Nonlinear Dynamic Modeling of the Voiced Excitation for Improved Speech Synthesis

Karthik Narasimhan, Jose C. Principe, Donald G. Childers

Computational NeuroEngineering Laboratory, EB 451,
University of Florida, Gainesville, FL 32611
principe@cnel.ufl.edu

ABSTRACT

This paper describes the implementation of a waveform-based global dynamic model with the goal of capturing vocal folds variability. The residue extracted from speech by inverse filtering is pre-processed to remove phoneme-dependence and is used as the input time series to the dynamic model. After training, the dynamic model is seeded with a point from the trajectory of the time series, and iterated to produce the synthetic excitation waveform. The output of the dynamic model is compared with the input time series. These comparisons confirmed that the dynamic model had captured the variability in the residue. The output of the dynamic models is used to synthesize speech using a pitch-synchronous speech synthesizer, and the output is observed to be close to natural speech.

1. INTRODUCTION

Naturalness of synthetic speech is dependent upon how well the voiced excitation is modeled [12]. Present day speech synthesizers approximate the voiced excitation by periodic functions of time, leading to a poor quality of synthetic speech, because the naturalness of speech depends primarily on the variations in the period of the excitation (jitter), [12 which is lost when periodic approximations are made. Hence, there exists a need for a model of the vocal folds that can approximate the glottal flow dynamics more closely.

Existing models of the vocal folds are based on analyzing anatomical and physiological features of the vocal folds and creating mechanical and electrical equivalents that mimic the workings of the vocal folds as closely as possible. Research has shown that the variations exhibited by the glottal flow waveform are chaotic rather than random [9]. The objective of this research is to use techniques developed for dynamic modeling [2], [7] and apply them to the synthesis of naturally sounding speech. In this paper we only address the synthesis of voiced speech.

2. DYNAMIC MODELING

Dynamic modeling [4] is defined as the identification of the mapping f: $\Re^d \rightarrow \Re$ that describes an unknown dynamical system of dimensionality d that produced the time series under analysis. The motivation behind this procedure is to obtain an input-output model for a time series, without the need for a detailed mathematical analysis of the dynamics underlying the time series generation. [2] The major difference between dynamic modeling and classical time series modeling is that in dynamic modeling the system is assumed autonomous. In other words, we seek a nonlinear oscillator which captures the complexity of the time series through nonlinear interactions [4]. A detailed mathematical background of dynamic modeling can be found in references [2], [4] and [11].

Dynamic modeling comprises two distinct components [2]. The first step is to transform the observed time series into a trajectory, according to Takens' embedding theorem [8]. Here we utilized a delay embedding of size 30 with a τ = 3. The next step is to model the trajectory either with global or local models.

Similar to linear modeling, the objective of dynamic modeling is to minimize the mean squared value of the error between the predicted output of the model and the desired response (which is the time series advanced by one step).

$$\varepsilon(n) = x(n+1) - \hat{f}(x(n), w)$$

The minimization can be done in two ways: [11]

In global dynamic modeling, all the points in the reconstruction space are approximated by a single global predictive map f(.).

In local dynamic modeling, the predictive map is composed of R local maps, with each one of the R maps fitting only the neighbors of the current point in the reconstruction space. The complete map then is the union of all the R

$$\text{maps } f(x) = \bigcup_{i = 1, \ldots R} f_i(x) .$$

After training, the dynamic model is seeded with a point from the trajectory of the input time series and its output is the predicted value of the next point in the time series. The predicted output is fed back to the input of the model and the model is iterated autonomously (oscillator). If the dynamic invariants of the synthesized time series coincide with the original, then the model is said to have captured the dynamics of the system that produced the time series [4]. Here we will be comparing the quality of the dynamic modeling by means of the spectrum of the original and synthesized time series, as well as their jitter.

3. INVERSE FILTERING AND PROCESSING OF THE RESIDUE

The input time series to the dynamic model is the glottal flow or excitation waveform, which is obtained from digitized speech (10 KHz, 12 bits) by inverse filtering using linear predictive modeling [5]. The speech signal to be analyzed is divided into 256 samples frames, and for each frame short-term LPC analysis (14 order model) is done to compute the coefficients. The error signal from each frame is determined and stored. The inverse filtering is done pitch synchronously to improve performance [12]. The residual waveform has a lot of high frequency components with an amplitude that tends to be phoneme dependent. However in dynamic modeling of the excitation, the objective is to obtain an output from the model that can be used to synthesize any phoneme, which implies that the input to the model should be phoneme-independent.

To accomplish this, the residue obtained by inverse filtering is subjected to a two step processing. The residue is first low pass filtered to remove some of the noise and high-frequency components that would complicate dynamic modeling. The cut-off frequency of the filter used is fixed at 1KHz as a compromise between simplicity and jitter preservation. A digital Butterworth filter [8] of order 6 is used for the low-pass filtering. The filter selected had almost linear phase.

The next step is to normalize the residue so that it has a uniform amplitude envelope. This is also done pitch synchronously. The filtered residue is segmented, and the minimum amplitude in each segment is determined. The minimum is subtracted from the samples in each segment and each segment is normalized by the maximum. This leads to the residue having flat amplitude envelope on the positive and negative sides.

Jitter is calculated as the difference in samples between the current period and the mean period (using a peak picking algorithm). Quantification of the jitter in the original and filtered waveforms verified that the variability of the original residue is reduced but not lost due to the low-pass filtering. Histograms of the jitter in the original and the amplitude normalized and filtered residue waveforms for the vowel “o” are shown in Figure 1. From the values for the variance, shown in table 1, it is clear that there is minimal loss of variability due to the filtering and normalization operations.

TABLE 1. Jitter in original and normalized residues

	Original Residue	**Normalized Residue**
Mean	3.9e-14	-1.6e-12
Variance	0.913	0.7906

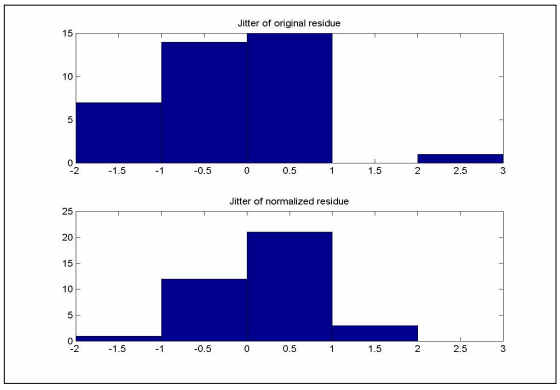

Figure 1 Histograms of jitter (in samples) obtained from the original and filtered residues.

4. IMPLEMENTATION AND RESULTS

Implementation of the Global model

A Time-Delay Neural Network (TDNN) with global feedback [4], trained as a one step predictor, is implemented on

NeuroSolutions, an object-oriented icon based simulator for neural networks [3]. The TDNN topology included a 30 tap delay line at the input with a delay of 3 between taps (implementing the embedding), and two hidden layers of 15 and 10 sigmoidal PEs. The first hidden layer PEs were augmented with a delay line with 5 taps and a delay of 5. The output PE was linear and its output was fed back to the input through a switch to implement the Teacher Forcing procedure [1]. The network was trained using Back Propagation Through Time (BPTT) [1]. A step size of 0.6 and a momentum constant of 0.5 were used to train the network. The percentage of output samples fed back to the output through Teacher Forcing was scheduled during learning. Initially, the network was trained completely with the input time series. After every 20 epochs of training, the percentage of samples per exemplar fed back from the output to the input was bumped up by 10%. The network was trained for 200 passes through the input data, which consisted of about 1500 samples of the filtered and normalized residue, and the error stabilized at 0.005. The trained data was the residue from a single speech segment of a male speaker voicing the vowel “o”.

Results

The output from the network in autonomous mode (i.e. as an oscillator) resembles the original time series exactly in shape (Figure 2). The spectra of the original time series and the network output are similar, indicating that the long term time structure had been captured. Figure 3 shows a zoomed version of the 256 sample FFT of the original and synthesized time series.

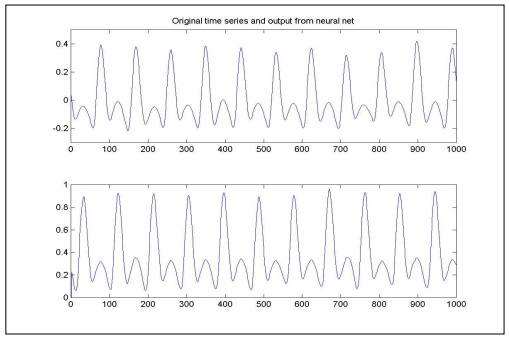

Figure 2. Original and synthetized voiced excitation

The distributions of jitter in the original residue and the output of the network are compared using an histogram and statistics. Table II compares the mean and variance of the original and synthesized time series which shows that the variance is very similar.

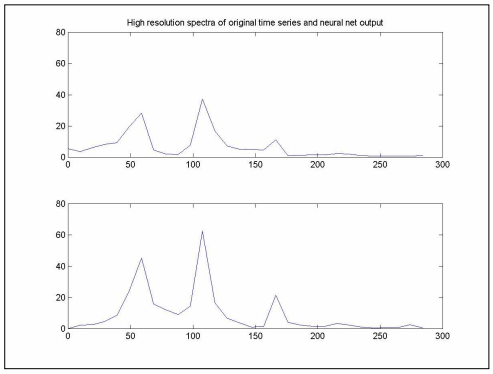

Figure 3 A zoomed (low frequency) spectra of original tim series and synthesized output (Hz).

TABLE 2. Jitter in original and synthesized signals

	Original Residue	**Normalized Residue**
Mean	-1.6e-12	5.4e-12
Variance	0.7906	0.7566

Another consideration in dynamic modeling is the stability of the model. A stable model implies that the quality of the output will not degrade over time. The stability of this non-linear oscillator can be confirmed by comparing the jitter plots of two one thousand point segments from the output time series of the oscillator. The first segment is from the beginning of the synthesis, and the next segment is chosen from the same time series after 50,000 points have elapsed. The two histograms are identical, indicating the stability of the oscillator (figure 4).

Synthesis of speech using the output of the network

An *entire voiced sentence* 'We were away a year ago' was chosen for synthesis. The LPC coefficients and gain for each segment were computed using the autocorrelation method. The sentence was segmented pitch synchronously to obtain the model parameters. The output from the global dynamic model trained with a single voiced phoneme as described above was used as the excitation to this LPC model. A segment of the excitation was synchronized with each speech segment. Care was taken to ensure that the

peak of the excitation coincided with the start of the speech segment. (Pitch Synchronous Analysis). The excitation segment was scaled by the gain and passed through the LPC model to obtain the speech output for the segment.

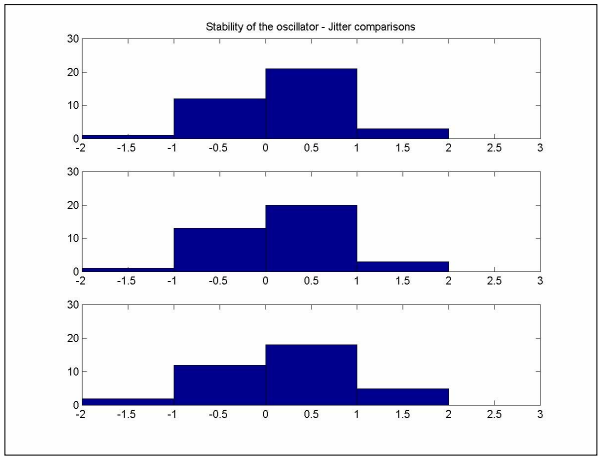

Figure 4. Jitter comparisons between input time series, initial part of output and output after 50,000 generated samples.

Listening tests showed that the quality of synthesized speech is very close to the original speech signal, preserving the naturalness of speech better that the government standard LPC-10 [10] algorithm. The synthesized speech signals are available for listening on the World Wide Web at the Computational NeuroEngineering Laboratory website http://www.cnel.ufl.edu.

5. CONCLUSIONS

Dynamic modeling was used successfully to model the dynamics of the vocal folds, using a waveform-based approach. Outputs that closely resembled the original inverse filtered residue were obtained from the model. It was seen that the jitter distribution remained the same in the outputs from both the models over time, indicating that the dynamic models developed were stable autonomous oscillators.

In spite of the fact that the neural model was trained with a short voiced vowel, the model output was used to generate the vocal excitation for an entire voiced sentence with very good results. The quality of the synthetic speech was judged of better quality than speech generated using an impulse excitation. This indicates a possible application of dynamic modeling in speech compression and high quality speech synthesis. We plan to apply the same technique to generate unvoiced excitations.

ACKNOWLEDGMENTS:

This work was partially supported by NSF grant # IRI-9526049.

REFERENCES

[1]Haykin, S. (1994). "Neural Networks", Macmillan, New York.
[2]Haykin,S., and Principe,J. (1998). "Dynamic Modeling of Chaotic Time Series with Neural Networks", IEEE DSP Magazine, May 1998.
[3]NeuroSolutions User's Manual, NeuroDimensions Inc., Gainesville, Fl 32601.
[4]Kuo, J. M., (1993). "Non-linear dynamic modeling with Artificial Neural Networks", Ph.D. dissertation, University of Florida, Gainesville.
[5]Makhoul, J. (1975). "Linear Prediction : A tutorial review", Proc. IEEE, 63(4), 561-580.
[6]Oppenheim, A., and Schafer, R. (1990) "Discrete-Time Signal Processing", Prentice Hall Inc., Englewood Cliffs, NJ.
[7]Principe, J.C., Rathie, A. and Kuo, J.M. (1992). "Prediction of chaotic time series with neural networks and the issue of dynamic modeling", International Journal of Bifurcations and Chaos, 2(4), 989-996.
[8]Takens, F., (1981) "On the numerical determination of the dimension of an attractor" Ed. Rand, D.A. and Yang L.S., Lecture Notes in Mathematics, 898, 365-381, Springer-Verlag.
[9]Titze, I.R., Baken, R., and Herzel, H., (1993). "Evidence of chaos in vocal fold vibration", Vocal fold physiology: Frontiers in basic science, Singular Publishing Group, San Diego, CA.
[10]Tremain, T.E. (1982). "The Government Standard Linear Predictive Coding Algorithm: LPC 10", Speech Technology (4), 40-49.
[11]Wang, L., (1996). "Local dynamic modeling with Self-Organizing Feature Maps", Ph.D. dissertation, University of Florida, Gainesville.
[12]Wu, C., (1996). "A flexible 3-D model of vocal fold vibrations", Ph.D. dissertation, University of Florida, Gainesville.