Echo State Networks for Real-Time Audio Applications

Stefano Squartini, Stefania Cecchi, Michele Rossini, and Francesco Piazza

A3Lab, DEIT, Universitá Politecnica delle Marche,
Via Brecce Bianche 31, 60131 Ancona, Italy
sts@deit.univpm.it
a3lab.deit.univpm.it

Abstract. This paper deals with the employment of Echo State Networks for identification of nonlinear dynamical systems in the digital audio field. The real contribution of the work is that such networks have been implemented and run in real-time on a specific PC based software platform for the first time, up to the authors knowledge. The nonlinear dynamical systems to be identified in the audio applications here addressed are the mathematical model of a commercial Valve Amplifier and the low-frequency response of a loud-speaker. Experimental results have shown that, at a certain frequency sampling rate, the ESNs considered (after the training procedure performed off-line) are able to tackle the real-time tasks successfully.

1 Introduction

Neural Networks [1] have been extensively employed in the literature to face different problems in several application fields, likely related to the Digital Signal Processing (DSP) area. Numerous distinct architectures and learning algorithms have been proposed on purpose, in dependence on the task under study. In particular, it often happens that the learning system is asked to have dynamical mapping capabilities, i.e. the ability of storing and updating context information occurring at arbitrarily distant time instants. Common static networks, as the FeedForward Neural Networks (FFNN), are not well-suited to tackle the problem and typically the focus is directed to the Recurrent Neural Networks (RNN) because they have an internal state that can represent context information. Gradient based algorithms are widely used for their simplicity and low computational cost as learning algorithms: back-propagation through time (BPTT) and real-time recurrent learning (RTRL) are well-known examples [1].

However, those types of algorithms have shown to be not sufficiently powerful to discover contingencies spanning long temporal distances. Indeed, as a consequence of the vanishing gradient effect [1], [2], either the system gets information latching being resistant to noise or, alternatively, it is efficiently trainable by gradient descent learning algorithm, but not both. Several solutions have been proposed to mitigate this effect [2], as the Recurrent Multiscale Architecture (RMN) [3], which significantly reduces the impact of the vanishing gradient

D. Liu et al. (Eds.): ISNN 2007, Part III, LNCS 4493, pp. 731–740, 2007.

even maintaining the usage of BPTT algorithm. However, the Echo State Network (ESN) [4], [5], recently appeared in the literature, seems to be the most effective solution from this perspective. Indeed the approach followed in ESNs consists in providing a large set of basis functions through a network of fixed recurrent connections and in combining them through a static linear or nonlinear adaptive mapper for optimal input representation. This allows dealing with dynamical properties of the input time series avoiding training the network feedback synapses, and so resulting in a strongly simplified learning procedure with immunity to the vanishing gradient effect. Such a property has been experimentally verified by some of the authors in a recent paper [3], by comparing ESNs, RMNs and common globally RNNs performances when applied to a specific benchmark.

ESNs properties have been also tested in the literature on more complicated and realistic tasks, as identification of NARMA systems [6], neural activity mapping [7], mobile robot modeling and control [8], Q-function modeling in reinforcement learning [9], speech recognition [10]. However, up to author's knowledge this work represents the first effort to evaluate their capabilities in real-world audio tasks, where we can experience nonlinear and dynamical systems to identify, taking also real-time constraints into account. Here, the modeling of a commercial Valve Amplifier and the identification of a loud-speaker low-frequency response are the audio applications addressed and ESNs have been employed for their fulfillment. Once performed the training procedure offline, we implemented the adapted networks on the Nu-Tech framework, a suitable SW platform for real-time audio processing directly on the PC hardware. As expected, the related real-time constraints result in some restrictions on the network parametrization, which, however, does not affect the effectiveness of the approach in the tasks under study, as shown by the computer simulations carried out.

2 Echo State Networks

The basic working principle of ESNs is that, under certain conditions, its activation state $\mathbf{x}(n)$ is a function of past input values $\mathbf{u}(n), \mathbf{u}(n-1), ...$, so it can be interpreted as an echo of the input history. Let us introduce an adequate terminology to describe such a kind of network. It has K input lines, N internal neurons and L output units. There are four types of synaptic weights: input, internal, output, output-internal. They are described by the corresponding weight matrices W^{in}, W, W^{out}, W^{back}, whose dimensions are respectively $N \times K$, $N \times N$, $L \times (K + N + L)$, $N \times L$. Connections between input and output lines and among output units are allowed. There are no specific assumptions on the topology of internal neural block, namely reservoir; in particular we are not constrained to consider a layer architecture. However it is expected that the internal connections form recurrent paths in order to have a state space behavior. The block diagram of an ESN is depicted in Fig.1. The activation state of the reservoir is given by:

$$x(n+1) = f(W^{in}u(n+1) + Wx(n) + W^{back}y(n)), \tag{1}$$

where $f = (f_1, ..., f_N)$ are the activation functions of the internal units (usually sigmoidal). T he out-put equation is:

$$y(n+1) = f^{out}(W^{out}(u(n+1), x(n+1), y(n)), \tag{2}$$

where $f^{out} = (f_1^{out}, ..., f_L^{out})$ are the activation functions of the output units (usually sigmoidal). In other words, it can be said that the echo functions are the basis functions that the output static mapper has to select for an optimal input representation. Therefore, the Echo State Property has to be satisfied. If we want the state to depend on the past inputs W^{back} must be neglected first. Then, as shown in [4], [5], a sufficient condition is contractivity of W. Nevertheless a weaker operative condition holds in practice: the spectral radius $|\lambda_{\max}|$ of W is less than unity. Sparseness and randomness of W connections are two important requirements to have sufficiently rich dynamics, for the final network to yield the desired mapping. Concerning the learning algorithms, it must be underlined that the reservoir weights are fixed. This allows getting a relevant simplification of the adaptation process, since we do not have to worry about adapting the recurrent connections, the main reason of vanishing gradient occurrence in gradient based algorithms. The only part of ESN subject to learning is the static mapper, for which we can use methods developed in the literature for static NNs. In particular, if the output lines have no feedback weights and the related nonlinearities are invertible, linear regression algorithms might be employed, avoiding iterative procedure based on gradient calculation. According to these assumption, and neglecting the direct input-output synapses, (2) becomes

$$\tilde{y}(n) = \left(f^{out}\right)^{-1}(y(n)) = W^{out}\mathbf{x}(n), \tag{3}$$

where $\mathbf{x}(n)$ is the state vector. If we consider a training observation range equal to $[1 \cdots T_{tr}, \ ]$, (3) becomes:

$$\tilde{\mathbf{y}} = W^{out}\mathbf{X}. \tag{4}$$

By applying Singular Value Decomposition (SVD) we can obtain the optimal W^{out} in terms of the available observations.

3 Implementation Issues

A graphical tool for dealing with ESNs has been developed in C++ (Fig.2), and it can be basically seen as composed by three different parts. The first is related to the determination of the Echo State Network , i.e. the number of internal units, the spectral radius, the internal matrix connectivity, the output-internal weight presence and the type of activation function. The second part refers to the training algorithm that could be based on the gradient based algorithms (like the conjugate gradient), or, as aforementioned, on the linear regression approach, performed through the Singular Value Decomposition (SVD).

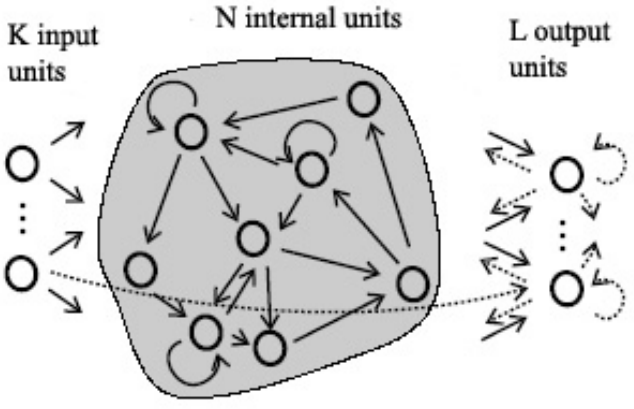

Fig. 1. Echo State Network block diagram: the input line, the reservoir and the output line (static mapper).

The last part gives out the values of the ESN parameters in different operating conditions (initialization, training, generalization). In the learning phase, beginning from the input and target signal, it is important to define the samples number requested to improve the forgetting time of the starting state. In order to have satisfying performances, the samples number should be greater than the network dimensions and the spectral radius. Furthermore it is possible to add white noise in order to avoid instability problems and generally improve the achievable results. At the end of the training phase, the output weight matrix and the correlation matrix of the internal states could be displayed and analyzed to evaluate the generalization performances of the network. It must be said, that in all computer simulations performed, the linear regression method has been employed.

Once trained, the ESN can be suitably saved in a proper format and then used for real time applications. This has been accomplished through the Nu-Tech Platform [11]. This software allows to implement and test real time DSP algorithms in multi-channel scenarios: the Nu-Tech framework is basically composed by two elements, i.e. the interface to the PC sound card and the PlugIn architecture. The former allows handling the audio streams (frame-by-frame) from the I/O sound card channels also through an accurate management of the latency times. The latter lets the user develop his own C/C++ algorithms within the graphical routing scheme reproducing the sound-card MIMO structure.

In our case study, an ESN Nu-Tech PlugIn has been realized as a standard C++ *dll* file able to operate within the Nu-Tech interface (Fig.3). Such a PlugIn can process the audio streaming according to the parametrization related to the trained Echo State Network contained in the proper file coming from the aforementioned C++ based tool.It must be remarked that the combination of the graphical tool and the Nu-Tech framework presents significant pros from

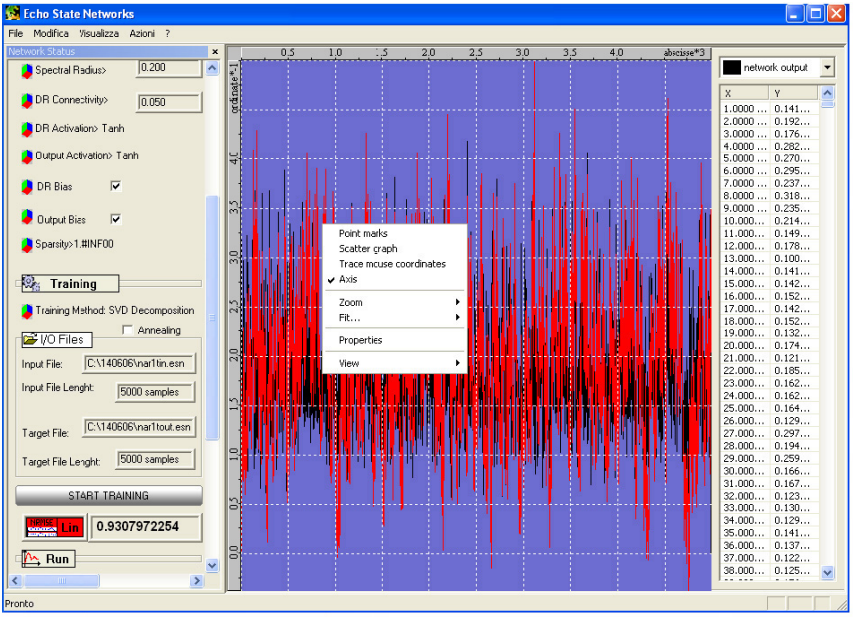

Fig. 2. Graphical tool for ESN generation, initialization, training and testing

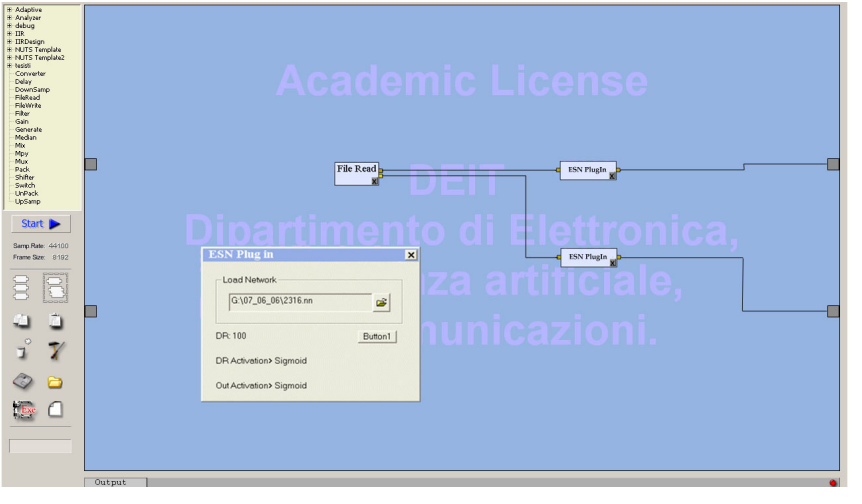

Fig. 3. Nu-Tech Platform with ESN PlugIn

a pure technological point of view: indeed we can easily adapt and run suitable ESNs for real-time applications by just dealing with the available friendly user-interfaces.

Looking at the real-time processing constraints, it must be said that they induce some restrictions on the ESN parametrization allowed. Looking at them from the perspective of the applications described in the following, we can say that the activation function of the internal units must be a sigmoidal function in order to have a lower computational cost and the maximum limit of dynamic reservoir dimension must be 200 units with a connectivity factor of 2.5% to avoid clicks during streaming.

4 Experimental Results

In this section some experimental results related to the field of audio processing will be presented and analyzed. The Echo State Networks have been created, initialized, and trained by using the C++ based graphical tool. Then, for the generalization phase mono wave files (sampled at 44100 Hz) have been used as the inputs feeding the trained networks running in real-time on the Nu-Tech platform. To evaluate the algorithm performances, a normalized mean square error has been defined as follows

$$NRMSE = \sqrt{\frac{\Sigma_{i=1}^{T} \left(y\left[i\right] - d\left[i\right]\right)^2}{T \cdot \sigma^2}}, \tag{5}$$

where y is the output of the network, d is the desired response of the system to identify, $\sigma^2$is the target variance and T the observation time range. For each experimental results a table will be shown with the ESN parameters and NRMSE calculated for the training and testing phases.

4.1 Linear Dynamical Systems

As starting case study, we consider a linear system identification problem. In this case we have considered the behavior of a signal filtered by a FIR filter of order M. To simulate the filter behavior, the Echo State Network has to calculate M+1 parameters storing M past input values. The filter lengths are 50, 80, 100 samples and the dynamical reservoir dimension strictly depends on this. The activation function can be linear or sigmoidal taking into account its linear zone and scaling the input values. The spectral radius is very high (0.97, 0.99) to improve the store capacity of the network. The results are shown in the Table 1. As we can see the best results are achieved for 200 internal units with a connectivity of 5.2% and high spectral radius.

4.2 Modelling of a Commercial Valve Amplifier

Almost a century after their introduction, vacuum tube amplifiers are still appreciated for their special sound qualities. It is known indeed that valve amplifiers are highly rated by audiophiles and musicians, and often preferred to the their digital counterpart [12]. So recently several works have been orientated to

Table 1. Linear system identification experimental results. FO is the filter order, DR is the dimension of the dynamical reservoir, CP is the connectivity percent, SR is the spectral radius, NRMSEtr is the normalized mean square error calculate for the training phase and NRMSE for ESN application.

FO	*DR*	*CP*	*SR*	*NRMSEtr mean*	*NRMSEtr st.dev*	*NRMSE mean*	*NRMSE st.dev*
50	120	8%	0.97	0.0240	$9.78\ 10^{-4}$	0.048	$9.65\ 10^{-4}$
80	200	5.2%	0.99	0.0054	$1.4\ 10^{-3}$	0.0079	$1.42\ 10^{-3}$
100	330	3.3%	0.99	0.0143	$3.4\ 10^{-3}$	0.0217	$3.2\ 10^{-3}$

the non linear digital modeling of a Tube Amplifier. There are mainly two approaches: the former refers to the application of a mathematical model derived from the study of the equivalent circuit, the latter is based on the characterization of a real Valve Amplifier through non linear identification techniques. In this work, as a term of comparison, we have used a free commercial VST PlugIn [13] that implements the behavior of a real tube amplifier. First of all the training phase has been based on a target signal filtered by the VST PlugIn, then the behavior of the ESN has been tested. The results are shown in Table 2. As we can seen, the best results have been achieved using the sigmoidal activation function both for the internal and the output weights. Moreover, the generalization performances do not improve if we increase the size of the dynamical reservoir.

Table 2. Modelling a commercial Valve Amplifier: experimental results. AF is the type of the activation function, DR is the dimension of the dynamical reservoir, CP is the connectivity percent, NRMSEtr is the normalized mean square error calculate for the training phase and NRMSE for ESN application.

AF	*DR*	*CP*	*bias*	*NRMSEtr mean*	*NRMSEtr st.dev*	*NRMSE mean*	*NRMSE st.dev*
Ell	150	5.2%	*Low*	0.0022	$5.23\ 10^{-4}$	0.0057	$1.8\ 10^{-4}$
Ell	100	8%	*Low*	0.0040	$8.54\ 10^{-4}$	0.0069	$1.6\ 10^{-3}$
Ell	100	8%	*Med*	0.0037	$1.40\ 10^{-3}$	0.0104	$3.3\ 10^{-3}$
Ell	100	8%	*High*	0.0041	$2.20\ 10^{-3}$	0.0195	$1.2\ 10^{-3}$
Atan	100	8%	*Low*	0.0116	$4.99\ 10^{-3}$	0.0294	$2.6\ 10^{-3}$
Tanh	100	8%	*Low*	0.0218	$1.40\ 10^{-3}$	0.1392	$1.09\ 10^{1}$

4.3 Identification of a Loudspeaker Low-Frequency Response

In the last decade, several efforts have been made to model the non linear response of loudspeaker in order to reduce the non linear distortion especially at low frequency. The principal causes of non linearities in loudspeaker include non linear suspension and non-uniform flux density. The main results refer

to a model derived from an equivalent circuit of a loudspeaker system. In this work we have considered a mathematical model of a Loudspeaker to analyzed its Low Frequency response as in [14]. The loudspeaker has the following parameters:

$$x(k+1) = \begin{bmatrix} -0.1 & 0 & -0.2 \\ 0 & 1 & 1 \\ 0.6 & -0.5 & -0.15 \end{bmatrix} x(k) + \begin{bmatrix} 0.4 \\ 0 \\ 0 \end{bmatrix} u(k) + \begin{bmatrix} -0.04x_2(k)\,x_3(k) - 0.05x_2^2(k)\,x_3(k) \\ 0 \\ -0.08x_2^3(k) + 0.01x_1(k)\,x_2(k) + 0.02x_1(k)\,x_2^2(k) \end{bmatrix}, \tag{6}$$

$$y(k) = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T x(k). \tag{7}$$

This relation derived from two differential equation associated to the mechanical and electrical equivalent circuit of the loudspeaker taking into account the distortions constraints. In the training phase, noise signal low pass filtered at 1kHz has been used. The reservoir dimension vary from 120 to 200 units with a connectivity of 2-1.5%; the activation functions are sigmoidal while the radius spectrum varies from 0.8 to 0.97. The results are shown in table 3. After the training, the neural network has been tested with a white noise signal and sweep signal (20Hz - 1kHz), played in the wave format within Nu-Tech. Again, the usage of sigmoidal activation functions both for the internal and the output weights with higher spectral radius allows to achieve the best results. Furthermore increasing the dimension of the dynamical reservoir does not yield better results.

Table 3. Identification of a Loudspeaker Low-Frequency response: experimental results . AF is the type of the activation function, DR is the dimension of the dynamical reservoir, CP is the connectivity percent, rs is the spectral radius, NRMSEtr is the normalized mean square error calculate for the training phase, NRMSEt for ESN application with noise input and NRMSEs for ESN application with sweep input.

AF	*DR*	*CP*	*rs*	*NRMSEtr mean/std*	*NRMSEtr mean/std*	*NRMSEt mean/std*	*NRMSEs mean/std*
Ell	120	2%	0.97	0.0049 $2.8\ 10^{-3}$	0.0482 $4.9\ 10^{-3}$	0.0181 $9.1\ 10^{-3}$	$2.521\ 10^{16}$ $2.01\ 10^{16}$
Ell	120	2%	0.8	0.0027 $6.36\ 10^{-4}$	0.0562 $1.6\ 10^{-3}$	0.0112 $4.8\ 10^{-3}$	$7.49\ 10^{17}$ $1.03\ 10^{16}$
Tanh	120	1.8%	0.8	0.00078 $1.17\ 10^{-4}$	0.0618 $9.23\ 10^{-4}$	0.0067 $9.64\ 10^{-4}$	$2.63\ 10^{17}$ $3.05\ 10^{17}$
Ell	200	1.5%	0.92	0.0038 $3.1\ 10^{-3}$	0.0579 $1.8\ 10^{-3}$	0.0138 $2.2\ 10^{-3}$	$2.653\ 10^{15}$ $3.753\ 10^{15}$

5 Conclusions

In this paper we have faced the problem of implementing the Echo State Networks in the Nu-Tech framework for real-time audio applications. Up to the authors' knowledge this represent the first attempt in this direction, and the achieved results seem to be encouraging: indeed, even though the real-time constraints induce some restrictions on the ESN parametrization, the performed ESNs (once adequately trained off-line) are able to solve the tasks successfully. Deep studies are actually ongoing on the possibility of implementing the training phase also in real-time, paying attention to the further to the ESN parametrization limitations which inevitably arise. Moreover, future work could be done to evaluate the applicability of other types of Neural Networks with memory (as RNNs) and compare them to the ESNs in terms of performances in real-time audio applications.

Acknowledgments. This work was supported by the European Commission as sponsor of the hArtes Project number 035143.

References

1. Haykin, S.: Neural network - A Comprehensive foundation. Englewood Cliffs, NI: Prentice Hall (1999)
2. Hochreiter, S., Bengio, S., Frasconi, P., Schmidhuber, J.: Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. A Field Guide to Dynamic Recurrent Network, Chapter 7, J.F. Kolen and S.C. Kremer, Eds. IEEE Press (2001)
3. Squartini, S., Paolinelli, S., Piazza, F.: Comparing Different Recurrent Neural Architectures on a Specific Task from Vanishing Gradient Effect Perspective. IEEE Int. Conf. Networking, Sensing and Control, Miami, USA (2006) 380-385
4. Jaeger, H.: The Echo State Approach to Analysing and Training Recurrent Neural Networks. German National Research Center for Information Technology, Fraunhofer Institute for Autonomous Intelligent Systems, Tech. Rep. (2001) GMD Report 148
5. Jaeger, H.: Short Term Memory in Echo State Networks. GMD report 152, German National Research center Information Technology (2002)
6. Jaeger, H.: Adaptive Nonlinear System Identification with Echo State Networks. Advances in Neural Information Processing Systems, 2002, Becker, S., Thrun, S., Obermayer, K. Cambridge, MA: MIT Press (2003) 593-600
7. Rao, Y.N., Kim, S.P., Sanchez, J.C., Erdogmus, D., Principe, J.C., Carmena, J.M., Lebedev, M.A., Nicolelis, M.A.: Learning Mappings in Brain Machine Interfaces with Echo State Networks. IEEE Int. Conf. Acoustics, Speech and Signal Processing **5** (2005) 233-236
8. Ploger, P.G. , Arghir, A., Gunther, T., Hosseiny, R.: Echo State Networks for Mobile Robot Modeling and Control. Lecture Notes in Artificial Intelligence of the 7th Robot Soccer World Cup **3020** (2003)
9. Bush, K., Anderson, C.: Modeling Reward Functions for Incomplete State Representations via Echo State Networks. Proc. Int. Joint Conf. Neural Networks, Montreal, Canada (2005) 2995-3000

10. Skowronski, M.D., Harris, J.G.: Minimum Mean Squared Error Time Series Classification using an Echo State Network Prediction Model. IEEE Int. Symp. Circuits and Systems, Island of Kos, Greece (2006) 3153-3156
11. Squartini, S., Ciavattini, E., Lattanzi, A., Zallocco, D., Bettarelli, F., Piazza, F.: NU-Tech: Implementing DSP Algorithms in a Plug-in based Software Platform for Real Time Audio applications. Presented at the 118th AES Convention, Barcelona, Spain (2005)
12. Van der Veen, M.: Universal System and Output Transformer for Valve Amplifiers. Presented at the 118th AES Convention, Barcelona, Spain (2005)
13. http://www.voxengo.com/product/tubeamp/ Voxengo Tube Amplifier VST PlugIn.
14. Gao, F.X.Y., Snelgrove, W.M.: Adaptive Linearization of a Loudspeaker. Int. Conf. Acoustics, Speech and Signal Processing **5** (1991) 3589-3592