Neural Networks Learning with L1 Criteria and Its Efficiency in Linear Prediction of Speech Signals

Munehiro Namba, Hiroyuki Kamata and Yoshihisa Ishida

School of Science and Technology
Department of Electronics and Communications
Meiji University, Japan

ABSTRACT

The classical learning technique such as the back-propagation algorithm minimizes the expectation of the squared error that arise between the actual output and the desired output of supervised neural networks. The network trained by such a technique, however, does not behave in the desired way, when it is embedded in the system that deals with non-Gaussian signals. As the least absolute estimation is known to be robust for noisy signals or a certain type of non-Gaussian signals, the network trained with this criterion might be less sensitive to the type of signals. This paper discusses the least absolute error criterion for the error minimization in supervised neural networks. We especially pay attention to its efficiency for the linear prediction of speech. The computational loads of the conventional approaches to this estimation have been much heavier than the usual least squares estimator. But the proposed approach can significantly improve the analysis performance, since the method is based on the simple gradient descent algorithm.

1. INTRODUCTION

The back-propagation (BP) algorithm has been a basic but successful training method for a lot of applications in neural networks. The traditional BP method minimizes the expectation of the squared error that arises between the actual output and the desired output of neural networks. This is because BP algorithm regards the error in neural networks is a Gaussian signal. As a matter of fact, this assumption is acceptable for many cases, hence the criterion for the error minimization adopted in the BP algorithm may be the most appropriate one. But in some applications of neural networks for signal processing, the signal to be handled is frequently a non-Gaussian signal. The original BP algorithm intrinsically has no capability to deal with such a process.

As for the speech signal processing, it is known that the maximum-likelihood estimator is asymptotically efficient in the linear prediction of speech, since its distribution of the prediction errors nearly follows a double exponential Laplace distribution [1]. In other words, the linear prediction of speech fails at interval, so that large errors much arise over the whole within the voiced part. The voiced speech contains several excitation signals corresponding to the pitch of speech. Consequently, the distribution of speech signals seems to be more sharp rather than the Gaussian distribution.

The least absolute (L1) estimator is known as one of the robust estimators. Even if the desired signals are corrupted by the unknown process, it tends to be impervious to the unexpected large noises such as a spike signal [2]. The traditional least squares (L2) estimator is rather sensitive to such large noises. The reason for that, needless to say, is its Gaussian error assumption mentioned above. Moreover, the L1 estimator is also known to be able to produce approximately the maximum-likelihood estimation, accordingly it can do robust and effective linear prediction of speech.

This paper discusses the implementational aspects of L1 criteria in neural networks, and its capability of the robust linear prediction of speech. Although lots of researchers have pointed out the importance of L1 criteria in the linear prediction of speech, their issues have been theoretical, and those computational loads are far from the practical realization compared with the least squares method. Because our idea is to simply expand the absolute expression, and take a stochastic gradient descent approach like the BP method, the analysis performance can be significantly improved. Simulation results for both of synthesized speech and practical speech are presented in the later section.

2. LEAST ABSOLUTE ESTIMATOR

2.1. Traditional Approach

Linear prediction technique is widely used in a wide variety of fields. With this technique, the signal can be represented by a few parameters that possess the important feature of the linear process. The well known L2 criterion for the linear prediction is defined as

$$L_2 = \sum_n \left\{ \sum_i a(i)s(n-i) \right\}^2 \tag{1}$$

where $\{a(i)\}$ is the unknown coefficient but $a(0)=1$, and $\{s(n)\}$ is the signal at each n. Because this equation can be differentiated by the unknown coefficient $\{a(i)\}$, the solution is easily derived from a matrix equation. Neural-like stochastic gradient method also works well. On the other hand, the L1 criterion defined by

$$L_1 = \sum_n \left| \sum_i a(i)s(n-i) \right| \tag{2}$$

As this expression cannot be differentiated, another approach must be devised.

Traditional approaches to solve this problem are mainly based on the "linear programming" theory. This theory has been developed for the optimization of production plans, economical surveys, and mathematical problems, in which an objective function should be minimized or maximized under the constraint that is formed by some linear expressions. Because arbitrary expressions are allowable for the constraints unless they are linear, the least absolute minimization problem can be returned to the following linear programming problem.

Objective function to be minimized:

$$L_1 = \sum_n y(n) \tag{3}$$

Under the linear constraints:

$$-y(n) \le \sum_i a(i)s(n-i) \le y(n) \tag{4}$$

where,

$$a(0)=1, a(i)=b(i)-\alpha, b(i)\ge 0, \alpha \ge 0, y(n)\ge 0 \tag{5}$$

Although this approach is comprehensive, and generally certify the unique solution (except for the case that the solutions inherently form a hyper-plane), the highly complexity in computation is a serious disadvantage.

2.2. Neural Networks Based Approach

Let us define the neural network as shown in Fig.1. In this network, the outputs of the hidden-layer $\{Hj(n)\}$ and the output-layer $\{O(n)\}$ are expressed respectively as follows.

$$H_j(n) = f\left(\sum_{i=0}^{P-1} W_{ij} I_i(n)\right), \quad O(n) = f\left(\sum_{j=0}^{Q-1} V_j H_j(n)\right) \tag{6}$$

where $\{Wij\}$ and $\{Vj\}$ are the coupling coefficient between the input-layer and hidden-layer, and between the hidden-layer and output-layer respectively. $f(\cdot)$ represents a certain linear or non-linear function. The L1 criterion in this network learning is formulated by

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

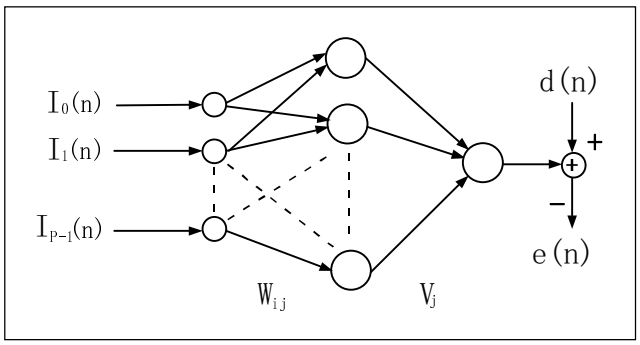

Figure 1. Neural Network Structure:

We rewrite above expression as,

$$E = \sum_{n=0}^{N-1} sgn[e(n)]\, e(n), \quad sgn[e(n)] = \begin{cases} 1 & if\ e(n)>0 \\ -1 & if\ e(n)<0 \end{cases} \tag{8}$$

By assuming the error never becomes zero, the partial derivatives with regard to each coupling coefficient can be obtained.

$$\begin{aligned} \frac{\partial E}{\partial V_j} &= -\sum_{n=0}^{N-1} sgn[e(n)]\, f'\left(\sum_{m=0}^{Q-1} V_m H_m(n)\right) H_j(n) \\ \frac{\partial E}{\partial W_{ij}} &= -\sum_{n=0}^{N-1} sgn[e(n)]\, f'\left(\sum_{l=0}^{P-1} W_{lj} I_l(n)\right) \\ &\qquad f'\left(\sum_{m=0}^{Q-1} V_m H_m(n)\right) V_j I_i(n) \end{aligned} \tag{9}$$

where $f'(x)$ means $df(x)/dx$. The gradient in Eq.(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. In this paper, we try to achieve the convergence toward the improved value by taking the incremental changes ΔVj and ΔWij, that is

$$\begin{aligned} \Delta V_j &= -\eta_1 \frac{\partial E}{\partial V_j} \approx -\eta_1 \frac{\partial e(n)}{\partial V_j} \\ \Delta W_{ij} &= -\eta_2 \frac{\partial E}{\partial W_{ij}} \approx -\eta_2 \frac{\partial e(n)}{\partial W_{ij}} \end{aligned} \tag{10}$$

where η_1 and η_2 are the additional variables that determine the convergence rate. This approach is very similar to the BP algorithm, but the different problem occurs in the decision of the above learning coefficients η_1 and η_2. The gradient of the L1 criterion has always certain value as its own nature (See Fig.2). Therefore, the incremental learning rule in Eq.(10) sometimes results an unstable solution. In order to avoid such a situation, the learning coefficients must be set small enough with the convergence.

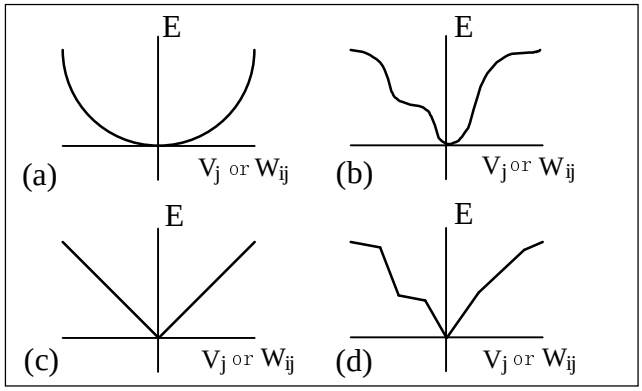

Figure 2. Criterion's Typical View: (a) L2 surface; (b) more complex L2 surface; (c) L1 surface; (d) more complex L1 surface

The decrease manipulation of the learning coefficients has to be arranged in advance. A monotonic reducing that is utilized in the self-organizing map neural networks is a good choice, or a randomly reducing is also works in a case by case. But the observation of the L1 criterion at the current training phase is often required, because the learning coefficients should be determined according to the degree of convergence. A phase means that an entire training of the analyzed signals, in this paper. If the value of the L1 criterion at the present phase is too close to the value at the previous phase, the decrease of the learning coefficients enables the further convergence, especially in the linear activation case (See Fig.3). In the case of non-linear activations (e.g. the sigmoidal functions), however, the increase of the learning coefficients might be occasionally required in order to avoid settling down a local minimum of the criterion.

3. SIMULATION RESULTS

Because the primary purpose of this paper is to examine the efficiency of L1 criteria for the linear prediction of speech, only the case of the linear network is simulated in this paper. Non-linear cases are still under research.

3.1. A Synthesized Signal Case

The first experiment analyzes the 256-points synthesized signal. Three impulses with appropriate pitch drive a given ARMA model that is shown in Table 1. The network consists of 12 input-nodes, 2 hidden-nodes, and 1 output-node. The rule for decreasing the values of the learning coefficients is arranged as follows.

Present Criterion: cL1 Previous Criterion: pL1 Learning Rate: lr (Initial value=1e-3)		
If	cL1/pL1>1.001 and lr>1e-6	then lr=lr×0.8
else if	cL1/pL1>1.0001 and lr>1e-8	then lr=lr×0.8
else if	cL1/pL1>1.00001 and lr>1e-10	then lr=lr×0.9

The estimated results by both the general LPC method as the most popular L2 estimator, and the proposed network are shown in Table 2. The analysis is performed as if the system is an AR model that consists of 8 poles. As they are shown, the L2 method fails to estimate the original poles precisely, and besides, the additional poles locate somewhat close to the unit circle, so they cannot be disregarded. On the contrary, the proposed network with the L1 criterion successes to precisely estimate the original poles. The additional poles, moreover, locate nearly at the center of the unit circle, thus they can be omitted. The difference between two methods is more clear in the prediction residuals that are shown in Fig.4. Fig.4(a) is the residual of the LPC, and Fig.4(b) is the residual of the proposed method. The residual of the proposed method perfectly matches the input signal of three impulses, and the zeros of the ARMA model. But in the LPC residual, the odd signal lasts after each of excitation impulses. This is the most critical problem of the L2 estimator in the linear prediction of speech. The L2 estimator definitely tries to minimize the excitations as possible as it can, because the L2 estimation is a weighted estimation on the larger error. As a result, the odd errors exist in the residual.

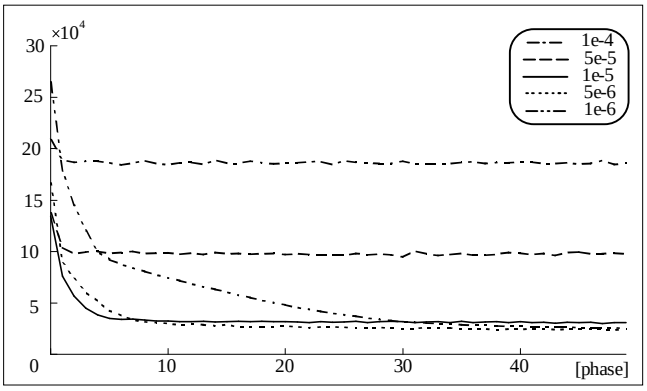

Figure 3. Affection of the Learning Rate: If the learning coefficient is too large as 1e-4 or 5e-5, the L1 criterion is vibrated too largely in a sense (i.e. abs[+e]=abs[-e]), and the search for the minimum point is unexpectedly converges.

Poles	0.7 ± j 0.6	-0.5 ± j 0.7
Zeros	± j 1.5	

Table 1. Poles and Zeros Location

L2 Norm	0.6920 ± j 0.6108	-0.5129 ± j 0.7189
	0.5803 ± j 0.3082	-0.5576 ± j 0.2747
L1 Norm	0.7000 ± j 0.6000	-0.5000 ± j 0.7000
	-0.0294 ± j 0.0296	0.0293 ± j 0.0292

Table 2. Estimated Poles and Zeros Location

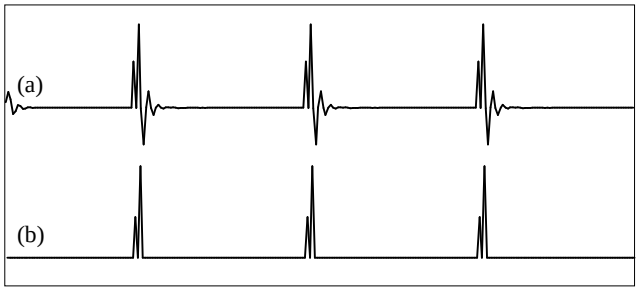

Figure 4. The Residual Difference in a Synthesized Signal: (a) L2 residuals; (b) L1 residuals

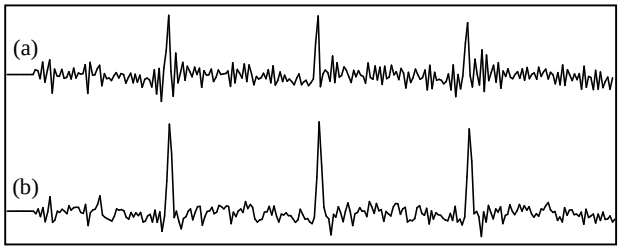

Figure 5. The Residual Difference in a Practical Speech: (a) L2 residuals; (b) L1 residuals

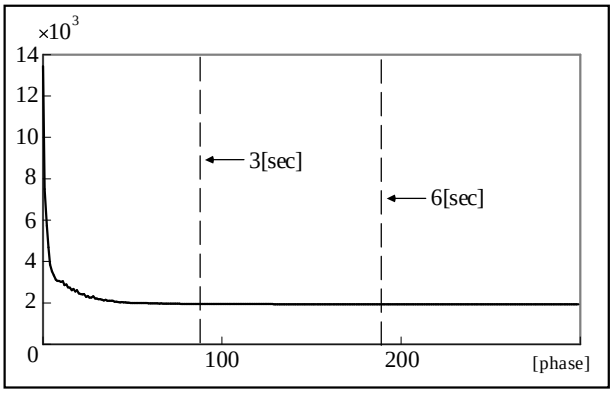

Figure 6. Convergence and Required Time:

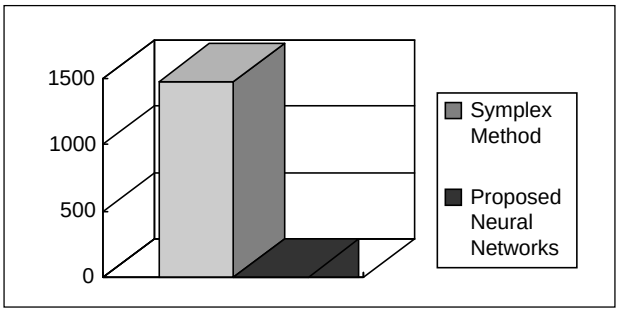

Figure 7. The Performance Comparison: The linear programming method requires 1000 to 1500[sec] with Sparc Station 10; The proposed neural network requires only 3 to 12[sec] with the same computer.

3.2. A Practical Speech Case

The second experiment analyzes the 256-points practical speech signal that is the Japanese vowel /a/. The network structure and the learning rule are set equivalently to the preceding experiment. The analysis is performed to estimate an AR model that consists of 24 poles, including the conjugates. The speech signal is pre-emphasized +6dB due to a physical consideration in speech production mechanism.

Fig.5(a) and Fig.5(b) show the residuals of the LPC and the proposed methods, respectively. Not only in the case of synthesized signal, but also in the case of practical speech signal, the excitation signals corresponding to the pitch frequency can be readily perceived with the proposed neural network.

The value of the L1 criterion at each phase in learning is shown in Fig.6. In this case, just 3 seconds are needed to search the optimal solution. The symplex method that is a common technique of the linear programming takes 1469 seconds to provide the result. The comparison is illustrated in Fig.7.

4. CONCLUSION

We have presented an approach to the linear prediction of speech with the L1 criterion that is based on a neural network. Different from the traditional approach using the linear programming technique, the proposed method is very simple, and can reduce both the computational complexity and the required memories, remarkably.

Regardless of the class of the activation functions in neural networks, the network learning with L1 criteria ought to make a network that behaves more functionally. This means that the network should be robust to the unexpected, irregular training sequences. There is also a possibility to avoid a local minimum in the convergence. But non-linear cases needs further research.

REFERENCES

1. Etienne Denoël and Jean-Philippe Solvay, "Linear Prediction of Speech with a Least Absolute Error Criterion", *IEEE Trans. ASSP* 33: 1397-1403, 1985.
2. Michael S. O'Brien, Anthony N. Sinclair and M. Kramer, "Recovery of a Sparse Spike Time Series by L_1 Norm Deconvolution", *IEEE Trans. ASSP* 42: 3353-3365, 1994.
3. John Makhoul, "Linear Prediction: A Tutorial Review", *Proc. IEEE* 63: 561-580, 1975.
4. Pierre F. Baldi and Kurt Hornik, "Learning in Linear Neural Networks: A Survey", *IEEE Trans. ASSP* 6: 837-858, 1995.
5. R. V. Hogg, "An introduction to robust procedures", *Commun. Statist. Theory Meth.* , vol. A6, pp. 789-794, 1994.
6. H. T. Hu, "Linear Prediction using L_1 norm in orthogonal vector space", *Electron. Lett.* , vol. 31, pp. 430-431, 1995.
7. R. Garcia Gómez, J.M. Alcázar Fernández and A. R. Figueiras Vidal, "On The Minimization of Pulse Density in Multipulse Coding," *Signal Processing III: Theories and Applications*, pp. 473–476, 1986.