Neural modeling of speech coding prediction

Sn.Pleshkova- Bekjarska [1] M.Momtchedjikov[2]

Abstract: **The neural networks are used in many applications as picture processing, computer vision, control system, signal processing etc. It is possible to use the neural networks for speech recognition. The goal of this article is to use the neural networks modeling in speech coding prediction.**

Keywords: **speech, speech coding, speech prediction, neural networks, predictive neural networks.**

I. INTRODUCTION

There are two general methods for using neural networks in speech coding: direct and hybrid. In the first case speech signal is applied to neural networks and it is supposed that the neural networks whole coding algorithm after an effective learning [1, 2]. In this case it is possible to achieve a compression of speech information if the numbers of neural network outputs are less than the inputs in a speech frame. The advantages of these methods are that the coding depends only from neural networks structure and his performances. The disadvantages are that it is not possible to use the specific characteristics of speech signals (sample frequency, pitch, voiced/unvoiced separation etc.).

These advantages are not present in the second group of hybrid methods, in which it is used the classical speech coding algorithms (LPC, CELP etc.), but some (or all) of the parts of these algorithms are modeling with a suitable neural network structure. The advantages in these cases are that it is possible to use the specific characteristics of speech signals. The disadvantages are the more complex and not regular structure of speech coders and decoders.

The goal of this article is to apply the second hybrid method for speech coding prediction, which is a part of many classical speech coding methods (LPC, CELP, FS-1016 etc.).

[1]Sn.Pleshkova is with the Faculty of Communications and Communications Technologies, Technical University, Kliment Ohridski 8, 1000 Sofia, Bulgaria,
E-mail: snegpl@tu-sofia.bg.
[2]M.Momchedjikov is with the Faculty of Communications and Communications Technologies, Technical University, Kliment Ohridski 8, 1000 Sofia, Bulgaria,
E-mail: snegpl@tu-sofia.bg.

II. NEURAL NETWORK STRUCTURE

The proposed neural networks structure is shown in Figure 1.

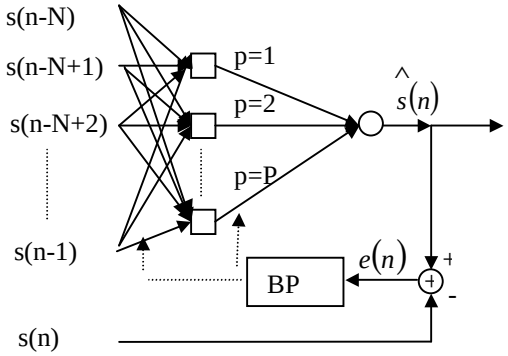

Fig.1. Neural networks structure.

It is chosen as a neural networks structure a two layers perception. The first layer is a hidden layer, which serve for a transformation of each input speech frame $(s(n-1) \div s(n-N))$ with dimension N with correspondence of a given numbers of $p = 1 \div P$ prediction elements (coefficients).

The second layer calculate the current value of predicted speech sample $\hat{s}(n)$, which is compared with real speech sample $s(n)$. The error $e(n)$ is used in learning algorithm. It is chosen the back propagation (BP) as a learning algorithm.

III. NEURAL NETWORK ANALYSIS

The neural network characteristics definitions are made in correspondence with speech prediction conditions. The number N of input samples in speech signal $s(n)$ depend from speech frame dimensions NF, which is used in speech coding method. Usually from coding standards sample frequency speech signal statistics etc., the speech frame dimension is chosen $NF = 160 \div 240$ samples. For a right neural network learning of first hidden layer with sufficient representatives in the input learning sequence, it is necessary to satisfy the condition:

$$N \le NF . \qquad (1)$$

The value of N chosen from (1) must satisfy the following contradicting requirements: the reducing of N lead to increasing of number of representatives in learning sequence, but the prediction accuracy decrease and the necessary time of learning increase. In practical cases N is chosen from (1) with estimation of real prediction error. It can be shown from

experimental tests, that a chosen suitable value of number *N* of input samples is $N = 20$.

The number *P* of neurons (cells) in second layer also depends from speech coding standards and is chosen usually in range of:

$$P = 10 \div 16. \qquad (2)$$

The analysis of proposed neural network structure from Fig.1 can be made with representing of the current speech frame to neural network as a vector:

$$\vec{S}_n = [s(n-1), s(n-2), \ldots, s(n-N)]^T. \qquad (3)$$

As a result of neural network action the output of neural network $\hat{s}(n)$ is the current predicted samples from the input vector $\vec{S}_n$:

$$\hat{s}(n) = F(\vec{S}_n), \qquad (4)$$

where

F is whole activation function of neural network.

Since the structure of neural network is composed from two layers the whole activation function *F* can be represent with the functions of activation of the first the second layer:

$$F = G_W . H_a, \qquad (5)$$

where

G_W is activation function of first hidden neural network layer with vector $\vec{w}$ of weights;

H_a - activation function of second neural network layer with vector $\vec{a}$ of weights.

The transformation of input vector $\vec{S}_n$ can be represented with respect of (4) end (5):

$$\vec{Z}_n = G_W(\vec{S}_n), \qquad (6)$$

$$\vec{s}_n = H_a(Z_n), \qquad (7)$$

where

Z_n is the output vector of the first layer of neural network $\vec{Z}_n = [z_1(n), z_2(n,), \ldots, z(n)]$.

From the expressions (5), (6) and (7) it can be shown, that the current predicted speech sample $\hat{s}(n)$ depend from whole neural network activation function *F* and from corresponding activation functions G_W and H_a of the first and second neural network layer.

The particular activation functions G_W and H_a can be chosen linear or nonlinear (for example sigmoid function) and corresponding neural network, work as a linear or nonlinear predictor of speech signal. This is an advantage of neural network prediction in correspondence of classical methods for speech predictions, which are all only linear methods.

The weights in the vectors $\vec{w}$ and $\vec{a}$ for corresponding first and second layer of neural network can be calculated in the learning phase of neural network by using minimization of quadratic prediction error $e(n)$:

$$e(n) = \sum_n \left(s(n) - \hat{s}(n) \right)^2. \qquad (8)$$

The error $e(n)$ from expression (8) is calculated as a sum of all possible representatives in learning sequence in each given speech frame. The number of these representatives in accordance with expression (1) is:

$$n = 1 \div (NF - N) \qquad (9)$$

The error expression can be generalized for all frames $i = 1 \div L$, for which the learning is made:

$$e_i(n) = \sum_i \sum_n \left(s_i(n) - \hat{s}(n) \right)^2. \qquad (10)$$

It is possible to make a substitution of the expressions (6) and (7) in the expression (10):

$$e_i(n) = \sum_i \sum_n (s_i(n) - H_{ai} . G_W(s_i(n))). \qquad (11)$$

The expression (11) can be used in the first phase of learning for calculating the weights $\vec{w}$ of the first hidden layer of neural network.

At this phase of learning the goal of second layer learning of weights in vector $\vec{a}$ is to improve the precisions of calculating the weights in vector $\vec{w}$ for the first hidden neural network layer using all presented speech frames in learning sequences.

The second phase of learning algorithm used the calculating in the first phase weights in vector $\vec{w}$ unchanged. Only the weights in vector $\vec{a}$ are changed of updated in this phase of learning. The second phase of learning is made only with the representatives in the current speech frame. This second learning can serve as prediction in speech coding of current speech frame. The reason for this are the expressions (6) and (7) shows that current predicted speech sample $\hat{s}(n)$ is directly calculated from the activation function H_a of the second layer of neural network. In H_a are the weights of vector $\vec{a}$ and this weights can be tread as prediction weights or coefficients and the prediction can be linear or nonlinear depending of choose of activation functions H_a and G_W, respectively for second and first layer of the neural network.

The calculation of error in the second learning phase is made only for the representatives in the current speech frame. It is possible to use the equation (8) modified with index *i* to show the number of speech frame:

$$e_i(n) = \sum_n (s_i(n) - \hat{s}_i(n)), \qquad (12)$$

but the sum is only in the current speech frame.

The effectiveness of the prediction with this proposed neural network can be calculated and estimated with the definition of the so called prediction gain:

$$G_i = 10\log_{10}\left(\sum_n s_i(n)^2 / \right.$$
$$\left. / \sum e_i(n)^2 \right). \qquad (13)$$

IV. NEURAL NETWORK TESTING

The purposed neural networks have been simulated with Matlab. The input speech signal (Fig.2) is a pronunciation of letter "a" as a fest signal for neural network. It is shown at the Fig.3 an input speech frame. The sample frequency is $f_s = 11025Hz$, the frame dimension *NF*=240 samples.

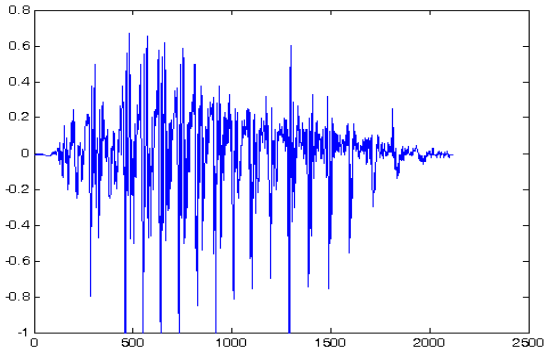

Fig.2. Input speech signal.

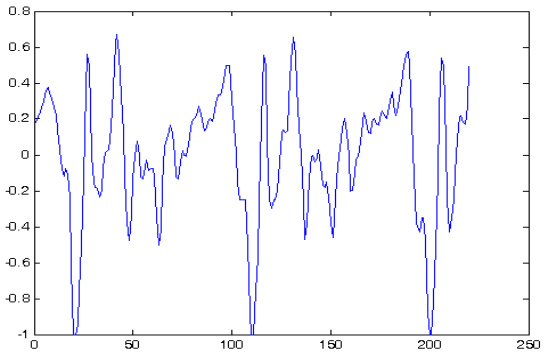

Fig.3. A frame of input speech signal.

At the Fig.4 is an assembly plot of input frame (solid line) and predicted frame (_. line) as an output of neural network. At the Fig.5 is an assembly plot with the same signal as at Fig.4, but here it is added the plot of error (x line).

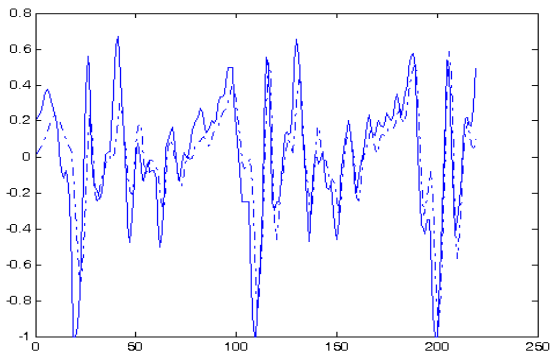

Fig.4. Input and predicted frame as an output of neural network.

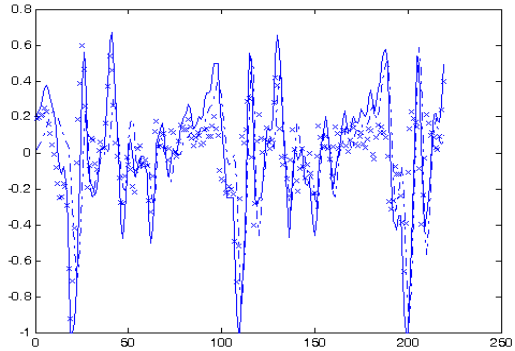

Fig.5. Input, predicted frame and error signal.

V. CONCLUSION

The test of proposed neural network show the possibility to modeling of the speech prediction, which is frequently used in speech coding methods. As a future work it is necessary to investigate the whole properties of this neural network: the error minimization, the complexity of software or hardware realization etc.

REFERENCE:

[1] S. Morishima, H. Harashima and Y. Katayama. Speech Coding based on a multiplayer neural network. IEEE Transac-tion on Signal processing. Vol.40, No.4, pp.429-432, 1990.

[2] Br. Verma and Vallipuram Muthukkuma-rasamy. Speech Compression for VOIP: Neural Network Vs. G723.1. School of Information Technology. Griffith University, Australia.

[3] J. Thyssen, D.S. Hansen Using Neural Networks for Vector Quantization in low Rate Speech Coders, IEEE Transactions on Signal Processing, Vol.4, 1993.

[4] S.C. Ahalt, A.R. Krishnamurthy, D.E. Melton, and P. Chen Neural Networks for Vector Quantization. IEEE Yournal of Selected Areas in Communications, Vol. 8, No. 8, pp. 1449-1457, 1990.

[5] E. Shlomot, V. Cuperman, and A. Gresho Hybrid Coding: Combined Harmonic and Waveform Coding of Speech at 4kb/s. IEEE Transactions on Speech and Audio Processing. Vol. 9, pp. 632-646, 2001.