A neurally inspired musical instrument classification system based upon the sound onset

Michael J. Newton[a)]
School of Music, University of Edinburgh, City of Edinburgh EH9 3JZ, United Kingdom

Leslie S. Smith
Institute of Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, United Kingdom

(Received 13 September 2011; revised 29 March 2012; accepted 6 April 2012)

Physiological evidence suggests that sound onset detection in the auditory system may be performed by specialized neurons as early as the cochlear nucleus. Psychoacoustic evidence shows that the sound onset can be important for the recognition of musical sounds. Here the sound onset is used in isolation to form tone descriptors for a musical instrument classification task. The task involves 2085 isolated musical tones from the McGill dataset across five instrument categories. A neurally inspired tone descriptor is created using a model of the auditory system's response to sound onset. A gammatone filterbank and spiking onset detectors, built from dynamic synapses and leaky integrate-and-fire neurons, create parallel spike trains that emphasize the sound onset. These are coded as a descriptor called the onset fingerprint. Classification uses a time-domain neural network, the echo state network. Reference strategies, based upon mel-frequency cepstral coefficients, evaluated either over the whole tone or only during the sound onset, provide context to the method. Classification success rates for the neurally-inspired method are around 75%. The cepstral methods perform between 73% and 76%. Further testing with tones from the Iowa MIS collection shows that the neurally inspired method is considerably more robust when tested with data from an unrelated dataset.
[http://dx.doi.org/10.1121/1.4707535]

PACS number(s): 43.75.Xz, 43.66.Jh, 43.75.Cd, 43.64.Bt [TRM] Pages: 4785–4798

I. INTRODUCTION

Relating human perception of sound to quantifiable acoustic parameters has driven much of the timbre research over the past few decades. Studies have sought spectral and temporal quantities that show promise from a signal processing perspective and that might be relatable to evidence from psychoacoustic and physiological research. Such studies can be traced back to Helmholtz's[1] suggestion that human timbre perception arises from the instantaneous spectral shape as decoded by the cochlea. Timbre research remains a very active field with many applications such as voice recognition, hearing disorder research, and music information retrieval.

An increasingly common application of timbre research has been to build automatic classifiers that can distinguish between musical instruments using calculable acoustic features. A related and larger branch of research has been automatic speech recognition (ASR). ASR has informed many musical instrument classifiers, most notably through the use of mel-frequency cepstral coefficients (MFCCs) as easily calculable acoustic features with a degree of biologically inspired motivation.[2]

We propose a musical instrument classification system based exclusively upon a neurally inspired description of the sound onset, and we compare it to a more classical system based upon MFCCs. The decision to work with the onset alone is based both upon the physiological evidence of its prominence in the early auditory coding of sound and upon the psychological evidence of its importance for perception. The premise is not that the onset contains all the relevant information for musical tone perception but that a sound onset representation may be useful for musical instrument classification.

A. Descriptions of the sound onset

The nature of the sound onset, for example its duration and spectro-temporal evolution, may be considered either physically or perceptually. Often there is a strong correlation between these viewpoints, but this need not always be the case. For example, relative movement of the listener and sound source may modify the perceived onset without change to the onset's physical production.

The physical sound onset results from the sound generation mechanism. At the start of a pitched trombone note, for example, there is an initial injection of air into the instrument, followed by a short period of time when the player's lips vibrate independently of instrument feedback. The acoustic result of this is a mixture of noise from the initial air pulse, and a periodic waveform from the autonomous lip vibrations. After some time, a steady state is reached where a pitched note due to acoustically reinforced lip vibrations dominates the instrument output.[3] The transition from onset to steady state and/or offset is continuous, so that isolating the physical onset from the rest of the sound requires some calculable metric. In the example, it could be argued that the time between the start of the initial air injection and the

a)Author to whom correspondence should be addressed. Electronic mail: michael.newton@ed.ac.uk

commencement of acoustically reinforced lip vibrations represents the *physical sound onset*. In musical acoustics, this interval is often referred to as the attack transient.[4] Similar mechanisms exist for many of the acoustic instruments.

Some sounds do not fit into such a clear onset-steady state–offset regime. Impulsive sounds, for example, may reach a maximum amplitude almost instantaneously, followed by a rapid decay. Nonetheless, the concept of "onset" remains important, regardless of the nature of the steady state and offsets. This is because the onset always represents a transfer of signal energy from a lower (possibly the noise floor) to a higher level, and there is considerable evidence that the auditory system can extract significant information about the sound source from the nature of this energy change. This leads to the concept of a *perceptual sound onset*, which is broadly defined as a significant increase in signal energy perceived by the sound receptor, which in this case is the cochlea.

The onset is one of the most strongly represented sound features within the early auditory system[5,6] along with amplitude modulation.[7] The auditory nerve responds most intensely at the onset of sustained sounds. Within the cochlear nucleus there are further neurons (octopus and some bushy and stellate cells) that are thought to specifically code the stimulus onset.[8–10] The precise mechanisms that govern these neural encodings, which may include specialized ion channels, neuron leakiness, and synapse quality and/or innervation, remain unclear. It is also unclear exactly how the low level onset coding is used by higher-order parts of the auditory system. There is, however, some evidence that the onset coding may be important for certain sound recognition tasks[11] and that the onset plays an important role in direction finding through representation of interaural time and level differences.[10]

B. Psychoacoustic timbre studies and the sound onset

Numerous psychoacoustic studies have shown that the onset provides an important cue for timbre perception and thus musical instrument identification, particularly in the case of isolated tones. In the mid 1960s, both Saldanha and Corso[12] and Clark *et al.*[13] found that the onset transient was a salient feature for timbre perception. Risset and Mathews[14] also showed that the temporal properties of the onset transient were important for the perception of trumpet tones.

Many other studies have investigated the relationship between timbre perception and acoustical properties of sound. Grey and Moorer[15] and Charboneau[16] presented listeners with original and modified versions of musical tones. Thresholds of timbral discriminability were evaluated that showed the negative effect of smoothing the patterns of spectro-temporal variation within complex tones. The removal of the onset transient also led to particularly high discriminability between the original and modified sounds and by implication suggested its importance in tone perception.

McAdams and Rodet[17] demonstrated that vibrato could be an important cue for certain types of musical sound, while Kendall[18] found that steady-state portions of the sound could sometimes be as important as the onset transients. More recently, McAdams and Bigand[19] concluded in their thorough review of timbre research that it is likely that the onset transient contains the most important cues for identification.

C. Musical instrument classifiers based on timbral considerations

There have been numerous prior attempts to build instrument classifiers, some using onset descriptors alongside others from the steady state. Typically the approach has been to calculate a vector of descriptors for each tone based upon its spectro-temporal evolution. Large numbers of tones are analyzed and used to train and test a classification system such as a neural network.

An early attempt to use neural networks to classify musical sounds can be found in De Poli and Tonella.[20] The concept of a "timbre space" first suggested by Grey[15] was replicated within a neural network, and clustering was used to categorize sounds. A similar approach involving a self-learning neural network was adopted by Cosi *et al.*,[21] capturing tone quality with MFCCs.

Feiten and Günzel[22] used supervised neural network learning to map timbral qualities to human verbal descriptions. Spevak and Polfreman[23] studied the suitability of a range of auditory model sound representations, including MFCCs, as timbral descriptors using neural networks. Both studies used the sound descriptors to build temporal representations of the dynamic sound development. Similar neural networks were used with a nearest neighbor classifier by Kaminskyj and Materka[24] to distinguish between four types of musical instrument.

In a widely cited conference paper, Martin and Kim[25] described 31 acoustic parameters, including the average pitch, the average spectral centroid, and the onset slope. A corpus of 1023 sounds, over five instrument families (classes), were used to build a Gaussian classifier model. The instrument families were identified with about 90% success.

Brown[26] used 18 cepstral coefficients as tone descriptors in a two-class musical instrument classification problem. A *k*-means algorithm was used to cluster training data in combination with a Bayesian decision rule for assigning instrument class. Success rates of around 85% compared favorably with human trials based on a subset of the data. A further study[27] based on the same cepstral tone descriptors also showed promising results.

Herrera *et al.*[28] classified impulsive drum sounds using a variety of acoustic descriptors and classifiers. Particular attention was paid to descriptors that captured details of the onset transient. Success rates between 90% and 99% were reported.

Recently Barbedo and Tzanetakis[29] described a system based upon the mapping between individual partials in a complex sound mixture. The system calculated features based upon key partials and provided a probability for the most likely source instrument. Reports of numerous other musical instrument classifiers can be found, mostly from conference proceedings. See review publications[30,31] for a detailed summary.

This study investigates the suitability of using the sound onset in isolation to form descriptors useful for musical

instrument classification. This was achieved by constructing two competing musical instrument classification systems and testing them with a common task based on sounds from the McGill dataset[32] (Sec. II). Further testing involving sounds from both the McGill and University of Iowa Musical Instrument Samples[33] collection provided an even more challenging task.

The novel classification system was called Strategy A and used a biologically inspired neural-like coding of the perceptual sound onset to form sound descriptors (Sec. III). The key feature was that the descriptors remained as time-domain signals and so required the use of a temporal recurrent neural network, the echo state network, as a classifier. Strategy B was a more classical instrument classifier based upon MFCCs, either over the whole tone or the onset alone (Sec. IV). A standard multilayer perceptron neural network was used as a classifier. This approach was broadly similar to previous systems described by Brown[26,27] and others[34–36] and provided context to the novel strategy. Results and discussion of both classifiers are provided in Sec. V.

II. THE CLASSIFICATION TASKS

The main classification task was based upon a corpus of 2085 tones drawn from the McGill dataset. All were sampled at 44.1 kHz with 24 bit resolution. The tones were split equally across five broadly hierarchical musical instrument families, or classes, over octaves 1–6 (Table I).

The instrument classes were sorted according to the physics of the initial tone generation mechanism. Each class involved a unique, although often related, set of tone generation physics. For example, while both the brass and reed classes involve vibrating air valves, the stiffness-dominated behavior of the reed instruments contrasts with the variable stiffness-mass dominated behavior of the brass (lip-reed) instruments.[3] For the bowed and plucked string classes, although both groups involve the same instruments, the method of playing differs (bowed versus pizzicato). The sound generation physics thus differ, leading to markedly different tone qualities. Broadly similar hierarchical grouping approaches have been adopted by several previous studies.[15,25,37]

The classification task was the same for both Strategies A and B. The dataset was randomly split into training (70%) and test data (30%), a method known as bootstrapping.[38] The training data was used to train the appropriate classifier (different for each strategy), which was then tested with the unseen test data to give success rates for each class, expressed as a confusion matrix. Multiple independent randomizations of the train/test split were computed for each classifier configuration, and a mean overall score calculated. The classification systems were thus tested for generality rather than their ability to simply classify known data. Both strategies are summarized in Table IV.

An additional classification task, described in Sec. V D, was designed to further test the generality of the classification systems. This was achieved by training each strategy using all 2085 sounds from the McGill corpus and testing the trained classifiers using 1000 sounds (200 per class) drawn from the publicly available University of Iowa Musical Instrument Samples collection.[33] This equated to a 67.6%/32.4% train/test split. These new sounds were also of good quality, sampled at 44.1 kHz and 16 bit depth but were obtained under completely different conditions (microphones, recording space, outboard gear, etc.) than the McGill sounds. They thus provided an ideal test of the ability of each strategy to deal with genuinely new data.

III. CLASSIFIER STRATEGY A

A. Biologically inspired tone descriptor based on the onset

For the Strategy A tone classification system, a neural-like coding of the perceptual sound onset[39,40] was used as the tone descriptor. The onset detection technique was based on a simple model of the mammalian auditory system, illustrated in Fig. 1. The cochlea response was modeled with the ubiquitous passive gammatone filterbank[41] [Fig. 1(A) and Sec. III A 1]. The output from each gammatone filter was then spike-encoded to give a low-level simulation of the auditory nerve's (AN) early response to sound stimuli [Fig. 1(B) and Sec. III A 2]. The strong spiking onset response observed by certain neurons within the cochlear nucleus[8–10] was then modeled using an array of leaky integrate-and-fire neurons stimulated by the simulated AN signal [Fig. 1(C) and Sec. III A 3]. Example outputs from each of these processing stages are shown in Fig. 2 and further details are provided in the following text. Finally, the raw onset spikes were coded into "onset fingerprints," a reduced space for use with the classifier [Fig. 1(D) and Sec. III A 4]. The key auditory model parameters are summarized in Table II.

1. Gammatone filtering

The first order response of the basilar membrane was modeled with a 15-channel gammatone filterbank. Channel center frequencies were spaced between 200 Hz and 5 kHz. Using only 15 channels was a clear abstraction from the 3000

TABLE I. Summary of instrument classes used in the classification task. There were 2085 tones in total (417 per class). The mean onset duration interval as detected by the auditory model used by Strategy A (see Sec. III) is shown.

Class label	Class description	Instruments included in class	Mean onset duration (ms)
Bs	Brass	Cornet, trumpet, french horn, trombone, tuba	80
Rd	Reed	Clarinet, bassoon, oboe, saxophone	110
SB	Bowed string	Cello, viola, violin, double bass (bowed)	120
SP	Plucked string	Cello, viola, violin, double bass (pizzicato)	45
SS	Struck string	Piano	46

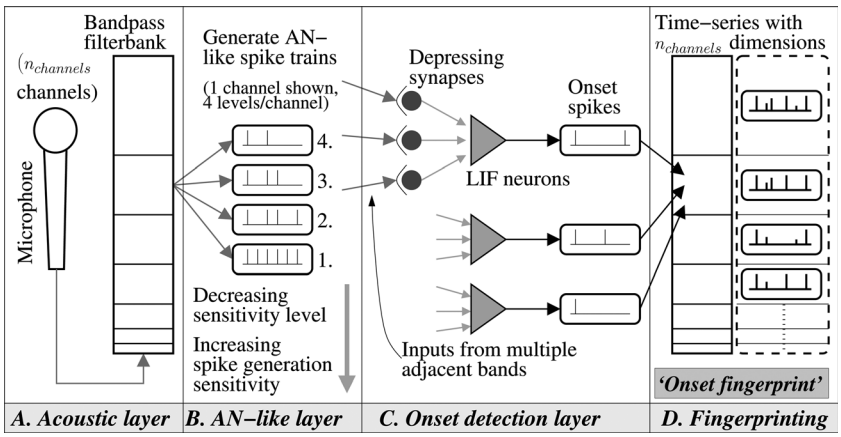

FIG. 1. Schematic of the auditory model used to form tone descriptors for Strategy A. AN spike generation is shown for one channel (of 15) and four sensitivity levels (of 20) and onset neurons/depressing synapses for one sensitivity level (of 20).

or so inner hair cells (IHCs) that make up the cochlear filter. The purpose was to obtain a tone descriptor dimensionality and frequency range that was broadly comparable to the 15 MFCCs used for Strategy B. This allowed a more reasonable comparison between the two methods that was not skewed by a frequency resolution advantage and so better isolated the novel nature of the onset coding of Strategy A.

2. AN-like spike encoding

The outputs from the filterbank channels were coded in a manner inspired by the phase-locked spiking behavior observed in low-to-mid frequency sensitive neurons that innervate the cochlea's IHCs.[42] The output from each channel was encoded as 20 spike trains (sensitivity levels), resulting in 300 spike trains to describe each sound.

Spikes were produced at positive-going zero-crossings of the filtered signals. For each zero-crossing i, the mean signal amplitude during the previous quarter cycle E_i was calculated and compared to the values $S_{j=1:20}$ of 20 sensitivity levels with a difference δ_{levels} of 3 dB between levels. Sensitivity level 1 was the most sensitive level with a low signal amplitude required to produce a spike. If $E_i > S_j$, then a spike was produced at the sensitivity level j. For any spike produced at level k, a spike was necessarily produced at all levels $j < k$. This representation was similar to that employed by Ghitza[43] where it was noted that it led to an improvement in automatic speech recognition in a noisy environment. The use of multiple sensitivity levels per channel allowed both temporal and dynamic level information to be retained.

There was information redundancy due to the parallel nature of the spike coding, but this was necessary for the onset detection system. Redundant spikes were later removed by a reduction to the 15 sensitivity-level-normalized channels of the *onset fingerprint* coding (Sec. III A 4).

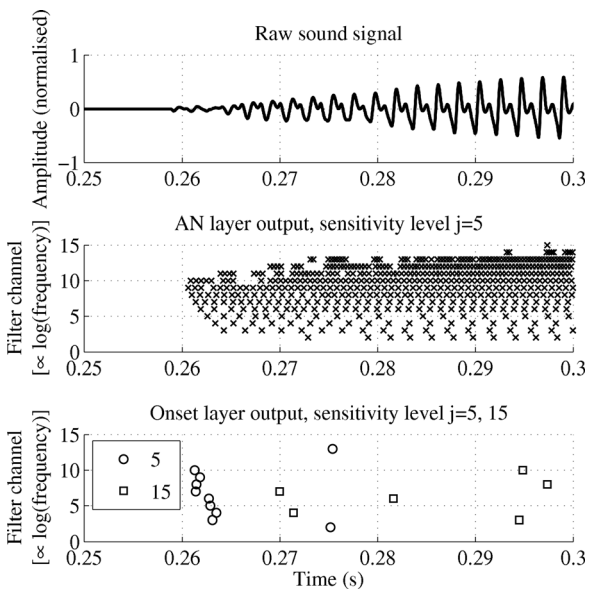

FIG. 2. Example raw sound signal, AN-coded spikes and onset spikes, clustered near the start of the signal, for an isolated trombone tone at sensitivity level 5 (15 also shown for onset spikes). The onset spikes over multiple sensitivity levels are coded into a single 15 channel time-series signal called the onset fingerprint (see Fig. 3, and Sec. III A 4).

3. Onset detection

The AN-like representation in the preceding text does not emphasize onsets in the encoded sound signal unlike the real mammalian AN.[6] However, its parallel coding makes it suitable for use with a secondary onset detection system.[39,40] This system was inspired by the onset response behavior exhibited by octopus, and some bushy and stellate cells,[10] cells within the cochlear nucleus.

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, giving 300 onset neurons in total. Each encoded

TABLE II. Summary of parameter values and variables used in the spiking auditory model and perceptual onset detector used for Strategy A.

Symbol	Description	Value
n_{channels}	Number of filterbank channels	15
n_{levels}	Number of sensitivity levels	20
δ_{levels}	Inter-sensitivity level difference (dB)	3
$S_{j=1}$	Lowest sensitivity level (sampled at 24 bits)	0.002
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

the behavior of a specific spectral and dynamic range of the sound signal during the onset.

The synapse model was based on the three-reservoir model used by Hewitt and Meddis[44] in the context of IHC-to-AN fiber transduction. A similar model has also been used by Tsodyks and Markram[45] 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 the following three first order differential equations

$$\frac{dM}{dt} = \beta R - \gamma M, \quad \frac{dC}{dt} = \gamma M - \alpha C, \quad \frac{dR}{dt} = \alpha C - \beta R \quad (1)$$

where α and β were 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 modeled, and the amount of post-synaptic depolarization was assumed to be directly proportional to the value of 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

$$I_{b,j}(t) = \sum_{h=b-n_{\text{adj}}}^{h=b+n_{\text{adj}}} wC_{h,j}(t) \quad (2)$$

where w was the weight of each synapse (the same for all inputs) 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. An n_{adj} value of 1 was used so that each onset neuron was innervated by three parallel AN channels at the same sensitivity level.

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, sensitivity level j, would receive at least F_b spikes per second (where F_b was the center frequency of the channel). This rate would normally be larger due to contributions from adjacent channels. However, depletion of the available neurotransmitter reservoir M, in conjunction with a slow reservoir recovery rate, meant that an evoked postsynaptic 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 set to ensure that a single EPSP was insufficient to cause the onset neuron to fire. Thus multiple ESPSs from adjacent synapses were required for the onset neuron to fire. The neurons employed were also leaky,[40,46] meaning that the ESPSs needed to be close to concurrent for an action potential, or "onset spike," to be produced. The overall aim was to ensure that onset spikes were only produced by sudden, cross-frequency rises in signal energy.

4. Onset fingerprint coding

It would be possible to use the raw onset spike trains as a time-domain tone descriptor. However, a condensed form was used that reduced the number of inputs, and computational load, to the classifier (see Sec. III B). It also made the coding dimensionality more comparable to the 15 MFCCs used by Strategy B. The 300 onset spike trains, each of which coded a specific frequency channel and signal level over time, were converted into 15 spike trains (one per frequency channel) normalized by the highest sensitivity level.

The single onset feature that corresponded to the start of the musical note was first identified. Certain sounds, such as some bowed instrument (SB) tones, produced secondary onset spikes during the steady state due to large amplitude variations caused by vibrato. Groups of onset spikes separated by more than 25 ms were treated as separate onset events. Only the first onset event grouping was picked out as the tone descriptor.

The onset grouping was further processed to reduce the sample rate and the number of parallel spike trains. To do this, the raw onset signal was time-sliced into 1 ms windows. For each channel, each time-sliced signal portion was examined to find the highest intensity onset spike s_j, and the value of the sensitivity level j used to label the time slice, normalized by the highest possible spike intensity $S_{j=20}$. If no spikes occurred, a zero was recorded. Thus each 1 ms time window of the signal was described by a single 15-element vector, with one element per channel. The signal over all time slices we call the *onset fingerprint*, $\mathcal{T}$. Two examples are shown in Fig. 3.

B. Temporal recurrent reservoir network classifier (echo state network)

To investigate the usefulness of the onset fingerprint coding, a suitable classifier was required. A range of classifiers has been used in previous musical instrument classification systems. An early study by Cosi *et al.*[34] used neural networks with timbral descriptor vectors. Martin and Kim[25] used Gaussian models in combination with Fisher

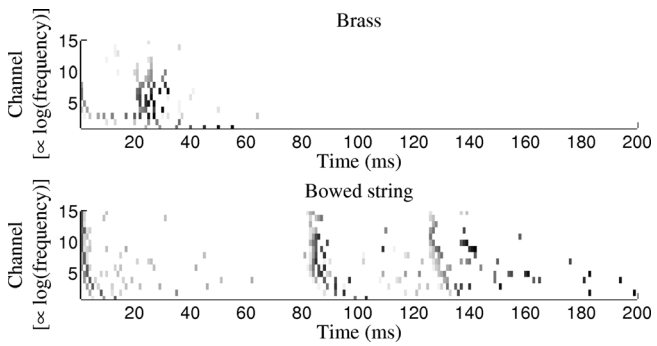

FIG. 3. Example onset fingerprint signals for brass (trombone, 64 ms duration) and bowed string (violin, 200 ms duration) classes. Signal intensity is normalised to the lowest sensitivity level used for the AN spike coding (Sec. III A 2).

multi-discriminant analysis. Brown[26] used a broadly similar method, building a classifier from Gaussian probability density functions acted upon by a Bayesian decision rule. Agostini *et al.*[47] experimented with classifiers built from support vector machines (SVM). Many other approaches can be found in the literature.[30]

The temporal onset coding of Strategy A called for a classifier capable of operating in the time domain. There are a range of tools available for performing such tasks, a number of which have been subsumed under the general category of reservoir computing.

1. Reservoir computing

Reservoir computing represents a general category of recurrent neural networks within which there are a number of related implementations.[48] Jaeger's echo state network[49] was used here. Such networks have commonly been used for time-series prediction. Recently they have also been applied to time-series classification in areas such as speech recognition,[50–52] and it is within this framework that the current application resides.

Reservoir computing networks are related to SVMs, where an input signal is more easily separated by translation to a higher-dimensional space. The basic structure of most reservoir networks is broadly the same with three principal layers as illustrated in Fig. 4.

A large, sparsely interconnected mass of simulated neurons, the reservoir layer, is stimulated by one or more input layer nodes. Each neuron in the reservoir has a nonlinear activation, the most common varieties being sigmoidal and LIF functions. The interconnection weights are randomized at the start of the task (most of them being set to zero) and do not change. Output layer nodes are connected to each reservoir node via a trainable weight. These weights form the network's learning framework.

Reservoir networks are designed to receive time-varying signals. At each time step, the current value of the input signal is projected through the reservoir layer to the output layer. However, the new reservoir neuron activations depend both upon the new input stimulation and upon a (tunable) number of their previous states. The network thus has a memory, which is sensitive to both the amplitude and timing of input stimulations. The principle is that there may exist some characteristic set of reservoir layer activations that better distinguish different types of input signal than can be found in the space of the raw input signals themselves.

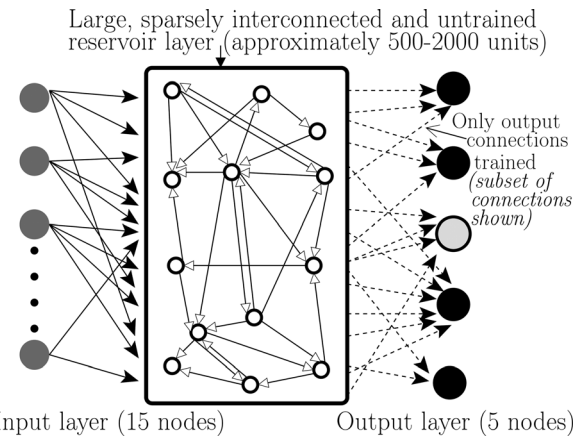

FIG. 4. Schematic of the structure of the echo state network (Ref. 53) used as a classifier for Strategy A. An input layer of 15 nodes (one per onset fingerprint filterbank channel) connects into a large, interconnected and untrained reservoir layer. Only connections from the reservoir layer to the output layer, which has one node per instrument class, are trained (dashed).

The echo state network used here was intended to exploit both the spectral and the temporal information contained within the onset fingerprint of Strategy A. The hypothesis was that the unique generation physics of each instrument class, encoded by the onset fingerprinting, could excite the nonlinear reservoir in some characteristic manner, regardless of pitch or individual instrument. By training the network on 70% of the dataset, the learned connections to the output classification layer would be optimized so that when presented with a new tone (belonging to one of the previously observed classes) the network would reach an appropriate characteristic state. Projection of the reservoir state to the output layer nodes would then indicate the appropriate instrument class.

2. Echo state network setup

The echo state network (ESN) implementation of reservoir computing was used to build a classifier for Strategy A. An open-source MATLAB toolbox released by Jaeger[54] was adapted to suit this application.

A number of parameters and configuration choices were required to configure the ESN. The first configuration choice was the type of neurons to use within the reservoir: Leaky sigmoidal neurons were chosen. There were then five key network parameters to assign, summarized in Table III. Parameter sweeps were performed to explore network classification performance.

The reservoir size determined the size of the space into which the input signal was projected, as well as the number of trainable output weightings. The ESN principle depends on a suitably large reservoir with appropriate temporal dynamics, but it should not be so large as to permit overfitting. The reservoir sizes of between 500 and 2000 units used here were typical of similar implementations in speech recognition.[55] Connections between reservoir neurons were randomly assigned using a sparse weighting matrix, with the connectivity fraction set to $10/R$ where R was the reservoir size.

The spectral radius was a critical parameter which defined the time scale of the reservoir dynamics.[53] Values closer to zero are useful for short sequences, values closer to

TABLE III. Summary of echo state network parameter ranges investigated for Strategy A. The optimal configuration is based on the mean of ten repetitions (see Sec. V A). Parameter explanations in Sec. III B 2.

Parameter	Range explored	Optimal value
Reservoir size	500–2000	1000
Spectral radius	0–1	0.18
Neuron leakage	0–1	0.14
Input scaling	0.5–10	1
Ignored states fraction	0.1–0.9	0.7

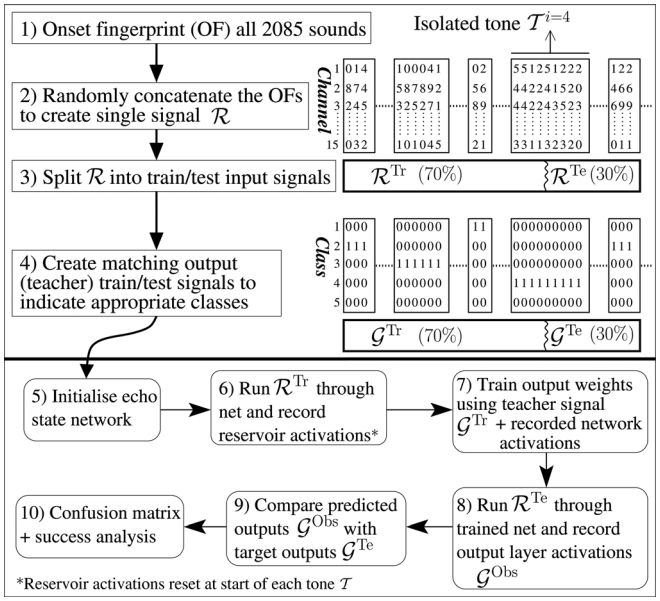

FIG. 5. Flowchart showing the principal steps involved in training and testing an echo state network with onset fingerprints as input signals. The upper half shows the formation of the train/test input/output signals from the individual tones in the dataset (each tone produces an onset fingerprint $\mathcal{T}^i$). The lower half shows the network training and testing routine.

one for tasks requiring longer network memory. A compromise value suitable for both short (e.g., plucked string) and long (e.g., bowed string) onset fingerprints was sought.

The neuron leakage parameter determined the leakiness of the reservoir neurons. The input scaling acted between the input and reservoir layers, either enhancing or reducing the amplitude of input stimulation received by the reservoir. This had the effect of increasing or decreasing the degree of nonlinearity displayed by the reservoir neurons.

The initial portion of each onset fingerprint was disregarded as the reservoir layer required a period of warm-up time to overcome statistical fluctuations caused by starting from zero activation.[53] The ignored states fraction determined this split (extension to the ESN toolbox[54]).

3. ESN training and testing routine

Figure 5 shows a flowchart of the main steps involved in creating, training and testing the ESN for Strategy A. Each sound file was first analyzed to produce its onset fingerprint $\mathcal{T}$. An training input signal $\mathcal{R}^{\mathrm{Tr}}$ was then created by randomly sorting together 70% of the fingerprints. The remaining 30% of the data was used to form the test signal $\mathcal{R}^{\mathrm{Te}}$. Matching output train ($\mathcal{G}^{\mathrm{Tr}}$) and test ($\mathcal{G}^{\mathrm{Te}}$) signals were then formulated that recorded the instrument class of each of the onset fingerprints in the input signals.

For each parameter combination (see Table III), a new ESN was initialized. The training input signal $\mathcal{R}^{\mathrm{Tr}}$ was run through the network and the reservoir activations recorded. At the start of each onset fingerprint within $\mathcal{R}^{\mathrm{Tr}}$, the reservoir layer activations were reset to zero to prevent overlap between network states belonging to consecutive fingerprints (this was an extension to the default ESN toolbox functionality). After the reservoir activations for all training fingerprints were recorded, the weights to the output layer were trained against the target output signal $\mathcal{G}^{\mathrm{Tr}}$.

The test input signal $\mathcal{R}^{\mathrm{Te}}$ was then passed through the trained network. For each onset fingerprint in the test signal, the predicted output signal $\mathcal{G}^{\mathrm{Obs}}$ was compared to the target output signal $\mathcal{G}^{\mathrm{Te}}$. The most commonly predicted class in $\mathcal{G}^{\mathrm{Obs}}$, indexed by the output node with the highest signal amplitude, was taken as the classification decision. This was compared the the actual class stored in $\mathcal{G}^{\mathrm{Te}}$ to deduce the classification success.

The routine was performed independently ten times for each ESN parameter set. This corresponded to 10 different initial reservoir layer randomizations and train/test input signal randomizations. The mean and standard deviation success rates were recorded in a Strategy A confusion matrix $\mathcal{C}_{\mathrm{A}}^{McGill_x}$, where x recorded the network parameters and dataset description. This ensured that the classification results were robust for each network configuration and were not simply a fluke of a particularly well-matched network and dataset.

IV. CLASSIFIER STRATEGY B

A. Classical MFCC-based tone descriptor

A separate musical instrument classification system, Strategy B, was sought for comparison with Strategy A (see Table IV for summary of classification methods for each strategy). The most common tone descriptors used in the literature, in line with speech recognition research, have been MFCCs.[26,27] These describe the spectral content of a tone in a manner inspired by the roughly logarithmic coding used by the cochlea.

The MFCC implementation used for Strategy B is summarized in Fig. 6. The onset portion of the audio signal,

TABLE IV. Summary of tone descriptor and classification methods used by Strategies A and B (details in Secs. III and IV).

Method	Tone descriptor	Classifier technique	Key features of the method
A	Simulated neural onset coding (the "onset fingerprint") with 15 signal channels, 0.2–5 kHz, 20–60 time steps per channel, (determined by onset duration).	Time-domain recurrent neural network [echo state network, using open source MATLAB toolbox by Jaeger (Ref. 54), customized for this application].	Tone descriptor captures spectral and temporal information during onset. Neurally inspired classifier works in time domain and allows tone descriptor to retain spectral and timing information.
B-1	15 MFCCs, 133–6854 Hz, single descriptor vector per tone.	Multilayer perceptron neural network [open source WEKA toolbox (Ref. 58)].	Mean MFCCs evaluated over whole signal. Classifier is non-temporal
B-2	15 MFCCs, 133-6854 Hz, single descriptor vector per tone.	Multilayer perceptron neural network [open source WEKA toolbox (Ref. 58)].	Mean MFCCs evaluated during onset only. Classifier is non-temporal.

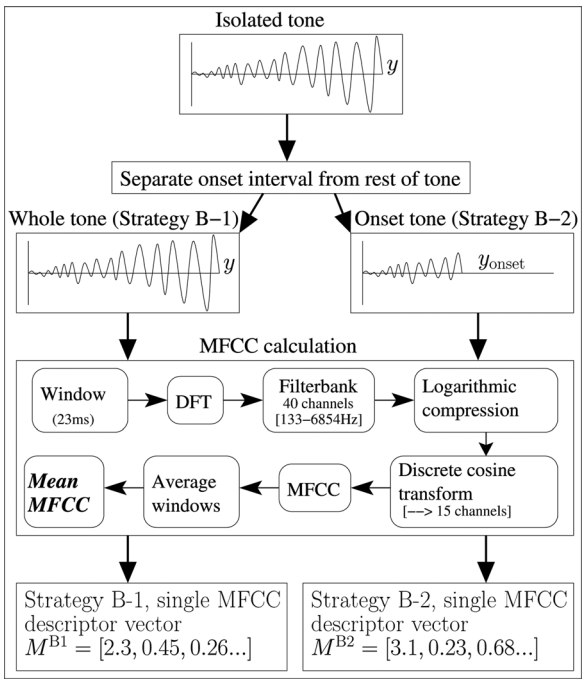

FIG. 6. Flowchart showing the calculations used to form the 15 element MFCC descriptor vectors required for the two different versions of Strategy B.

y_{onset} was first identified from the overall signal, y. The onset timing determined by the auditory model used for Strategy A (see Sec. III A) was used to set the timing of the onset duration. The whole tone and the onset section were then processed separately to produce two alternative MFCC descriptor vectors, M^{B1} (used in Strategy B-1) and M^{B2} (used in Strategy B-2).

The MFCC calculations were based on the formulation of Slaney[56] and were performed independently for each signal portion/sub-strategy (y or y_{onset}). The signal was first Hamming windowed into 23 ms chunks with an inter-chunk overlap of 11.5 ms. Windowed signals with a mean amplitude of less than 0.2% of the peak were ignored (16 000 units at 24 bits). For each remaining windowed signal portion the discrete Fourier transform was calculated, the output of which was passed through the filterbank array. There were 40 filters, the lower 13 of which were linearly spaced at 66 Hz intervals, starting at 133 Hz. The upper 27 filters, all located above 1 kHz with an upper limit at 6854 Hz, were logarithmically spaced at interval factors of approximately 1.07. The filterbank outputs were logarithmically compressed to give set of parallel signals X_i where i indexed the filterbank channel number. Finally, the signals X_i were passed through a discrete cosine transform to reduce dimensionality and provide the vector of MFCCs M_j for the windowed signal segment. The cosine transform was computed as

$$M_j = \sum_{i=1}^{C} X_i \cdot \cos\left[j \cdot \left(i - \frac{1}{2}\right) \cdot \frac{\pi}{C}\right], \quad \text{for } j = 1, 2, 3, \ldots, J \tag{3}$$

where specifying $J = 15$ resulted in 15 MFCCs, and $C = 40$ was the total number of filterbank channels.

The sound signal (y or y_{onset}) was thus described by an array of 15 MFCC vectors with each vector calculated from 23 ms of the signal. The final step was to reduce this down to a single, mean MFCC vector that represented the average distribution of MFCCs over the whole tone (M^{B1}) or over the onset section of the tone (M^{B2}). This averaging was somewhat less subtle than the clustering method used by, for example, Brown,[26] but was justified as a straightforward approach suitable for forming a functional comparison system.

Two alternative sub-strategies were specified for two reasons. First, it was important to have a descriptor vector formed in a manner reasonably comparable to previous studies,[26,27,34–36] in this case M^{B1}. This provided a broadly standardized classification score for the dataset. Second, an MFCC based descriptor vector M^{B2} was sought that was more directly comparable to the onset fingerprint coding of Strategy A. The overall aim was to ensure that the dataset was thoroughly explored with a variety of methods to provide a clear context for the novel contribution of Strategy A.

B. Multilayer perceptron neural network classifier

The two varieties of Strategy B used a multilayer perceptron neural network[38] (MLP) as a classifier. This classifier has been used in many machine learning tasks, including for musical instrument classification.[57] The MLP implementation used the open-source WEKA toolbox.[58] The processing and classification methodology was the same for both strategies but was performed independently to produce two alternative MFCC-based classification systems, B-1 and B-2 (see Sec. IV A).

For each strategy, the sounds from the corpus of 2085 isolated tones were first analyzed to produce feature vectors of 15 MFCCs. These vectors were randomly sorted into a single large dataset array $M^{\text{B1/B2}}_{i=1:2085}$ for each strategy, together with a note of the corresponding instrument class of each entry. The dataset arrays were then split into train (70%) and test (30%) dataset arrays, and passed to the MLP classifier.

A range of MLP sizes (10–1000 units), configurations (1–5 layers), and training rates was explored with parameter sweeps. It was found that a single layer arrangement with 100 neuron (hidden) units was generally optimal for both versions of Strategy B. An MLP with only 50 units performed approximately 3% below this level.

Ten differently randomized train/test splits were thus run through a 100 unit MLP for each version of Strategy B, and mean and standard deviation success rates calculated. These provided optimal mean confusion matrices for each strategy, $\mathcal{C}^{\text{McGill}}_{\text{B1}}$ and $\mathcal{C}^{\text{McGill}}_{\text{B2}}$.

V. RESULTS AND DISCUSSION

A. Configuration of the ESN for Strategy A

Parameter sweeps are essential to determine the most suitable network configuration for an ESN and dataset.[55] The explored parameter space of the ESN is summarized in Table III.

It was found that a reservoir size of 1000 units provided optimal results for test data. Larger reservoir sizes

provided increased performance of up to 100% on training data without increased success on test data. This was likely due to overfitting caused by the higher learning capacity of such large networks. The smaller 500 unit network performed approximately 3% below the level of the 1000 unit system on test data.

The input scaling parameter is known to be quite robust,[53] and indeed changing its value did not greatly affect the classification rate. Optimal performance occurred at a value of 1, that is, no additional input signal amplification was applied. With the other parameters set optimally, classification performance decreased by a maximum of approximately 6% over the range of input scaling values.

With the reservoir size and input scaling optimally set, the key parameters were the spectral radius and reservoir neuron leakage values. Both parameters were explored between normalized values of 0 (short time scale, no neuron leakage) and 1 (long time scale, large neuron leakage), in increments of 0.01.

The spectral radius had to be tuned so that the network memory time scale was suitable for the duration of a typical onset fingerprint (approximately 50 time steps). The neuron leakage affected the time scale of the individual neuron dynamics and, by implication, the relative importance of the input stimulation timing. Together these parameters controlled the temporal properties of the network and thus its suitability for onset fingerprints.

Figure 7 summarizes the variation of classification success as a function of the two timing parameters. Mean classification success rate, averaged over 10 repetitions, is plotted against spectral radius with multiple lines to show different values of the neuron leakage. For both parameters, the best results occurred between 0.1 and 0.3 with the training data (dashed lines) quite robust around these values. A maximum 75% mean success rate occurred on test data at values of 0.18 and 0.14, respectively, with a standard deviation of 1.8% between trial repetitions. This optimal network configuration was used to produce a mean (over 10 trial repetitions) confusion matrix $\mathcal{C}_{\mathrm{A}}^{\mathrm{McGill}}$ to describe the best performance of Strategy A.

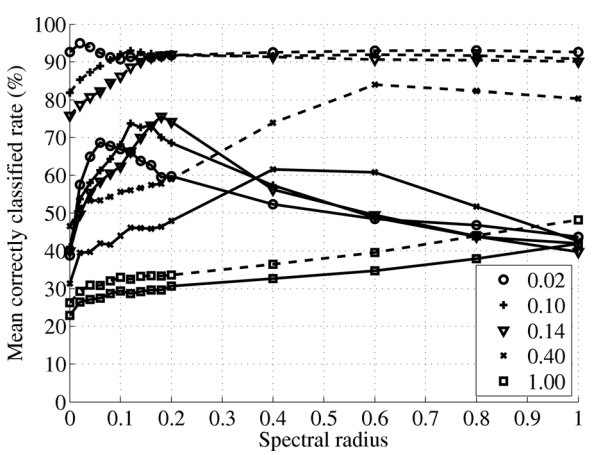

FIG. 7. Plot of the mean correct classification rate against spectral radius of the reservoir layer for multiple reservoir neuron leakage values and a reservoir size of 1000 units (Strategy A). Test data are solid, train data are dashed. Data are the mean of 10 repetitions with the same network parameters. The optimal test data configuration is listed in Table III.

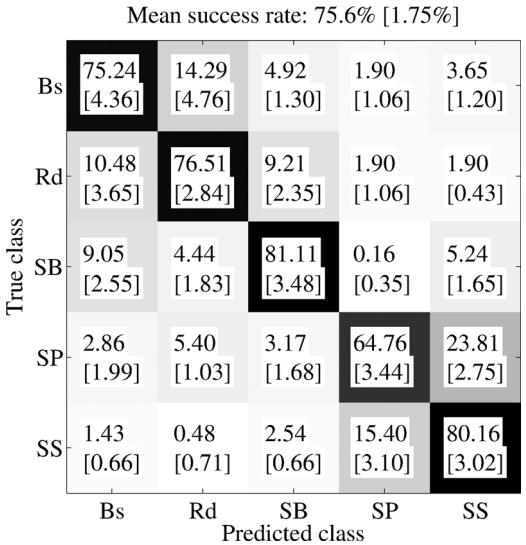

FIG. 8. Normalized optimal confusion matrix $\mathcal{C}_{\mathrm{A}}^{\mathrm{McGill}}$ for Strategy A based on 10 trials with different initial network and data randomizations. All data from the McGill dataset. Standard deviation over 10 trials shown in brackets.

B. Comparison between Strategies A and B

Confusion matrices which describe the performance of Strategies A and B-1 and B-2 are shown in Figs. 8, 9, and 10, respectively. Such matrices allow visualization of not only the overall classification but also the confusion between classes. The scores have been scaled to percentages, and standard deviations over multiple classification trials are included in brackets.

1. Onset fingerprinting vs whole tone MFCCs

Strategy A used a tone descriptor derived from a simulation of the perceptual sound onset. This onset fingerprint

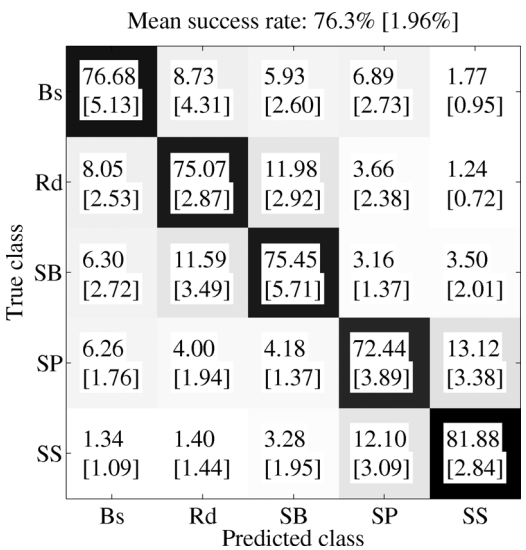

FIG. 9. Normalised optimal confusion matrix $\mathcal{C}_{\mathrm{B1}}^{\mathrm{McGill}}$ for Strategy B-1 (trials and data randomizations as for Fig. 8).

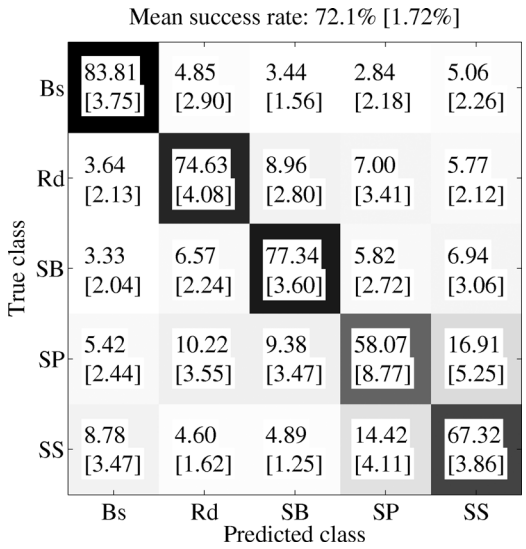

FIG. 10. Normalized optimal confusion matrix $\mathcal{C}_{B2}^{McGill}$ for Strategy B-2 (trials and data randomizations as for Fig. 8).

encoded the timing and intensity of the spectral energy changes during the onset transient. Strategy B-1 was based upon mean MFCCs evaluated over the whole tone and was inspired by a number of previous musical instrument classifiers.[21,26,30] It formed a reduced-space representation of the average timbre of the tone.

The maximum mean classifier performance of Strategy B-1 (76.4%) was marginally higher than that of Strategy A (75%). However, taking into account trial repetitions, this difference was not statistically significant at a confidence level of 5% ($P = 0.12$).

2. Onset fingerprinting vs onset-only MFCCs

Strategy B-2 was based upon mean MFCCs calculated during the onset only. Its tone descriptors thus captured the mean timbre of the onset transient, regardless of the spectral energy change timing. Its performance was approximately 3% below that of Strategy A. This difference was statistically significant over the multiple trial repetitions at a confidence level of 5% ($P = 0.002$).

3. Whole tone MFCCs vs onset-only MFCCs

Comparing the results of Strategies B-1 and B-2 it is clear that, overall, the whole tone version performed slightly better. This is perhaps to be expected as the mean MFCC vectors for Strategy B-1 included more information about the tones. The principle reason for its superior performance was better discrimination between the plucked string (SP) and struck string (SS) classes. It is interesting to note that Strategy B-2 in fact performed slightly above Strategy B-1 for the bowed string (SB) and brass (Bs) classes, but this was within the range of error between the trial repetitions.

4. Analysis and discussion

For all strategies, the most common confusion was between the SP and SS classes. This confusion was highest, 33% of all errors, for the onset-only based technique of Strategy A. It was 21% for Strategy B-1 and 23% for Strategy B-2. Given N percent correct classifications made, the chance rate for confusion pairs such as SP-SS was $((100 - N)/4) \times 2$. This was approximately 12.5% for Strategy A, meaning that the actual SP-SS confusion was almost three times the expected chance rate.

The significant SP-SS confusion, particularly for Strategies A and B-2, can be attributed to the close similarity in the tonal excitation mechanism. For both classes, a tensioned string was impulsively brought into vibration. This represents the most similar pair of excitation mechanisms for the instrument classes studied here, evidenced by the similarity in mean onset duration detected by the auditory model, shown in Table I. The result suggests that discrimination between these classes is significantly aided by tonal information after the sound onset.

The second most common confusion for Strategy A was between the brass (Bs) and reed (Rd) classes. This accounted for 17% of all errors. Once again, this confusion likely reflected the broad similarity in excitation mechanism between these classes, whereby an air valve was brought into periodic vibration by interaction with an airflow and instrument bore.

It was interesting to note that the two most common confusions for Strategy A accounted for over 50% of all errors, more than four times the expected chance rate. Recalculation of the Strategy A performance by consecutively excluding these confusion errors increased the classification success rate to 83.3% and 88.3%, respectively. For Strategy B-1, the corresponding calculations produced success rates of 81.3% and 86.0%, respectively. For Strategy B-2, the results were 78.6% and 81.7%. These modified values reveal the extent to which the overall success rate of Strategy A was impeded by its relatively poor performance in distinguishing between tones with very similar excitation mechanisms. However, they also show that such an onset-based technique, based upon only 2%–10% of the whole tone, performs rather well as a classifier for instrument families where the tone generation mechanisms are more distinct.

It should be further noted that the design of the auditory model for Strategy A used only 15 filters, a rather coarse frequency resolution, to produce a tone descriptor dimensionality broadly comparable with the 15 MFCCs of Strategies B-1 and B-2. It is possible that a larger filterbank and finer-resolution onset coding may have captured some of the more subtle differences between very similar class pairs such as SP-SS.

C. Results compared to other studies

Results reported by Martin and Kim[25] attained classification rates of around 90% for five instrument families. Their tone descriptors involved numerous spectro-temporal features not captured by any of the techniques presented here, and the hierarchy of the instrument classes did not distinguish between plucked and struck string instruments.

In this work as in previous studies, MFCCs have proven to be a rather robust technique for capturing salient features from musical sounds, and the MLP useful as a classifier.

Brown's study[26] based on MFCCs examined a two class problem and achieved success rates of around 85%. A more sophisticated MFCC processing routine than the overall mean calculation used for Strategies B-1 and B-2 was employed. This identified the most useful MFCC time slices to use for classification. Considering the larger number of instrument classes in this study, the MFCC success rates between 73% and 76% suggest that the technique was well-suited to the dataset.

Despite the relative success of tone descriptor techniques such as onset fingerprints and MFCCs, care must be taken when making comparisons between them and the actual function of the human auditory system in perceiving and classifying sounds. In particular it is not reasonable to draw a direct comparison between the design and implementation of Strategy A and the real neural mechanisms involved in processing sound onsets.

Rather, it has been shown that the use of a tone descriptor broadly based on the neural processing of the sound onset can capture sufficient information for use in a successful instrument classifier. While such a description is not literal, it is closer to the underlying physiology than more conventional tone descriptors based on standard signal processing metrics and so provides a viable alternative framework. As with the auditory system, this framework is readily expandable to include other tonal features.

The time-domain approach of the ESN classifier is also somewhat closer in design and function to some neural circuitry[59] than standard techniques such as the MLP and Gaussian classifiers. The results show that a combination of these systems can be used to create a classifier capable of performing at least as well as the more established methods such as MFCCs. In so doing, they further demonstrate the capacity of the onset transient to encode useful information about the sound source.

D. Further strategy testing with an alternative dataset

1. Salient features across different datasets

In machine learning tasks involving multiple degrees of freedom, it can be difficult to determine which aspects of the given input signal coding are most salient for a particular classification result. One of the most dangerous pitfalls in this regard is the possibility that the learning algorithm may latch on to unexpected, and sometimes persistent, features contained in the dataset. In the worst case, such features may be coded according to class so that the algorithm may appear to learn a dataset very well. However, when presented with data from a different dataset that does not contain such unseen but persistent features, the classification success will likely suffer considerably.

Any consistent factor that affects the dataset could cause such an effect. In the present study, such factors would likely be related to the nature of the original sound recording. They could include, for example, the imprinting of a particular frequency response on the recorded sound due to the characteristics of the microphones, recording environments, outboard gear, or even some factor relating to the particular set of instruments used in the McGill corpus. Livshin and Rodet[60] have previously drawn attention to this problem for the case of sound classification.

We therefore further tested the classifiers built in this study with data obtained under completely different environmental conditions. The extensive and publicly available University of Iowa Musical Instrument Samples[33] corpus formed an ideal dataset for this purpose.

From the Iowa dataset 1000 new sounds, split evenly over the five instrument classes, were obtained and processed exactly like the McGill sounds for Strategies A and B-1. Only version B-1 of Strategy B was considered as it had proven the most reliable during the main classification task. To provide the strategies with the greatest challenge, only the McGill data were used for training the classifiers with the new Iowa data exclusively forming the test set. These conditions meant 2085 training sounds (67.6% of the new combined dataset) and 1000 testing sounds (32.4%). It would thus be highly unlikely that a given classifier score was attributable to unforeseen but salient features unique to the McGill dataset.

The two strategies were optimized using parameter sweeps as for the main classification task (see Secs. III B 2 and IV B). Optimal confusion matrices were obtained for each strategy, $\mathcal{C}_{\mathrm{A}}^{\mathrm{McGill/Iowa}}$ and $\mathcal{C}_{\mathrm{B1}}^{\mathrm{McGill/Iowa}}$, each of which was the mean result of 10 classification trials using the same network parameters but different initial network randomizations. These matrices are shown in Figs. 11 and 12, respectively. It is important to note that fixing the training and testing data as described meant that the dataset could not be randomly sorted for each trial as for the main task. Thus the variation in classifier performance between trial repetitions was relatable only to the different initial network randomizations. A summary of the results for the various strategies and datasets, both McGill and McGill/Iowa, used in the study is presented in Table V.

2. Onset fingerprinting vs whole tone MFCCs tested with the Iowa data

The key result from the additional testing was that the onset fingerprinting and ESN classifier approach of

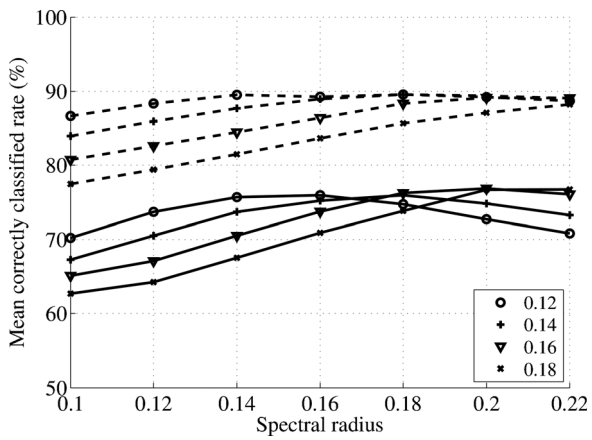

FIG. 11. Normalized optimal confusion matrix $\mathcal{C}_{\mathrm{A}}^{\mathrm{McGill/Iowa}}$ for Strategy A based on training with all 2085 McGill sounds from the main task outlined in Sec. II, and testing with 1000 new and unseen sounds from the University of Iowa collection. Figure shows the mean (standard deviation in brackets) of ten repetitions with different initial network randomizations.

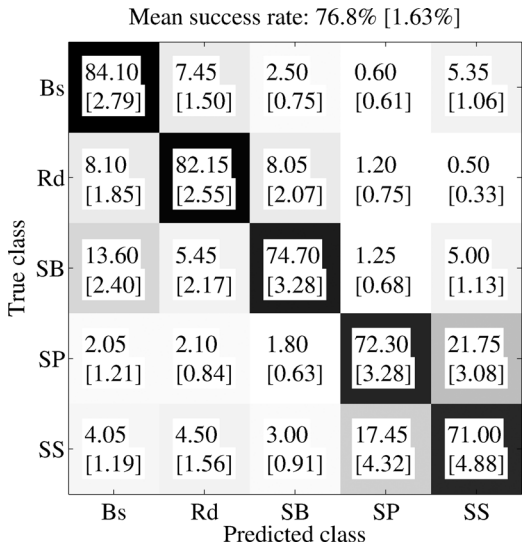

FIG. 12. Normalized optimal confusion matrix $C_{B1}^{McGill/Iowa}$ for Strategy B-1 based on the same data split as Fig. 11. MLP network parameters were the same as for Figs. 9 and 10. Figure shows the mean (standard deviation in brackets) of 10 repetitions with different initial network randomizations.

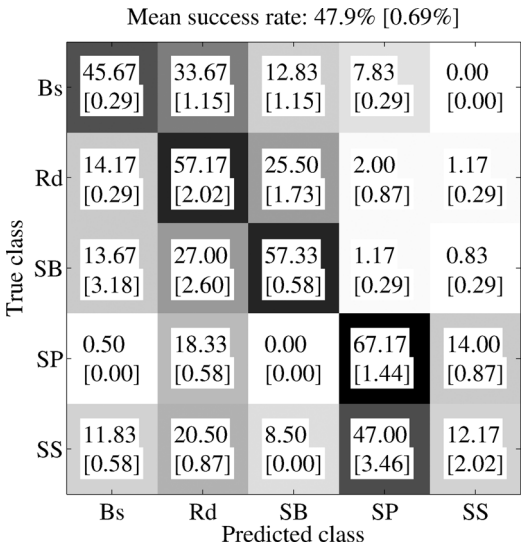

FIG. 13. Plot of the mean correct classification rate against spectral radius of the reservoir layer for multiple reservoir neuron leakage values and a reservoir size of 1000 units (Strategy A). Train data exclusively from the McGill dataset (dashed), test data exclusively from the Iowa dataset (solid). Data are the mean of 10 repetitions with the same network parameters.

Strategy A was much more robust when presented with the new and unlearned Iowa data than the MFCC and MLP-based approach of Strategy B-1. The overall performance of Strategy A was in fact slightly higher than during the main testing with the McGill data. This could be attributed to the 30% increase in the quantity of training data from 1460 to 2085 sounds. Not evident from the confusion matrix is the peak individual performance, which was 80.4%.

Conversely, Strategy B-1 shows a dramatic drop in performance, appearing to be much poorer at generalizing to the new data. It should be noted that increasing the size of the MLP used for Strategy B-1 beyond the established 100 neurons in a single layer, as used in the main task, did not greatly improve the performance with the Iowa data. Trials involving up to three layers, each with 100 neurons, and with a resulting increase in the network's number of degrees of freedom by a factor of 10^5, did not increase the overall score beyond 50%. However, when the standard 100-unit MLP was trained and tested on a randomized 70%/30% mixture of both datasets, it was able to recover back to around 78% overall success.

3. Analysis and discussion

Figure 13, which shows the classification performance as a function of reservoir neuron leakage and spectral radius, provides an important insight into the behavior of the ESN classifier of Strategy A when trained and tested with the McGill and Iowa data, respectively. The optimal performance was achieved, as with the main classification task, using a reservoir size of 1000 units with no further performance increase obtained from using larger reservoir sizes. Apparent from the figure is the similarity in the peak classifier performance as a function of the crucial network parameters, spectral radius, and neuron leakage, when compared to the results obtained in the main classification task (see Sec. V A and Fig. 7). The absolute difference between the optimal values for each of these parameters was low (a few hundredths of a unit) relative to the broad range over which the scores for both classification tasks were above 70% (more than a tenth of a unit). The broadly consistent values of the optimized network parameters, together with the consistently high classification score with the Iowa data, suggest a reasonable degree of underlying robustness in the technique.

As with the main task, it is possible that further improvement could be gained from using a finer-resolution (in both time and frequency) onset fingerprint coding. This would probably require a corresponding increase to the network size and a further dramatic increase in computational load to take advantage of the increased number of degrees of freedom in the input signal. It is emphasized that the degree to which such an improvement might be achieved is outside the scope of the current paper, where the fundamental principle of the technique is the point at stake.

TABLE V. Summary of the classification performance of all strategies and train/test data combinations. Standard deviations over 10 trial repetitions with the same network configuration, but different initial randomizations, are shown in brackets.

Strategy	Training data	Testing data	Train/test split (%)	Score (%)
A	McGill	McGill	70/30	75.6 [1.75]
B-1	McGill	McGill	70/30	76.3 [1.96]
B-2	McGill	McGill	70/30	72.1 [1.72]
A	McGill	Iowa	67.6/32.4	76.8 [1.63]
B-1	McGill	Iowa	67.6/32.4	47.9 [0.69]

VI. CONCLUSIONS

The aim of this study was to explore the usefulness of a neurally inspired representation of the sound onset for musical tone classification. This was achieved by constructing a musical instrument classification system based upon an auditory model of the perceptual sound onset. A time domain neural network, the ESN, acted as classifier. The system was

trained and tested using 2085 tones drawn from the McGill dataset. Within trial repetition error this system, Strategy A, performed as successfully (75% mean success rate) as a more conventional system, Strategy B-1 (76.4%), based upon mean MFCCs evaluated over the whole tone and classified with a multilayer perceptron. The key feature of Strategy A was that the tone descriptor was derived from the onset transient alone, an interval which lasted for 2–10% of a typical isolated musical tone. A further strategy, B-2, based upon MFCCs evaluated only during the onset transient, performed slightly more poorly than either Strategy A or B-1.

Further testing of the strategies was carried out with tones obtained from the University of Iowa Musical Instrument Samples collection. This provided a more rigorous test of the classifier performance as there was no chance that the resulting classification score was only attainable with the particularly high quality McGill dataset. It was important to evaluate this as any classification system can be susceptible to persistent, unexpected features buried within a particular dataset. The results of this testing showed that Strategy A performed significantly better (76.8%) than the MFCC based method of Strategy B-1 (47.9%). It did so with optimized network parameters that were almost identical to those required for the main classification task based exclusively on the McGill dataset. The classification success rate of Strategy A did not increase by using a larger neural network, suggesting that the neurally inspired approach adopted was robust within the limits of the onset fingerprint coding resolution used in the study. This result is broadly in line with other neurally inspired signal processing systems used as speech front ends that tend to be rather robust to noisy input data. It thus provides a useful framework for tone description and classification.

There remain numerous possible further directions of study for the technique. An obvious enhancement of the onset fingerprinting method would be to include information about the steady state and offset. Numerous improvements could also be made to the MFCC based approaches, such as incorporating temporal information by retaining the original 23 ms MFCC time slices and using them directly with a classifier like the ESN. This approach was not tested in the current work as the method of combining a single MFCC vector and classifier was closer to what has appeared previously in the literature. It would also be interesting to test the classifier systems on deliberately, and perhaps extremely, noisy data.

While an ultimate goal could be to fully replicate the human sound processing and perception system, this is not within the scope of the current work. Rather we have simply sought to explore the use of a neurally inspired signal processing technique for the practical application of musical tone classification. The work has demonstrated that, as has been shown in the literature through psychoacoustic and physiological evidence, the onset can be a very useful cue for musical sound identification.

ACKNOWLEDGMENTS

The authors thank Herbert Jaeger and Kevin Swingler for useful discussions about the classifier configurations and the two anonymous reviewers for helpful comments about an earlier version of the paper. This work was funded by the Engineering and Physical Sciences Research Council, UK. Grant EP/G062609/1.

[1]H. L. F. Helmholtz, *On the Sensations of Tone*, 2nd ed. (Dover, New York, 1954), pp. 129–151.
[2]L. R. Rabiner and B. H. Huang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ, 1993), Chap. 3.
[3]D. M. Campbell, "Nonlinear dynamics of musical reed and brass wind instruments," Contemp. Phys. **40**, 415–431 (1999).
[4]N. H. Fletcher, "The nonlinear physics of musical instruments," Rep. Prog. Phys. **62**, 723–761 (1999).
[5]R. L. Smith, "Adaptation, saturation and physiological masking in single auditory nerve fibers," J. Acoust. Soc. Am. **65**, 166–178 (1979).
[6]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. **88**, 857–864 (1990).
[7]W. S. Rhode and S. Greenberg, "Encoding of amplitude modulation in the cochlear nucleus of the cat," J. Neurophysiol. **71**, 1797–1825 (1994).
[8]C. Darwin, "Speech perception," in *The Oxford Handbook of Auditory Science: Hearing*, edited by D. R. Moore (Oxford University Press, Oxford, UK, 2010), Vol. 3, Chap. 9, pp. 207–230.
[9]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 Commun. **41**, 135–149 (2003).
[10]E. Rouiller, "Functional organization of the auditory pathways," in *The Central Auditory System*, edited by G. Ehret and R. Romand (Oxford University Press, Oxford, UK, 1997), Chap. 1.
[11]M. A. Akeroyd and L. R. Bernstein, "The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses," J. Acoust. Soc. Am. **110**, 2516–2526 (2001).
[12]E. L. Saldanha and J. F. Corso, "Timbre cues and the identification of musical instruments," J. Acoust. Soc. Am. **36**, 2021–2026 (1964).
[13]M. Clark, P. T. Robertson, and D. Luce, "A preliminary experiment on the perceptual basis for musical instrument families," J. Audio Eng. Soc. **12**, 199–203 (1964).
[14]J.-C. Risset and M. Mathews, "Analysis of musical-instrument tones," Phys. Today **22**, 32–40 (1969).
[15]J. M. Grey, "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**, 1270–1277 (1977).
[16]G. R. Charbonneau, "Timbre and the perceptual effects of three types of data reduction," Comput. Music J. **5**, 10–19 (1981).
[17]S. McAdams and X. Rodet, "The role of FM-induced AM in dynamic spectral profile analysis," in *Basic Issues in Hearing*, edited by H. Duifhuis, J. W. Horst, and H. P. Wit (Academic, London, 1988), pp. 359–369.
[18]R. A. Kendall, "The role of acoustic signal partitions in listener categorization of musical phrases," Music Percept. **4**, 185–214 (1986).
[19]S. McAdams and E. Bigand, "Recognition of auditory sound sources and events," in *Thinking in Sound: The Cognitive Psychology of Human Audition* (Oxford University Press, Oxford, UK, 1993), pp. 146–195.
[20]G. De Poli and P. Tonella, "Self-organizing neural networks and Grey's timbre space," in *Proceedings of the International Computer Music Conference (ICMC)* (University of Michigan Library, Ann Arbor, MI, 1993), pp. 441–444.
[21]P. Cosi, G. De Poli, and G. Lauzzana, "Auditory modelling and self organising neural networks for timbre classification," J. New Music Res. **23**, 71–98 (1994).
[22]B. Feiten and S. Günzel, "Automatic indexing of a sound database using self-organizing neural nets," Comput. Music J. **18**, 53–65 (1994).
[23]C. Spevak and R. Polfreman, "Analyzing auditory representations for sound classification with self-organizing neural networks," in *Proceedings of COST G-6 Conference of Digital Audio Effects* (DAFX-00) (Universita degli Studi di Verona, Verona, Italy, 2000), pp. 119–124.
[24]I. Kaminskyj and A. Materka, "Automatic source identification of monophonic musical instrument sounds," in *IEEE International Conference on Neural Networks* (IEEE, Washington, DC, 1995), Vol. 1, pp. 189–194.
[25]K. D. Martin and Y. E. Kim, "Musical instrument identification: A pattern-recognition approach," J. Acoust. Soc. Am. **104**, 1768 (1998).
[26]J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," J. Acoust. Soc. Am. **105**, 1933–1941 (1999).
[27]J. C. Brown, "Feature dependence in the automatic identification of musical woodwind instruments," J. Acoust. Soc. Am. **109**, 1064–1072 (2001).

[28]P. Herrera, A. Yeterian, R. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," in *Proceedings of the 2nd International Conference on Music and Artificial Intelligence* (Springer-Verlag, London, 2002), pp. 69–80.
[29]J. Barbedo and G. Tzanetakis, "Musical instrument classification using individual partials," IEEE Trans. Audio, Speech, Lang. Process. **19**, 111–122 (2011).
[30]P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," J. New Music Res. **32**, 3–21 (2003).
[31]P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal Processing Methods for Music Transcription*, edited by A. Klapuri and M. Davy (Springer, New York, 2006), p. 34–54.
[32]F. Opolko and J. Wapnick, *The McGill University Master Samples Collection on DVD (3 DVDs)*, McGill University, Montreal, Quebec, Canada (2006).
[33]The University of Iowa Musical Instrument Samples at http//:theremin.-music.uiowa.edu/ (Last viewed 3/24/2012).
[34]P. Cosi, G. De Poli, and P. Prandoni, "Timbre characterization with mel-cepstrum and neural nets," in *Proceedings of the International Computer Music Conference (ICMC)* (University of Michigan Library, Ann Arbor, MI, 1994), pp. 42–45.
[35]G. De Poli and P. Prandoni, "Sonological models for timbre characterization," J. New Music Res. **26**, 170–197 (1997).
[36]A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Washington, DC, 2000), Vol. 2, pp. 753–756.
[37]S. McAdams, J. W. Beauchamp, and S. Meneguzzi, "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," J. Acoust. Soc. Am. **105**, 882–897 (1999).
[38]S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice-Hall, Englewood Cliffs, NJ, 1998), pp. 156–255.
[39]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 Trans. Audio, Speech, Lang. Process. **15**, 2278–2286 (2007).
[40]L. S. Smith and D. S. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses," IEEE Trans. Neural Netw. **15**, 1125–1134 (2004).
[41]R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Technical Report No. 2341, Applied Psychology Unit (APU)*, Cambridge (1988).
[42]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," Hear. Res. **24**, 1–15 (1986).
[43]O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Comput. Speech Lang. **1**, 109–130 (1986).
[44]M. J. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," J. Acoust. Soc. Am. **90**, 904–917 (1991).
[45]M. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," Proc. Nat. Acad. Sci. USA **94**, 719–723 (1997).
[46]C. Koch, "Simplified models of individual neurons," in *Biophysics of Computation* (Oxford University Press, Oxford, UK, 1999), Chap. 14, pp. 324–331.
[47]G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," in *Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing* (IEEE, Washington, DC, 2001), pp. 97–102.
[48]M. Lukosevicius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," Comput. Sci. Rev. **3**, 127–149 (2009).
[49]H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless telecommunication," Science **304**, 78–80 (2004).
[50]D. Verstraeten, B. Schrauwen, and D. Stroobandt, "Reservoir-based techniques for speech recognition," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (IEEE, Washington, DC, 2006), pp. 106–112.
[51]M. H. Tonga, A. D. Bicketta, E. M. Christiansen, and G. W. Cottrella, "Learning grammatical structure with echo state networks," Neural Netw. **20**, 424–432 (2007).
[52]S. Scherer, M. Oubbati, F. Schwenker, and G. Palm, "Real-time emotion recognition from speech using echo state networks," in *Perception in Multimodal Dialogue Systems, Lecture Notes in Computer Science* (Springer, Berlin, 2008), pp. 200–204.
[53]H. Jaeger, "Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the 'echo state network' approach," Technical Report No. 159, Fraunhofer Institute AIS, St. Augustin, Germany (2002).
[54]H. Jaeger, "ESN toolbox for matlab," http://www.faculty.jacobs-university.de/hjaeger/pubs/ESNtools.zip (Last viewed: 9/7/2011).
[55]D. Verstraeten, B. Schrauwen, D. Stroobandt, J. V. Campenhout, "Isolated word recognition with the liquid state machine: A case study," Inf. Process. Lett. **95**, 521–528 (2005).
[56]M. Slaney, "AUDITORY TOOLBOX (version 2)," Technical Report No. 1998-010, Interval Research Corporation, Palo Alto, CA (1998).
[57]R. Loughran, J. Walker, M. O'Neill, and M. O'Farrell, "Musical instrument identification using principal component analysis and multi-layered perceptrons," in *Proceedings of the IEEE International Conference on Audio and Language Image Processing* (IEEE, Washington, DC, 2008), pp. 643–648.
[58]M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," SIGKDD Explor. **11**, 10–18 (2009).
[59]P. F. Dominey, "Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning," Biol. Cybern. **73**, 265–274 (1995).
[60]A. Livshin and X. Rodet, "The importance of cross-data evaluation for sound classification," in *Proceedings of the International Symposium on Music Information Retrieval*, Baltimore, MD (October 27–30, 2003).