Separation and Recognition of multiple sound source using Pulsed Neuron Model

Kaname Iwasa, Hideaki Inoue, Mauricio Kugler,
Susumu Kuroyanagi, Akira Iwata

Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan

Abstract. Many applications would emerge from the development of artificial systems able to accurately localize and identify sound sources. However, one of the main difficulties of such kind of system is the natural presence of multiple sound sources in real environments. This paper proposes a pulsed neural network based system for separation and recognition of multiple sound sources based on the difference on time lag of the different sources. The system uses two microphones, extracting the time difference between the two channels with a chain of coincidence detection pulsed neurons. An unsupervised neural network processes the firing information corresponding to each time lag in order to recognize the type of the sound source. Experimental results show that three simultaneous musical instruments' sounds could be successfully separated and recognized.

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. With the aim of developing such kind of device, a sound localization and recognition system using Pulsed Neuron (PN) model [2] have been proposed in [3]. PN models deal with input signals on the form of pulse trains, using an internal membrane potential as a reference for generating pulses on its output. PN models can directly deal with temporal data, avoiding unnatural windowing processes, and, due to its simple structure, can be more easily implemented in hardware when compared with the standard artificial neuron model. The system proposed in [3] can locate and recognize the sound source using only two microphones, without requiring large instruments such as microphone arrays [4] or video cameras [5].

However, the accuracy of the system deteriorates when it is used in real environments due to the natural presence of multiple sound sources. Therefore, an important feature of such system is the ability of identifying the presence of multiple sound sources, separating and recognizing each of them. This would enable the system to define an objective sound source type, improving the sound localization performance.
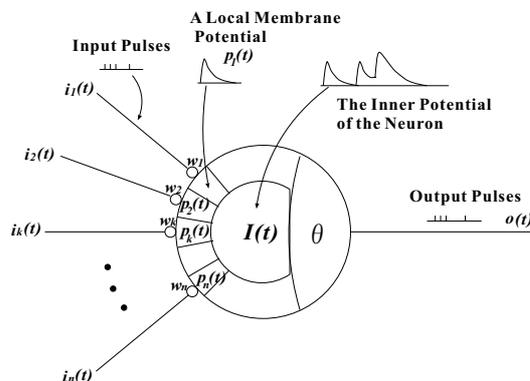
Fig. 1. A pulsed neuron model

In order to extend the system proposed in [3], this paper proposes a PN based system for separation and recognition of multiple sound sources, using their time lag information difference. Based on the time lags' firing information, the sound sources are recognized by an unsupervised pulsed neural network.

2 Pulsed Neuron Model

When processing time series data (e.g., sound), it is important to consider the time relation and to have computationally inexpensive calculation procedures to enable real-time processing. For these reasons, a PN model is used in this research.

Figure 1 shows the structure of the PN model. When an input pulse $i_k(t)$ reaches the kth synapse, the local membrane potential $p_k(t)$ is increased by the value of the weight w_k. The local membrane potentials decay exponentially with a time constant τ_k across time. The neuron's output $o(t)$ is given by

$$o(t) = H(I(t) - \theta) \qquad\qquad I(t) = \sum_{k=1}^{n} p_k(t) \qquad\qquad (1)$$

where n is the total number of inputs, $I(t)$ is the inner potential, θ is the threshold and $H(\cdot)$ is the unit step function. The PN model also has a refractory period t_{ndti}, during which the neuron is unable to fire, independently of the membrane potential.

3 Proposed System

The basic structure of the proposed system is shown in Fig. 2. This system consists of three main blocks, the frequency-pulse converter, the time difference extractor and the sound recognition estimator. The time difference extractor and sound recognition estimator blocks are based on a PN model.
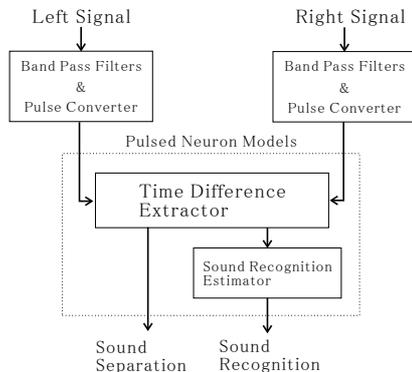
Fig. 2. Basic structure of the proposed system

The left and right signals' time difference information is used to localize the sound source, while the spectrum pattern is used to recognize the type of the source.

3.1 Filtering and Frequency-Pulse Converter

In order to enable pulsed neuron based modules to process the sound data, the analog input signal must be divided on its frequency components and converted to pulses. A bank of band-pass filters decomposes the signal, and each frequency channel is independently converted to a pulse train, which rate is proportional to the amplitude of the correspondent signal. The filters' center frequencies were determined in order to divide the input range (100 Hz to 16 kHz) in 72 channels equally spaced in a logarithm scale.
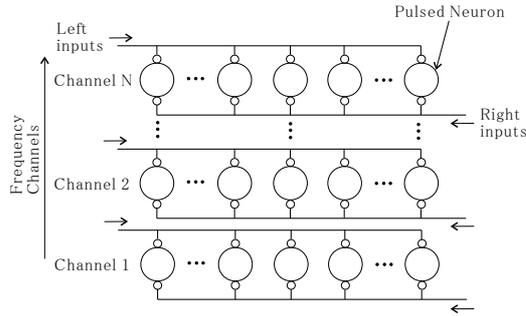
3.2 Time Difference Extractor

Each pulse train generated at each frequency channel is inputted in an independent time difference extractor. The structure of the extractor is based on Jeffress's model [7], in which the pulsed neurons and the shift operators are organized as shown in Fig. 3. The left and right signals are inputted in opposed sides of the extractor, and the pulses are sequentially shifted at each clock cycle. When a neuron receives two simultaneous pulses, it fires. In this research, the neuron fires when both input's potentials reach the threshold θ_{TDE}. The position of the firing neuron on the chain determines the time difference.

This work uses an improved method, initially proposed in [8], which consists on deleleting the two input pulses when a neuron fires for preventing several false detections due to the matching of pulses of different cycles, as shown in Fig. 4.

3.3 Sound Recognition Estimator

The sound recognition estimator is based on the Competitive Learning Network using Pulsed Neurons (CONP) proposed in [6]. The basic structure of CONP is shown in Fig.5.

Fig. 3. Time Difference Extractor

(a) time t (b) time $t+1$
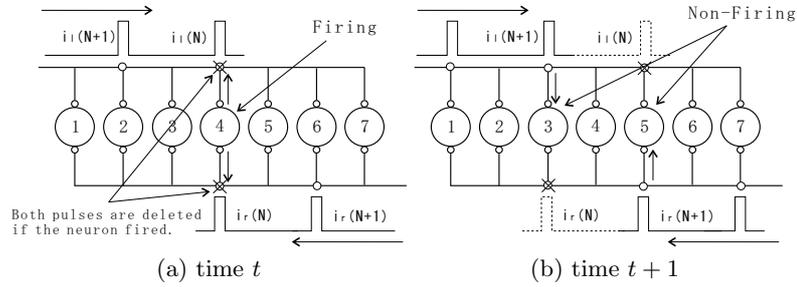
Fig. 4. Pulse deleting algorithm in Time Difference Extractor

In the learning process of CONP, the neuron with the most similar weights to the input (winner neuron) is chosen for learning in order to obtain a topological relation between inputs and outputs. For this, it is necessary to fire only one neuron at a time. However, in the case of two or more neurons firing, it is difficult to decide which one is the winner, as their outputs are only pulses, and not real values. In order to this, CONP has extra external units called control neurons. Based on the output of the Competitive Learning (CL) neurons, the control neurons' outputs increase or decrease the inner potential of all CL neurons, keeping the number of firing neurons equal to one. Controlling the inner potential is equivalent to controlling the threshold. Two types of control neurons are used in this work. The No-Firing Detection (NFD) neuron fires when no CL neuron fires, increasing their inner potential. Complementarily, the Multi-Firing Detection (MFD) neuron fires when two or more CL neurons fire at the same time, decreasing their inner potential.

The CL neurons are also controlled by another potential, named the input potential $p_{in}(t)$, and a gate threshold θ_{gate}. The input potential is calculated as the sum of the inputs (with unitary weights), representing the frequency of the input pulse train. When $p_{in}(t) < \theta_{gate}$, the CL neurons are not updated by the control neurons and become unable to fire, as the input train has a too small potential for being responsible for an output firing. Furthermore, the inner
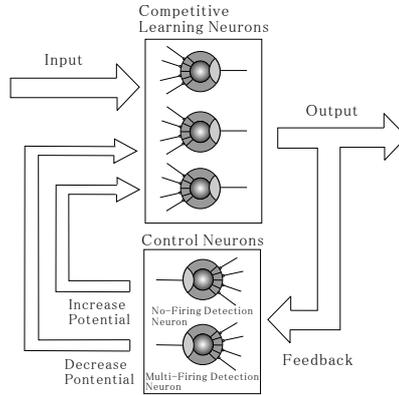
Fig. 5. Competetive Learning Network using Pulsed Neurons (CONP)

potential of each CL neuron is decreased by a factor β, in order to follow rapid changes on the inner potential and improving its adjustment.

Considering all the described adjustments on the inner potential of CONP neurons, the output equation (1) of each CL neurons becomes:

$$o(t) = H\left(\sum_{k=1}^{n} p_k(t) - \theta + p_{nfd}(t) - p_{mfd}(t) - \beta \cdot p_{in}(t)\right) \qquad (2)$$

where $p_{nfd}(t)$ and $p_{mfd}(t)$ corresponds respectively to the potential generated by NFD and MFD neurons' outputs, $p_{in}(t)$ is the input potential and β $(0 \leq \beta \leq 1)$ is a parameter.

4 Experimental Results

In this work, several sound signals generated by computer were used: three single frequency signals (500 Hz, 1 kHz and 2 kHz), and five musical instruments' sounds ("Accordion", "Flute", "Piano", "Drum" and "Violin"). Each of these signals were generated with three different time lags: -0.5 ms, 0.0 ms and $+0.5$ ms, with no level difference between left and right channels.

4.1 Separation of Multiple Sound Sources

Initially, the time difference information is extracted as described in section 3.2. The used parameters for the signal acquistion, preprocessing and time difference extraction are shown in Table 1. The 48 kHz sampling frequency causes the pulse train to shift 20.83 μs at each clock cycle (Fig.3), resulting in output time lags of 41.66 μs for each neuron.

Figure 6(a) shows the output of the time difference extractor for an input composed by the 500 Hz single frequency signal ($+0.5$ ms lag) the 1 kHz signal

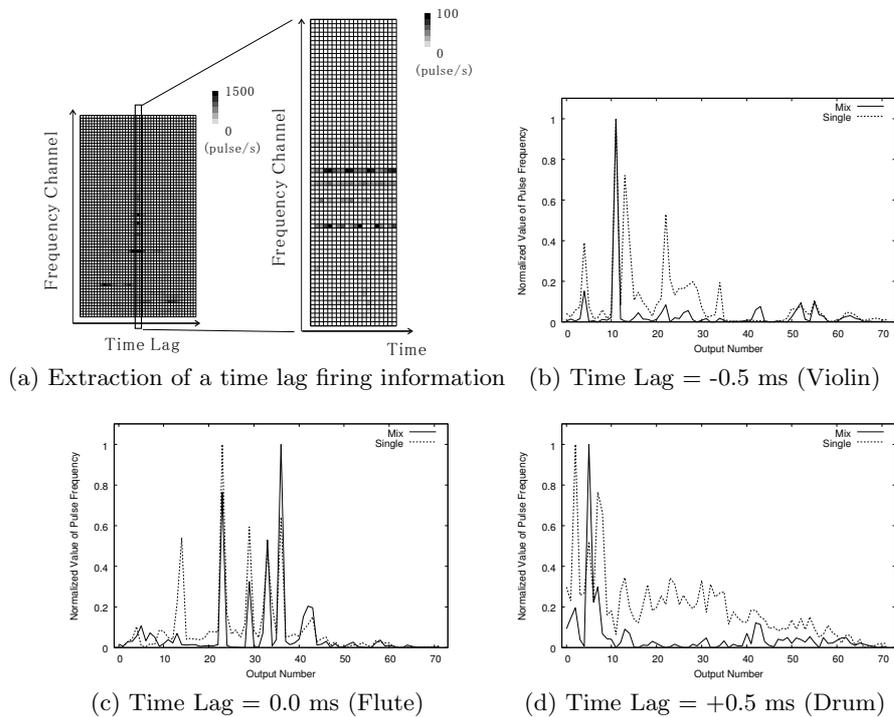**Table 1.** Parameters of each module used on the experiments

Input Sound	
Sampling frequency	48 kHz
Quantization bit	16 bit
Number of frequency channels	72
Time Difference Extractor	
Total number of shift units	121
Number of output neurons	41
Threshold θ_{TDE}	1.0
Time constant	350 μs

(a) single frequency signals   (b) musical instruments
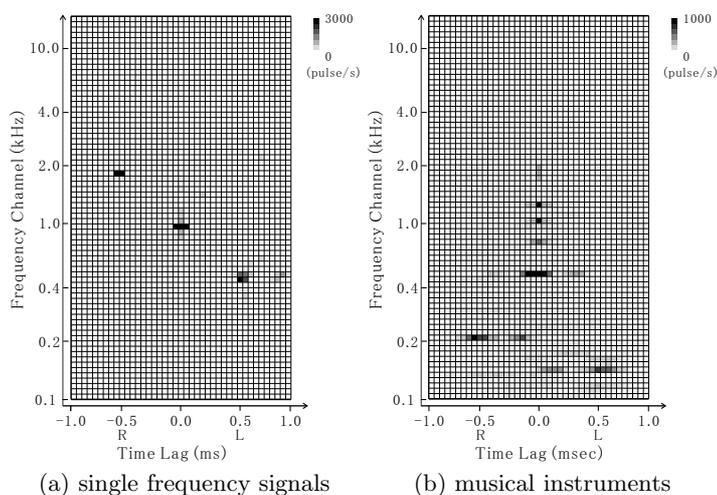
Fig. 6. Output of Time Difference Extractor for three different signals

(0.0 ms lag) and the 2 kHz signal (-0.5ms lag) The x-axis corresponds to the time lag (calculated from the firing neuron in the time difference extractor) and the y-axis corresponds to the channels' frequency. The gray-level intensity represents the rate of the output pulse train. Figure 6(b) shows the output relative to the musical instruments' sounds "Drum" ($+0.5$ ms lag), "Flute" (0.0 ms lag) and "Violin" (-0.5 ms lag). Again, each time lag shows a different firing pattern in each position.

Figure 7(a) shows the extraction of the firing information for each of the identified instruments in Fig. 6. It can be seen that the frequency components are constant along time. Furthermore, Fig 7(b) to (d) show the output firing information of each sound (Mix), together with the original firing information for the independent sounds with no time lag (Single). All data is normalized for comparison, showing that important components are similar. As both results present firing in different frequency components for each time lag, it is possible to recognize the type of sound source for each time difference.

(a) Extraction of a time lag firing information   (b) Time Lag = -0.5 ms (Violin)

(c) Time Lag = 0.0 ms (Flute)   (d) Time Lag = +0.5 ms (Drum)

Fig. 7. Extraction of the independent time lags firing information

4.2 Recognition of Independent Sound Sources

Each time lag's firing information is recognized by the CONP model described in section 3.3. Initially, the firing information of each type of sound source is extracted with no time lag. This data is used for training CONP, according to the parameters shown in Table 2.

The five musical instruments' sounds were applied to the CONP in all combinations of three simultaneous sounds with the three time lags (60 combinations). Table 3 shows the average accuracy of the CONP model for each instrument in each position. The recognition rate is calculated by the ratio between the number of firings of the neuron corresponding to the correct instrument and the total number of firings.

In this result, the accuracy of "Piano" was particularly bad at the central position. Figure 8 shows the weights of the neurons corresponding to the sounds of "Accordion", "Flute" and "Piano" after learning. Not only the "Piano" neuron does not present any relevant weight but also some of the highest weights are very similar to the weights of other instruments' corresponding neurons (e.g., inputs 4 and 23). The reason for this pour performance is that the "Piano" sound is not constant, presenting a complex variation along a short period of time. This characteristic makes this kind of sound difficult to be learned by

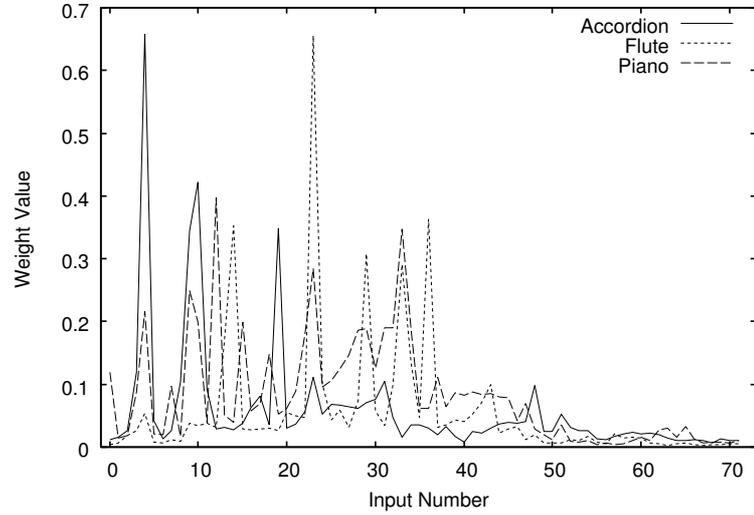**Table 2.** Parameters of CONP used on the experiments

| Competitive learning Neuron	
Input Number of CL neurons	72
Number of CL neurons	5[units]
Threshold θ	1.0×10^{-4}
Gating threshold θ_{gate}	100.0
Rate for input pulse frequency β	0.11785
Time constant τ_p	20[msec]
Refractory period t_{ndti}	10[msec]
Learning coefficient α	2.0×10^{-7}
Learning iterations	1000
No-Firing Detection Neuron	
Time constant τ_{NFD}	0.5[msec]
Threshold θ_{NFD}	-1.0×10^{-3}
Connection weight	
to each CL neurons	0.8
Multi-Firing Detection Neuron	
Time constant τ_{MFD}	1.0[msec]
Threshold θ_{MFD}	2.0
Connection weight	
from each CL neurons	1.0

Table 3. Results of sound recognition

Input \ Time Lag	Recognition Rate[%]		
	$-0.5ms$	$0.0ms$	$+0.5ms$
Acordion	89.9	88.1	88.8
Flute	92.3	94.4	92.4
Piano	62.5	32.9	64.0
Drum	90.3	89.1	88.6
Violin	79.7	78.4	79.0

the CONP model. Nevertheless, other instruments' sounds could be correctly identified in all positions with accuracies higher than 78%. This confirms the efficiency of the proposed system on identifying multiple sources based on the time lag information.

Similarly to the human being, the proposed system cannot distinguish between two simultaneous similar sound sources. For instance, the results shown in Fig. 9(a) show the output of the Time Difference Extractor for signal composed by the "Violin" sound coming from the left and central directions (-0.5 ms and 0.0 ms lags) and the "Flute" sound in the right direction (+0.5 ms lag). For reference, Fig. 9(b) shows a single "Violin" signal on the central position. As expected, only two firing patterns can be observed, on corresponding to the "Flute" sound at +0.5 ms and another corresponding to the "Violin" sound at -0.25 ms. This is, however, an unrealistic situation, as in applications on real environments the occurrence of two identical simultaneous sounds is very improbable, not compromising the applicability of the system.

Fig. 8. The weights about three sound source

(a) two identical "Violin" signals
on the left and central positions
and "Flute" signal on the right position

(b) single "Violin" signal
on the central position
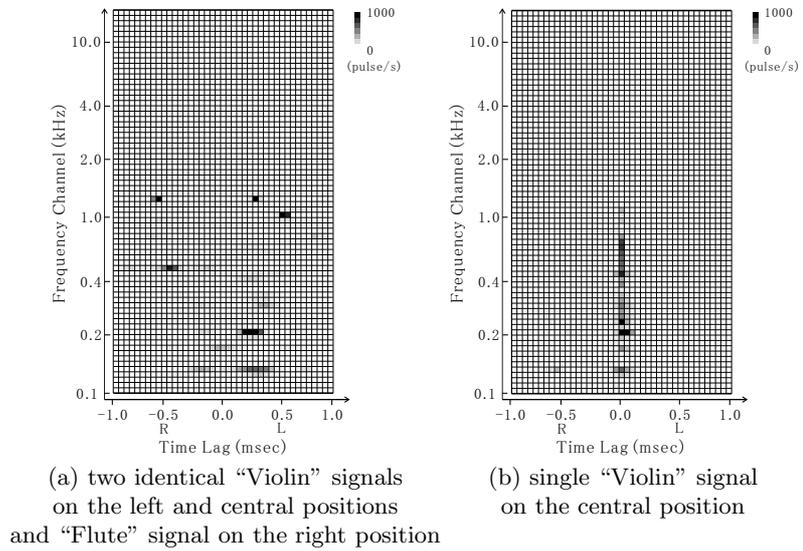
Fig. 9. Output of Time Difference Extractor for two identical signals

5 Conclusions

This paper proposes a system for multiple sound source recognition based on a PN model. The system is composed of a time difference extractor, which separates the spectral information of each sound source, and a CONP model which recognizes the sound source type from the firing information of each time lag.

The experimental results confirm that the PN model time difference extractor can successfully separate the spectral components of multiple sound sources. Using the time lag firing information, the sound source type could be correctly identified in almost all cases. The proposed system can separate the multiple sound sources and classify the each sound.

Future works include the application of the proposed system to real sound signals, and also the use of the information of the sound sources type for locating this source with high precision. The implementation of the current system in hardware using an FPGA device is also in progress.

Acknowledgment

References

1. Pickles J.O.: "An Introduction to the Physiology of Hearing" , ACADEMIC PRESS, 1988
2. Maass W., and Bishop C.M., : "Pulsed Neural Networks", MIT Press , 1998
3. Kuroyanagi, S. , Iwata, A. : "Perception of Sound Direction by Auditory Neural Network Model using Pulse Transmission – Extraction of Inter-aural Time and Level Difference", Proceedings of IJCNN 1993, pp.77-80, 1993.
4. Valin J.M. , Michaud F. , Rouat J. , Letourneau D. : "Robust Sound Source Localization Using a Microphone Array on a Mobile Robot", Proceedings of IROS 2003, pp.1227-1233 , 2003.
5. Asoh H. , et al : "An Application of a Particle Filter to Bayesian Multiple Sound Source Tracking with Audio and Video Information Fusion" , Proceedings of The 7th International Conference on Information Fusion , pp.805-812 , 2004.
6. Kuroyanagi, S. , Iwata, A. : "A Competitive Learning Pulsed Neural Network for Temporal Signals", Proceedings of ICONIP 2002, pp.348-352, 2002.
7. Jeffress, L.A,: "A place theory of sound localization", J.Comp.Physiol.Psychol. , 41 , pp.35-39(1948).
8. Iwasa, K., et al : "Improvement of Time Difference Detection Network using Pulsed Neuron model", Technical Report of IEICE NC2005-150, pp151-156 , 2006.