RBF Networks for Synthesis of Speech and Music Signals

RBF Netze zur Synthese von Sprach– und Musiksignalen

Dr. A. Röbel
GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
Tel: 030 - 6392 1900 FAX: 030 - 6392 1805 email: roebel@first.gmd.de

Kurzfassung

Ein aktuelles Gebiet der Anwendung Neuronaler Netze besteht in der Vorhersage von Zeitreihen. Nach einer kurzen Einführung in die zugrundeliegende Theorie, wird die iterierte Vorhersage eines dynamischen Systems als ein Modell der Systemdynamik eingeführt. Unter Verwendung von RBF-Neuronalen Netzen wird eine Ergänzung der Methodik erläutert, die die Modellierung instationären Systemverhaltens ermöglicht. Der Ansatz wird am Beispiel der Modellierung eines Musik- und eines Sprachsignals demonstriert. Dabei steht der Einsatz der Modelle zur Synthese von Signalen im Mittelpunkt der Untersuchungen.

Abstract

Time series prediction is one of the major applications of neural networks. After a short introduction into the basic theoretical foundations we argue that the iterated prediction of a dynamical system may be interpreted as a model of the system dynamics. By means of RBF neural networks we describe a modeling approach and extend it to be able to model instationary systems. As a practical test for the capabilities of the method we investigate the modeling of musical and speech signals and demonstrate that the model may be used for synthesis of musical and speech signals.

1 Introduction

Since the formulation of the reconstruction theorem by Takens [7] it has been clear that a nonlinear predictor of a dynamical system may be directly derived from a systems time series. The method has been used extensively and with good success for the prediction of time series of nonlinear systems. Especially the combination of reconstruction techniques and neural networks has shown good results [9].

In our work we extend the ideas of predicting nonlinear systems by the more demanding task of building system models, which are able to resynthesize the systems time series. In the case of chaotic or strange attractors the identical resynthesis of time series is known to be impossible.

However, in this case the modeling of the underlying attractor leads to the possibility to resynthesis time series which are consistent with the system dynamics [4]. Moreover, the models may be applied to the analysis of the system dynamics, for example the estimation of the Lyapunov exponents [5]. In the following we investigate the modeling of music and speech signals, where the system dynamics are known to be instationary. Therefore, we develop an extension to the modeling approach, such that we are able to resynthesize instationary systems.

In the following, we first give a short review concerning the state space reconstruction from time series by delay coordinate vectors, a method that was introduced by Takens [7] and later extended by Sauer et al. [6]. Then we explain the structure of the neural networks we used in the experiments and the enhancements necessary to be able to model instationary dynamics. As an example we apply the neural models to a saxophone tone and a speech signal and demonstrate that the signals may be resynthesized using the neural models. Moreover, we give some interpretation of the results and explain further developments of the applications.

2 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[6] ensures that, for $D > 2d$, the set of all *delayed coordinate vectors*

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

with an arbitrary delay time T, 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 an embedding preserves characteristic features of A, especially it is one to one, it may be employed for building a system model. For this purpose the reconstruction of the attractor is used to uniquely identify the systems state thereby establishing the possibility of uniquely predicting the systems evolution. The prediction function may be represented by a hyperplane over the attractor in an $(D + 1)$ dimensional space. By iterating this prediction function we obtain a vector valued system model which, however, is valid only at the respective attractor. For the reconstruction of instationary systems dynamics we confine ourselves to the case of slowly varying parameters and model the instationary system using a sequence of attractors.

3 RBF neural networks

There are different topologies of neural networks that may be employed for time series modeling. In our investigation we used radial basis function networks which have shown considerably better scaling properties, when increasing the number of hidden units, than networks with sigmoid activation function. As proposed by Verleysen et. al [8] 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 factor of about ten in prediction error compared to the standard RBF network approach [2].

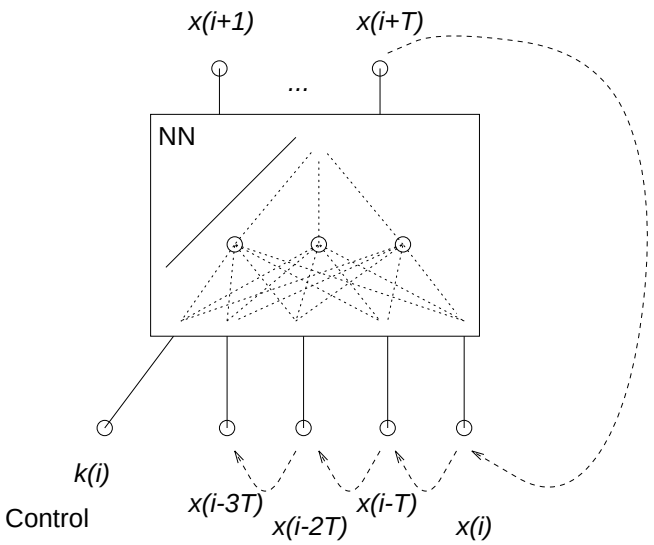

Figure 1: Input/Output structure of the neural model.

The resulting network function for m-dimensional vector valued output is of the form

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

where σ_j stands for the standard deviation of the Gaussian, the input $\vec{x}$ and the centers $\vec{c}$ are n-dimensional vectors 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 $R^n \rightarrow R^m$ [3]. This universal approximation property is the foundation of using neural networks for time series modeling, where they are referred to as *neural models*. In the context of the previous section the neural models are approximating the systems prediction function.

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

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

From the universal approximation properties of the RBF-networks stated above it follows, that eq. (3) with appropriate control sequence $k(i)$ is able to approximate any sequence of functions. In this setting i represents the sample time. The control sequence may be optimized during training [1] or, with lower computational demands, may be appropriately chosen in advance. Selecting $k(i)$ to be monotonically increasing with i will often work, as long as the number of training samples is high enough to fix the different network functions. In our investigation we select $k(i)$ to be a linear, increasing function of i.

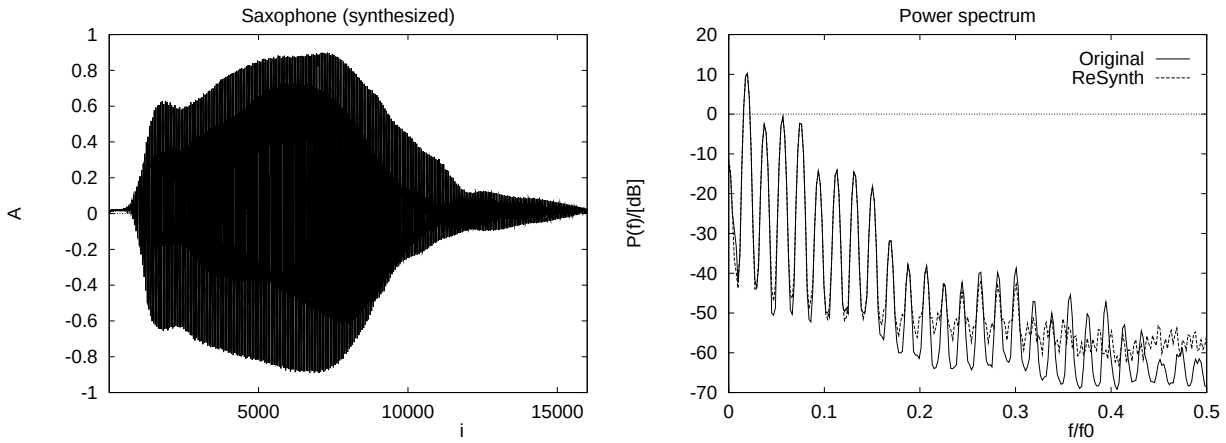

Figure 2: Synthesized saxophone signal and power spectrum estimation for the original (solid) and synthesized (dashed) signal.

4 Neural models

As is shown in figure 1 we use the delayed coordinate vectors and a selected control sequence to train the network to predict the sequence of the following T time samples. The vector valued prediction avoids the need for a further interpolation of the predicted samples. The interpolation would be necessary to obtain the original sample frequency, but, because the Nyquist frequency is not regarded in choosing T, is not straightforward to achieve.

After training we initialize the network input with the first input vector $(\vec{x}_0, k(0))$ 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 may be copied from the training phase to resynthesize the training signal or may be varied to get emulate another sequence of system dynamics.

The question that arises in this context is the question of stability. Due to the networks approximation errors the iterated model will soon leave the attractor embedding. It is not guaranteed to stay close to it because there is no training data in the neighborhood of the attractor. Nevertheless, as we will see in the examples, the neural models are stable for at least some parameters D and T. It may be the case for other time series, however, that there exist no parameters for stable models.
One method to increase the model stability is to use noisy input data, thereby smoothing the prediction function in the surrounding of the attractor. Due to the additional effect of higher prediction errors, the overall effect is not always positive. Recurrent backpropagation has been tried to learn some iterated prediction steps, but this did not improve the results. The stability of the models is subject of further research.

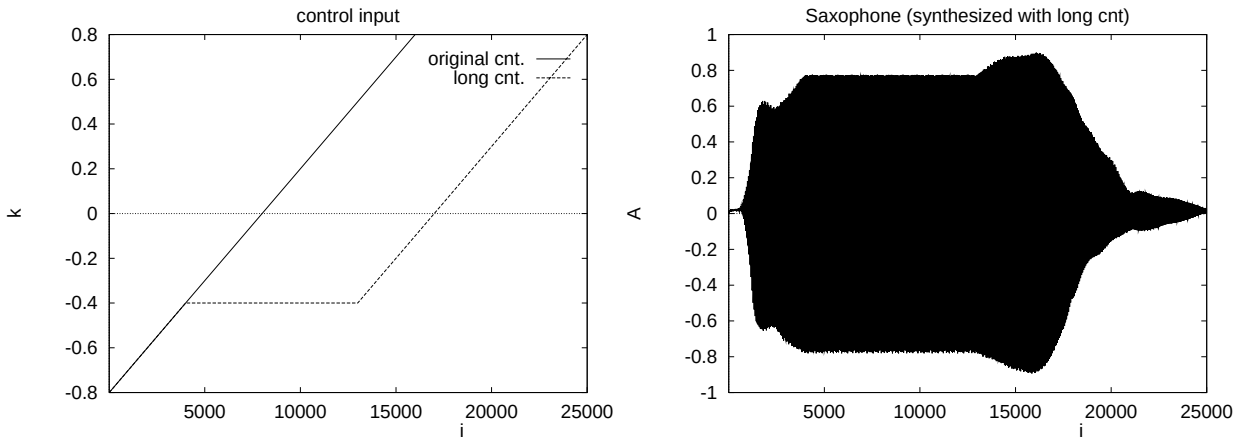

Figure 3: Varying the synthesized tone by varying the control input sequence.

5 Practical results

We have applied our method to two acoustic time series, a single saxophone tone, consisting of 16000 samples sampled at 32kHz and a speech signal of the word *manna*[1]. The latter time series consists of 23000 samples with a sampling rate of 44.1kHz. Both time series have been normalized to stay within the interval $[-1, 1]$. The estimation of the dimension of the underlying attractors yields a dimension of about 2-3 in both cases.

We chose the control input $k(i)$ to be linear increasing from -0.8 to 0.8. Stable models we found for both time series using $D > 5$. Namely for the parameter T we observed considerable impact on the model quality. While smaller T results in better one step ahead prediction, the iterated model often becomes unstable. This might be explained by the decrease in variation within the prediction hyperplane, that has to be learned. For small T the model tends to become linear and does not capture the nonlinear characteristics of the system. Therefore the iteration of those models failed.
To large values of T results in an insufficient one step ahead prediction error, which pushes the model far away from the attractor also producing unstable behavior.

5.1 Modeling a saxophone

In the following we consider the results for the saxophone model. The model we present consists of 10 input units, 200 hidden units and 5 output units and was trained with additional Gaussian noise at the input. The standard deviation of the noise is 0.0005 and the RMS training error obtained is 0.005. The resulting saxophone model is able to resynthesize a signal which is nearly indistinguishable from the original one. The resynthesized time series is shown in figure 2. The time series follows the original one with a small phase shift, which stems from a small difference in the onset of the model. Also in figure 2 the power spectrum of the saxophone signal and the neural model is shown. From the spectrum we see the close resemblance of the sound.

[1]The name of our parallel computer

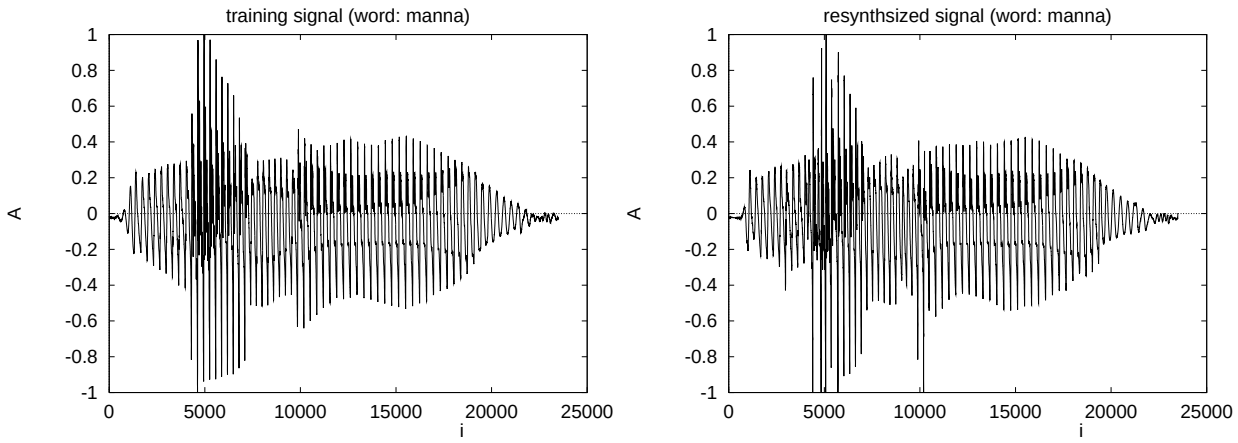

Figure 4: Original and synthesized signal of the word *manna*.

One major demand for the practical application of the proposed musical instrument models is the possibility to control the synthesized sound. At the present state there exists only one control input to the model. Nevertheless, it is interesting to investigate the effect of varying the control input of the model. We tried different control input sequences to synthesize saxophone tones. It turns out that the model remains stable such that we are able to control the envelope of the sound. An example of a tone with increased duration is shown in figure 3. In this example the control input first follows the trained version, then remains constant to produce a longer duration of the tone and then increases to reproduce the end of the tone from the trained time series.

5.2 Modeling a speech signal

For modeling the time series of the spoken word *manna* we used a similar network compared to the saxophone model. Due to the increased instationarity in the signal we needed an increased number of RBF units in the network. The best results up to now has been obtained with a network of 400 hidden units, delay time $T = 8$, output dimension 8 and input dimension 11.

In figure 4 we show the original and the resynthesized signal. The quality of the model is not that high as in the case of the saxophone. Nevertheless, the word is quite understandable. From the figure we see, that the main problems stem from the transitions between consecutive phonemes. Therefore we assume that a more sophisticated control sequence may solve the problem.

6 Further developments

There are two practical applications that directly follow from the presented results. The first one is to synthesize music signals. To consider musicians demands, we need to enhance the control of the synthesized signals. Therefore, in the future we will try to enlarge the models, incorporating different flavors of sound into the same model and adding additional control inputs. Especially we plan to build models for different volume and pitch.

As a second application we will investigate the possibility to use the neural models in a speech synthesizer. In this case we also have to enhance the control, to be able to represent intonation and coarticulation. The single models, however, will represent smaller entities, for example phonemes or diphones. Due to the physically oriented representation of the single entities, we expect a more natural result with compacter models compared to existing speech synthesis systems.

7 Summary

The article describes a methodology to build instationary models from time series of dynamical systems. We give theoretical arguments for the universality of the models and discuss some of the restrictions and actual problems. As practical test for the models we apply the models to the demanding task of the synthesis of musical and speech signals. It is demonstrated that the models are capable to resynthesize the trained signals. At the present state the envelope and duration of the synthesized signals may be controlled. Intended further developments have been shortly described.

References

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

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

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

[4] Axel Röbel. *Neural Models of Nonlinear Dynamical Systems and their Application to Musical Signals.* PhD thesis, Technical University of Berlin, 1993. In German.

[5] Axel 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.

[6] Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *Journal of Statitstical Physics*, 65(3/4):579–616, 1991.

[7] 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.

[8] Michel Verleysen and Katerina Hlavackova. An optimized RBF network for approximation of functions. In *Proceedings of the European Symposium on Artificial Neural Networks, ESANN'94*, 1994.

[9] Andreas S. Weigend and Neil A. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley Pub. Comp., 1993.