A RECURRENT NEURAL NETWORK SPEECH PREDICTOR BASED ON DYNAMICAL SYSTEMS APPROACH

Ekrem Varoglu and Kadri Hacioglu *

Electrical and Electronic Engineering Department
Eastern Mediterranean University, Magusa, Mersin 10, TURKEY

** He is on leave at the Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado, USA*

ABSTRACT

A nonlinear predictive model of speech, based on the method of time delay reconstruction, is presented and approximated using a fully connected recurrent neural network (RNN) followed by a linear combiner. This novel combination of the well established approaches for speech analysis and synthesis is compared to traditional techniques within a unified framework to illustrate the advantages of using an RNN. Extensive simulations are carried out to justify the expectations. Specifically, the networks' robustness to the selection of reconstruction parameters, the embedding time delay and dimension, is intuitively discussed and experimentally verified. In all cases, the proposed network was found to be a good solution for both prediction and synthesis.

1. INTRODUCTION

In the traditional speech prediction, the present speech sample is approximated as a linear or nonlinear function of a fixed number of previous consecutive samples. That is, the prediction of a speech sample at time n is

$$\hat{s}(n) = F(s(n-1), s(n-2), \cdots, s(n-p)) \qquad (1)$$

where p is called the prediction order.

In linear predictive (LP) analysis F(·) is assumed to be linear. LP methods are now well understood and very popular because of their relatively good performance and computational efficiency [1]. However, their success is limited by the degree of linearity among speech samples.

In nonlinear predictive (NLP) analysis F(·) is assumed to be nonlinear. Both theoretical and practical advances in the field of neural networks have activated research on realizing F(·) using a Time-Delay Neural Network (TDNN) [2], a Radial Basis Function Network (RBFN) [3] and a Recurrent Neural Network (RNN) [4].

An alternative nonlinear predictive model based on the Takens' embedding theorem [5] was introduced in [6,7]. Here, speech is assumed to be the output of a deterministic nonlinear, autonomous, dynamical system whether it is voiced or unvoiced. Takens stated that there exists an exact predictive model given by

$$\hat{s}(n) = F(s(n-1), s(n-\tau_E-1), \cdots, s(n-1-(d_E-1)\tau_E)) \quad (2)$$

where τ_E is the embedding time delay and d_E is the embedding dimension, provided that $d_E \geq 2d+1$; d is the dimension of the attractor on that the system evolves. Note that (1) is a special case of (2) with probably suboptimally selected $\tau_E=1$ and $d_E=p$. As a result, the dynamical systems approach provides a more general framework. Here, the problem is the determination of τ_E, d_E and F(·) in some optimal sense. It should be noted that Takens' embedding theorem is an existence theorem and tells nothing about how to find (2).

To the best of our knowledge, in the context of nonlinear speech processing based on the dynamical systems approach, only MLP and RBF networks were used for realizing F(·) in (2). A detailed list of related work recently conducted in this field can be found in [7]. Here, in contrast, we selected a fully connected RNN followed by a linear combiner [4] motivated by the fact that a recurrent network introduces an internal (or implicit) memory of infinite length but of fading nature in addition to the external memory with a size determined by the embedding. With more past information relevant to prediction, we expect a better performance. In addition the implicit network memory is also expected to make the predictor more robust to the improper selection of the embedding parameters.

The joint optimization of d_E, τ_E and F(·) is a rather difficult task, if not impossible. In this paper we adopt the following frame-by-frame analysis approach. Despite the limited size of the analysis frame (due to stationarity requirements) the time delay τ_E is taken as the first minimum of a nonlinear measure of the time series called the mutual information [8] and the embedding dimension d_E is selected using the correlation dimension as an estimate for d [9]. After the choice of the embedding parameters (the optimal choice is still an open problem) we approximate F(·), in the sense of least squares, using neural networks.

2. APPROXIMATION OF THE NONLINEAR MAPPING F(·) USING A RECURRENT NEURAL NETWORK

The recurrent neural network used to approximate F(·) in (2) consists of three layers; the input layer, the processing layer and the output layer. The input vector is the concatenation of L external inputs, a biased input and delayed signals fed back from the processing layer which consists of N units with bipolar sigmoid activation function. The output layer has a single unit

which linearly combines the processing layer outputs. If the network is fully connected, it has a total of $N^2+(L+1)N$ connections from the input layer to the processing layer and N connections from the processing layer to the output layer. A RNN predictor with L=2 and N=3 is exhibited in Figure 1. Each delay unit introduces a delay of τ_N samples. This form of predictor has been extensively studied in [10] for both formant and pitch prediction using the traditional approach. The external inputs were formed by taking p (typically 8-10) successive speech samples for formant prediction and 1-3 samples at a distance equal to the pitch period for pitch prediction. In the context of (2), these predictors are probably suboptimal and require a larger dimension than that is suggested by the embedding theory. The following equations describe the operation of the network:

$$Y(n)=f(W_F\, Y(n-\tau_N)+W_i\, R(n)) \qquad (3)$$

$$\hat{s}(n) = W_o Y(n) \qquad (4)$$

where $Y(n)=[y_1(n),y_2(n),\cdots y_N(n)]$ is the state vector, $R(n)=[1,s(n-1),s(n-\tau_E-1),\cdots,s(n-(d_E-1)\tau_E-1)$ is the augmented input vector, W_f is the feedback weight matrix, W_i is the input weight matrix, W_o is the output weight matrix and τ_N is the network time delay.

Adaptation of W_f and W_i is performed using the real-time recurrent learning (RTRL) algorithm [11] and W_o is adapted using the well known LMS algorithm [12]. At the processing layer, being hidden, error signals are not available and they are generated by backpropogating the output error through the linear combiner.

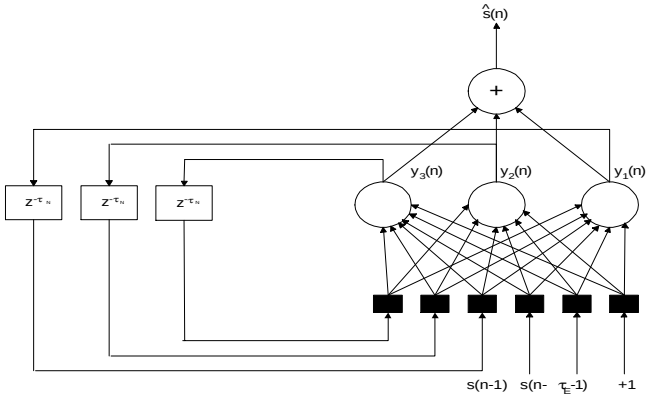


Figure 1. Proposed Recurrent Neural Network.

It is clear that for $\tau_N=\tau_E$,

$$\hat{s}(n+1) = F(s(n), s(n-\tau_E), \cdots, s(n-(d_E-1)\tau_E), \cdots) \qquad (5)$$

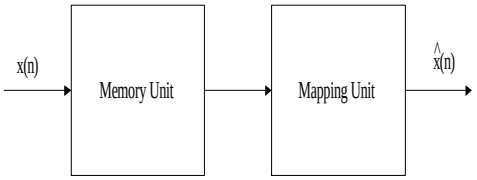
Due to the feedback, the memory of the network is infinite but of fading nature. The authors conjecture that this makes the network more robust to the slightly false estimations of d_E. On the other hand, this may allow selection of d_E smaller than that required. It is also conjectured that the selection of $\tau_N \neq \tau_E$ increases the robustness of the network to the estimation errors in τ_E by providing interleaved samples that are used in prediction.

3. COMPARISON OF METHODS WITHIN A UNIFIED FRAMEWORK

3.1 Speech Analysis

For a systematic comparison of the approaches introduced above we adopt the framework illustrated in Figure 2 for prediction [13]. The framework consists of two parts; i) The memory unit, ii) The mapping unit.

Figure 2. A framework for prediction



The success of prediction depends on the following:

(a) the amount of information kept in the memory unit relevant to prediction

(b) the ability of the mapping unit to realize the actual relation between the information in the memory and the predicted value.

Firstly we discuss the amount of past information that can be considered relevant to prediction in the case of speech signals. As is well known, in voiced speech prediction we distinguish between two types of correlations; i) short term correlations and ii) long term correlations. The former indicates the dependency of adjacent samples and the latter is a result of the periodicity in speech signals. The periodicity implies similarity among samples which are one period apart. Therefore, for a successful speech prediction the memory unit should span a time interval of length at least one pitch period.

Secondly we discuss the structure of the memory in all approaches. For this purpose, we define the memory depth, τ_D, as the duration of the signal history (in samples) stored in the memory unit and the memory resolution, τ_R, as the reciprocal of the delay between the signal samples stored in the memory unit. In the traditional prediction approach (equation (1)) the memory unit can be considered as a tapped delay line containing the most recent p speech samples. That is $\tau_D=p$ and $\tau_R=1$. On the other hand, in the dynamical systems approach (equation (2)) the memory unit consists of d_E samples each separated by τ_E samples. Here, $\tau_D=d_E\tau_E$ and $\tau_R=1/\tau_E$. In the pitch-formant approach, the memory unit is made of two blocks. The first block contains the p most recent speech samples each separated by a unit delay. The second block consists of a few samples around exactly one pitch period away from the predicted sample again each separated by a unit delay. As a result the overall memory depth, τ_D, is slightly greater than the pitch period and each block has resolution $\tau_R=1$.

We conclude that to meet the requirement (a) mentioned above in the traditional approach ($\tau_E=1$) we need a relatively large value of p (or d_E) depending on the pitch period of the speech

signal. However, a smaller value of d_E is possible in the dynamical systems approach by using a relatively large τ_E but at the expense of a lower resolution. It appears that, for a given memory depth, the number of samples in the memory unit and the resolution are two conflicting parameters in the sense that our aim is to select d_E as small as possible (to avoid the curse of dimensionality), but yet keep the resolution as high as possible (not to miss samples relevant to prediction). In the traditional pitch-formant approach, with the use of two different blocks the resolution is kept high (at least in each block) without sacrificing too much from dimensionality. Since information corresponding to the short and long term correlations is included in the memory unit with a high resolution, that structure is expected to show the best performance as a predictor.

Thirdly we demonstrated the validity of our expectations through some simulations using voiced speech frames taken from male and female speakers. Speech waveforms were low-pass filtered at 3.4 kHz cut-off frequency, sampled at 8 kHz and stored at 16 bits. Each analysis frame consisted of 256 samples. The results were presented for the memory structures with a linear mapping unit by plotting the mean value of the prediction error (MSE) in dB with respect to the memory depth.

Figure 3 shows the results corresponding to a female speaker. Here, the pitch period of the analysis frame is 24 samples. The MSE of the pitch-formant approach is shown as a baseline. The so-called formant memory block consists of the most recent 8 samples and the so-called pitch memory block consists of 3 consecutive samples centered at the pitch period. The other two plots are for memory resolutions τ_R=1 and τ_R=1/3. The poor performance of the lower resolution and the good performance of the pitch-formant approach are obvious. However, all memory structures have comparable performances when all relevant information is included in the memory as illustrated for τ_D > pitch period.

Finally we extend the discussion to the nonlinear mapping unit and present several experimental results. As mentioned, the possible networks that can be used are MLP, RBF and RNN. The first two are feedforward networks. Being static, they do not contribute to the memory content of the predictor. The RNN, in contrast, has two implicit effects on the memory content:

1) the effective memory depth is increased because of the infinite fading memory (compensates for relatively small d_E).

2) the missing samples in the memory are partially supplied by properly selecting the network delay, τ_N different from the embedding delay,τ_E (compensates for low resolution).

Thus a predictor using an RNN as the mapping unit is expected to outperform the predictors using MLP and RBF in cases where;

(i)the memory depth does not cover a full pitch period

(ii)the memory depth covers a pitch period but at a low resolution.

In other words, for a given performance RNN allows a relatively small d_E with a relatively large τ_E, provided that $\tau_N \neq \tau_E$ is properly selected. However, in cases where all relevant information is in the memory unit, we expect all structures to give similar performances in terms of the prediction error. However, in case of linear mapping the prediction performance is expected to be worse when compared to its nonlinear counterparts.

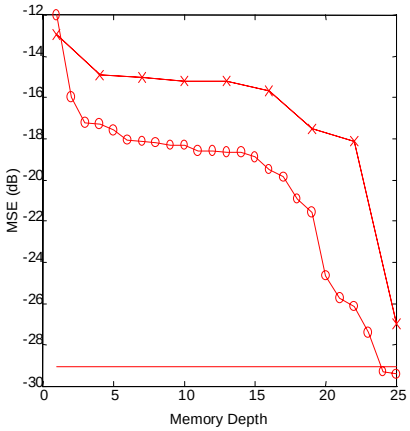


Figure 3. Comparison of different memory structures with the linear mapping unit;female speech.
o: τ_E=1, x: τ_E=3, _: pitch-formant memory block (p=8, M=3).

Figure 4 shows the performances of RNN predictor (with a unity network delay τ_N=1), MLP predictor and a linear predictor with respect to the memory depth. Note that all of the predictors have memory resolution τ_R=1. In both RNN and MLP networks the number of neurons was set to 4. In all simulations the learning rate was set to 0.1 and kept fixed and the networks were trained for 1000 epochs. The results were averaged over 10 trials. The performance of the RBF network at comparable complexities was found to be very poor with the learning algorithm that was implemented. So, the RBF network results are not shown to avoid overcrowded plots. In Figure 4 the baseline corresponds to the performance of the pitch-formant RNN [8] with 8 most recent samples and 3 samples around the pitch period. Its better performance at relatively lower complexity (11 input samples) is obvious. Note that the RNN significantly outperforms the MLP network for very short memory depths (1-5 input samples). Their performances becomes comparable as all short term correlated samples are included in the memory. This remains until the implicit memory of RNN starts to capture the samples around the pitch period, though, the external memory depth is still less than the pitch period. Their performances meet again as the external memory covers the full pitch period. Again this behaviour is common to all speech frames.

Nevertheless, because of its infinite memory, the RNN outperforms the MLP network if it is not stuck at a local minimum with a relatively bad performance. It is a general belief that learning algorithms in the RNN explores a more complex surface. To check the frequency of occurrence of the above phenomena we run the RNN and the MLP networks over 200 frames of speech taken from 4 speakers. In all simulations τ_E was chosen as the first minimum of the mutual information function and d_E was chosen as the smallest integer greater than the estimated attractor dimension (d_E>d+1), even though the estimates are not expected to be accurate due to small size of

analysis frame. The results are presented in Figure 5. Here, the solid line represents the boundary where the performances of the two predictors are equal. Note that the RNN performs better than the MLP for almost all the frames. So, we conclude that it is very safe to use the RNN.

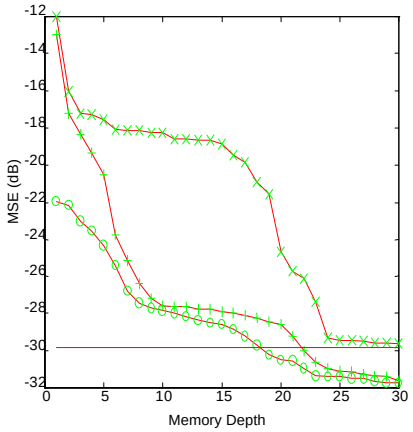


Figure 4. Comparison of RNN vs. MLP (N=4, τ_E=1, τ_D<Pitch Period).
o: RNN, +: MLP, x: Linear, _: Pitch-formant RNN.

Another set of simulations for the male speaker with pitch period 56 was performed to check the effect of the memory resolution on the performance. For this purpose, we fixed the memory depth to a value slightly larger than the pitch period(τ_D=60). We changed the embedding delay (equivalently, the memory resolution) from 1 to 10. Since $\tau_D=(d_E-1)\ \tau_E$, for each value of τ_E we have a different value of d_E. Figure 6 illustrates the performances of linear, MLP and RNN (for both $\tau_N=\tau_E$ and $\tau_N\neq\tau_E$) predictors. The irregular behaviour of all predictors is due to the different combinations of samples used for embedding with varying degrees of relevance to prediction. To get a better understanding of this phenomena consider the following cases. When τ_E=7, the samples used for prediction are n-1, n-8, n-15, n-22, n-29, n-36, n-43, n-50, n-57. Here, the last sample (n-57) is exactly one pitch period away from the predicted sample. When τ_E=6, the samples used for prediction are n-1, n-7, n-13, n-19, n-25, n-31, n-37, n-43, n-49, n-55. Notice that although more samples are used for prediction at a slightly higher resolution the performance is slightly worse due to the fact that the sample which is exactly one pitch period away is not included in the prediction. According to Figure 6, the variations in the performance of RNN with $\tau_N\neq\tau_E$ is smaller than those of others, indicating the robustness of RNN to the selection of embedding delay and dimension parameters. Although we have ysed τ_N=1 here, the choice of the optimal network delay for a given embedding is still an open problem.

3.2 Speech Synthesis

In the preceding section, the one step ahead prediction paradigm is applied to obtain the nonlinearity F(·). In this section, assuming that F(·) has captured the underlying dynamics of the attractor, we use it for speech synthesis. The synthesizer is operated in an autonomous manner by seeding it with an all zero state vector and feeding the output to the delay line which is constructed the same as the memory structure employed in prediction. Here, the fundamental question is "Which predictor implementation is the most suitable for synthesis ?" We claim that the RNN with time delay embedding appears the most promising since it offers a lower dimension for the state space. Lower dimensionality helps

a) to avoid curse of dimensionality.

b) to decrease the degree of mismatch in the analysis and synthesis models.

c) to have a faster convergence to the reconstructed attractor.

d) to reduce the computational complexity

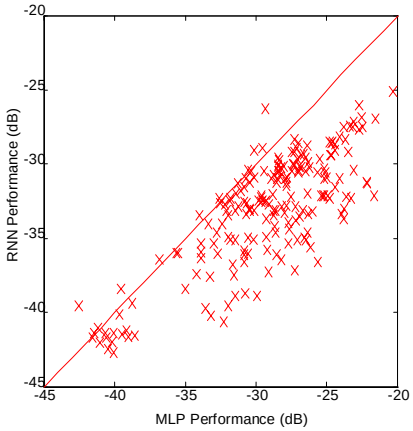


Figure 5. Robustness of RNN compared to the other schemes (N=4)

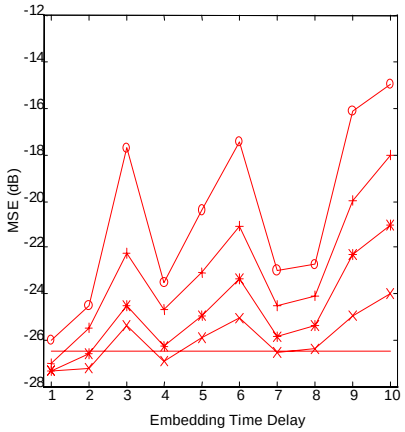


Figure 6. Effect of Resolution on the performance of all structures (N=4)
x: RNN ($\tau_E\neq\tau_N$), *: RNN ($\tau_E=\tau_N$), +: MLP,
o: Linear, _: pitch-formant RNN.

The mismatch concept in b), reflects the difference among the samples used for the prediction and synthesis. In the prediction, the time delay line is fed using samples taken from the original signal. However, during synthesis the reconstructed samples are used.

We did several simulations to justify our claims. The first set of simulations were carried out with clean speech. In the second set, a noisy speech at 15dB was used. Noisy speech was generated by adding zero mean Gaussian noise with a variance accordingly adjusted. The test with noisy speech is aimed to show which network exhibits good generalization. The networks used are an RNN with N=4, d_E=4, τ_E=2 and τ_N=1, and two MLP networks with N=4, d_E=8,12 and τ_E=1. After the predictive analysis, the networks were used to synthesize speech.

Figure 7 shows the noise free waveform together with the synthesized waveforms.. The following conclusions can be reached from the results:

(a) The RNN has converged very fast to the attractor that is very similar to the original.

(b) The MLP network with d_E=8 has showed fast convergence but with an attractor dissimilar to the original (failed to capture the underlying dynamics)

(c) The MLP network with d_E=12 has converged slowly to an attractor that looks better than the attractor in (b) but still worse than the attractor in (a).

In addition results obtained in the noisy case has shown us that the MLP network has captured the specific details contributed by the noise(overfitting) and hence is not a very general network structure for synthesis.

The results here clearly illustrate that the RNN with the delay embedding is an appealing network for future research in nonlinear speech processing that includes speech analysis, speech synthesis and even speech enhancement.

4. CONCLUSIONS

In this study, a nonlinear predictive model of speech, based on the method of time delay reconstruction, has been presented. A fully connected recurrent neural network followed by a linear combiner has been proposed to realize the model. Its prediction performance has been compared to other structures in a unified framework. These comparisons have shown that the proposed network offers satisfactory prediction at relatively lower input dimensions and shows more robust behaviour to the selection of embedding parameters, namely the embedding time delay and the embedding dimension, compared to MLP and, equivalently, to RBF networks. The advantages of a lower input dimension has been stressed in the context of speech synthesis and justified via some simulations for both clean and noisy speech. As a result the proposed approach has been found promising for applications that might employ nonlinear speech processing.

5. REFERENCES

[1]Kondoz, A. M. *Digital Speech: Coding for low bit rate communication systems*. J. Wiley, Chichester, New York, 1994) .

[2]Thyssen, J., Nielsen, H., and Hansen, S.D "Non-linear short-term prediction in speech coding ," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-94, April 1994, vol.1, pp. 185-188.

[3] Diaz-de Maria F. and Figuerias Vidal, A. R. "Nonlinear prediction for speech coding using radial basis functions," in Proc.IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-95, May 1995, pp. 788-791.

[4]Wu, L. and Niranjan, M. "On the design of nonlinear speech predictors with recurrent nets," in Proc.IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-94, April 1994, VOL.2, pp. 529-532.

[5] Takens, F. "Detecting Strange Attractors in Turbulence" in *Dynamical systems and turbulence*, vol. 898 of Springer, Lecture Notes in Mathematics, Berlin: Springer, 1981 , pp. 366-381.

[6]Tishby, N. "A dynamical systems approach to speech processing," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-90, April 1990, pp. 365-368.

[7] Kubin, G. "Nonlinear Processing of Speech" in *Speech Coding and Synthesis*, 1995,(Elsevier Science B.V), pp. 557-610.

[8]Fraser, A.M. and Swinney, H. L. "Independent coordinates for strange attractors from mutual information," *Physical Rev. A*, 1986, **33** (2), pp. 1134-1140.

[9]Grassberger, P. and Proccacia, I. "Characterization of strange attractors," *Physical Rev. Letters*, January 1983, **50** (5), pp. 346-349.

[10] Varoglu E., and Hacioglu K "Nonlinear formant-pitch prediction using recurrent neural networks," in Proc. VIII European Signal Processing Conference, EUSIPCO-96, September 1996, pp. 463-466.

[11] Williams, R.J. and Zipser, D. "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, 1989, pp. 270-280.

[12] Widrow B. and Stearns, S. D.: "Adaptive Signal Processing,".(Prentice Hall, Englewood Cliffs, N.J, 1985.

[13] Mozer, C.M. "Neural Net Architectures for Temporal Sequence Processing," in *Time Series Prediction: Forecasting the Future and Understanding the Past*, vol. XV of Sciences of Complexity Proc., 1993, Addison Wesley, pp. 243-261.

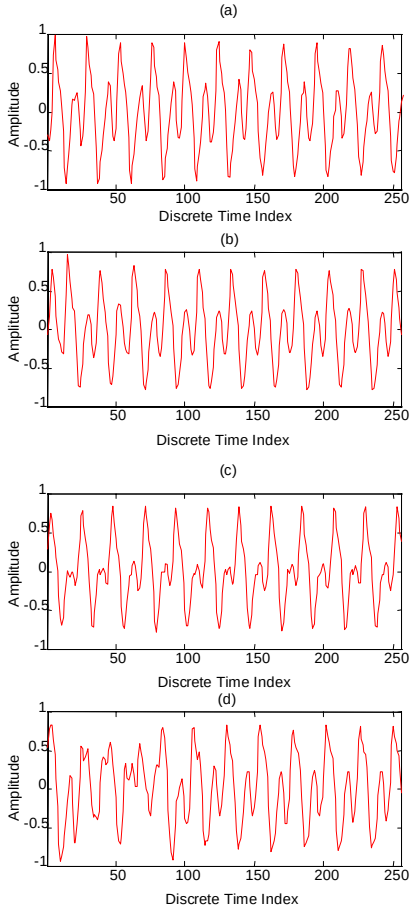


Figure 7. Synthesized speech waveforms.
(a) Original waveform (b) RNN with N=4, d_E=4, τ_E=2, τ_N=1
(c) MLP with N=4, d_E=8, τ_E=1 (d) MLP with N=4, d_E=12