Journal of Shanghai Jiaotong University, Vol. E-3, No. 2, 1998, 56～60

SPEECH ENHANCEMENT BASED ON SECOND ORDER ARCHITECTURE AND INFORMATION MAXIMIZATION THEORY*

Yu Xiao(虞　晓)　*Hu Guangrui*(胡光锐)　*Chen Wei*(陈　玮)
(Dept. of Electronic Engineering, Shanghai Jiaotong Univ., China)

Abstract Based on the idea of adaptive noise cancellation (ANC), a second order architecture is proposed for speech enhancement. According as the Information Maximization theory, the corresponding gradient descend algorithm is proposed. With real speech signals in the simulation, the new algorithm demonstrates its good performance in speech enhancement. The main advantage of the new architecture is that clean speech signals can be got with less distortion.

Key words speech enhancement; blind signal separation; information maximization; ANC

Introduction

The broad use of speech recognition demands enhancement of noisy speech. The speech enhancement system attempts to improve the perceptual aspects (e.g. quality, intelligibility) of noisy speech signals. One of the speech enhancement algorithms is based on the Minimizing Mean Square Error (MMSE) method, such as the LMS algorithm in adaptive noise cancellation (ANC)[1] system. But in the ANC, if all inputs are the mixture of speech and noise, the algorithms based on MMSE will not perform well and the output of the speech anhancement system will be distorted by so-called "music noise"[1]. On the other hand, noise cancellation in speech enhancement can be seen as a Blind Source Separation (BSS) problem[2～8], which has received a lot of attention since 90's. In BSS, the problem is how to recover a number of original stochastic independent sources when only linear or convolved mixtures are available. Neither the mixing coefficients nor the probability distribtuions of the original sources are known. Most of the works on BSS addressed the case of instantaneous and linear mixture[2～5], where the mixture model is assumed as $\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{S}(t)$. But in the speech enhancement problem, because of the propagation delay in the medium and the filter response of the observed sensors, there will be time delay or phase difference among the observed inputs. So the speech enhancement problem shall be seen as a convolved BSS problem. Some preliminary researches have been done on this[6,7]. In this paper, a second order architecture is proposed for speech enhancement based on the idea of ANC. And based on the information maximization theory[3], the gradient descent algorithm of a new architecture is also proposed. From computer simulations of real speech signals, it can be concluded that the new algorithm works well with less distortion in speech enhancement.

1　The Second Order Architecture

In this paper, the speech enhancement problem is modeled as a convolved BSS problem, i.e., how to separate the clean speech signal from the observed mixture of the mutually independent sources, one of which is the clean speech signal and others are the noise signal. Without loss of generality, a BSS forward architecture of two

Received 1998-02-09
* Supported by the National Natural Science Foundation of China (No. 69672007)

sources is illustrated in Fig. 1(a). Two sensors at different channels are just like the two ears of a human being. The inputs received by the two sensors can be formulated as Eq. (1) in the Z-transform domain, where $A_{ij}(Z)$ is the Z-transform of the transfer filter response from the ith source to the jth sensor, and $S_1(Z)$ is the clean speech signal and $S_2(Z)$ is the noise signal.

$$\left.\begin{aligned} X_1(Z) &= S_1(Z) + A_{21}(Z)S_2(Z) \\ X_2(Z) &= A_{12}(Z)S_1(Z) + S_2(Z) \end{aligned}\right\} \quad (1)$$

The system in the forward architecture can be formulated as Eq. (2) in the Z-transform domain:

$$\left.\begin{aligned} Y_1(Z) &= X_1(Z) + X_2(Z)W_{21}(Z) \\ Y_2(Z) &= X_2(Z) + X_1(Z)W_{12}(Z) \end{aligned}\right\} \quad (2)$$

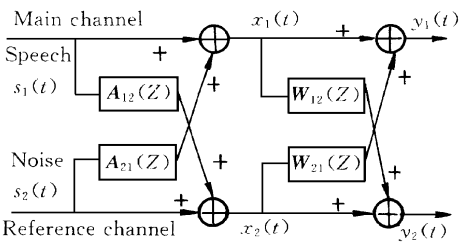

Fig. 1(a) The forward BSS architecture in speech enhancement

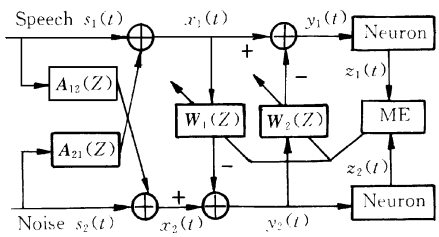

Fig. 1(b) The second order architecture in speech enhancement using the ME principle

In this paper, it is assumed that the speech source is close to the main channel sensor and the noise source is close to the reference channel sensor. Then if the outputs are mutually independent, the best BSS solutions can be given as Eq. (3). It is clear that the output would be distorted by a filter whose Z-transform is $[1 - A_{12}(Z)A_{21}(Z)]$. Postprocessing needs to be done in order to get the original speech signal. Without knowing the characters of mixture filters $A_{ij}(Z)$, it is difficult to recover the original speech in the general forward architecture. On the other hand, Kari Torkkola proposed a feedback network architecture to solve the BSS problem[6]. The main drawback of the feedback structure is the stability of the BSS system because the cross filters $A_{12}(Z)A_{21}(Z)$ should be assumed strictly to be causal.

$$\left.\begin{aligned} Y_1(Z) &= [1 - A_{12}(Z)A_{21}(Z)]S_1(Z) \\ Y_2(Z) &= [1 - A_{12}(Z)A_{21}(Z)]S_2(Z) \end{aligned}\right\} \quad (3)$$

Here, a second order architecture is proposed for the speech enhancement problem as in Fig. 1(b). The main idea is motivated by the principle of ANC. The first order ANC architecture is used to cancel the mixed speech signal in the reference channel to get $y_2(t)$, which will be a function of only noise signal. While getting the original speech signal, the second order ANC architecture is used to cancel the noise in the main channel by using $y_2(t)$. The second order BSS architecture can be formulated as Eq. (4) in Z-transform domain.

$$\left.\begin{aligned} Y_1(Z) &= X_1(Z) - Y_2(Z)W_2(Z) \\ Y_2(Z) &= X_2(Z) - X_1(Z)W_1(Z) \end{aligned}\right\} \quad (4)$$

If it is assumed that the observed sensors are close to the corresponding independent sources, the best speech enhancement solutions of the new second order BSS architecture will be like Eq. (5). It is clear that the main advantage of the new architecture is that the original clean speech can be separated from the noisy inputs without filter distortion.

$$\left.\begin{aligned} Y_1(Z) &= S_1(Z) \\ Y_2(Z) &= [1 - A_{12}(Z)A_{21}(Z)]S_2(Z) \end{aligned}\right\} \quad (5)$$

2 Information Maximization Theory and Maximum Entropy Algorithm

In speech enhancement, it is assumed that the original speech signal and the noise signal are mutually independent. So in the second order BSS architecture, if the adaptive BSS algorithm can successfully separate the original speech signal and the noise signal, it is expected that the mutual information between the outputs $y_1(t)$ and $y_2(t)$ shall be equal to zero. According as

the information theory, it is clear that the mutual information between two random variables would involve all orders of their statistics. So only by using the second order statistics, the algorithms based on MMSE can not solve the speech enhancement problem in the proposed architecture. Maximum Entropy (ME)[3] and Minimum Mutual Information (MMI)[4] or Independent Component Analysis (ICA)[5] are two major approaches to deduce the blind source separation algorithms. The Mutual Information (MI) in the MMI algorithm is one of the best contrast functions since it is invariant under the transforms such as permutation, scaling and componentwise nonlinear transform[4]. But the MMI algorithm needs a lot of computation to estimate the high-order statistics or cumulants in practice. On the other hand, the ME algorithm has not been rigorously justified except for the case when the sigmoid function in ME algorithm happens to be the cumulative density function of the unknown sources. If all sources are zero mean signals, it has been proved that the ME algorithm can give local solutions of BSS. The BSS algorithm derived by the ME approach is often very effective in some practical cases[3], where the ME algorithm has a better performance than the MMI algorithm.

According to the information maximization principle depicted by A. J. Bell[3], the outputs $y_i(t)$ will be transformed to $z_i(t)$ by a nonlinear function $g(\cdot)$ in the proposed architecture. $z_i(t)$ can be regarded as the output of several analog neurons as in Fig. 1(b). If $z_i(t)$ is considered as a random variable at time t, the joint entropy of $z_i(t)$ will be written as

$$H(\boldsymbol{Z};\boldsymbol{W}) = -\int f(\boldsymbol{Z};\boldsymbol{W})\log f(\boldsymbol{Z};\boldsymbol{W})\,\mathrm{d}\boldsymbol{Z} \quad (6)$$

If the sources are all super-gaussian signals, it has been proved that maximizing the joint entropy $H(\boldsymbol{Z};\boldsymbol{W})$ with respect to weight vectors would always minimize the mutual information between the outputs. In practice, many real-world signals (including speech signals) are super-gaussian signals[3]. The joint probability density function (pdf) of the sources $x_1(t), x_2(t)$ is assumed as below at time t:

$$f_x(x_1(t)x_2(t)\,|\,x_1(t-1)x_1(t-2)\cdots x_1(t-M)x_2(t-1)x_2(t-2)\cdots x_2(t-M))$$

Here M is the order of FIR weight filters $W_1(Z)$ and $W_2(Z)$. The joint pdf of the neuron outputs $z_1(t), z_2(t)$ can be formulated as

$$f_z(z_1(t), z_2(t)) = f_x(x_1(t), x_2(t))/|\boldsymbol{J}| \quad (7)$$

where $|\boldsymbol{J}| = \left|\frac{\partial z_1}{\partial x_1}\frac{\partial z_2}{\partial x_2} - \frac{\partial z_1}{\partial x_2}\frac{\partial z_2}{\partial x_1}\right|$ is the absolute value of the determinant of the Jacobean matrix. Then Eq. (6) can be written as

$$H(z_1, z_2; \boldsymbol{W}) = \mathrm{E}[\ln|\boldsymbol{J}|] - \mathrm{E}[\ln f_x(x_1, x_2)] \quad (8)$$

The mutual information between outputs $y_1(t)$ and $y_2(t)$ will be minimized by maximizing the entropy of the neuron outputs $Z_i(t)$, which is equivalent to maximizing $\mathrm{E}[\ln|\boldsymbol{J}|]$.

3 Gradient Descend Algorithm

Rewriting Eq. (4) in time domain as

$$\left.\begin{aligned} y_1(t) &= x_1(t) - \sum_{k=0}^{M} w_2(k)y_2(t-k) = \\ & x_1(t) - w_2(0)y_2(t) - \sum_{k=1}^{M} w_2(k)y_2(t-k) \\ y_2(t) &= x_2(t) - \sum_{k=0}^{M} w_1(k)x_1(t-k) \end{aligned}\right\} \quad (9)$$

The the Jacobean matrix $\boldsymbol{J}$ can be deduced as

$$\boldsymbol{J} = \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_1}\frac{\partial z_2}{\partial y_2}\frac{\partial y_2}{\partial x_2} - \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_2}\frac{\partial z_2}{\partial y_2}\frac{\partial y_2}{\partial x_1} = \frac{\partial z_1}{\partial y_1}\frac{\partial z_2}{\partial y_2} \quad (10)$$

If the sigmoid function is used as the nonlinear function $g(y_i(t))$, then $\frac{\partial z}{\partial y} = \frac{\mathrm{e}^{-y}}{(1+\mathrm{e}^{-y})^2} > 0$ can be got. That is to say, $|\boldsymbol{J}| = \boldsymbol{J}$. In the general forward architecture[3,6], it can be noticed that one item D, which is the function of the weight vectors, will be in the Jacobean matrix $\boldsymbol{J}$. It is difficult to determine the sign of D to get the absolute of $\boldsymbol{J}$. As one advantage of the proposed second order architecture for the speech enhancement problem, there will be no such ambiguous problem here.

Instead of the expectation of the stochastic gradient with the instantaneous value, by maximiz-

ing $\ln(\boldsymbol{J})$, the learning rule of the proposed architecture can be deduced as Eq. (11).

$$\left.\begin{aligned}\Delta w_1(k) &\propto [(1-2z_1(t))w_2(0) - (1-2z_2(t))]x_1(t-k)\\ \Delta w_2(k) &\propto [-(1-2z_1(t))]y_2(t-k)\\ k &= 0, 1, \cdots, M\end{aligned}\right\} \tag{11}$$

Compared with Eq. (11) of the general forward architecture[3], the learning rule of the proposed architecture is easy to compute, because there is no such item as the inverse of the weight matrix $\boldsymbol{W}$.

4 Simulation

We illustrate the performance of the proposed architecture and the BSS adaptive algorithm by using real speech signals in the speech enhancement problem. In our simulation, the mixture model is as

$$\left.\begin{aligned}x_1(t) &= s_1(t) - 0.6s_2(t-2)\\ x_2(t) &= s_2(t) + 0.8s_1(t-3)\end{aligned}\right\} \tag{12}$$

where $s_1(t)$ is a clean speech signal, $s_2(t)$ is a color noise which is generated by the computer. The waves of the original sources and the mixture inputs are displayed in Fig. 2. In Fig. 3, the separate result of the proposed algorithm is displayed, when the algorithm is convergent. In this case, the input signal to noise ratio (SNR) is -9.614 dB on the main channel, and the output SNR is 33.559 dB, where the SNR is computed by

$$10\log\left(\frac{\sum_{t=1}^{N}[s_1(t)]^2}{\sum_{t=1}^{N}[s_1(t)-y_1(t)]^2}\right)$$

The order of weight FIR filters $\boldsymbol{W}_i$ is chosen as 16 in the simultion. In Fig. 4, the convergent curve of the second weight of the weight filter $W_2(Z)$ is shown when the stepsize is 0.001. From the results of simulation, it is clear that the new architecture can fulfil the task of speech enhancement successfully. When listening to the output speech signal, there is little distortion between the output and the original speech signal. In the simulation, it can be found that a big stepsize will accelerate the convergence rate and a small stepsize will minimize the fluctuation of the convergence weight vectors. So in practice, the proposed algorithm can get better performance by using an adaptive stepsize instead of the fixed stepsize described in this paper.

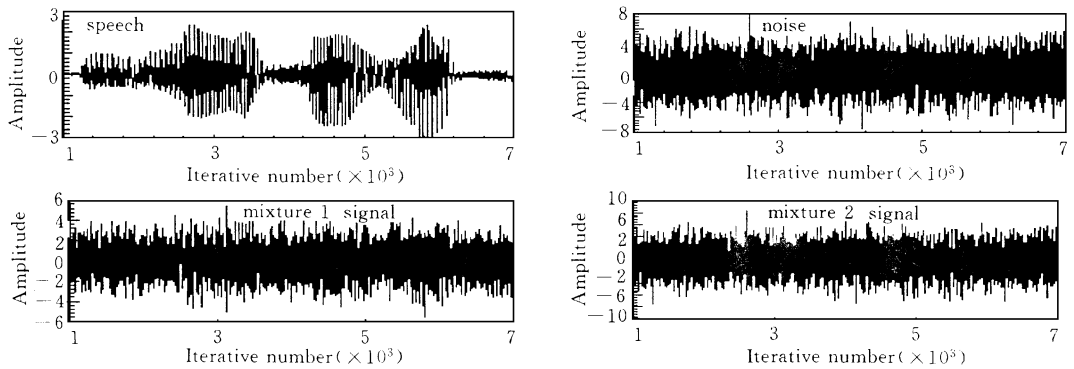

Fig. 2 The waves of original speech signal and noise signal and the mixture observed signals

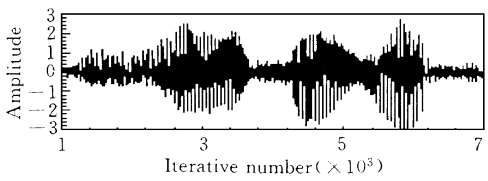

Fig. 3 The output wave of the proposed architecture

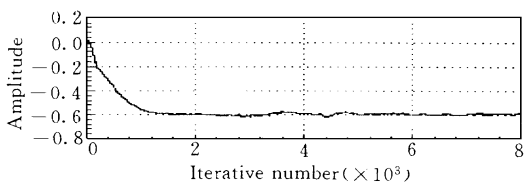

Fig. 4 The convergence curve of weight $W_2(Z)$

5 Conclusion

From the computed results of simulation by using real speech signals, it can be concluded that the proposed speech enhancement algorithm is concise and effective in practice. The main advantage of the new architecture is that the clean speech signal can be got with less distortion. In this paper, the FIR filters are used in second order architecture. Future work will include investigating the use of nonlinear filter to improve the performance of the speech enhancement algorithm, especially in the nonlinear mixture model in real world.

References

1 Widrow B, Stearns S. Adaptive signal processing. New York: Prentice-Hall, 1985.

2 Amari S, Cichocki A, Yang H H. A new learning algorithm for blind signal separation. In: Touretzky D, Mozer M eds, Hasselmo M. Advances in Neural Information Processing Systems 8. MIT Press, MA, 1996. 757～763

3 Bell A J, Sejnowski T J. An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 1995, 7(6): 1129～1159

4 Comon P. Independent component analysis, a new concept? Signal Processing, 1994, 36(3): 287～314

5 Jutten C, Herault J. Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing, 1991, 24(1): 1～10

6 Torkkola K. Blind separation of delayed sources based on information maximization. Proceeding of ICASSP, Atlanta, GA, USA, May 7-10, 1996

7 Lee TW, Orglmeister R. Blind source separation of real-world signals. Proceeding of ICNN Houston, USA, 1997

8 Cardoso J F, Laheld B. Equivariant adaptive source separation. IEEE Trans on Signal Processing, 1996, 45(2): 434～444

(**Continued from page 55**)

theory, people are now turning to the research of the applications of the nonlinearity theory in speech processing. The fractal theory is a promising nonlinearity science. Fractal dimension can be used quantitatively to analyze the self-similarity of speech. At the same time, to overcome the drawback of the use of Hausdorff-Besicovitch dimension, the generalized fractal is introduced to improve the description of inner information of speech signals and the analysis of speech signals. We have got better speech segmentation results from the use of it. We think that the fractal theory is very important for the improvement of speech signal processing.

References

1 Parker T S, Chua L O. Practical numerical algorithms for chaotic systems. New York: Springer, 1989.

2 Mandelbrot B B. The fractal geometry of nature. San Francisco: Freeman, 1982.

3 Pickover C, Khorasani A. Fractal characterization of speech waveform graphs. Comp & Graphics, 1986, 10(1): 55～61

4 Barnsley M F, Elton J H, Hardin D P. Recurrent iterated function systems. Constructive Approximation, 1989, 5(1): 3～31