Neural networks for modeling time series of musical instruments

Axel Röbel
GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
Tel: 030 - 6392 1900 email: roebel@first.gmd.de

ABSTRACT: We demonstrate the use of neural networks to model the attractors of musical instruments from time series. We show how the neural networks have to be extended to be able to model instationary behavior. To give an impression of the use of the models we summarize some results of analyzing the instrument dynamics and we show that the neural models are stable in the sense that they allow resynthesis of natural sounds.

Introduction

There has been a number of interesting results concerning the modeling of time series from nonlinear dynamical systems. Since the formulation of the reconstruction theorem by Takens (1981) it has been clear that a nonlinear model of a system may be derived directly from a systems time series. The method of state space reconstruction has already been used for modeling and analysis of musical sounds (Monro, 1993; Pressing *et al.*, 1993). For other time series prediction tasks the combination of reconstruction techniques and neural networks has shown good results (Weigend and Gershenfeld, 1993). In our work we extend this ideas by the more demanding task of building models, which are able to resynthesize the systems time series. Because we are interested in modeling musical instruments we extended the standard neural network predictors such that they are able to model instationary dynamics.
In the following, we first give a short review concerning the state space reconstruction from time series by delay coordinate vectors. Then we explain the neural networks we used and the modification necessary to model instationary dynamics. As an example we investigate the neural models of a saxophone tone. In the end of the paper we describe further results and applications.

Reconstructing attractors

Assume an n-dimensional dynamical system $f(\cdot)$ evolving on an attractor A. A has fractal dimension d, which often is considerably smaller then n. The system state $\vec{z}$ is observed through a sequence of measurements $h(\vec{z})$, resulting in a time series of measurements $y_t = h(\vec{z}(t))$. Under weak assumptions concerning $h(\cdot)$ and $f(\cdot)$ the fractal embedding theorem(Sauer *et al.*, 1991) ensures that, for $D > 2d$, the set of all *delayed coordinate vectors*

$$Y_{D,\tau} = \{t > t_0 : (y_t, y_{t-\tau}, \ldots, y_{t-(D-1)\tau})\}, \tag{1}$$

with an arbitrary delay time $\tau$and arbitrary t_0, forms an embedding of A in the D-dimensional *reconstruction* space. We call the minimal D, which yields an embedding of A, the *embedding dimension* D_e. Because the embedding preserves characteristic features of A, it may be employed for building a system model.

In the case of instationary systems the concept of attractors, which does rely on the behavior of the system for $t \to \infty$, does not take over immediately. If, however, the instationarity is due to slowly varying system parameters, it is possible to model the system dynamics using a sequence of attractors. As it turns out this approach leads to reasonable results in the case of musical instruments.

RBF neural networks

There are different topologies of neural networks that may be used for time series modeling. In our investigation we used radial basis function networks. As proposed by Verleysen et. al (1994) we initialize the network using a vector quantization procedure and then apply backpropagation training to finally tune the network parameters. The tuning of the parameters yields an improvement of factor ten in prediction error compared to the standard RBF networks approach (Moody and Darken, 1989). The resulting network function for m-dimensional vector valued output and n-dimensional input vector $\vec{x}$ is of the form

$$\vec{N}(\vec{x}) = \sum_j \vec{w}_j \exp(-\frac{1}{2}(\frac{\vec{c}_j - \vec{x}}{\sigma_j})^2) + \vec{b}, \tag{2}$$

where σ_j stands for the standard deviation of the Gaussian, the centers $\vec{c}$ are n-dimensional and $\vec{b}$ and $\vec{w}_j$ are m-dimensional parameters of the network. Networks of the form eq. (2) with a finite number of hidden units are able to approximate arbitrary closely all continuous mappings $R^n \to R^m$ (Park and Sandberg, 1991). This universal approximation property is the foundation of using neural networks for time series modeling, where they are referred to as *neural models*.

To be able to represent instationary dynamics, we extend the network function to have an additional input t, that enables us the control of the actual mapping

$$\vec{N}(\vec{x}, t) = \sum_j \vec{w}_j \exp(-\frac{1}{2}(\frac{\vec{c}_j - \vec{x}}{\sigma_j})^2 - \frac{1}{2}(\frac{t_j - t}{\sigma_{tj}})^2) + \vec{b}. \tag{3}$$

From the universal approximation properties of the RBF-networks stated above it follows that eq. (3), with appropriate control sequence t, is able to approximate any sequence of functions. The control input sequence is used to discriminate between the different system attractors that generate the overall instationary dynamics. The control sequence may be optimized during training (Levin, 1993), or may be chosen a-priori fixed, in which case we have to ensure the discriminating behavior. For training the network it is necessary to supply enough training data for each of the generating attractors.

Neural models

Using the delayed coordinate vectors of a music signal and a fixed control sequence we train the network to give a vector valued prediction of the following τ time samples. To ensure a discriminating control input, we chose a control sequence linear increasing with the sample time. After training we initialize the network input with the first input vector $(\vec{x}_0, t_0)$ of the time series and iterate the network function shifting the network input and using the latest output unit to complete the new input. The control input sequence may be copied from the training phase to resynthesize the training signal or may be varied to get a variation of the musical sound.

The question that arises in this context is the question of stability. As we will see in the example, the neural models are stable for selected parameters D and τ. Due to the embedding of the attractor, however, this stability, which depends on its neighborhood, is not guaranteed and it may be the case for other time series, that there are no parameters for stable models. A method for controlling the stability of the models is subject of further research.

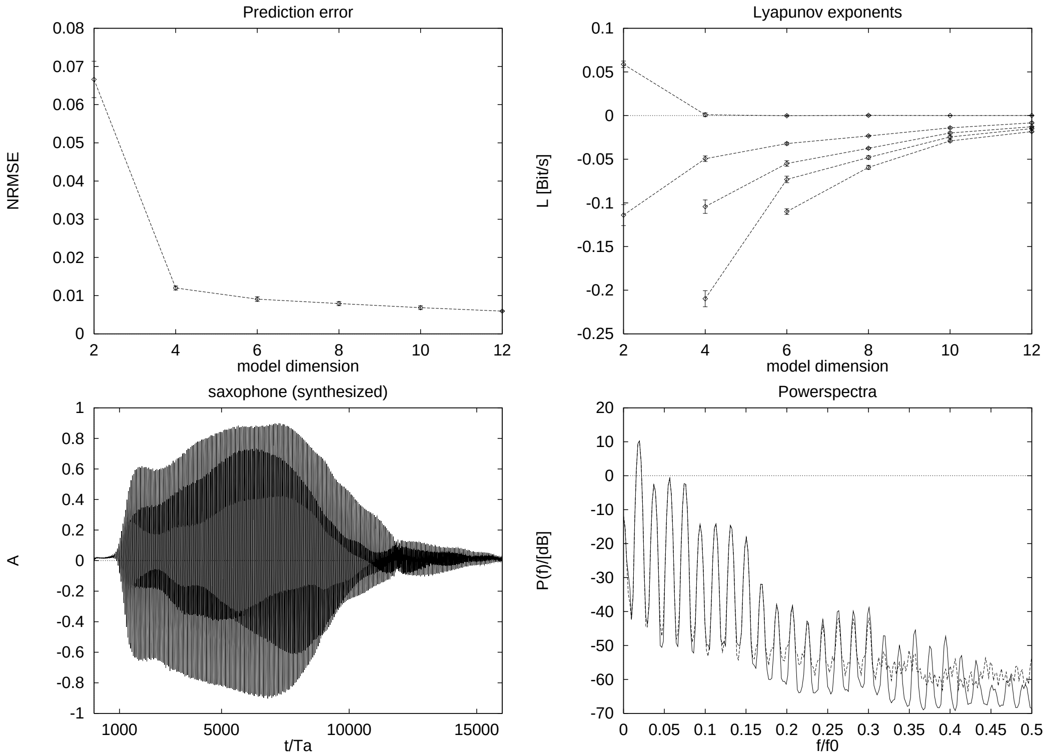

Fig. 1. Prediction error and the largest 5 Lyapunov exponents of the saxophone models with varying dimension D. Synthesized saxophone signal and the powerspectrum estimation for the original (solid) and synthesized signal (dashed).

Results

We have applied our method to various music and speech signals and present the results for a single saxophone tone, consisting of 16000 samples sampled at 32kHz. For training the signal is normalized within the range $[-1, 1]$ and the control input t is linear increasing from -0.8 to 0.8. For the following results $\tau = 5$ has been used. The time series has been analyzed to estimate the fractal dimension of the underlying attractors (Grassberger and Procaccia, 1983). We obtain a fractal dimension $d \approx 3$, which due to the instationary dynamics has to be interpreted as an upper limit for the dimension of the generating attractors. In fig. 1 we show the relation between input dimension and prediction error* of the models. According to the reconstruction theorem the prediction error nearly remains constant for $D \geq 4$.

Besides the fractal dimension of the attractor its Lyapunov exponents are important to describe the dynamics. The Lyapunov exponents measure the sensitivity of the trajectories of the system to small perturbations. They are mainly used to analyze wether the system dynamics are chaotic, which is indicated if at least the largest Lyapunov exponent is positive (Eckmann and Ruelle, 1985). Similar as in (Röbel, 1995) we estimate the Lyapunov exponents of the saxophone models. Due to the instationarity of the models the results estimate the average Lyapunov exponents for the sequence of attractors. In fig. 1 we show the largest 5 Lyapunov exponents and realize that for training on an embedding of the attractors, $D \geq 4$, the largest Lyapunov exponent is zero. Therefore, we conclude that the dynamics generating the tone at hand are not chaotic.

The most demanding task for the models is the resynthesis of the music time series. From our systematic investigations we found that for synthesis purposes the saxophone models need higher input dimension $D = 9$. With 200 hidden units these models are capable to resynthesize the input signal with high quality and,

*The root mean squared error for one step ahead prediction of the music signal.

by variation of the control sequence, does even allow considerable variations of synthesized sounds. The resynthesized time series and the power spectra of the original and resynthesized signal are shown in fig. 1. From the spectrum we see the close resemblance of the sound. As an example for possible sound control we may invert the control sequence, such that the sound is synthesized reverse in time, or fix the control input for some time to generate longer duration of the tone. At the conference we will give acoustical demonstration of the synthesized signals.

Further work

To consider musicians demands, the next step towards a suitable music synthesis is to enhance the control of the synthesized signals. To achieve this we will enlarge the models, incorporating different flavors of sounds into the same model and adding additional control inputs. For example it may be possible to build a model representing different volume and pitch of a tone.
As a second application we investigate the possibility of speech synthesis. Our results concerning the synthesis of spoken words show that the neural models are capable to resynthesize natural speech.

References

J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Review Mod. Physics*, 57(3):617–656, 1985.

P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983.

E. Levin. Hidden control neural architecture modelling of nonlinear time varying systems and its applications. *IEEE Transactions on Neural Networks*, 4(2):109–116, 1993.

G. Monro. Synthesis from attractors. In *Proc. of the International Computer Music Conference*, pages 390–392, Tokyo, 1993.

J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294, 1989.

J. Park and I. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, 1991.

J. Pressing, C. Scallan, and N. Dicker. Visualization and predictive modelling of musical signals using embedding techniques. In *Proc. of the International Computer Music Conference*, pages 110–113, Tokyo, 1993.

A. Röbel. Using neural models for analyzing time series of nonlinear dynamical systems. In *Proceedings of the 5th International IMACS-Symposium on System Analysis and Simulation*, 1995.

T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal of Statitstical Physics*, 65(3/4):579–616, 1991.

F. Takens. *Detecting Strange Attractors in Turbulence*, volume 898 of *Lecture Notes in Mathematics (Dynamical Systems and Turbulance, Warwick 1980)*, pages 366–381. D.A. Rand and L.S. Young, Eds. Berlin: Springer, 1981.

M. Verleysen and K. Hlavackova. An optimized RBF network for approximation of functions. In *Proceedings of the European Symposium on Artificial Neural Networks, ESANN'94*, 1994.

A. S. Weigend and N. A. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley Pub. Comp., 1993.