Biologically-inspired neural coding of sound onset for a musical sound classification task

Michael J. Newton and Leslie S. Smith

***Abstract*—A biologically-inspired neural coding scheme for the early auditory system is outlined. The cochlea response is simulated with a passive gammatone filterbank. The output of each bandpass filter is spike-encoded using a zero-crossing based method over a range of sensitivity levels. The scheme is inspired by the highly parallellised nature of the auditory nerve innervation within the cochlea. A key aspect of early auditory processing is simulated, namely that of onset detection, using leaky integrate-and-fire neuron models. Finally, a time-domain neural network (the *echo state network*) is used to tackle the *what* task of auditory perception using the output of the onset detection neurons alone.**

I. Introduction

The mammalian auditory system performs a diverse range of signal processing tasks in near real time. Presented with a raw sound field, analysis is carried out to extract meaningful features, which may or may not be buried along with contributions from other sound sources. Such useful features include the direction from which a particular sound arrived (the *where* task)[1], [2], the nature of a individual sound (the *what* task)[3], interpreting the meaning of the sound (as in speech perception)[4] and decomposing a many-source sound field into seperable *audio streams*[5], [6]. In many cases several of these tasks must be performed at the same time.

The processing of sound within the auditory system is highly integrated, involving neural processes at all levels, from the cochlea to the cortex. The system is two-way, with information passed both upwards to the cortex[7], and back downwards towards the sensory units through the efferent system[8], [9]. A key feature is that certain kinds of processing occur early on, even in advance of the brain stem[10].

In this work a biologically-inspired scheme for sound onset representation within the auditory system is investigated. There is strong evidence to suggest that mammalian auditory systems are particularly attuned to the detection of sound onsets, even from the earliest stages of the auditory processing chain[11], [12]. The auditory nerve itself is known to respond more strongly to the start of a stimulus, and there are neurons within the cochlear nucleus which spike strongly at stimulus onset[4], [13], [14]. Sound onsets may be important for sound source location[1], sound identification[15], [16], and are thought to play a role in the segregation of auditory streams[5], [6], [17].

Michael J. Newton and Leslie S. Smith are with the Institute of Computing Science and Mathematics, University of Stirling, UK (email: {lss, mjn}@cs.stir.ac.uk).

This work was supported by EPSRC (UK) Grant EP/G062609/1.

From an ecological perspective the sound onset is potentially useful because its location at the start of a sound may aid in priming a response. The initial onset also tends to be relatively untainted by reverberation, as it usually arrives at the listener via a direct path from the source. For most tasks later reflections are ignored in favour of the initial onset[18].

Every sound begins with an onset. However, the precise definition of what constitutes the 'sound onset' is less clear[19]. It is possible to analyse a sound onset based on the physics of the sound production mechanism. In the case of a trumpet blowing a pitched note, for example, there is a short period of time at the beginning of the note when the vibrating lips of the player are not influenced by the acoustics of the instrument. At some later time a coupled interaction begins, which leads to the steady-state pitched note. It may be argued that the onset portion of the note occurs before full coupling between instrument and player, and the steady-state portion follows coupling. However, such a physical process is not necessarily perceived in the same clear order by the auditory system. A number of further factors, such as reverberant reflections, may contribute to the final waveform which reaches the ear.

The precise meaning of 'onset' in the context of perception can thus only be properly explored by studying the response of the auditory system to real sounds. What is clear is that the temporal fine-structure and frequency evolution of sound onsets varies widely, both in terms of perception[13] and from a generative standpoint. A drum hit, for example, clearly involves a different kind of physical onset than a slowly bowed violin string, and would be expected to produce a different sensation of 'onset' in a listener. We henceforth refer to the perceptual onset simply as the *onset*, and, in seeking to explore it with an auditory model, define it as a sudden and rapid rise in signal energy as seen by the sound receptor (in this case the cochlea). This may be a rise from a zero-level, or a pronounced increase from one level to a higher level.

In this work the perceptual onset is simulated using a spiking time-domain auditory model, based on the gammatone filterbank[20]. Section II provides an overview of the model and the coding scheme. In section III a method is outlined which uses the simulated spiking onset response as a descriptor for a musical sound classification task. Musical samples are sourced from the McGill dataset[21]. The classification is performed using a time-domain reservoir neural network known as the echo state network[22], which is outlined in section IV. Section V provides an overview of some initial classification results.

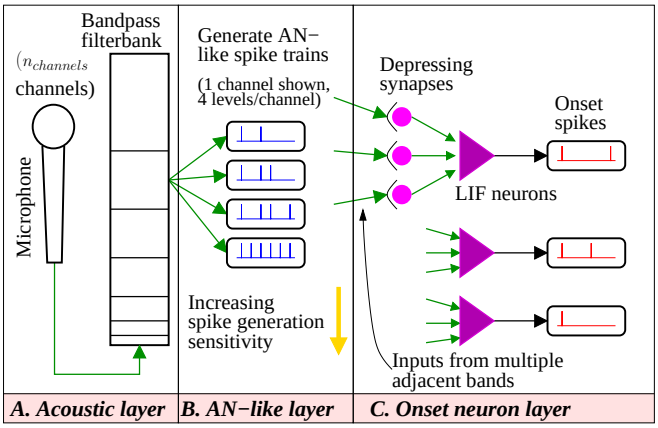

Fig. 1. A schematic diagram outlining the auditory model. Note that AN spike generation is shown for only one channel, and onset neurons/depressing synapses for a single sensitivity level and three input channels.

II. Biologically-inspired coding of the early auditory signal

There have been numerous attempts to design a sound onset detector[23], [24], [25], [26]. The most common uses for such a detector have been in automatic music transcription[27], sound segmentation[28], [17], [10], lip synchronisation[29], monaural sound-source separation[5], [30] and sound direction finding[1]. In this work an attempt was made to use a neural-like coding of the sound onset as a descriptor in a musical sound classification problem.

The onset detection technique was based on a biologically-inspired model of the mammalian auditory system, illustrated schematically in Fig. 1. The cochlea response was modelled with the ubiquitous passive gammatone filterbank (*A*)[20]. The output from each gammatone filter was spike-encoded (*B*) using a zero-crossing based technique[31], the design of which was inspired by the phase-locked spiking behaviour observed in neurons which innervate the cochlea's inner hair cells (IHC)[32]. This encoding thus provides a crude simulation of the auditory nerve's (AN) early response to sound stimuli. The strong spiking onset response observed by certain neurons within the cochlear nucleus[14] was then modelled using an array of leaky integrate-and-fire (LIF) neurons, innervated by the simulated AN signal (*C*), as implemented in [31]. Example outputs from these processing stages are shown in Fig. 2.

A. Gammatone filtering

The first processing step of the auditory model was to filter the sounds using a gammatone filterbank[20]. This filterbank was comprised of $n_{channels} = 100$ bandpass filters, the (roughly logarithmic) spacing and bandwidths of which are designed to mimic the first order response of the basilar membrane. The 6dB down point bandwidth is approximately 20% of the centre frequency of the channel. Using 100 channels between 0.1kHz and 10kHz ensured considerable overlap between adjacent filters, as is the case with the cochlea filter. All sound samples used were sampled at 44.1kHz and 16bits.

B. AN-like spike encoding

The outputs from the filterbank channels were coded in a manner inspired by the neural coding within the mammalian auditory nerve. The output from each channel was spike-encoded over $n_{levels} = 16$ sensitivity levels, leading to a total of $n_{channels} \times n_{levels} = 1600$ individual spike trains. The use of multiple sensitivity levels per channel provided information about the dynamic level changes of the signal across frequency and time.

Spikes were produced at positive-going zero-crossings of the filtered signals. For each detected zero-crossing i, the mean signal amplitude during the previous quarter cycle E_i was calculated and compared to the values $S_{j=1:16}$ described by the $n_{levels} = 16$ sensitivity levels. If $E_i > S_j$ then a spike was produced at the j^{th} sensitivity level. The sensitivity levels ran from small values at $j = 1$ (high sensitivity, low signal level required to produce a spike) to large values at $j = 16$ (low sensitivity, large signal level required to produce a spike), with a difference δ_{levels} of 3dB between levels. For any spike produced at level $j = k$, a spike was necessarily produced at all levels $j < k$. This representation is similar to that employed in [33], where Ghitza noted that it led to an improvement in automatic speech recognition in a noisy environment.

C. Onset detection

The AN-like representation described above does not emphasise onsets in the encoded sound signal, unlike the real mammalian auditory nerve[12]. However, its highly parallelised design makes it suitable for use with a secondary onset detection system. This system was inspired by the onset response behaviour exhibited by certain cells within the cochlea nucleus (octopus, and some bushy and stellate cells)[14].

The AN-like spike trains were passed through depressing synapses to a leaky integrate-and-fire (LIF) neuron layer. There was one LIF neuron per filterbank channel per sensitivity level (i.e. $n_{channels} \times n_{levels} = 1600$ onset neurons), and each neuron was innervated by AN spike-trains from n_{adj} adjacent frequency bands (at the same sensitivity level) on either side of its centre frequency.

The synapse model was based on the 3-reservoir model used in [34] in the context of IHC-to-AN fibre transduction. A similar model has also been used in [35] to model rat neocortex synapses. The model employed three interconnected reservoirs of neurotransmitter. Reservoir M represented the available presynaptic neurotransmitter, reservoir C was the neurotransmitter currently in use, and reservoir R contained neurotransmitter in the process of reuptake (i.e. used, but not yet available for reuse). The reservoir quantities were related by three first order differential equations as follows:

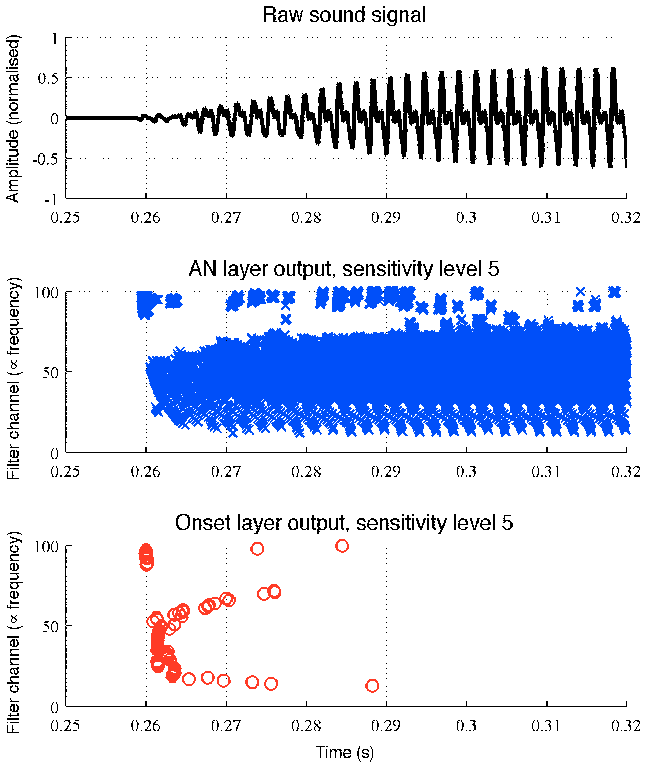

Fig. 2. Example plots showing the raw sound signal, AN-coded spikes and onset spikes for a single tone produced by a brass instrument. Onset spikes are clustered at the start of the note. Only the 5^{th} sensitivity level is shown here. We call the overall pattern of onset spikes, across all channels and sensitivity levels, the *onset fingerprint* of the sound.

$$\frac{dM}{dt} = \beta R - \gamma M \quad (1)$$

$$\frac{dC}{dt} = \gamma M - \alpha C \quad (2)$$

$$\frac{dR}{dt} = \alpha C - \beta R \quad (3)$$

where α and β are rate constants, and γ was positive during an AN-spike, and zero otherwise. The differential equations were calculated for each time sample as the AN spike train signals were fed to the onset layer through the depressing synapses. The loss and manufacture of neurotransmitter was not modelled, and the amount of post-synaptic depolarisation was assumed to be directly proportional to C.

Innervation of each onset neuron in channel b and sensitivity level j from n_{adj} adjacent channels resulted in a total input to the neuron of

$$I_{b,j}(t) = \sum_{h=b-n_{adj}}^{h=b+n_{adj}} wC_{h,j}(t) \quad (4)$$

where w was the weight of each synapse (the same for all inputs here) and $C_{h,j}$ was the neurotransmitter currently in use in the cleft between the AN input from channel h, at sensitivity level j and the onset neuron. A n_{adj} value of 3 was used, so that each onset neuron was innervated by 7 AN channels.

TABLE I
SUMMARY OF PARAMETERS USED IN THE SPIKING AUDITORY MODEL

Parameter	Description	Value
n_{adj}	Number of co-innervating AN channels on each onset neuron	3
α	Rate constant, neurotransmitter reservoir C	100
β	Rate constant, neurotransmitter reservoir R	9
γ	Value during an AN-spike	1100
w	Synapse weight (all synapses)	1
$n_{channels}$	Number of filterbank channels	100
n_{levels}	Number of sensitivity levels	16
δ_{levels}	Inter-sensitivity level difference	3dB
E_i	Mean signal level over quarter cycle before detected zero-crossing	N/A
S_j	Value of sensitivity level j	N/A
b	Filter channel index	N/A
$\mathcal{T}$	Time-series representation of the initial onset fingerprint signal	N/A N/A
δt_{step}	Duration of time windows used for time-series	2ms
δt_{group}	Timing threshold for onset event groupings	20ms

Assuming the signal in a given bandpass channel b was strong enough to produce AN spikes at sensitivity level j, the corresponding onset neuron for channel b, at sensitivity level j, would receive at least F_b spikes per second (where F_b is the centre frequency of the channel). In the case of multiple co-innervating adjacent channels on each onset neuron ($n_{adj} > 0$), as used in this study, this rate would normally be greater due to contributions from higher frequency channels. However, depletion of the available neurotransmitter reservoir M, in conjunction with a slow reservoir recovery rate, meant that an evoked post-synaptic potential (EPSP) would only be produced for the first few incoming AN spikes. The recovery rate was purposefully set low to ensure that synapses did not continue to produce EPSPs much beyond the initial sound onset.

The synapse weights w were further set to ensure that a single EPSP was insufficient to cause the onset neuron to fire. This ensured that multiple ESPSs from adjacent synapses were required for the neuron potential to be large enough to fire. The neurons employed were also leaky[36], [31], meaning that the ESPSs needed to be close to concurrent for an action potential, or 'onset spike', to be produced.

III. ONSET FINGERPRINTS AND TIME-SERIES REPRESENTATIONS OF THE ONSET

The problem of musical sound classification has been the subject of extensive study. The most common approach to the task has been to calculate a range of descriptors for a sound based on its Fourier components[15], [16], [37], [38], [39]. Such descriptors may be based on analysis of the whole sound, or upon just the steady state and/or the initial transient portion of the sound. Cepstral coefficients are a particularly popular quantity, and have shown good performance with certain classification tasks[40]. A mixture of frequency and time-domain quantities has also been proposed and shown to be up to 90% successful in a 15 class task[41].

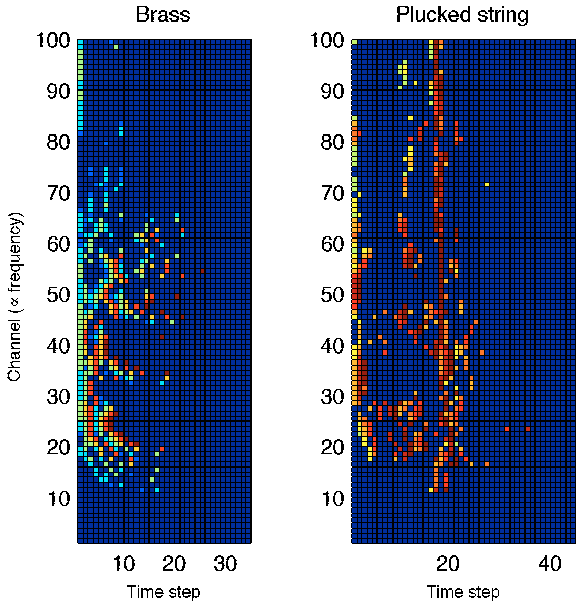

Fig. 3. Example plots showing *initial onset fingerprint time-series* (IOFTS) $\mathcal{T}$ representations of a brass note (trombone, at left) and plucked string (guitar, at right). Each raw onset fingerprint (described by onset spikes across channel, sensitivity level and time) has been reduced to 2-dimensional time-series. The number of time steps depends on the nature of the individual onset. Local signal intensity is represented by colour (red is high), and each time step is 2ms. This is the representation used as a sound descriptor for the classification problem with the echo state network (see section IV).

Most of the outlined approaches have used standard signal processing techniques to calculate a large number of descriptors ($\sim 30-50$), which form a one-dimensional feature vector D. Many sounds can be quickly processed, and a standard feed-forward learning framework may be employed to classify the sounds based on their D vector[41], [40]. Although such techniques can be remarkably successful, their underpinnings are somewhat removed from the spiking, highly parallelised nature of the mammalian auditory perception and learning 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 learning and classification itself.

The auditory model described in section II takes raw sound as an input and provides a simplified representation of the onset response within the cochlea nucleus as an output. The objective of this work was to use this onset response as a descriptor in a musical sound classification task similar to that presented in [41]. The key feature of this approach is that it operates purely in the time domain, and produces onset spikes which are also in the time domain. If the principle advantage of the method was to be exploited, namely the retention of precise timing and frequency information during the sound onset, then the standard feed-forward classification procedures (such as back-propagating or radial basis function neural networks) were unsuitable.

It was thus proposed that the classification descriptor be based entirely on the pattern of onset spikes, which we call the *onset fingerprint*, and that the descriptor should remain as a time-domain representation of the onset. In order to exploit such a temporal representation, a neural network which operated in the time-domain was required. The echo state network approach, though originally developed for time-series prediction[22], has also proven to be a popular choice for similar classification tasks[42], [43], and was employed here. Its recurrent, temporal nature was also appropriate for the biologically-inspired framework of the present study.

It would be possible to use the raw onset fingerprint as the time-domain onset descriptor. However, in the present study a simplified form was used which reduced the large 3-dimensional onset array (over multiple channels and sensitivity levels) to a smaller 2-dimensional time-series matrix. This was done to reduce the number of input channels required by the echo state classification network (see section IV). For each processed sound, the entire onset fingerprint was first grouped to identify the onset feature which corresponded to the start of the musical note. This was important as certain sounds, such as a flute tone played with heavy vibrato, can produce onset spikes during the steady state due to the rather large amplitude variations introduced by the vibrato. Here the onset descriptor was limited exclusively to the initial onset transient. The grouping procedure examined the time separation between onset spikes within each onset fingerprint. Groups of onset spikes separated by more than a critical time period δt_{group} (here set to 20ms) were treated as separate onset events. Only the first onset event grouping was picked out as the descriptor, which we term the *initial onset fingerprint* (IOF, see Fig. 2).

The IOF was further processed to reduce the number of temporal sample points. This was achieved by time-slicing the IOF into windows of duration δt_{step} (2ms used here). The spiking onset behaviour (across all sensitivity levels) within each time window of each filter channel b was examined, and only a single spike at the least sensitive sensitivity level (corresponding to the maximum signal level) retained. If no spikes occurred, a zero was recorded. In this manner the onset data within each time window was reduced to a single vector, as shown in Fig. 3. For an IOF lasting 24ms, this resulted in a 12-step *initial onset fingerprint time-series* (IOFTS) $\mathcal{T}_{i=1:12}$ (where i indexes time-step), with each step comprised of an $n_{channels}$-sized vector. Without this reduction to a 2D time-series, the same IOF, in its raw onset state, would require a $1058 \times n_{channels} \times n_{levels}$ 3-dimensional matrix (where 1058 is the original number of time samples). More sophisticated methods for performing this step, such as PCA, are currently under investigation. The key outcome was that the raw IOF, while coded in a reduced space to give the IOFTS, remained as a time-domain representation of the sound onset. This was the signal used as input to the time-domain neural network (see section IV-A) to solve the classification task outlined in section IV-B.

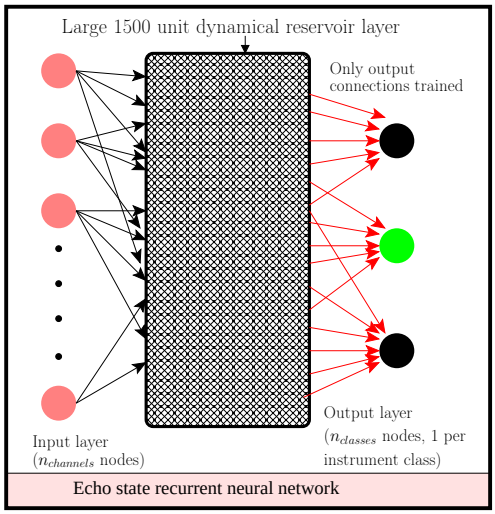

Fig. 4. A schematic diagram outlining the structure of the echo state network[22]. A single input layer consisting of $n_{channels}$ nodes connects directly into a large 1500 unit untrained reservoir layer. Only connections from the reservoir layer to the output layer are trained. The green node in the output layer illustrates the manner in which the network is trained to flag a certain class according to the current input time sequence.

IV. Recurrent neural networks for time-series classification

A. The echo state network approach

The echo state network (ESN) approach to recurrent neural networks has grown in popularity over the past decade[22]. It represents an implementation of reservoir computing, where a large, fixed and interconnected reservoir layer is perturbed by an input signal(s), as illustrated in Fig. 4. A trained linear combination of the nonlinear responses of the reservoir units is used as the learning framework. This approach is related to the support vector machine techniques, which transform data from an input space into a (much) higher dimensional feature space[44], within which the data is easier to separate. Such networks have proved particularly effective at certain kinds of time-series learning problems[44]. Time domain classification problems have also been addressed using the ESN approach, in particular for speech recognition[42], [43], [45].

In this work an ESN was used as a framework for classifying the time-series' (IOFTS, see section III) obtained from the onset fingerprints of single musical instrument notes. A Matlab implementation of the ESN method developed by Jaeger *et al* was employed[46]. The input layer consisted of 100 nodes, one for each of the $n_{channels}$ filter channels (see Fig. 4), each fed by the corresponding channel of the IOFTS illustrated in Fig. 3. The reservoir layer consisted of 1500 randomly connected and weighted additive sigmoid neurons, the interconnections of which were untrained. The output layer consisted of $n_{classes}$ nodes, one for each of the musical instrument classes (5 in this case). Only connections between the reservoir layer and the output nodes were trained, using a Delta-rule method.

TABLE II
Summary of instrument types used in the classification task

Class label	Description	Examples	Number in class	Mean onset duration
Bs	Brass	Trumpet, cornet	456	80ms
Rd	Reed	Clarinet, oboe	469	110ms
SB	Bowed string	Cello, violin	524	120ms
SP	Plucked string	Cello, violin	438	45ms
SS	Struck string	Piano	510	46ms

B. The classification task

The classification task used musical instrument samples drawn from the McGill Master Samples dataset[21]. This dataset is comprised of high quality recordings of orchestral musical instruments, generally playing isolated notes. Rather than trying to classify individual instruments, the present work attempted to use the initial onset fingerprint time-series descriptor to differentiate between instruments based on their excitation technique, as in [41]. This is an easier task, but as outlined in section I it is both physically and taxonomically relevant, and so provides a useful test case for the method.

The instrument categories chosen each involve a different excitation mechanism[47], and so may be expected to produce a different kind of perceptual onset. The five instrument categories ($n_{classes}$) used were brass, reed (both single and double reed), plucked string, bowed string and struck string. These classes are summarised in Table II, together with a note of their mean onset durations. A total of 2397 individual sounds were used, with approximately 450 sounds per class. The data was split into a 70%/30% training/testing ratio, with 10 different random permutations run through the echo state network classifier and analysed separately.

To assemble the dataset, the sounds were first processed individually to obtain a set of stand-alone *initial onset fingerprint time-series'* $\mathcal{T}_i^s$, where s indexes each sound, and thus runs from $1:2397$ in the present study. The temporal duration of each $\mathcal{T}_i^s$, i.e. the number of time-steps i, varied between approximately 10 and 50 (20-100ms) depending on the nature of the onset (see Fig. 3 for an illustration of the IOFTS). The training dataset was assembled by randomly picking 70% of the individual time-series sequences $\mathcal{T}_i^s$. These sequences were concatenated together one after the other in a random order, separated by short periods of 'silence', to create a single overall input training sequence $\mathcal{R}^{Tr}$.

A teaching sequence $\mathcal{G}^{Tr}$ was created with the same number of time-steps as $\mathcal{R}^{Tr}$, and consisting of $n_{classes}$ parallel sequences (one for each instrument class). At each time-step i, the teacher signal $\mathcal{G}_i^{Tr}$ was zero everywhere except in the sequence index of the current IOFTS, which had a value of unity.

Periods of 'silence', composed of 10 time-steps with zeros across all channels and outputs, were inserted between each time-series $\mathcal{T}^s$ within the training/teachings sequences $\mathcal{R}^{Tr}$ and $\mathcal{G}^{Tr}$. Inserting silence in this manner was found to aid the classification success, most likely because it allowed the

network to revert back to a 'rest state' between being stimulated by the successive time-series' $\mathcal{T}^s$. Further study into the nature of this effect are ongoing. The testing sequences, composed of the input signal $\mathcal{R}^{Te}$ and the target signal $\mathcal{G}^{Te}$, were similarly assembled from the remaining 30% of the onset data.

C. Analysis of the ESN classification task

The echo state network described in section IV-A was trained/tested ten times with different 70%/30% splits of the five-class musical instrument dataset outlined in section IV-B and Table II. During the training phase the network was stimulated with the training sequence $\mathcal{R}^{Tr}$. At each time-step the difference between the measured signal in the output layer $\mathcal{M}^{Tr}$ and the target output $\mathcal{G}^{Tr}$ was used to refine the weights between the reservoir layer and the output layer.

During the testing phase the trained network was stimulated with the testing sequence $\mathcal{R}^{Te}$ and the observed output layer signal $\mathcal{M}^{Te}$ recorded. No training was performed. The network's classification at each time-step i was deduced by identifying the largest component of the $\mathcal{M}^{Te}$ signal.

The comparison between target and measured output class was performed separately for each $\mathcal{T}^s$ (IOFTS) within the overall testing sequence $\mathcal{R}^{Te}$. In order to allow the network time to compute the class of the current $\mathcal{T}^s$, this evaluation was performed only during the final 50% of each IOFTS. The class computed by the network for each 50% portion of an IOFTS was taken as the most frequently occurring class in the output signal $\mathcal{M}^{Te}$ during this time period. This class was then treated as the network's prediction of the instrument class for the current IOFTS, and was directly compared with the true class recorded in the teacher signal $\mathcal{G}^{Te}$ for the testing data.

V. Results of the Classification Task

A. Classification success rates

Fig. 5 shows a mean confusion matrix, produced from a ten-fold cross-validation of the testing data, for the five class problem described in section IV-B (and Table II). It is important to note that this classification result was based exclusively on the initial onset fingerprint representation produced by the auditory model outlined in section II. No information regarding the steady state timbre was used.

The reed (Rd), bowed string (SB) and struck string (SS) classes were all identified correctly more frequently than not. The bowed string, which produced the longest mean onset duration of all five classes (see Table II), was particularly well identified (more than three-quarters of the time). The reed, which also featured a relatively long mean onset duration, was also identified with reasonable success.

However, the confusion matrix revealed an overall classification success of approximately 45%. This low value was clearly influenced in part by the poor performance of the network in identifying the brass (Bs), plucked string (SP) and struck string (SS) classes. Such a result suggests that in its present condition the onset fingerprint/echo state

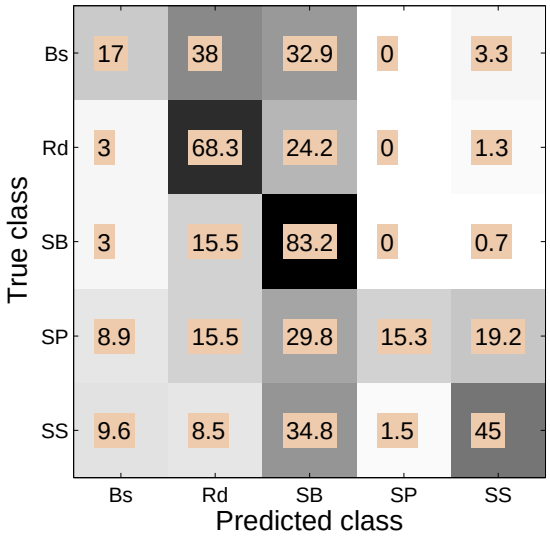

Fig. 5. A mean confusion matrix calculated from testing data (30% of the total data), using ten cross-validations of the 5 class musical instrument type identification task (see Table II). The predictions were obtained from the echo state network, with time-series onset fingerprints (IOFTS) as sound descriptors. Values are expressed as percentages of the total number of examples in each class. Shading is scaled from white (0%) to black (100%).

network method under-performed, at least with instrument classification problems, relative to results reported elsewhere which use the entire audio signal[41]. However, the work reported here represents an initial pilot study, the results of which may improve with further refinement of both the onset fingerprinting method and the implementation of the echo state network.

It may also be the case that there are limits as to the feasibility of identifying musical instrument types based on such a reduced representation of their initial transients. Indeed, while there is much evidence in the literature which reports on the significance of the onset for sound identification tasks[15], [16], this has always been in combination with other features of the sound. The fact that it can be difficult to identify the musical instrument type in absence of the original sound onset does not necessarily imply that the sound onset alone may permit a successful identification. However, the relative success of the technique with three of the five instrument classes does suggest that the technique, with suitable refinement, may prove useful.

B. Analysis of the network learning strategy

It is interesting to note that the two instrument classes with the longest mean onset durations, the brass and the reed groups (see table II), were the most accurately identified. There may be two factors at play here. Firstly, a longer onset means more average training time spent by the network in learning the onset fingerprints from these groups. Secondly, the way in which the network was configured to learn may have favoured longer onsets. During the training phase the network was set to learn continuously, including during the

'silent' periods between the successive $\mathcal{T}^s$ sequences within the overall $\mathcal{R}^{Tr}$ signal. A key feature of the ESN lies in its memory of previous states. By allowing the network to continue to train during the silent periods, regardless of the immediately previous target class, it is possible that it's ability to accurately identify the most recent time steps of the previous $\mathcal{T}^s$ sequence was disrupted. This would be proportionally less significant for onsets of longer duration. This theory is further supported by the fact that progressively reducing the testing classification period (see section IV-C) towards the end of the onset did not increase the success rate. This is in direct contrast to what would be expected if the network was consistently and successfully classifying the current $\mathcal{T}^s$. The 50% quotient used here was found to be about as good as could be obtained within the current framework.

Work is currently under way to alter the network's learning pattern during the training phase. In particular, the network will stop learning during all silent portions of the training signal $\mathcal{R}^{Tr}$. It will also be prevented from learning during the first part of each $\mathcal{T}^s$ within $\mathcal{R}^{Tr}$, in a further attempt to prevent echoed states from recent $\mathcal{T}^s$ sequences from interfering with the current input signal.

VI. Conclusions and further work

The technique of onset fingerprinting was used to form sound descriptors for a five class musical instrument identification task. The initial results presented here suggest that such a method may provide useful as an initial classifier which, in combination with further parameterisation of the auditory signal, could allow a robust biologically-inspired classification framework to be developed.

On their own the results here appear relatively poor in terms of the method's overall success rate. Further work will be required to determine if such a representation of the auditory signal, based on the onset fingerprint technique, can prove to be robust in isolation from the steady-state period of a sound. A more advanced implementation of a reservoir network, such as an echo state network with periodically engaged learning, may prove to be more suited to highly variable onset fingerprint patterns than the continuous online learning employed so far. Development of a more traditional Fourier-based description of the sound onset is also ongoing. This will allow a more detailed comparative picture to emerge of the true success of the present biologically-inspired technique.

Alternative methods for reducing/encoding the full onset fingerprint as a time-series are currently under investigation. In particular, principle component analysis (PCA) may prove to be a more useful technique for capturing the detail of the onset fingerprint than the simple time-windowing method employed here. Sound descriptors which involve aspects of the steady state timbre, in combination with the onset fingerprinting technique, are also under development. The objective throughout remains to develop descriptors which are biologically-inspired representations of the auditory signal. The remarkable success of the ear with all auditory tasks remains a high benchmark at which to aim.

References

[1] L. Smith and S. Collins, "Determining ITDs using two microphones on a flat panel during onset intervals with a biologically inspired spike-based technique," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2278–2286, 2007.

[2] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[3] S. K. Scott and D. G. Sinex, "Cortical processing of speech," in *The Oxford Handbook of Auditory Science: The Auditory Brain*, D. R. Moore, Ed. Oxford University Press, 2010, vol. 2, ch. 9, pp. 200 – 214.

[4] C. Darwin, "Speech perception," in *The Oxford Handbook of Auditory Science: Hearing*, D. R. Moore, Ed. Oxford University Press, 2010, vol. 3, ch. 9, pp. 207–230.

[5] A. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[6] Y. I. Fishman and M. Steinschneider, "Formation of auditory streams," in *The Oxford Handbook of Auditory Science: The Auditory Brain*, D. R. Moore, Ed. Oxford University Press, 2010, vol. 2, ch. 10, pp. 215 – 245.

[7] J. H. Kaas and T. A. Hackett, "Subdivisions of auditory cortex and processing streams in primates," *Proc. Natl. Acad. Sci. USA.*, vol. 97, no. 11793-11799, 2000.

[8] M. C. Holley and J. F. Ashmore, "On the mechanism of a high-frequency force generator in outer hair cells isolated from the guinea pig cochlea," *Proc. R. Soc. Lond.*, vol. 232, pp. 413–429, 1988.

[9] R. F. Huffman and O. W. Henson Jr., "The descending auditory pathway and acousticomotor systems: connections with the inferior cells," *Brain Res. Rev.*, vol. 15, pp. 295–323, 1990.

[10] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[11] R. L. Smith, "Adaptation, saturation and physiological masking in single auditory nerve fibres," *J. Acoust. Soc. Am.*, vol. 65, pp. 166–178, 1979.

[12] T. C. Chimento and C. E. Schreiner, "Time course of adaptation and recovery from adaptation in the cat auditory-nerve neurophonic," *J. Acoust. Soc. Am.*, vol. 88, no. 857-864, 1990.

[13] I. Winter, A. Palmer, L. Wiegrebe, and R. Patterson, "Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus," *Speech Communication*, vol. 41, pp. 135–149, 2003.

[14] E. Rouiller, "Functional organization of the auditory pathways," in *The Central Auditory System*, G. Ehret and R. Romand, Eds. Oxford University Press, 1997.

[15] J. M. Grey and J. A. Moorer, "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Am.*, vol. 62, no. 2, pp. 454–462, 1977.

[16] S. McAdams, J. W. Beauchamp, and S. Meneguzzi, "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 882–897, 1999.

[17] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 396–405, 2007.

[18] J. Blauert, *Spatial Hearing*, revised edition ed. MIT PRESS, 1996.

[19] J. P. Bello, L. Daudet, S. Abdallah, D. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[20] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Paper presented at a meeting of the IOA: Speech Group on Auditory Modelling at RSRE*, 1987.

[21] F. Opolko and J. Wapnick. (2006) McGill University Master Samples. [Online]. Available: http://www.music.mcgill.ca/resources/mums/html/

[22] H. Jaeger, "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.

[23] L. Smith, "Sound segmentation using onsets and offsets," *Journal of New Music Research*, vol. 23, no. 1, pp. 11–23, 1994.

[24] X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients," in *In Proceedings of ICMC 2001*. Cuba: University of Michigan, Ann Arbor: International Computer Music Conference, 2001. [Online]. Available: http://hdl.handle.net/2027/spo.bbp2372.2001.105

[25] J. P. Bello, D. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, 2004.

[26] L. S. Smith, "Using depressing synapses for phase locked auditory onset detection," in *In Proceedings of ICANN 2001: Artificial Neural Networks*, G. Dorffner, H. Bischof, and K. Hornik, Eds., 2001.

[27] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*. Washington, DC, USA: IEEE Computer Society, 1999, pp. 3089–3092.

[28] L. S. Smith, "Using an onset-based representation for sound segmentation," in *Neural networks and their applications: Proceedings of NEURAP95, Marseilles, France*, 1995, pp. 274–281.

[29] C. Tait, "Wavelet analysis for onset detection," Ph.D. dissertation, University of Glasgow, UK, 1997.

[30] G. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, pp. 297–336, 1994.

[31] L. S. Smith and D. S. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses," in *IEEE Transactions on Neural Networks*, vol. 15, no. 5, 2004, pp. 1125–1134.

[32] A. R. Palmer and I. J. Russell, "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells," *Hearing Research*, vol. 24, no. 1, pp. 1–15, 1986.

[33] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech & Language*, vol. 1, no. 2, pp. 109 – 130, 1986.

[34] M. J. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," *J. Acoust. Soc. Am.*, vol. 90, no. 2, pp. 904–917, 1991.

[35] M. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proc. Nat Acad Sciences*, vol. 94, pp. 719–723, 1997.

[36] C. Koch, "Chapter 14: Simplified models of individual neurons," in *Biophysics of Computation*. Oxford University Press, 1999.

[37] A. Klapuri and M. Davy, "Signal processing methods for the automatic transcription of music," *Tampere University of Technology Publications*, vol. 460, 2004.

[38] J. C. Brown, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1064–1072, 2001.

[39] J. Barbedo and G. Tzanetakis, "Musical instrument classification using individual partials," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 111–122, 2011.

[40] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1933–1941, 1999.

[41] K. D. Martin and Y. E. Kim, "Musical instrument identification: A pattern-recognition approach," *J. Acoust. Soc. Am.*, vol. 104, no. 3, p. 1768, 1998.

[42] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[43] M. D. Skowronski and J. G. Harris, "Automatic speech recognition using a predictive echo state network classifier," *Neural Networks*, vol. 20, pp. 414–423, 2007.

[44] M. Lukosevicius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.

[45] D. Verstraeten, B. Schrauwen, and D. Stroobandt, "Reservoir-based techniques for speech recognition," in *International Joint Conference on Neural Networks (IJCNN)*, 2006.

[46] H. Jaeger, "ESN toolbox for Matlab," 2007. [Online]. Available: http://www.faculty.jacobs-university.de/hjaeger/pubs/ESNtools.zip

[47] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. USA: Springer, 1998.