Vol. 43 No. 2 IPSJ Journal Feb. 2002

Regular Paper

Recurrent Neural Network Synthesis

Ken'ichi Ohya†

This paper introduces a new technique of sound synthesis RNNS (Recurrent Neural Network Synthesis) by recurrent neural network. As a model of individual neuron, a continuous-time, continuous-variable neuron model is adopted. RNNS has two variations of sound synthesis: One variation of sound synthesis is "hand-tuning model", which is composed of relatively small network and is suitable for realtime sound synthesis. And another variation of sound synthesis is "resynthesis model", which is made of relatively large numbers of neurons. The latter is not only capable of learning of the original sound of the musical instruments, but is also capable of producing a new sound by modifying parameters of the neurons after learning process was finished.

1. Introduction

Study of sound synthesis has a long history in computer music and many models have been presented. One of the most famous algorithm is the FM (Frequency Modulation) synthesis in the early 1980s, and produced sounds by that algorithm were worldwidely used in so many tunes. In the 1990s, as memory prices going down, PCM sounds became more and more popular. In the late 1990s, software synthesis became possible because of the fast clock speed of CPU.

This paper introduces RNNS (Recurrent Neural Network Synthesis), a new technique of sound synthesis using recurrent neural network.

Recurrent neural network is a neural network that each neuron connects recurrently. RNNS belongs to both nonlinear sound synthesis and software sound synthesis. Followings are some features of RNNS:

(1) dynamics of neurons are directly used for synthesized waveform itself,
(2) RNNS resembles to the FM synthesis, each neuron corresponds to each operator in the FM synthesis model,
(3) complex waveforms are produced from relatively small numbers of neurons,
(4) resynthesis is also possible with relatively large number of neurons by a learning algorithm.

RNNS has two variations, one is "hand-tuning model", which is composed of relatively small network and is suitable for realtime sound synthesis. And another is "resynthesis model", which is capable of resynthesis of the sound of the musical instruments by a learning algorithm.

Sound synthesis by "hand-tuning model" is totally new. "Resynthesis model" was used as a speech synthesis model [7]. This paper shows resynthesis of the sound of the musical instruments by resynthesis model.

2. Single Neuron Model and Recurrent Neural Network

As a model of individual single neuron, a continuous-time, continuous-variable neuron model is adopted. Therefore output value from any single neuron can be directly used for waveform of a synthesized sound.

Equation of dynamics of each neuron in a recurrent neural network is given [2] as

$$\tau_i \frac{du_i}{dt} = -u_i + f(\sum_{j=1}^{n} W_{ij} u_j) + I_i$$

where $u_i(t)$ is the i-th unit output at a time t, τ_i a time delay constant, $f(x)$ a sigmoid function, I_i an external input of the i-th unit, W_{ij} a connection weight from the j-th unit to the i-th unit.

3. Sound Synthesis Variation 1: Hand-Tuning Model

Recurrent Neural Network can generate very complex dynamics pattern because of its recurrent connections even if it is composed of relatively small numbers of neurons.

As the 1st variation of RNNS, sound synthesis by hand-tuning model is shown. This model is useful for realtime software sound synthesis.

3.1 1 Pair Model

"1 pair model" is composed of one pair neuron and another output neuron. Each neuron

† Nagano National College of Technology

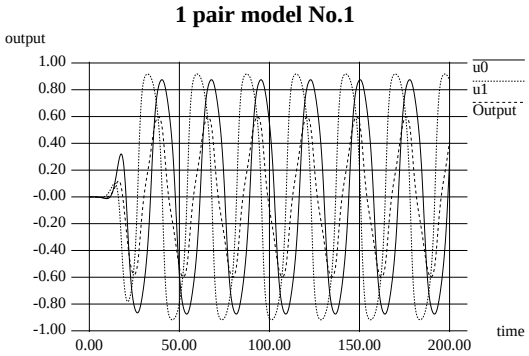

Fig. 1 An example of the produced waveform from 1 pair model.

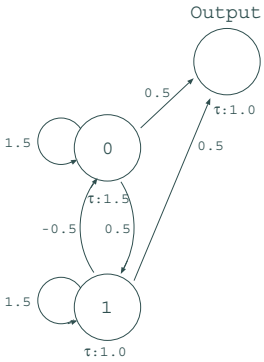

Fig. 2 Parameters to produce waveforms in Fig. 1.

of the pair neuron is connected each other and is also connected to itself. This pair neuron is the smallest recurrent neural network architecture. The output neuron receives two output values from the pair neuron and computes output value, which is waveform itself.

Dynamics of the pair neuron is completely described by eight parameters, 2 initial values, 2 time delay constants, 4 connection weight values. In an 8-dimension space of parameters, the region which gives this system lasting oscillation seems to be small. One of such region is given if weight values connecting each other are asymmetric. In this condition, 2 output values of a pair neuron spikes by turns, and this pair neuron functions as a kind of oscillator. It is also possible to use only one of the outputs from the pair neuron as waveform itself, but the third neuron, the output neuron, is used for producing more complex, rich waveform.

An example of a waveform of 1 pair model is shown in **Fig. 1**, where "u0" is the output from the neuron No.0, "u1" is the output from the neuron No.1, "Output" is the output from the output neuron, that is waveform itself. Parameters of neurons are described in **Fig. 2**.

Frequency of the waveform mainly depends on the time delay constant τ, and the total waveform chiefly depends on values of connection weight. Because values of connection weight and feedback loop do not directly produce waveforms, it is difficult to predict the produced waveform while hand-tuning except for the time delay constant τ.

Another example of a waveform of 1 pair model is shown in **Fig. 3**, and parameters are shown in **Fig. 4**. In this case, feedback terms are increased.

In this way, we can make various oscilators by changing many parameters, even in this simple 1 pair model.

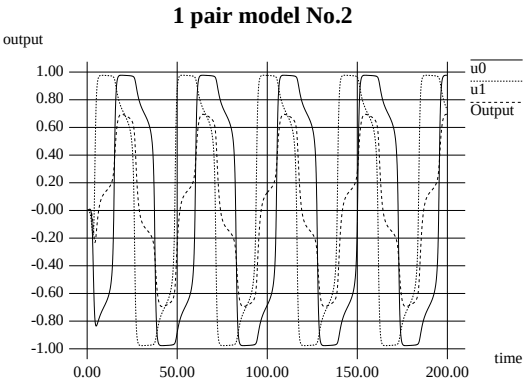

Fig. 3 Another example of the produced waveform from 1 pair model.

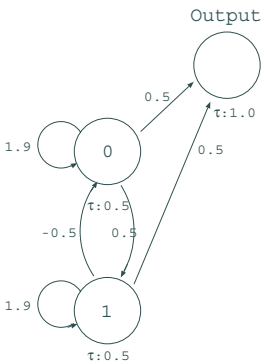

Fig. 4 Another set of parameters to produce waveforms in Fig. 3.

3.2 2 Pair Model and 3 Pair Model

2 pair model is composed of 2 pair neurons and 1 output neuron (**Fig. 5**). This model has 2 oscillators, frequency of each oscillator depends on its time delay constant τ. Therefore this model resembles 3 operator model in FM syn-

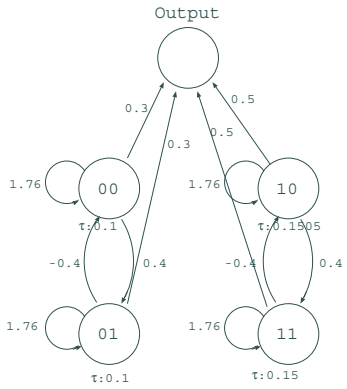

Fig. 5 Parameters to produce waveforms in Fig. 7.

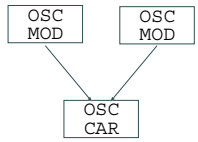

Fig. 6 3 operator model in FM.

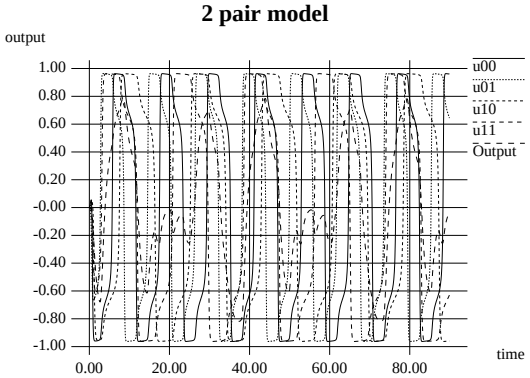

Fig. 7 An example of the produced waveform from 2 pair model.

thesis in a sense (**Fig. 6**). By modifying two time delay constants, various waveforms can be produced (**Fig. 7**).

As shown in Fig. 5, in this example, the ratio of the 2 time delay constants of 2 pair neurons is 1.5, thus harmonics of the sound can be controlled.

And more, by slightly modifying the time delay constant of one of neurons (ex. Neuron No.10 in Fig. 5), the output of the right pair neuron can be slightly fluctuated, and the produced sound also can be slightly fluctuated.

In the same way, we can produce more complex waveform by adding more pair neurons. **Figure 8** is an example of 3 pair model, and **Fig. 9** shows parameters to produce waveforms in Fig. 8.

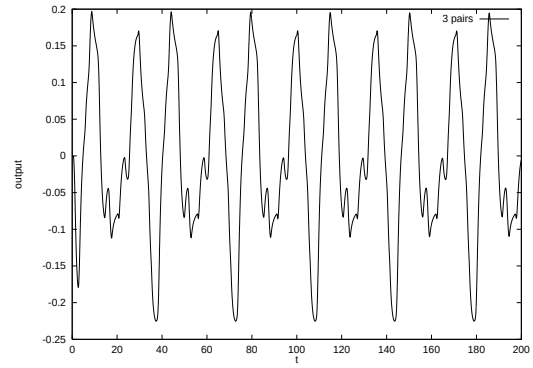

Fig. 8 An example of the produced waveform from 3 pair model.

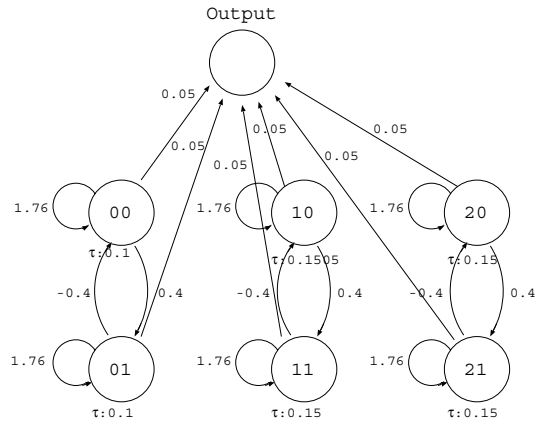

Fig. 9 Parameters to produce waveforms in Fig. 8.

4. Sound Synthesis Variation 2: Resynthesis Model

"Sound Synthesis Variation 2" in this section introduces resynthesis model by Recurrent Neural Network.

4.1 APOLONN

Some architectures of recurrent neural networks can be trained to learn spatiotemporal pattern [4),5)] and chaotic dynamics [6),7)].

Adaptive nonlinear pair oscillators with local connections, APOLONN [7)], is one of the architectures that was applied for speech synthesis of "Ah" sound (sampling ratio was 4 kHz) [7)]. An APOLONN consists of many pairs of oscillators. A pair of oscillators is locally connected with its neighboring pairs, and all neurons are connected to one neuron; the output neuron.

Each pair of oscillators generates various kinds of complex patterns depending on its parameters, such as τ_i or its weight connections. Since each pair is locally connected to the neigh-

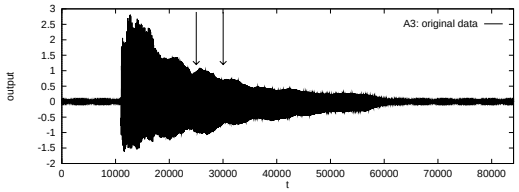

Fig. 10 Sampled data of the sound of the piano, and the training data (denoted between the 2 arrows).

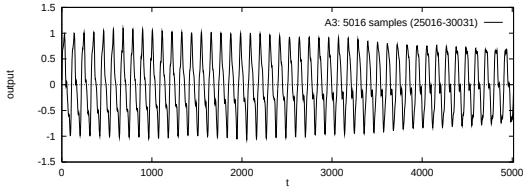

Fig. 11 Training data of the piano. 5,016 samples. The note is A3.

boring pairs, oscillations of a pair are not independent from another. The total system can produce further complex nonlinear patterns that are rich in frequencies.

4.2 Resynthesis of the Sound of the Piano

An APOLONN is trained to learn waveforms, including fluctuations of amplitude and periodicities, of an acoustic musical instrument (See Ref. 7) for learning algorithm details).

A waveform of a piano tone (A3, 440 Hz), known as a mixture of attack noise, simple vibrations and their fluctuations, is used for the teacher signal [2]. The data were sampled in 16-bit integer format at a sampling rate of 44.1 kHz.

Figure 10 is the sampled data of the sound of the piano, and the training data is denoted between the 2 arrows. This part, shown in **Fig. 11**, is relatively flat; 46 periods after the attack. 5,016 samples are used for the training. To look into the chaotic dynamics of the data, 3-dimensional phase space trajectory is also shown (**Fig. 12**).

20 pairs of oscillators were used in the simulation. τ_i of each pair was set to slightly different from the neighboring pair. The ratio of the τ_i between two neighboring pairs was 0.9.

After 990 iterations, error signal became sufficiently small. An output of the recurrent neural network, shown as **Fig. 13**, indicates that the APOLONN can learn some complex temporal pattern. 3-dimensional phase space trajectory, presented as **Fig. 14**, shows that the APOLONN has learned even fluctuations of the

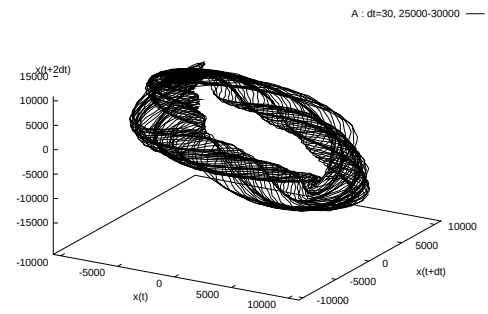

Fig. 12 Phase space trajectory of the training data of the sound of the piano.

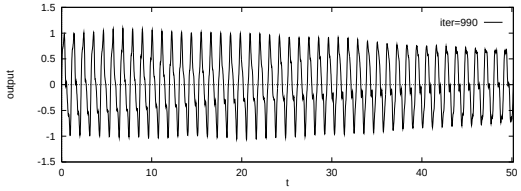

Fig. 13 Produced waveform by resynthesis of the sound of the piano by RNNS.

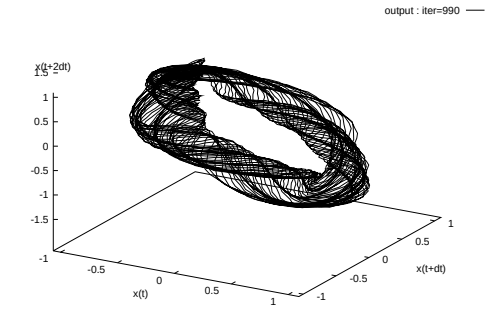

Fig. 14 Phase space trajectory of the data by resynthesis of the sound of the piano by RNNS.

original sound data.

By changing some parameters, such as τ_i and weight connections, new sound data is easily synthesized.

4.3 Resynthesis of the Sound of the Violin

The sound of the piano is a good example of "decaying sound". The other good example is a sound of the violin, which is "lasting sound". In this section, resynthesis of the sound of the violin is shown.

The data were sampled in 16-bit integer format at a sampling rate of 22.05 kHz [3]. The note was G, open 4th string. This part, shown between the two arrows in **Fig. 15**, is used for teacher signals. It has 3,792 samples, 34 peri-

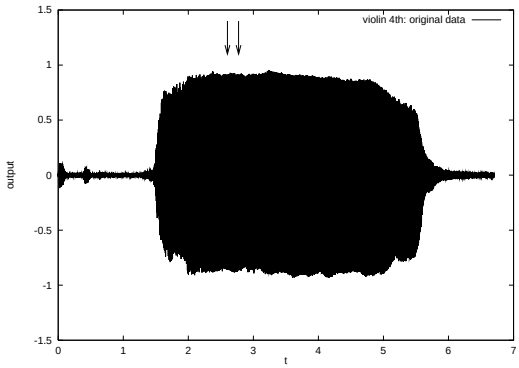

Fig. 15 Sampled data of the sound of the violin, and the training data (denoted between the 2 arrows).

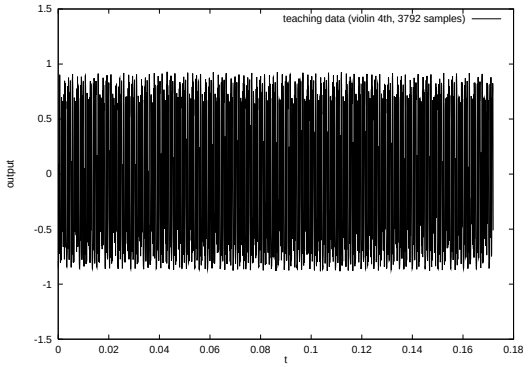

Fig. 16 Teacher signals of the sound of the violin. 3,792 samples (0.17 sec).

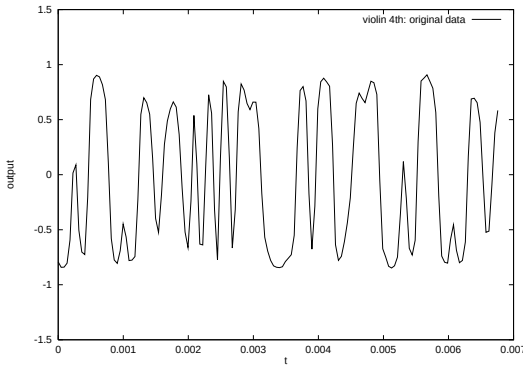

Fig. 17 The beginning of 300 samples of the teacher signals.

ods of 0.17 second (**Fig. 16**). **Figure 17** is the head part of this teacher signals.

To look into the dynamics of the data, 3-dimensional phase space trajectory is also shown (**Fig. 18**). Time lag constant is set to 30 samples.

As an architecture of recurrent neural network, an APOLONN is adopted [1),5)~7)].

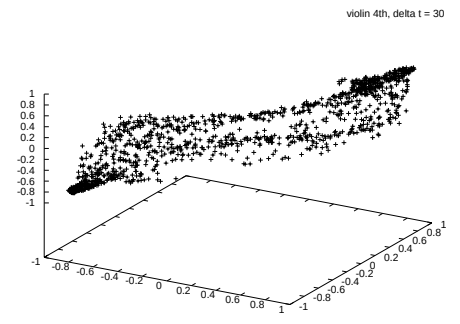

Fig. 18 Phase space trajectory of the teacher signals of the sound of the violin.

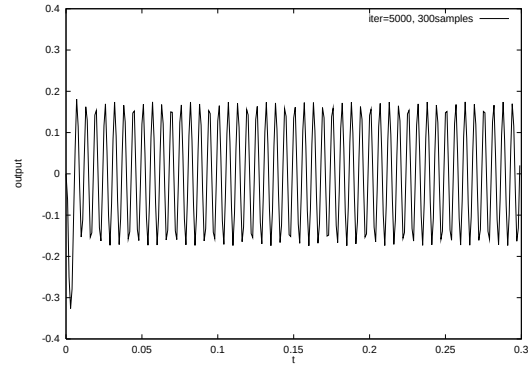

Fig. 19 Resynthesized waveform of the violin by RNNS (iter=5,000, 300 samples).

The learning process was done by a software simulation. 25 pairs of oscillators were used in the simulation. The ratio of the τ_i between two neighboring pairs was 0.9.

One of the leading feature of RNNS is the possibility of resynthesis of the sound using connection weight values. This system is fully described by many differential equations, time unlimited sound resynthesis is possible.

By reading connection weight after learning of 5,000 iterations, new sound is resynthesized. The head 300 samples of the sound is shown in **Fig. 19**.

On the other hand, **Fig. 20** is a waveform of another sound by reading connection weight during learning of only 100 iterations. This figure shows possibilities of a harsh decreasing sound resynthesis without any envelope filters.

5. Concluding Remarks

A new technique of sound synthesis, RNNS (Recurrent Neural Network Synthesis), is presented.

RNNS has two ways of sound synthesis, one

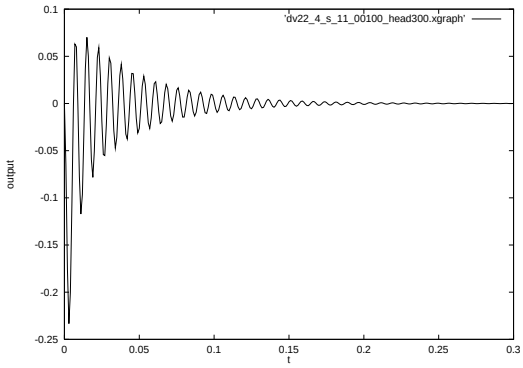

Fig. 20 Resynthesized waveform of the violin by RNNS (iter=100, 300 samples).

sound synthesis variation "hand-tuning model" is by modifying parameters of recurrent neural networks composed of several neurons. It resembles FM synthesis in a sense. It is shown that it is possible to synthesis new sound by recurrent neural network. Since this is composed of relatively small numbers of neurons, it is possible to synthesize realtime in software.

And as another variation, "resynthesis model", a sound synthesis by resynthesis of the sound by RNNS is shown. Using connection weight values after learning process or during learning process, both resynthesis of the original sound and synthesis of new sound are shown.

References

1) Murakami, Y. and Sato, M.: A recurrent network which learns chaotic dynamics, *Proc. ACNN'91*, pp.1–4 (1991).
2) Ohya, K.: A Sound Synthesis by Recurrent Neural Network, *Proc.1995 International Computer Music Conference*, pp.420–423 (1995).
3) Ohya, K.: Sound Variations by Recurrent Neural Network Synthesis, *Proc. 1998 International Computer Music Conference*, pp.280–283 (1998).
4) Pearlmutter, B.A.: Learning state space trajectories in the recurrent neural network, *Neural Computation*, Vol.1, No.2, pp.263–269 (1989).
5) Sato, M.: A learning algorithm to teach spatiotemporal patterns to recurrent neural networks, *Biological Cybernetics*, Vol.62, pp.259–263 (1990).
6) Sato, M., Murakami, Y. and Joe, K.: Chaotic dynamics by recurrent neural networks, *Proc. International Conference on Fuzzy Logic and Neural Networks*, pp.601–604 (1990).
7) Sato, M., Joe, K. and Hirahara, T.: APOLONN brings us to the real world: Learning nonlinear dynamics and fluctuations in nature, *Proc. International Joint Conference on Neural Networks*, San Diego, Vol.I, pp.581–587 (1990).

(Received June 15, 2001)
(Accepted December 18, 2001)

Ken'ichi Ohya received the B.S. degree in physics from the University of Tokyo in 1988. From 1988 to 1992 he worked in YAMAHA Corporation and engaged in research on computer music. He is currently a Lecturor at Nagano National College of Technology. His main research interests include neural networks and nonlinear dynamics, mainly for sound synthesis. He is a member of IPSJ, JNNS and ICMA.