Using spiking onset neurons and a recurrent neural network for musical sound classification

Michael Newton & Leslie Smith
{mjn,lss}@cs.stir.ac.uk
Dept. of Computing Science and Mathematics
University of Stirling, FK9 4LA, UK

161st Meeting of the ASA
Seattle, WA, USA
23-27 May, 2011
Session 2pPP (24th May, PM)
Poster ID: 2pPP8

EPSRC Engineering and Physical Sciences Research Council
UNIVERSITY OF STIRLING

Abstract

Physiological evidence suggests that specific neurons within the cochlear nucleus specialise in sound onset detection. Sudden increases in sound energy, (e.g. during the initial transient of a sound), result in a transient, increased firing rate in such onset neurons. Onset timing and spectral location are thought to play a key role in a number of auditory tasks including sound identification and direction finding. Onset neurons are modeled using leaky integrate-and-fire units, innervated by parallel spiking data streams produced using a passive gammatone filterbank followed by positive-going zero-crossing detection. Dynamic level is coded using multiple spike trains per filter channel. The model is presented with 2085 musical samples across five musical instrument categories from the McGill dataset. Clusters of onset spikes occur close to the beginning of each note, and these are used to produce a unique *onset fingerprint* signal for the sound. The objective of the study is to use these onset fingerprints as descriptors for classification. A recurrent neural network (echo state reservoir network), which allows the use of temporal signals, is used as a classifier. The results are compared with a regular, non-temporal sound classification scheme based on cepstral coefficients and a multilayer perceptron neural network.

Introduction

An onset is defined as a sudden and rapid rise in signal energy, often wideband, as perceived by the sound receptor (in this case the cochlea). A common example is the initial transient of an isolated musical note. This work investigates the role of the transient *sound onset* in providing a means for discriminating between different types of musical instrument.

There is evidence in the literature [1] which emphasises the onset as providing an important cue for sound identification in human subjects. Ecologically this is plausible because the onset, coming at the start of a sound, may aid in priming a response. It is also less likely to be corrupted by reverberation, and so may contribute a clean signal from which to perform a range of tasks such as direction finding and sound identification. This work is motivated by 2 questions:

1) **Is the sound onset, simulated using a biologically-inspired model of the early auditory system, useful as a descriptor in a classification task?**
2) **How does the biologically-inspired scheme compare to a simpler sound descriptor based on cepstral coefficients?**

Method

Experimental design

The experiment was based upon 2085 isolated musical note samples from the McGill[2] dataset, between octaves 1 and 6. The samples were split evenly between 5 instrument classes, which were chosen according to the physics of the sound generation mechanism[3]:

- *Brass (trumpet, trombone, cornet)*
- *Reed (clarinet, oboe)*
- *Bowed string (violin, viola, cello)*
- *Plucked string (violin, viola, cello)*
- *Struck string (piano)*

The hypothesis was that there may be correlation between the similar sound production physics within each class and the nature of the sound onset. The experimental work proceeded via two alternative strategies:

1. Biologically-inspired approach. Simulate the spiking onset response of certain specialised neurons within the cochlear nucleus[4] to each sample. Use the onset responses, with a time-domain recurrent neural network as a classifier, to attempt robust identification of the five instrument categories.
2. Classical approach. Classification based upon cepstral coefficients[6] evaluated over the same onset period, using a multilayer perceptron neural network[7] as a classifier.

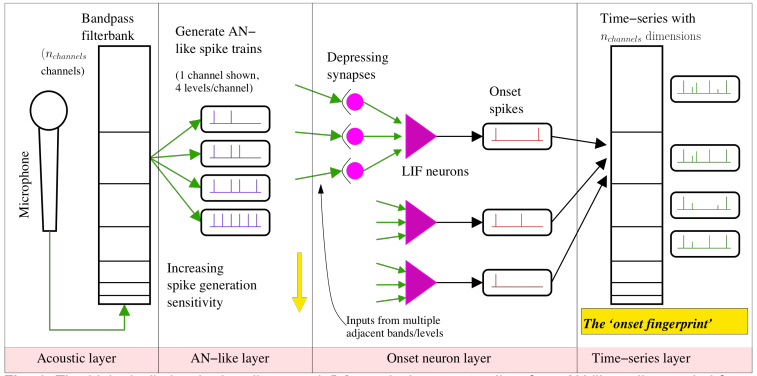

Fig. 1: The biologically-inspired auditory model[4], producing onset spikes from AN-like spikes coded from the raw sound. The onset fingerprint *is then coded as an* $n_{channels}$*-dimensional time-series.*

Biologically-inspired auditory model (strategy 1)

A spiking representation of the auditory nerve (AN) signal (Fig. 2) was generated using a gammatone filterbank (15 channels, 0.1-10kHz) coupled to a positive-going zero crossing detector over multiple sensitivity levels (Fig. 1). Parallel AN channels innervated (LIF) *onset detector neurons*[4], which produced spikes around sound onset. These spikes were then coded into an $n_{channels}$-dimensional time-series signal, the *onset fingerprint* (Fig. 3), which included both timing and dynamic level information (duration ~ 50ms).

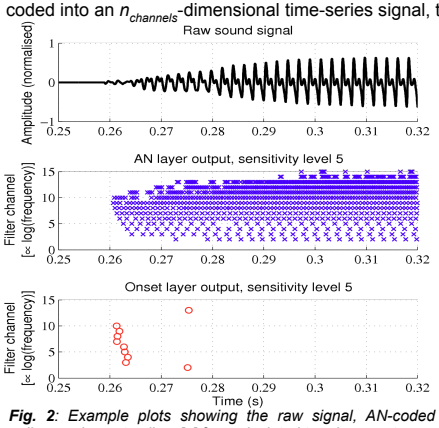

Fig. 2: Example plots showing the raw signal, AN-coded spikes and onset spikes [4] for an isolated trombone note.

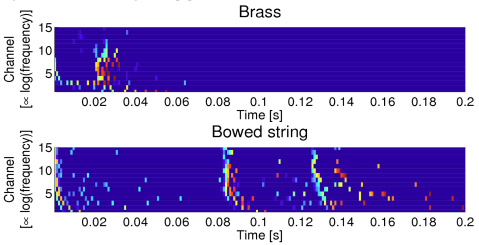

Fig. 3: Onset fingerprint time-series plots of brass and bowed string notes. Colour maps signal intensity (red is high, blue is low).

The echo state network classifier (strategy 1)

A recurrent neural network, the echo state network[5], was used as a classifier (Fig. 4) for the onset fingerprint signals.

- Operates in the time domain, allowing retention of the precise timing information included in the onset fingerprint.
- Training performed on 70% of the data. Classification success based on testing with the remaining 30% which have not been seen by the network.

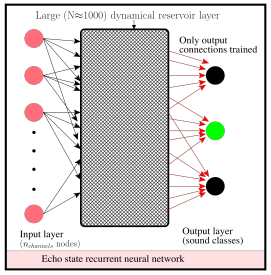

Fig. 4: Diagram of an echo state network[5], used to classify sounds based on their onset fingerprints.

Cepstral coefficient-based model (strategy 2)

As a comparison to the biologically-inspired spiking method, the dataset was also coded using 15 cepstral coefficients[6], from 0.13-8kHz, evaluated during the same onset period.

- The 15 coefficients formed a single descriptor vector for each sound.
- Classification was attempted using a standard multilayer perceptron neural network[7,8], with 15 input and 5 output (class) neurons.
- 70%-30% train/test split, with 5 repetitions which were averaged.

Results and discussion

A comparison between the success rates of the two methods is shown using confusion matrices in Fig. 5. Despite the use of only the transient sound onset as a descriptor (typically 50ms in duration), both methods achieved reasonable success rates.

- The onset fingerprinting and echo state neural network approach achieved consistently better success rates for all classes.

There were similarities in what the two methods found easy/difficult.

- The plucked and struck string classes were often confused. This seems intuitively reasonable as both groups involve impulsively setting a string into free vibration.
- The bowed string class was consistently the most distinguishable. This group tended to have the longest onset period (see Fig. 3) which may have provided a useful temporal clue for the onset fingerprinting method.

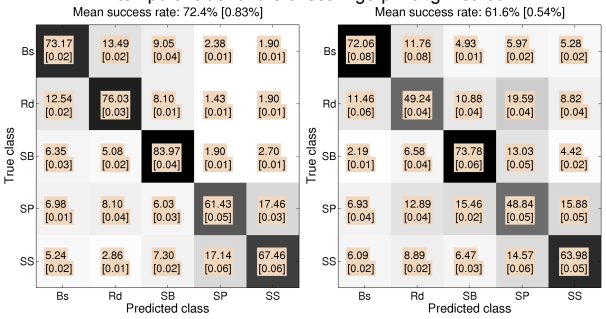

Fig. 5: [Left] Confusion matrix for the onset fingerprint/echo state network method. [Right] Confusion matrix for the cepstral coefficient method. Classification scores are shown as percentages. Standard deviations based on 5 repetitions with the same network parameters but different train/test splits shown in brackets.

Conclusions

This work has demonstrated that it is feasible to build a musical instrument category classifier based only upon the transient onset period of isolated notes.

Considering the small time period of the onset relative the the full note (generally 1/10 – 1/100) this tends to confirm existing theories which emphasise its importance for sound identification.

Future work will expand the methods to include information from the steady state signal alongside the onset.

This work was sponsored by the Engineering and Physical Sciences Research Council, UK. Grant EP/G062609/1.

References

1. J. M. Grey and J. A. Moorer (1977). J. Acoust. Soc. Am. 62(2), pp. 454–462.
2. F. Opolko and J. Wapnick (2006). *McGill University Master Samples*.
3. N. Fletcher and T. Rossing (1998). *The Physics of Musical Instruments*. Springer.
4. L. S. Smith and S. Collins (2007). IEEE Transactions of Audio, Speech and Language Processing, 15(8), 2278-2286.
5. H. Jaeger *et al* (2007). Editorial, Neural Networks, 20(3), 287–289.
6. J. C. Brown (1999). J. Acoust. Soc. Am. 105(3), pp. 1933–1941.
7. S. Haykin (1998). Neural Networks: A Comprehensive Foundation, Prentis Hall.
8. M. Hall et al (2009). *The WEKA Data Mining Software: An Update*, SIGKDD Explorations 11(1).