ECHO CANCELLATION BY GLOBAL OPTIMIZATION OF KAUTZ FILTERS USING AN INFORMATION THEORETIC CRITERION

Ching-An Lai, Deniz Erdogmus, Jose C. Principe

CNEL, Electrical Engineering Department, University of Florida, Gainesville FL, 32611, USA

ABSTRACT

In practical settings, the echo cancellation problem generally requires the adaptation of an IIR filter using some optimality criterion. This brings two problems: direct adaptation of numerator and denominator polynomial coefficients of IIR filters might result in unstable systems and/or the optimization might result in a suboptimal local minimum of the criterion. These two issues are addressed in this paper. To resolve the first problem, orthogonal Kautz filters are utilized for their stability is easily controlled through the pole locations. The second problem is addressed by employing an information theoretic optimality criterion, which has a parameter that is annealed to ensure global optimization.

1. INTRODUCTION

Echo cancellation is an important practical problem whose solution generally necessitates the optimization of an adaptive infinite impulse response (IIR) filter. If the actual channel is a finite impulse response (FIR) filter, the ideal inverse of the channel is guaranteed to be IIR. It might be possible to find an approximate FIR equalizer for an FIR channel in some cases. In that case, the solution, given by the Wiener-Hopf equations, is easy to obtain both analytically with on-line adaptation. Determining the model order, however, is a major problem in this case. The adaptation of IIR filters, on the other hand, result in two major difficulties: filter stability and suboptimal solutions. If the adaptive IIR system is parameterized in terms of the numerator and denominator polynomial coefficients of its transfer function, then maintaining the stability of the poles is difficult. If the IIR filter is expressed in terms of its zeros and poles, the gradient expressions for these become extremely complicated.

Kautz filters form an orthogonal set of basis impulse response functions so that any impulse response function could be approximated with arbitrarily small errors as higher order Kautz filters are utilized [1]. In addition, the Kautz filters are expressed explicitly in terms of their poles. Therefore, maintaining the stability is trivial. The derivatives with respect to these poles are less complicated than an arbitrary IIR filter. Hence, Kautz filters provide an ideal solution to the dilemma of filter stability.

Whatever IIR filter topology and optimality criterion is utilized, if the poles are adapted, the problem of suboptimal solutions will exist. Commonly, the mean square error (MSE) is the criterion of choice. Due to the mentioned difficulties in adapting the feedback parameters of generalized feedforward filters, in problems that require adaptive IIR filtering, such as echo cancellation, the poles of the filter are not adapted [2]. Adapting only the feedforward weight vector conveniently reduces to an LMS-type algorithm where some variant of the Wiener solution can be reached. Here, we propose a method to adapt both the feedforward and feedback parameters of an adaptive IIR filter to achieve global optimization, yet still use gradient descent.

Recently, we have proposed and experimented with an information theoretic alternative to MSE called minimum error entropy (MEE) [3]. In this paper, we will employ a Euclidean distance approach to the supervised training of IIR filters, where the new criterion will show some resemblance to the previously investigated Renyi's entropy measures. When this new criterion is estimated from samples with Parzen windowing, it is possible to achieve global optimization by annealing the kernel size.

We will propose an annealing scheme for the kernel size and the global optimization capability of the proposed algorithm will be investigated through Monte Carlo simulations. As a comparison, we will also provide results obtained using LMS variants, which are known to have improved chances of avoiding local minima. The comparison will be performed by investigating the L_p-norm of the error between the identified and ideal inverse impulse responses (truncated at a sufficiently large delay) for various choices of p.

2. LMS VARIANT ALGORITHMS

Extending LMS to the case of IIR filters is trivial. In fact, the global optimization capabilities of LMS-based algorithms are previously investigated. The two algorithms that we will focus on here are called LMS-

SAS, which is a slightly modified version of the algorithm by Srinivasan *et al.* [4], and NLMS, which is a straight forward extension of the normalized LMS in FIR training to the IIR filter case. Suppose we are given a training sequence $\{\mathbf{x}_k, d_k\}$, and an adaptive IIR filter whose parameters (weights) are collected in a vector $\boldsymbol{\theta}$ and that generates an output y_k. These stochastic MSE minimization algorithms are [5]:

$$\begin{aligned} &LMS-SAS: \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mu_k e_k \nabla_{\boldsymbol{\theta}} y_k + \mu_k e_k \eta_k \\ &NLMS \quad : \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \frac{\mu_k}{\left\| \nabla_{\boldsymbol{\theta}} y_k \right\|^2} e_k \nabla_{\boldsymbol{\theta}} y_k \end{aligned} \tag{1}$$

In (1), k is the sample/time index, μ_k is the possibly time-varying step size, $e_k = d_k - y_k$ is the output error, and $\nabla_{\boldsymbol{\theta}} y_k$ is the gradient of the output with respect to the weights.

3. INFORMATION THEORETIC LEARNING

Recently, we have investigated the performance of MEE in supervised learning, which provided generalization results favorable to MSE [3]. Our cost function was based on Renyi's entropy, which is, for a random variable e with probability density function (pdf) $p_e(.)$ was defined as [6].

$$H_\alpha(e) = \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} p_e^\alpha(\varepsilon) d\varepsilon \tag{2}$$

In this paper, we will concentrate on a Euclidean distance measure based on the error pdf. In supervised training, the purpose is to find the weight vector that makes the error as small as possible. One alternative way of enforcing this is to minimize the divergence between the error pdf and a Dirac-δ distribution located at zero.

$$\begin{aligned} I_{ED} &= \int_{-\infty}^{\infty} \left(p_e(\varepsilon) - \delta(\varepsilon) \right)^2 d\varepsilon \\ &= \int_{-\infty}^{\infty} p_e^2(\varepsilon) d\varepsilon - 2 p_e(0) + \int_{-\infty}^{\infty} \delta^2(\varepsilon) d\varepsilon \end{aligned} \tag{3}$$

The final expression in (3) contains three terms: the argument of Renyi's quadratic entropy (α = 2), the error pdf evaluated at zero, and a term that is independent of the filter coefficients. In essence, minimizing this criterion is maximizing the likelihood of achieving zero error while trying to maximize the quadratic error entropy (since minimizing the first term is equivalent to maximizing quadratic entropy). In a sense, the objective of this criterion corroborates Jaynes' maximum entropy principle [7]. This principle suggests selecting a distribution that best fits the available data, but that makes minimal commitment to unobserved data. This is mathematically formulated as finding the maximum entropy density that satisfies equality constraints regarding the data statistics.

Since the analytical form of the error pdf is not available in practice, it has to be estimated from the samples. Parzen windowing is a suitable pdf estimation method for our purposes [8]. Given a set of independent and identically distributed (iid) samples $\{e_1,\ldots,e_N\}$, the Parzen estimate of the underlying pdf is

$$\hat{p}_e(\varepsilon) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(\varepsilon - e_i) \tag{4}$$

where $\kappa_\sigma(.)$ is the kernel function. A commonly used kernel choice is the Gaussian pdf. In general, the kernel can be any valid pdf. Here, σ controls the kernel width. For Gaussian kernels, it is the standard deviation.

When the Euclidean distance in (3) is estimated from the samples of the error on the training set using Gaussian kernels, the following expression is obtained.

$$\begin{aligned} \hat{I}_{ED} &= \int_{-\infty}^{\infty} \left(\hat{p}_e(\varepsilon) - G_\sigma(\varepsilon) \right)^2 d\varepsilon \\ &= \int_{-\infty}^{\infty} \hat{p}_e^2(\varepsilon) d\varepsilon - 2 \int_{-\infty}^{\infty} \hat{p}_e(\varepsilon) G_\sigma(\varepsilon) d\varepsilon + \cancel{\int_{-\infty}^{\infty} G_\sigma^2(\varepsilon) d\varepsilon} \\ &\equiv \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} G_{\sigma\sqrt{2}}(e_j - e_i) - \frac{2}{N} \sum_{i=1}^{N} G_{\sigma\sqrt{2}}(e_i) \end{aligned} \tag{5}$$

In this derivation, we used the fact that the convolution of two Gaussians is another Gaussian. This formulation was previously used to estimate Renyi's entropy [9], as well as for testing the null hypothesis of whether two random vector sets are drawn from the same distribution [10]. Due to these strong ties with Renyi's entropy, IIR filter adaptation using the criterion in (5) is called information theoretic learning (ITL).

4. WEAK CONVERGENCE OF ITL

The global convergence properties of ITL are investigated in a probabilistic framework. This approach requires the calculation of the escape probability of ITL from the domain of a local minimum. Due to Parzen windowing, the estimated pdf will asymptotically converge to the convolution of the actual error pdf with the kernel.

$$\lim_{N \to \infty} \hat{p}_e(\varepsilon) = p_e(\varepsilon) * G_\sigma(\varepsilon) \tag{6}$$

This can be interpreted as the addition of independent Gaussian noise to the error signal, i.e., $\hat{\varepsilon} = \varepsilon + N$. After solving the Fokker-Plank equations that arise from Ito's integral, under the separability assumption, we obtain the escape probability for single-weight as (details are in [5]).

$$p(\theta, t \mid \theta_*, t_0) = Gaussian(\mu, \Sigma) \tag{7}$$

where $\mu = \theta - \theta_*$ and

$$\Sigma = \sigma^2 \int_{-\infty}^{\infty} (\hat{p}_e(\theta_*, \varepsilon) - \delta(\varepsilon))^2 d\varepsilon \int_{t_0}^{t} \mu(s) ds \tag{8}$$

Method	Global (%)	Local (%)
LMS	48	52
LMS-SAS	58	42
NLMS	96	4
ITL	100	0

Table 1. System identification using Kautz model.

p	1	2	3	4	10	∞
MSE	0.94	0.29	0.24	0.22	0.22	0.22
ITL	1.59	0.37	0.26	0.22	0.18	0.17

Table 2. L_p impulse response errors with MSE and ITL.

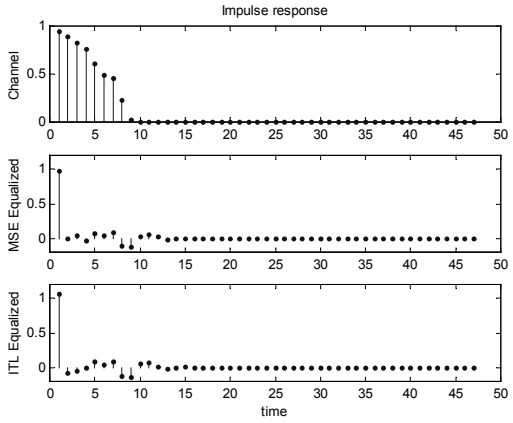

Figure 1. Channel and equalized channel impulse responses with NLMS and ITL.

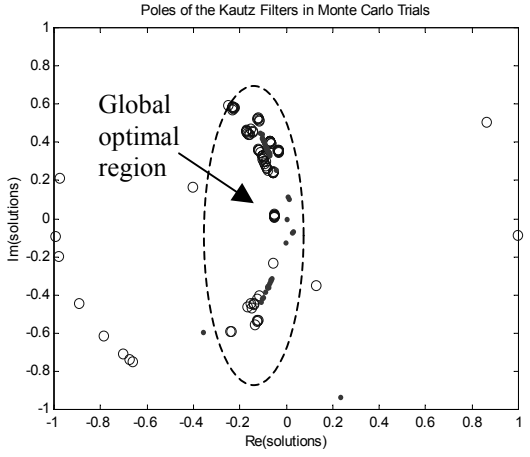

Figure 2. Poles of the Kautz filters after training.

Notice that the escape probability depends on the kernel size σ. In addition, the probability of escape from a local minimum is greater than the probability of escape from the global minimum, since the cost is smaller in the latter. However, the kernel size can be set to dominate the escape probability. Starting with a large kernel size, we improve the algorithm's chances of global optimization. However, the kernel size must be annealed slowly towards a small value to reduce bias. Designing an annealing schedule is an open problem.

An analogy between this method and the noise-injection approach in [11] can be formed. In [11], independent noise is added to the desired signal to facilitate global optimization, which is equivalent to adding noise to the error as in here. We have seen that the kernels effectively implement this idea. There is no physical noise, but the algorithm experiences an equivalent effect through the bias imposed by the kernel.

5. SIMULATION RESULTS

In order to demonstrate the effectiveness of the proposed approach, we first present the results of a Monte Carlo training example where the following 4th order *unknown* system is identified with a 2nd order Kautz filter. The details of the Kautz filter are in the Appendix [1,5].

$$H(z)=\frac{0.089-0.2199z^{-1}+0.2866z^{-2}-0.2199z^{-3}+0.089z^{-4}}{1-2.6918z^{-1}+3.5992z^{-2}-2.4466z^{-3}+0.8288z^{-4}} \quad (9)$$

Using standard LMS, LMS-SAS, NLMS, and ITL, we have performed 100 training sessions from random initial conditions for the Kautz filter parameters. The results are in Table 1 in the form of percentages of hitting the global and local minima. ITL achieves 100% global hits. In these experiments, the kernel size is annealed linearly according to the following formula.

$$\sigma^2 = 3(1-2.5\cdot 10^{-5}k)+0.5 \quad (10)$$

where k indicates the iteration index.

In order to understand what quality of the impulse response error ITL focuses on, we have computed the L_p norms of the impulse response error vectors corresponding to the global optimal solutions of NLMS and ITL algorithms. The impulse response error vector is defined as the difference between the actual impulse response of (9) and the impulse response of the final Kautz filter. Since these are IIR, the error is truncated at 100 taps, at which point the system impulse responses decay to insignificant values. These results are in Table 2. We observe that MSE emphasizes the lower norm indices (the L_2 of MSE is better) whereas ITL focuses on higher norms behaving like an L_∞ error minimizer.

Our second example addresses the echo cancellation problem. In general, echo can be modeled as the application of an FIR filter to the original source signal, which presumably comes through the line of sight (LOS) between the source and the sensor. Most probably, the power of the echo components will be smaller than the direct LOS component, mainly because of the additional distance traveled by the signal. In the following Monte Carlo simulations, we assume a randomly selected FIR channel model that conforms to these specifications. Since the ideal inverse of an FIR channel is IIR, again adaptive Kautz filters are used.

In this set of Monte Carlo simulations, the training is performed with recorded speech (note that the type of data has no significance in the procedure). For the channel whose impulse response is shown in Fig. 1, we have trained a second order Kautz filter (complex conjugate poles) using NLMS and ITL (with linear kernel annealing) with approximately 30000 samples. Starting from 100 randomly selected initial conditions for NLMS and 50 for ITL (due to simulation time constraints), the feedforward weights and the poles (usually the poles of an IIR filter are not adapted due to the mentioned difficulties) of the Kautz filter are optimized to determine the inverse of the channel. The equalization results for the best solutions obtained by the two algorithms are in Fig. 1.

To demonstrate the global convergence of ITL and NLMS, we present one of the complex conjugate poles obtained by these algorithms in the Monte Carlo trials in Fig. 2. We observe that NLMS failed to find a pole in the global optimum region 12/100 of the time, whereas for ITL, this ratio was 2/50. A slower annealing in these two simulations would result in global optimization.

6. CONCLUSIONS

The widespread use of adaptive IIR filters is prevented by two factors: the poor performance of currently available MSE-based training algorithms in avoiding local minima and the problems associated with maintaining the stability of the IIR filter during training.

In this paper, we demonstrated that a criterion that exhibits similarities to the information theoretic Renyi's quadratic entropy measure is useful in global optimization of IIR filters through annealing of its kernel parameter. Although we have provided a theoretical weak convergence result and demonstrated global convergence through Monte Carlo runs, the problem of setting the annealing schedule still remains an open problem.

Comparisons between the L_2 norms of the impulse response errors of the filters suggested by the optimization of MSE and ITL criteria resulted in favor of MSE. Therefore, if the global minimization of this L_2 norm is desired, then we suggest the use of a hybrid algorithm that starts with ITL and switches to MSE towards the end. This way the weights approach the global optimum with ITL and find the best L_2 solution through MSE training.

Acknowledgments: This work was partially supported by the NSF grant ECS-9900394.

APPENDIX

The Kautz filter output is a linear combination of outputs from a cascade of first-order Kautz units, $y_k = \mathbf{\Psi}_k^T \mathbf{w}$, where $\mathbf{w}$ is the weight vector and $\mathbf{\Psi}$ is the vector of outputs from individual Kautz units. The Kautz units are defined by the following transfer functions and gains.

$$K_i(z,\xi) = \gamma_0 \frac{z^{-1} - (-1)^i}{(1-\xi z^{-1})(1-\xi^* z^{-1})} \quad i = 0,1 \qquad \text{(A.1)}$$

$$K_i(z,\xi) = K_{i-2}(z,\xi) A(z,\xi) \quad i = 2,3,... \qquad \text{(A.2)}$$

$$A(z,\xi) = \frac{(z^{-1} - \xi)(z^{-1} + \xi^*)}{(1-\xi z^{-1})(1-\xi^* z^{-1})} \qquad \text{(A.3)}$$

$$\gamma_i = \left|1 + (-1)^i\right| \sqrt{\frac{1-\xi\xi^*}{2}} \quad i = 0,1 \qquad \text{(A.4)}$$

Here ξ is a complex pole of the filter, i.e., $\xi = \alpha + j\beta$.

REFERENCES

[1] W.H. Kautz, "Transient Synthesis in the Time Domain," *IRE Trans. Circuit Theory*, vol. 1, pp. 22-39, 1954.

[2] J.C. Principe, B. de Vries, P.G. de Oliviera, "The Gamma Filter - A New Class of Adaptive IIR Filters with Restricted Feedback," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 649-656, 1993.

[3] D. Erdogmus, J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *IEEE Trans. Neural Networks*, vol. 13, no. 5, pp. 1035-1044, 2002.

[4] K. Srinivasan, W. Edmonson, J.C. Principe, C. Wang, "A Global Least Square Algorithm for Adaptive IIR Filtering," *IEEE Trans. Circuits and Systems*, vol. 45, pp. 379-383, 1998.

[5] C.-A. Lai, "Global Optimization Algorithms for Adaptive Infinite Impulse Response Filters," Ph.D. Dissertation, University of Florida, Gainesville, Florida, 2002.

[6] A. Renyi, *Probability Theory*, American Elsevier Publishing Company, New York, 1970.

[7] E.T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review*, vol. 106, no. 4, pp. 620-630, 1957.

[8] E. Parzen, "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, Inc., San Diego, California, 1967.

[9] J.C. Principe, D. Xu, J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, vol I, S. Haykin (Ed.), Wiley, pp. 265-319, 2000.

[10] C. Diks, J. Houwelingen, F. Takens, J. deGoede, "Detecting Differences Between Delay Vector Distributions," *Phys. Rev. E*, vol. 53, pp. 2169-2176, 1996.

[11] C. Wang, J.C. Principe, "Training Neural Networks with Additive Noise in the Desired Signal," *IEEE Trans. Neural Networks*, vol. 10, no. 6, pp. 1511-1517, 1999.