Approximation Neural Network for Phoneme Synthesis

Marius Crisan

Polytechnic University of Timisoara,
Department of Computer and Software Engineering,
Blvd. V. Parvan 2, 300223 Timisoara, Romania
marius.crisan@cs.upt.ro

Abstract. The paper presents a dynamic method for phoneme synthesis using an elemental-based concatenation approach. The vocal sound waveform can be decomposed into elemental patterns that have slight modifications of the shape as they chain one after another in time but keep the same dynamics which is specific to each phoneme. An approximation or RBF network is used to generate elementals in time with the possibility of controlling the characteristics of the sound signals. Based on this technique a quite realistic mimic of a natural sound was obtained.

Keywords: Neural networks, Time series modeling, Phoneme generation, Speech processing.

1 Introduction

Speech synthesis remains a challenging topic in artificial intelligence since the goal of obtaining natural human-like sounds is not yet fully reached. Among the different approaches of speech synthesis, the closest to human-like sounds are obtained by the concatenation of segments of recorded speech. However, one of the main disadvantages of this technique, besides the necessity of a large database, is the difficulty of reproducing the natural variations in speech [1]. There are some attempts towards expressive speech synthesis using concatenative technique [2] but the lack of an explicit speech model in these systems makes them applicable mainly in neutral spoken rendering of text. The other techniques of speech synthesis such as formant/parametric synthesis and articulatory synthesis [3], although they employ acoustic models and human vocal tract models, cannot surpass the results of a robotic-sounding speech. The main difficulty resides in dealing with the nonlinear character of natural language phenomenon. Therefore the need of developing new models able to encompass the dynamics of speech became prominent in the recent years. More research works were invested in nonlinear analysis of speech signals in order to derive valid dynamic models [4], [5], [6], [7], [8]. A promising direction is given by the neural networks dynamic models. There are classic approaches in speech synthesis (text-to-speech synthesis) that use neural networks, but not for generating directly the audio signal [9], [10], [11]. However, neural networks did prove successfully to have the potential of predicting and generating nonlinear time-series [12], [13], [14].

S.C. Satapathy et al. (eds.), *Proc. of the 3rd Int. Conf. on Front. of Intell. Comput. (FICTA) 2014 – Vol. 1*, Advances in Intelligent Systems and Computing 327, DOI: 10.1007/978-3-319-11933-5_1

Different topologies have been studied starting from a feed-forward neural network architecture and adding feedback connections to previous layers [15], [16]. Applications in speech and sound synthesis have also been proposed [17], [18]. In a recent work [19] we have studied the possibility of training a feedback topology of neural network for the generation of three new periods of elemental patterns in phonemes. The phoneme sound was finally generated in a repetitive loop with promising results. In the present work we are interested in extending the ideas of dynamic modeling of speech sounds with neural networks, this time using approximation nets. The approximation or interpolation networks, also known as radial basis function (RBF) networks, offer a series of advantages for time-series prediction due to the nature of the non-linear RBF activation function.

The remainder of the paper consists of the following sections: A nonlinear analysis of the phoneme signals, the RBF network model and the experimental results. Finally, the summary and future researches conclude the report..

2 Nonlinear Analysis of Phoneme Signals

The purpose of nonlinear analysis of phoneme time series is to characterize the observed dynamics in order to produce new time series exhibiting the same dynamics. According to Takens' embedding theorem [20], a discrete-time dynamical system can be reconstructed from scalar-valued partial measurements of internal states. If the measurement variable at time t is defined by $x(t)$, a k-dimensional embedding vector is defined by:

$$X(t_i) = [x(t_i), x(t_i + \tau),\ldots, x(t_i + (k - 1)\,\tau)], \tag{1}$$

where τ is the time delay, and $k = 1\ldots d$, where d denotes the embedding dimension. The reason was to have samples from the original sound signal $x(t)$ delayed by multiples of τ and obtain the reconstructed d–dimensional space. The conditions in the embedding theorem impose $d \geq 2D + I$, where D is the dimension of the compact manifold containing the attractor, and I is an integer. According to the embedding technique, from sampled time series of speech phonemes the dynamics of the unknown speech generating system could be uncovered, provided that the embedding dimension d was large enough. The difficult problem in practical applications is finding the optimal length of the time series and the optimal time delay. However, there are some methods to estimate these parameters [21], [22]. We used the false nearest neighbor method and the mutual information method, for establishing the optimal embedding dimension and the time lag (embedding delay), respectively [23]. The optimal choice depends on the specific dynamics of the studied process. Prediction requires sufficient points in the neighborhood of the current point. As the dimension increases the number of such points decreases. Apparently a high embedding seems advantageous because a sufficiently large value ensures that different states are accurately represented by distinct delay vectors. When the value becomes unnecessarily high the data become sparse and each embedding dimension introduces additional noise. A technique that may provide a suitable embedding for

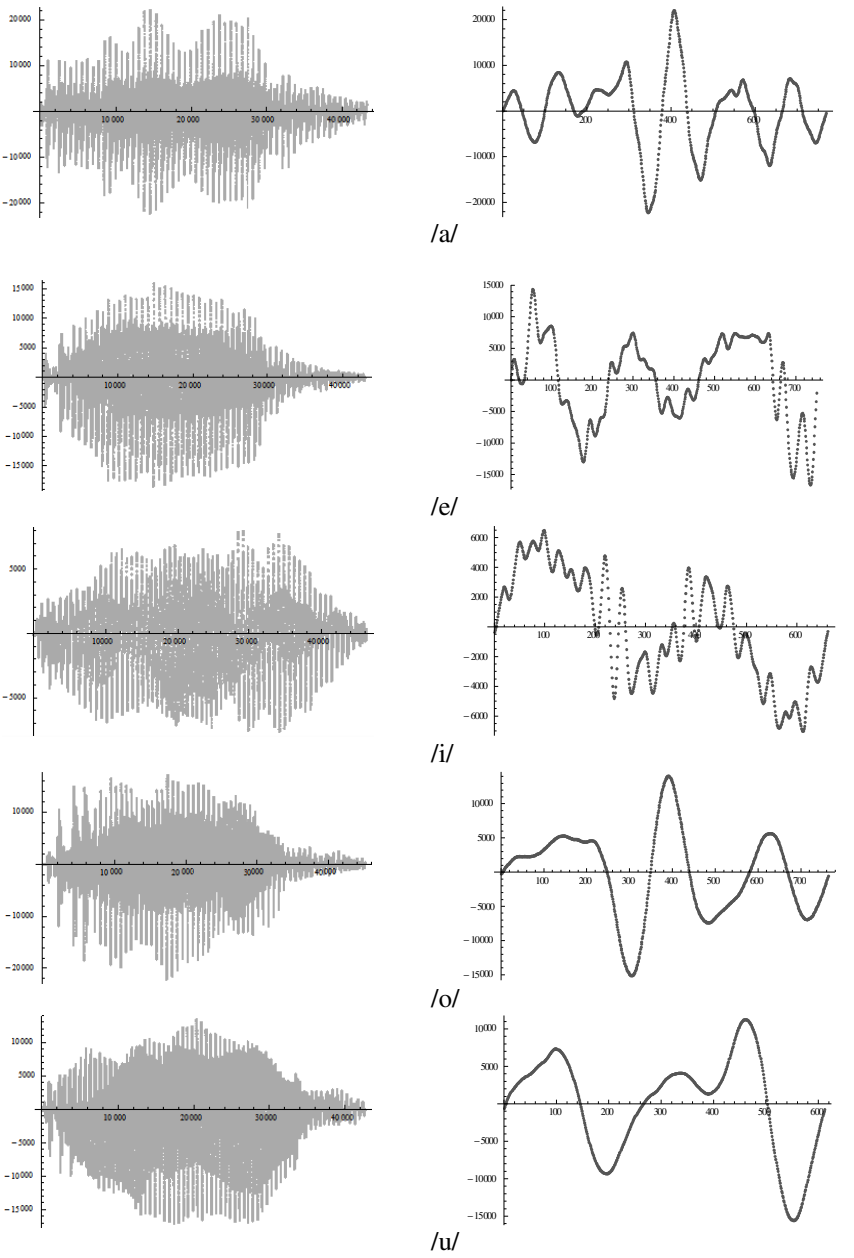

Fig. 1. The waveforms of phonemes /a/, /e/, /i/, /o/, /u/ along with a typical elemental pattern

most purposes is the false nearest neighbors (FNN) method. A provisional value for *d* is first assumed. Then the nearest neighbor of each delay vector is located using the

d-dimensional metric. The same pairs of vectors are then extended by adding one more delay coordinate and are compared, this time using $d + 1$-dimensional metric. If they become far apart then we may consider that nearest neighbor to be false. Regarding the selection of τ, if the value is too large the successive components in a delay vector are completely unrelated. At contrary, if the value is too small the components are nearly identical and therefore adding new components does not bring new information. Out of several techniques available to estimate τ, we selected the minimum mutual information method. Mutual information is a measure of how much one knows about $x(t + \tau)$ if one knows $x(t)$. It is calculated as the sum of the two self-entropies minus the joint entropy. The optimal time lag can be estimated for the point where the mutual information reaches its first minimum.

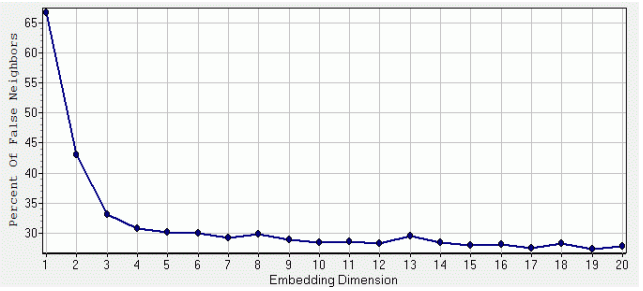

Fig. 2. The percentage of FNNs in dependence on the value of *d*

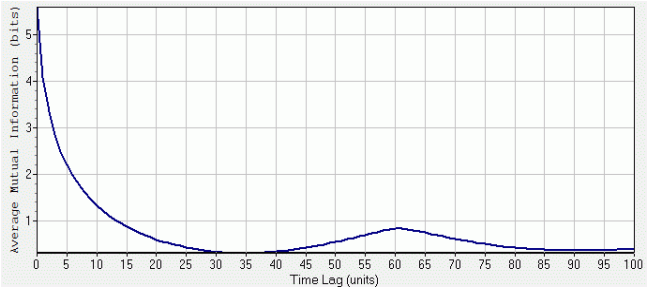

Fig. 3. The average mutual information in dependence on the embedding delay

For experimental purposes, we considered in this work the signals of the main vocal phonemes, /a/, /e/, /i/, /o/, and /u/ pronounced by a male person. The waveforms along with an instance of the corresponding elemental pattern are shown in Fig. 1. The vocal sound data were sampled at 96 kHz with 16 bits. The elemental pattern may be viewed as a basic component in constructing the phoneme signal. If this pattern is repeated in time (concatenated) the phoneme sound can be reconstructed. For these phonemes data we applied the FNN method and the mutual information method. As an example, the percentage of FNNs in dependence on the value of *d* can be seen in Fig. 2 for phoneme /a/. The percentage of FNN should drop to zero for the optimal global embedding dimension *d*. In Fig. 3 it is depicted the average mutual information in dependence on the embedding delay. The optimal time delays result in

the point where the average mutual information reaches the first minimum. The results obtained for the phonemes under study are presented in Table 1.

Table 1. Phonemes nonlinear analysis results

Phonemes/ Vocals	Sample length	Optimal embedding dimension	Optimal time delay	LE ≈
/a/	43304	20	34	42
/e/	42984	14	59	16
/i/	42240	9	24	167
/o/	45120	15	48	53
/u/	42624	10	62	54

After the estimation of d and τ we could proceed with the embedding process. For a convenient exploration of the reconstructed phase-space we constructed the following three–dimensional map:

$$\begin{aligned} x &= x(t) \\ y &= x(t + k) \\ z &= x(t + 2k), \end{aligned} \tag{2}$$

and we selected the points along the z axis for $k = d$. The samples are depicted in Fig. 4 for phoneme /a/. These samples constituted the input vector to the RBF neural network as will be detailed in the next section.

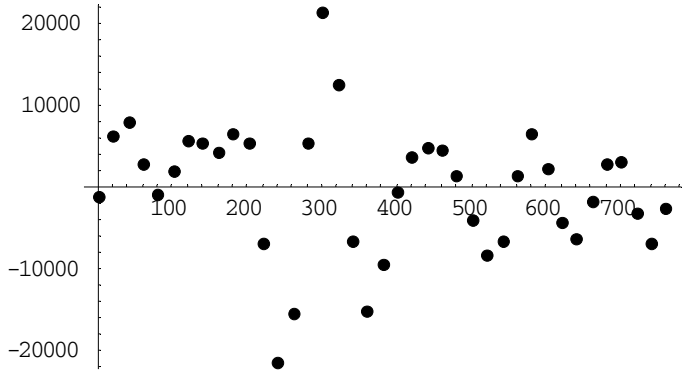

Fig. 4. Training samples taken out from the original time-series

3 The RBF Network Model and Experimental Results

By choosing the appropriate values for d and τ in the embedding process we may have a high degree of confidence that the intrinsic dynamics of the time-series was captured. If we apply the samples on a higher dimension as inputs to an interpolation neural network then we can have a good starting point in generating the phoneme sound. Using a RBF network in this case has several advantages. The network has three layers. The input and output layers have one neuron. The output of the network is a scalar function y of the input (x_1, x_2, …,x_n) and is given by

$$y(x_1, x_2, \dots, x_n) = \sum_{i=1}^{S} w_i g_i(||x - c_i||), \tag{3}$$

where S is the number of neurons in the hidden layer, w_i is the weight of the neuron i, c_i is the center vector for neuron i, and g_i is the activation function:

$$g_i(||x - c_i||) = exp^{-||x - c_i||^2/2\sigma} \tag{4}$$

The Gaussian function has the advantage of being controllable by the parameter σ. In this way, the Gaussian functions in the RBF networks centered on samples enable good interpolations. Small values of σ reflect little sample influence outside of local neighborhood, whereas larger σ values extend the influence of each sample, making that influence too global for extremely large values.

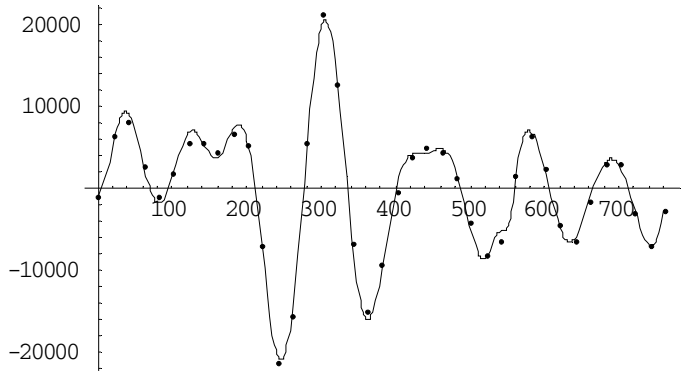

Fig. 5. The approximated elemental

In Fig. 5 it is shown the function approximation along with the data samples. The network consisted of 20 basis function of Gaussian type and was trained with 39 samples. The mean squared error decreased below 0.02 after only 5 iterations. We can observe a good match. The strength of this approach is given by the capability of controlling the width of the basis function and hence the final shape of the approximation. The dynamics of the elementals is still very well preserved if the

number of nodes is not too low. In the next stage, a series of different elementals were generated by varying the width of the basis function according to a random source (for instance a quadratic map). Finally these elementals were concatenated to generate the sound signal. If the concatenation was performed with the same elemental the resulting sound created an artificial impression even if the original elemental was used. The impression of naturalness is not given by the phoneme elemental alone, but by the temporal perception of the slight variations of the elementals in succession.

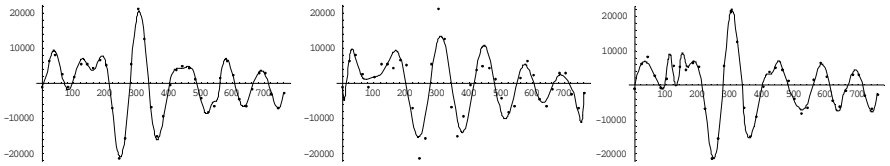

Fig. 6. Three elemental samples generated when σ was controlled by a random source

In Fig. 6, a series of three different elemental are depicted. Slight variations can be observed but with the preservation of the original dynamics as can be observed in the original signal. In conclusion, the method suggested proved to be simple and effective encouraging further researches.

4 Summary

A dynamic method for phoneme synthesis using an elemental-based concatenation technique was proposed. The phonemes' elementals could be generated by an approximation network and the signal parameters can be controlled, at every iteration, through the width of the Gaussian activation function. The resultant elementals were concatenated and finally assembled in the resultant phoneme sound. The sound impression was quite realistic. Future researches in this direction are encouraged by the positive results obtained in this work.

References

1. Edgington, M.: Investigating the limitations of concatenative synthesis. In: Proceedings of Eurospeech 1997, Rhodes/Athens, Greece, pp. 593–596 (1997)
2. Bulut, M., Narayanan, S.S., Syrdal, A.: Expressive speech synthesis using a concatenative synthesizer. In: Proceedings of InterSpeech, Denver, CO, pp. 1265–1268 (2002)
3. Balyan, A., Agrawal, S.S., Dev, A.: Speech Synthesis: A Review. International Journal of Engineering Research & Technology (IJERT) 2(6), 57–75 (2013)
4. Banbrrok, M., McLaughlin, S., Mann, I.: Speech characterization and synthesis by nonlinear methods. IEEE Trans. Speech Audio Process 7(1), 1–17 (1999)
5. Pitsikalis, V., Kokkinos, I., Maragos, P.: Nonlinear analysis of speech signals: Generalized dimensions and Lyapunov exponents. In: Proc. European Conf. on Speech Communication and Technology-Eurospeech-03, pp. 817–820 (September 2003)

6. McLaughlin, S., Maragos, P.: Nonlinear methods for speech analysis and synthesis. In: Marshall, S., Sicuranza, G. (eds.) Advances in Nonlinear Signal and Image Processing, vol. 6, p. 103. Hindawi Publishing Corporation (2007)
7. Tao, C., Mu, J., Xu, X., Du, G.: Chaotic characteristic of speech signal and its LPC residual. Acoust. Sci. & Tech. 25(1), 50–53 (2004)
8. Koga, H., Nakagawa, M.: Chaotic and Fractal Properties of Vocal Sounds. Journal of the Korean Physical Society 40(6), 1027–1031 (2002)
9. Lo, W.K., Ching, P.C.: Phone-Based Speech Synthesis With Neural Network And Articulatory Control. In: Proceedings of Fourth International Conference on Spoken Language (ICSLP 1996), vol. 4, pp. 2227–2230 (1996)
10. Malcangi, M., Frontini, D.: A Language-Independent Neural Network-Based Speech Synthesizer. Neurocomputing 73(1-3), 87–96 (2009)
11. Raghavendra, E.V., Vijayaditya, P., Prahallad, K.: Speech synthesis using artificial neural networks. In: National Conference on Communications (NCC), Chennai, India, pp. 1–5 (2010)
12. Frank, R.J., Davey, N., Hunt, S.P.: Time Series Prediction and Neural Networks. Journal of Intelligent and Robotic Systems 31, 91–103 (2001)
13. Kinzel, W.: Predicting and generating time series by neural networks: An investigation using statistical physics. Computational Statistical Physics, 97–111 (2002)
14. Priel, A., Kanter, I.: Time series generation by recurrent neural networks. Annals of Mathematics and Artificial Intelligence 39, 315–332 (2003)
15. Medsker, L.R., Jain, L.C.: Recurrent Neural Networks: Design and Applications. CRC Press (2001)
16. Kalinli, A., Sagiroglu, S.: Elman Network with Embedded Memory for System Identification. Journal of Information Science and Engineering 22, 1555–1568 (2006)
17. Coca, A.E., Romero, R.A.F., Zhao, L.: Generation of composed musical structures through recurrent neural networks based on chaotic inspiration. In: The 2011 International Joint Conference on Neural Networks (IJCNN), pp. 3220–3226 (July 2011)
18. Röbel, A.: Morphing Sound Attractors. In: Proc. of the 3rd. World Multiconference on Systemics, Cybernetics and Informatics (SCI 1999) AES 31st International Conference (1999)
19. Crisan, M.: A Neural Network Model for Phoneme Generation. Applied Mechanics and Materials 367, 478–483 (2013)
20. Takens, F.: Detecting strange attractors in turbulence. Lecture Notes in Mathematics 898, 366–381 (1981)
21. Small, M.: Applied nonlinear time series analysis: applications in physics, physiology and finance. World Scientific Publishing Co., NJ (2005)
22. Sprott, J.C.: Chaos and Time-Series Analysis. Oxford University Press, NY (2003)
23. Kononov, E.: Visual Recurrence Analysis Software Package, Version 4.9 (accessed 2013), http://nonlinear.110mb.com/vra/