \mathbf{z} and weight ω , during the training process, are converted to binary values only for distance calculation.

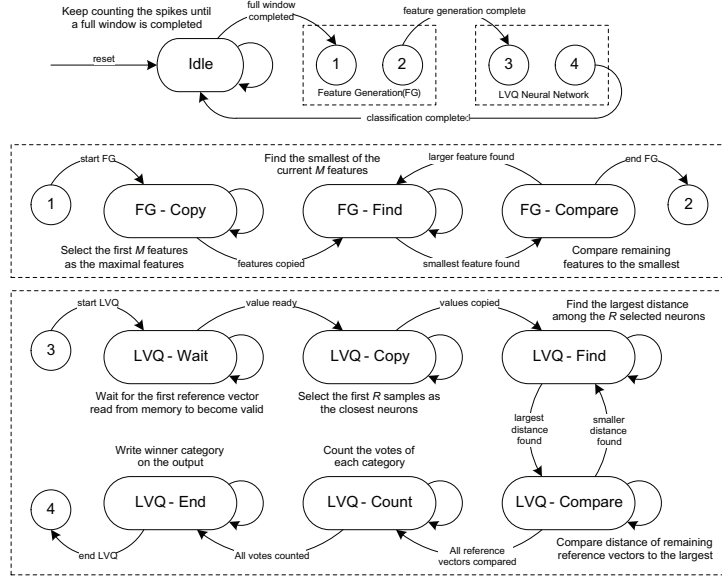


Fig. 4. Sound classification system's state machines

3.3 Time Potentials

In the feature generation process, no feature extraction or selection procedure is being used in order to “clean” the patterns. Although such techniques would avoid misclassifications by the LVQ neural network, they also would add more parameters to the system and increase the training complexity.

Up to the LVQ neural network, no time information had been used for the classification. It can be assumed that the sound sources being recognized will not present instant changes, i.e. they last for periods of time much larger than the size of the time windows. Thus, by the use of potentials similar to the membrane potential of spiking neurons, one can remove the instant errors from the LVQ neural network without modifying the training process. The time potentials are defined as:

$$u_k(t) = \begin{cases} \min(u_{\max}, u_k(t-1) + \gamma) & \text{if } k = y(t) \\ \max(0, u_k(t-1) - 1) & \text{if } k \neq y(t) \end{cases} \quad (5)$$

where u_k is the potential of the k^{th} category, γ is the increment for the winner category and u_{\max} is the maximal potential. Hence, the winner category at time t is the one with higher $u_k(t)$ value. It must be noted that, by setting the u_{\max} parameter, the increment γ does not need to be adjusted. In the experiments of this paper, γ was set fix to 2.

Table 1. LVQ neural network confusion matrix

| Original | Recognition Result | | | | | | | | |
|-------------------|--------------------|------------------|-------------------|-------------------|---------------|--------------|---------------|--------------|----------------|
| | <i>alarm</i> | <i>ambulance</i> | <i>fire truck</i> | <i>interphone</i> | <i>kettle</i> | <i>phone</i> | <i>police</i> | <i>voice</i> | <i>unknown</i> |
| <i>alarm</i> | 1512 | 0 | 0 | 1 | 13 | 0 | 0 | 2 | 0 |
| <i>ambulance</i> | 0 | 3543 | 1 | 870 | 0 | 0 | 18 | 1 | 91 |
| <i>fire truck</i> | 42 | 6 | 4349 | 7 | 16 | 28 | 96 | 0 | 40 |
| <i>interphone</i> | 0 | 121 | 6 | 1985 | 0 | 0 | 4 | 0 | 22 |
| <i>kettle</i> | 40 | 0 | 0 | 0 | 1479 | 0 | 0 | 0 | 1 |
| <i>phone</i> | 1 | 0 | 1 | 0 | 0 | 1765 | 0 | 0 | 1 |
| <i>police</i> | 11 | 106 | 147 | 565 | 13 | 46 | 4039 | 0 | 181 |
| <i>voice</i> | 59 | 902 | 202 | 1850 | 1 | 70 | 62 | 6931 | 387 |

4 Hardware Implementation

The state machine of the sound classification system is shown in Figure 4. The spike counting happens as an independent process, which transfer the total number of spikes to the feature generation process when a full window is completed. The feature generation module sequentially searches the M largest spike rates and send this result to the classification module, which searches for the cluster(s) with the smallest distance to the feature vector. Finally, a winner-takes-all voting is performed and the final label is sent to the output.

The circuit was implemented in an Altera Stratix II EPS260F484C4, which contains 48352 Adaptive Look-Up Tables (ALUT) and more than 2M bits of internal memory. The proposed method (with 9 features, 1000 clusters per class and a 1000 samples time window) uses a total of 2136 ALUTs, 1206 dedicated logic registers (DLR) and 184K bits of memory (for storing the reference vectors of the LVQ neural network). The size of the memory scales linearly with the number of clusters per class and the number of categories, while the number of ALUTs and DLRs presents small increases for larger values of R and number

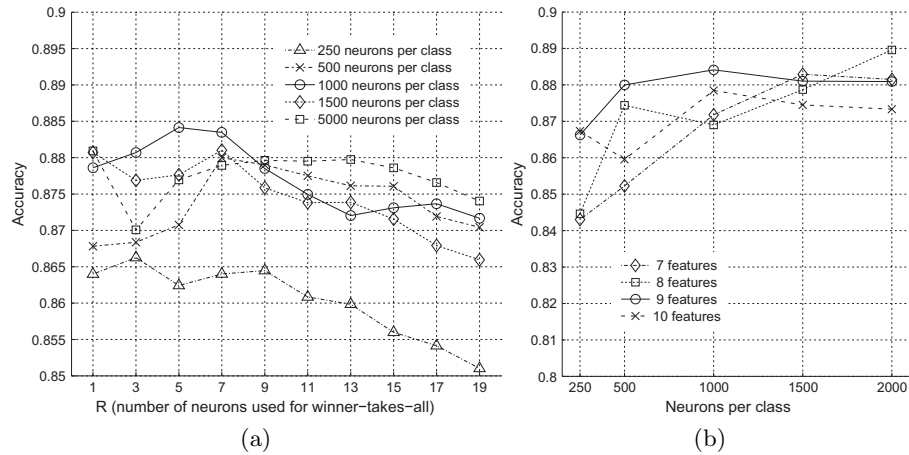


Fig. 5. LVQ accuracy (a) as function of R and (b) as function of the number of features

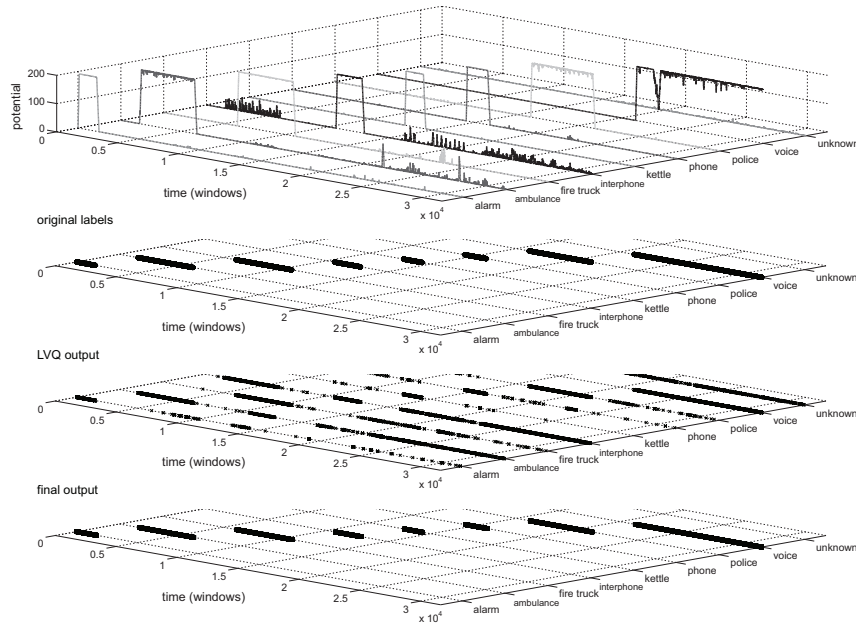


Fig. 6. Time potentials and final sound classification results

of categories. The number of clusters per class and number of features do not significantly increase the number of ALUTs and DLRs.

For this same configuration, the average processing time for the feature extraction and classification procedures are, respectively, is $3.74\mu\text{s}$ and $42.16\mu\text{s}$. The classification processing time grows linearly with the total number of clusters. Considering a 48kHz sampling rate, these timings enable the use of a much larger number of categories and reference vectors.

5 Experiments

Eight sound signals were used on the experiments: *alarm bell*, *ambulance*, *fire truck*, *interphone*, *kettle*, *phone ring*, *police car* and *human voice*. The databases were split in training and test sets in a 2:1 rate. The *voice* database contains several individuals' voice signals, thus, presenting a larger number of samples.

The LVQ network was trained with $\alpha_0 = 0.1$, $\beta = 0.995$ and a maximal of 1000 learning epochs. Figure 5 shows the test accuracy for several training parameters, and Table 1 shows the confusion matrix of the chosen parameters set (9 features, 1000 neurons per class and 1000 samples time window).

The accuracy values presented in Figure 5 and Table 1 are the raw classification result of the LVQ neural network. Figure 6 show the results when calculating the time potentials, for a maximal potential u_{max} equal to 192. All the misclassifications were eliminated, with the the drawback of a small delay introduced on the response of the system.

6 Discussion and Conclusions

This paper proposed a new method for implementing a sound recognition system in an FPGA device. A novel and robust feature generation approach permits the use of a very simple classifier, a standard LVQ neural network, while an independent temporal layer eliminates the misclassifications.

The obtained classification accuracy is encouraging. When running in the FPGA, the proposed model showed to be very robust and insensitive to background noise. On the case of the *human voice* database, several individual's voice not used on the training set were successfully recognized. An improved version of the time potential layer with a better response time is being developed.

Future works include the use of a larger filter bank in order to increase the system's accuracy. Several more sound sources will be used in order to determine the limits of accuracy of the proposed system. The implementation of the learning system in hardware would permit several new applications, as well as increasing the system's flexibility.

References

1. Pickles, J.O.: An Introduction to the Physiology of Hearing. 2nd. edn. Academic Press, London (1988)
2. Cowling, M., Sitte, R., Wysocki, T.: Analysis of speech recognition techniques for use in a non-speech sound recognition system. In: Proceedings of the 6th International Symposium on Digital Signal Processing for Communication System, Manly, TITR (January 2002) 16–20
3. Turk, O., Sayli, O., Dutagaci, H., Arslan, L.M.: A sound source classification system based on subband processing. In Hansen, J.H.L., Pellom, B., eds.: Proceedings of the 7th International Conference on Spoken Language Processing, Denver (September 2002) 641–644
4. Sakaguchi, S., Kuroyanagi, S., Iwata, A.: Sound discrimination system for environment acquisition. Technical Report NC99-70, Nagoya Institute of Technology (December 1999) pp. 61–68.
5. Iwasa, K., Kugler, M., Kuroyanagi, S., Iwata, A.: A sound localization and recognition system using pulsed neural networks on FPGA. In: Proceedings of the 20th International Joint Conference on Neural Networks, Orlando, IEEE Computer Society (August 2007) 1252–1257
6. Kugler, M., Iwasa, K., Benso, V.A.P., Kuroyanagi, S., Iwata, A.: A complete hardware implementation of an integrated sound localization and classification system based on spiking neural networks. In: Proceedings of the 14th International Conference on Neural Information Processing, Volume LNCS 4985 of Part II., Kitakyushu, Springer-Verlag Heidelberg (November 2007) 577–587
7. Fausett, L. In: Fundamentals of Neural Networks: architectures, algorithms and applications. In: Neural Networks Based on Competition. 1st edn. Fundamentals of Neural Networks, New Jersey (1994) 156–217
8. Friedman, M., Kandel, A. In: Introduction to Pattern Recognition: statistical, structural and fuzzy logic approaches. In: Classification by Distance Functions and Clustering. 1st edn. Imperial College Press, London (1999) 73–77