Predictive Neural Network Model for CELP Coding

Sn.Pleshkova-Bekjarska[1]

Abstract - **Linear prediction technique is widely used in speech coding methods and systems. The speech signal can be represented by a few parameters that possess the important feature of the linear process. Linear prediction coding (LPC) is in the base of different modifications of CELP standards of speech coding. It is the goal of this article to represent this part of CELP coder as a predictive neural network model and to investigate the theoretical and practical advances of this representation.**

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

I. INTRODUCTION

Linear prediction coding (LPC) is the essential part of CELP method [1].The implementation of LPC analysis in the standard CELP coders is based on the traditional LPC algorithms [2] for calculating the linear prediction coefficients using the correlation between speech samples in a short time sequence – frame in which it is possible to consider the speech signal as a stationary. These algorithms are well investigated and optimized for hardware or software applications [3]. But it is interesting and useful to consider the possibilities of Neural Network Model implementation as linear prediction part of the CELP coders. Such a possibility is based on the wide range of Neural Network applications and the ability of these networks to solve many of engineering tasks. It is necessary to analyzing the characteristics and effectiveness of different types of Neural Networks in sense of speech linear prediction. Such a study is shown in [4] for a single two layers perceptron. Now the goal of this article is to extend this study for other Neural Networks types such as Predictive Neural Networks.

II. PREDICTIVE NEURAL NETWORK MODEL OF LPC

The proposed Predictive Neural Network Model can be based on the structures of Multiple Adaptive Linear Neural Networks. These structures allow determination of the coefficients of prediction for all frame speech samples in comparison of a single two layers perceptron. The block structure of this proposed network is shown in Fig. 1.

First it is necessary to use a tapped delay line (TDL) to have all N frame speech samples p_k as the inputs of adaptive neural network $pd_1(k),..., pd_N(k)$. The input speech signal is entered from left, and passed through $N-1$ delay stages D.

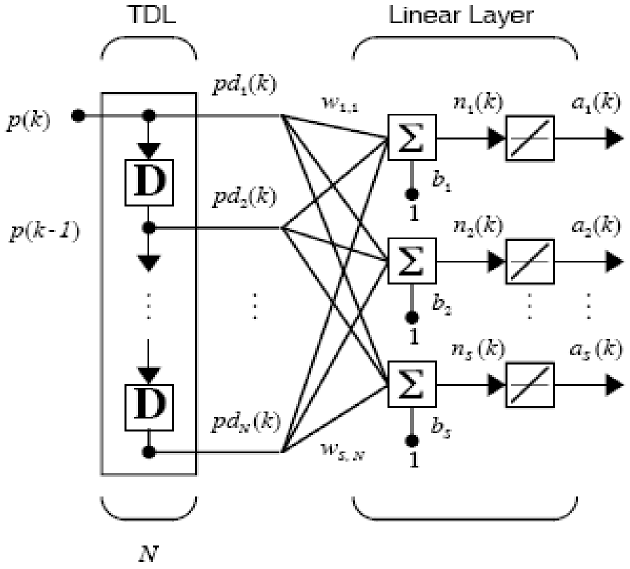

Fig.1. Neural networks structure.

The output of the taped delay line (TDL) is a N – dimensional vector, made up of the input speech signal at the current time, the previous input signal, etc. The weights of Linear Layer $w_{1,1},..., w_{S,N}$ represent the coefficients of linear prediction coding using in CELP method. The weights are sum up in blocks $\sum$ together with bias $b_1,...,b_S = 1$. The outputs of blocks $\sum$ are $n_1(k),..., n_S(k)$. The transfer functions are choose linear. The outputs $a_1(k) - a_S(k)$of neural network are the predicted speech samples if the network is trained to give these samples with a defined minimal error:

$$a_j(k) = \sum_{i=1}^{N} w_{j,i} a(k-i+1) + b_j , \qquad (1)$$

where

$a_j(k)$is the j-th output of network for $j = 1,..., S$;

N – number of past speech samples $pd_2(k),..., pd_N(k)$plus current speech sample $pd_1(k)$;

[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.

$w_{j,i}$ - weight from $i-th$ input to $j-th$ neuron for $i=1,...,N$ and $j=1,...,S$;

b_j - bias of $j-th$ neuron for $j=1,...,S$.

The advantage of this proposition is that the calculated weights $w_{j,i}$ after training of neural network are related with all N speech samples p_k in each frame.

If it is necessary to satisfy the CELP standard in which there are ten coefficients of linear prediction, then the number of the weights in each sum must be N=10 and must correspond to past speech samples according to $a_j(k)$. In the traditional linear prediction methods it is widely used the so called L2 criterion:

$$L_2 = \sum_n \left\{ \sum_n a(i) s(n-i) \right\}^2 , \qquad (2)$$

where

$\{a(i)\}$ is the unknown coefficient, only $a(0)=1$;

$\{s(n)\}$ - the n- th speech signal sample.

The solution of equation (2) is easily derived from a matrix equation. Neural-like stochastic gradient method also works well. On the other hand there is a so called L1 criterion defined by:

$$L_1 = \sum_n \sum_i a(i) s(n-i), \qquad (3)$$

As this expression cannot be differentiated, another approach be devised. It is possible to define an objective function to be minimized:

$$L_1 = \sum_n y(n). \qquad (4)$$

Under the linear constraints:

$$-y(n) \le \sum_i a(i) s(n-i) \le y(n), \qquad (5)$$

where:

$$a(0)=1,\ a(i)=b(i)-\alpha,\ b(i)\ge 0,$$
$$\alpha \ge 0,\ y(n) \ge 0.$$

This approach is comprehensive and give an unique solution, but highly complexity in computation, which is a serious disadvantage.

III. Neural Network Aproach

For the neural network defined in Fig.1 it is possible to describe the outputs of the hidden layer $\{H_j(n)\}$ and the output layer $\{O_j(n)\}$ respectively as follow:

$$H_j(n) = f\left(\sum_{i=0}^{P-1} W_{ij} I_i(n) \right),$$

$$O(n) = f\left(\sum_{i=0}^{Q-1} V_j H_j(n) \right), \qquad (6)$$

where $W_{i,j}$ and V_j are the coupling coefficient between the input layer and hidden layer, and between the hidden layer and output layer respectively. The function $f(.)$ represents a certain linear or not linear function. The L1 criterion for network learning can be formulated by:

$$E = \sum_{n=0}^{N-1} |e(n)| =$$

$$E = \sum_{n=0}^{N-1} |d(n) - O(n)| . \qquad (7)$$

The above expression can be rewrite as:

$$E = \sum_{n=0}^{N-1} \operatorname{sgn}[e(n)] e(n),$$

$$\operatorname{sgn}[e(n)] = \begin{cases} 1 & if \quad e(n) \rangle 0 \\ -1 & if \quad e(n) \langle 0 \end{cases} \qquad (8)$$

It is possible to assuming that the error never become zero, so the partial derivatives with regard to each coupling coefficient can be obtained:

$$\frac{\partial E}{\partial V_j} = -\sum_{n=0}^{N-1} \operatorname{sgn}[e(n)] f'\left(\sum_{m=0}^{Q-1} V_m H_m(n) \right) H_j(n)$$

$$\frac{\partial E}{\partial W_{ij}} = -\sum_{n=0}^{N-1} \operatorname{sgn}[e(n)] f'\left(\sum_{l=0}^{P-1} W_{ij} I_l(n) \right)$$

$$f'\left(\sum_{m=0}^{Q-1} V_m H_m(n) \right) V_j I_i(n), \qquad (9)$$

where $f'(x)$ means $df(x)/dx$.

The gradient in (9) might change suddenly from a present value to the following value, since the surface of E is not continuous. However, if the coupling coefficients are updated descending along this gradient, it is supposed to reach the optimal solution whereby the L1 criterion is minimized. It is proposed in this article to try to achieve the convergence toward the improved value by taking the incremental changes ΔV_j and $\Delta W_{i,j}$, that is:

$$\Delta V_j = -\eta_1 \frac{\partial E}{\partial V_j} \approx -\eta_1 \frac{\partial e(n)}{\partial V_j}$$
$$\Delta W_{ij} = -\eta_2 \frac{\partial E}{\partial W_{ij}} \approx -\eta_2 \frac{\partial e(n)}{\partial W_{ij}}, \qquad (10)$$

where η_1 and η_2 are the additional variables that define the convergence rate.

The proposed approach is very similar to the back propagation (BP) algorithm, but the different problem occurs in the decision of the learning coefficients η_1 and η_2. The gradient of L1 criterion has always certain value as its own nature, which is shown in Fig.2. In this figure (a) represent L2 surface, (b) – more complex L2 surface, (c) – L2 surface and (d) – more complex L2 surface.

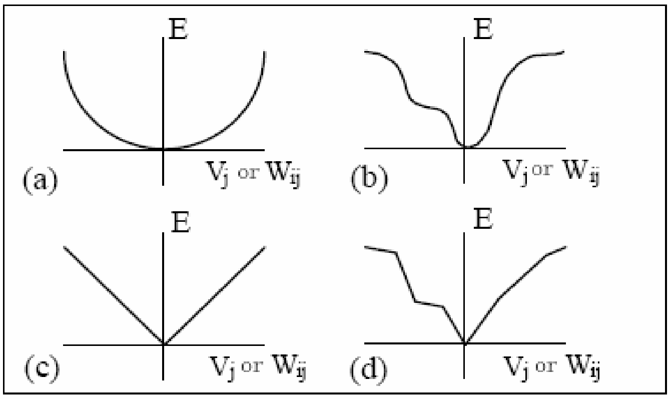

Fig.2a,b. A practical simulation in Matlab.

Therefore, the incremental learning rule in equation (10) sometimes results an unstable solution. In order to avoid such a situation, the learning coefficients must be set small enough with the convergence.

IV. Matlab Simulation

The block schema from Fig.1 represent the structure of the proposed perceptual neural network. It is possible to present this network in Matlab notations for a practical simulation.This representation is shown in Fig.3.

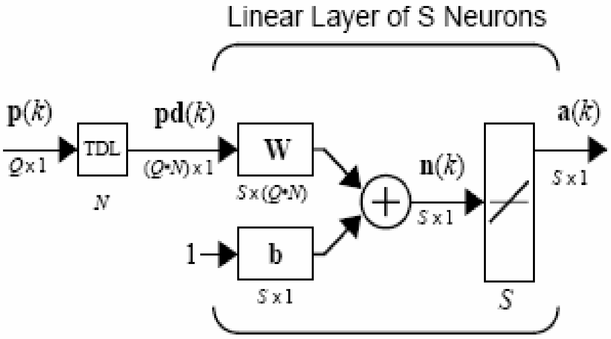

Fig.3. Block schema of practical simulation

The results of a practical simulation for a given real speech signal of vowel /a/ with 256 samples are present. Fig. 4(a) and Fig. 4(b) show L2 residuals difference and L1 residuals difference respectively with the proposed perceptual neural network.

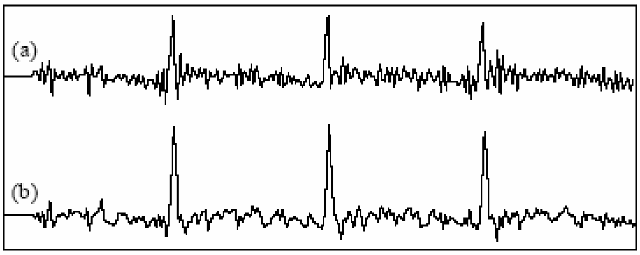

Fig.4 a,b. The results of practical simulation.

V. Conclusion

It is shown that the L1 learning criterion give a more realistic predicted speech signal, because the excitation signal perceived with the proposed neural network corresponding much more to the pitch frequency of the real speech signal. It is possible to use these results in practical implementation in a CELP coder.

Reference

[1] Atal B.S., Schroeder M.R. “Stochastic coding of Speech at Very Low Bit Rates”, Proceedings of ICC, pp. 1610-1630,1984.

[2] Draft GSM 06.20, version 0.3, Half - rate Speech Transcoding, European Telecommunications Standard Institute (ETSI), 1995.

[3] Atal, B.S.,Schroeder, M.R. “Code excited linear prediction (CELP): High quality at very low Bit Rates“, Proc. ICASSP, vol3, pp937-940, 1985.

[4] Pleshkova Sn., Momchedjikov M. “Neural Network Modeling of Speech Coding Prediction”, Proceedings of Papers ICEST 2004, vol.1, pp 95-97,Bitola, Macedonia, 2004