Proc. 19th Conv. Elec. Electron. Eng. in Israel Oct. 1995

D2-4

ADAPTIVE SPEECH SIGNALS DEREVERBERATION

Y. Ephraim and D. Malah
Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa, Israel

Abstract

In this presentation we introduce the use of sequential adaptive filtering to the problem of speech signal dereverberation. Two processing methods are considered. Direct adaptive filtering in the time domain and a new proposition for complex adaptive filtering in separate frequency bands. The underlying algorithm for weights adaptation is the well known LMS algorithm. The input signals are obtained from two microphones and the dereverberation is based on the correlation properties of the speech signal and the reverberation at the two inputs. An analytical model is used to study the performance and properties of the two systems and to choose its parameters. The optimal parameters were found in simulations by means of a weighted spectral distance measure and by informal listening. These means were also used for comparing the two methods showing the superiority of the frequency domain approach. A comparison to a method recently developed at Bell Laboratories is also provided.

I. Introduction

In several practical speech recognition applications, the acoustic signal received by the microphone suffers from room reverberation which usually have an adverse effect on the temporal and spectral shape of the signal, as dipicted in Fig. 1.

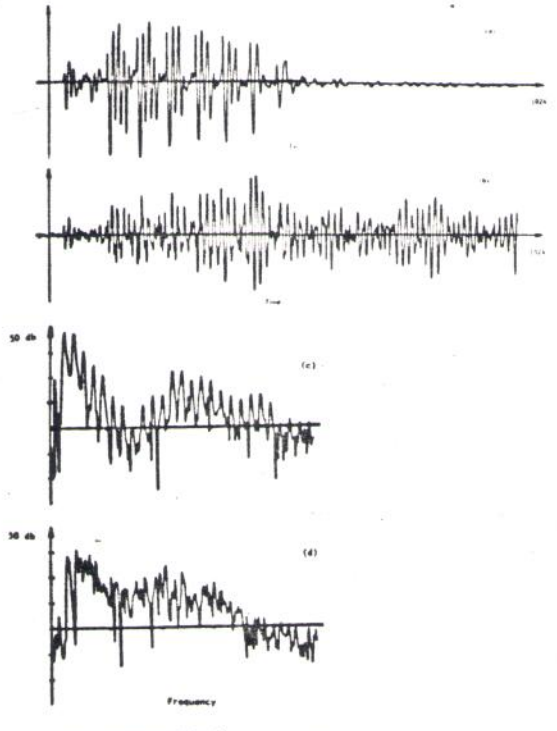

Fig. 1: (a) - source speech signal
(b) - reverberant signal, recorded in a room of (5.6, 6.0, 4.0) meters, radius vector to the source was (4.8, 4.0, 1.6) meters, radius vector to the microphone was (3.6, 3.9, 1.8) meters.
(c),(d)-Fourier transform (absolute value in log scale) of the signals in (a),(b) respectively.

This figure illustrates the breaking of the harmonic structure of the recorded speech, and also the spectral distortion. Clearly, the reverberation effects can cause great difficulties in the recognition process. For instance, any recognition process based on a spectral matching criterion, will not operate adequately. For efficient recognition, preprocessing is needed to remove room reverberation from the recorded signal.

In this paper, we address the problem of speech dereverberation, using sequential adaptive filtering based on the well known LMS (Least Mean Square) algorithm [1,2,3]. Two methods that utilize stereophonic recording are considered. Direct adaptive filtering and adaptive filtering is separate frequency bands.

Generally, reverberation are classified into two categories, depending on their time-delays. Short term reverberation, are characterized by a delay which is shorter than 60 msec, and they cause spectral distortion known as coloration [4]. Longer term reverberation, contribute time domain "tails" to the speech signals which are percieved as noise [5]. It is known [5,6], that long term reverberation recorded by two microphones in a closed room are uncorrelated. In addition, because of the nonstationary nature of speech signals and because of their spectral structure, the correlation between the source signal and the long term reverberation is low. For example, the source signal can be a voiced speech signal with a given pitch (fundamental frequency), while the long term reverberation can also be voiced but with another pitch (as a result of the usual change in pitch by the speaker).

Two dereverberation methods were previously proposed by Flanagan [4] and by Allen [5]. Flanagan's processor is effective for short term reverberation only. Allen's processor is effective for both reverberant degradations.

II. Direct Adaptive Filtering

Fig. 2 describes an adaptive noise canceller [7].

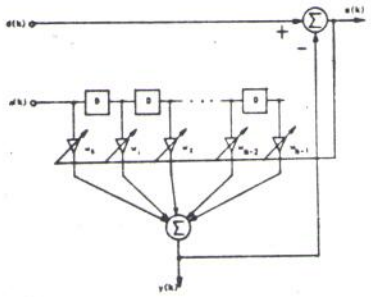

Fig. 2 Adaptive noise canceller

The adaptive filter operates on its reference input x(k), with a transversal filter whose impulse response is w(k), in order to minimize the mean square error (the expected value of $e^2(k)$). The filter w(k) is composed of a tapped delay line (TDL) and a set of variable weights controlled by the adaptation algorithm. The reference input x(k) feeds the TDL. The prediction output is the sum of the TDL outputs multiplied by the weights values. The error output (the difference between the primary input and the prediction output) is used by the adaptation algorithm to update the weights. The LMS algorithm is given by:

$$\underline{W}_{k+1} = \underline{W}_k + 2\mu e(k)\underline{X}_k \qquad (1)$$

where $\underline{W}_k$ and $\underline{X}_k$ are the weights and the reference input vectors, respectively, at the k-th iteration.

$$\underline{W}_k^T = [w_0(k), w_1(k), \ldots, w_{N-1}(k)] \qquad (2)$$

$$\underline{X}_k^T = [x_0(k), x_1(k), \ldots, x_{N-1}(k)] \qquad (3)$$

N - number of weights.

μ - a parameter (step-size factor in the adaptation algorithm) which controls convergence rate and stability.

The convergence rate is measured by the time constant, τ_{mse}, given by the time it takes the mean square error to decay to 1/e of its initial value (for stationary signals).

Fig. 3 describes the principle of using the direct adaptive filtering (DAF) to remove room reverberation, from speech recorded by two microphones.

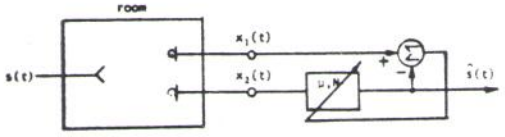

Fig. 3 Principle of using the direct adaptive filtering to dereverberate speech signals.

The output signals of the microphones $\{x_1(t), x_2(t)\}$ feed the two inputs of the filter. The dereverberated signal is taken from the prediction output. Since the two microphones are of equal distance from the source, $x_1(t)$ and $x_2(t)$ can be written as follows:

$$x_1(t) = s(t) + n_1(t) \qquad (4)$$

$$x_2(t) = s(t) + n_2(t) \qquad (5)$$

where $n_1(t)$ and $n_2(t)$ are the reverberation recorded by the two microphones. It is well known [7], that the adaptive filter provides the correlated part between its two input signals at its prediction output. Therefore, the prediction output in Fig. 3 will be composed of the source signal s(t) (common to the two inputs), if s(t), $n_1(t)$, $n_2(t)$ are mutually uncorrelated. This assumption can be made only if $n_1(t)$ and $n_2(t)$ are long term reverberation.

Actually, the prediction output will only be an estimate of s(t) for the following reasons:

1) The correlation between the source signal and the long term reverberation is not identically zero.

2) Any record will also contain short term reverberation, which will not vanish at the prediction output, since the short term reverberation at the two microphones are correlated.

3) Tracking problems of the adaptive filter, that arise due to the nonstationary nature of speech signals. Increasing the value of μ, improves the tracking ability [1,2], but will also increase the mean square error due to higher weights fluctuations [1,2]. In addition, as will be discussed later, increasing the value of μ result in increasing the reverberation power in the prediction output.

In order to find the optimal parameters of the adaptive filter, we assumed a mathematical model which is based on Glover's work [8], considering only voiced speech (which is quasi-periodic and can be assumed therefore to be harmonic).

According to Glover, if the reference input of the adaptive filter is composed of L sinusoidal components, $x(k) = \sum_{\ell=1}^{L} C_\ell \cos(\omega_\ell Tk + \theta_\ell)$, and assuming N is sufficiently large, the adaptive filter can be described as a linear time-invariant system. The transfer function of this system, from the primary input to the prediction output, is that of L band-pass filters (BPFs), centered at the reference frequencies $\{\omega_\ell\}_{\ell=1}^{L}$ and having a 3db bandwidth given by:

$$BW_\ell = \mu N C_\ell^2 / T \ [\text{rad/sec}]; \ \ell = 1, \ldots, L \qquad (6)$$

The amplitude of this transfer function at the frequencies $\{\omega_\ell\}_{\ell=1}^{L}$ is 1. The mean square error is composed of L geometric series, each with a time constant, τ_{mse_ℓ}, given by [9]:

$$\tau_{mse_\ell} \cong 1/BW_\ell \ [\text{sec}]; \ \ell = 1, \ldots, L \qquad (7)$$

Fig. 4 demonstrates the spectrum of the output signal from one microphone for two harmonic signals (source with the fundamental frequency ω_0 and long echo with the fundamental frequency ω_1)

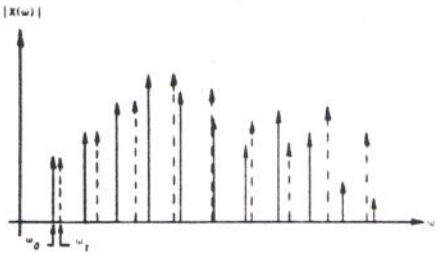

Fig. 4 Spectrum of the output signal from one microphone for two harmonic signals (source with fundamental frequency ω_0 and long echo with fundamental frequency ω_1).

For efficient filtering of this source signal, a "comb filter" with extremely narrow BPFs centered at the source harmonics is needed. This ideal filter should be able to change its center frequencies according to the changes in the source signal fundamental frequency (pitch). Feeding this filter from one microphone, will result in filtering out the source harmonic components and rejecting the echoes harmonic components.

The model introduced above, implies that the DAF can approximates this ideal filter. The adaptive filter will build a series of BPFs centered at its reference input harmonic components, that is, the harmonic components of the source signal and of the reverberation at one of the microphones. The output signal from the microphone that feeds the primary input, will pass through this set of BPFs. Since the source signal is common to the two inputs, and the amplitude of the transfer function of each BPF at its center frequency is 1, the source signal will pass unchanged to the prediction output. In contrast with that, the harmonic components of the long term reverberation at the primary input, whose frequencies are different from those at the reference input, will be attenuated at the prediction output. The amount of attenuation depends on the difference in pitch and on the adaptive filter parameters (μ,N). Decreasing the value of the product μN, will narrow each BPF (see (6)), but at the same time, will increase the convergence time-constant (see (7)). Since speech signals are quasi-stationary, there exists a lower bound for μN (to enable tracking), and therefore also for the bandwidth of each BPF. In light of this model, one should choose the parameters μ,N as follows: N should be sufficiently large so that Glover's model will hold. Once N is chosen, μ will be determined from (7) by the requirement of the achieving a desired value for τ_{mse}.

The main disadvantage of the DAF in this application, is its inflexibility in choosing the values of μ,N. Once these values are determined, the bandwidth of each BPF will be proportional to the power of the harmonic component at the center of the BPF. As a result, wide BPFs will be obtained at the formants regions (the peaks in the spectral envelop, with the first formant having the highest power content). This causes an unsufficient attenuation of the reverberation in the prediction output. An attempt to reduce μ will cause long time constants for the weak harmonics and hence distortion in the output signal.

Since the adaptive filter is a nonlinear system, its optimal parameters in a given application depends on the nature of its input signals. Therefore, the exact values of μ and N were determined by simulations. In these simulations, the DAF with various parameters pairs (μ,N) was tested for recorded speech signals with room reverberation. Because of the difficulty of choosing the best parameters, from a large number of trails, by listening to the filtered signals, a spectral matching measure (based on [10]) between the desired source signal and the filtered output signal was used. The values of μ,N, which minimize the "average distance" between the power spectrum envelops of the source and the filtered signals, were chosen [9]. These values were: μ = 0.03, N = 30. (the value of μ depends on the power of the reference signal to the filter [1]). Informal listening to the processed speech, with parameters values close to those determined by the spectral matching criterion, confirmed these values. It should be also remembered, that the operation of the adaptive filter depends on its inputs ordering. The result obtained with DAF approach are discussed in section IV.

Because of the inflexibility of the DAF as discussed above, a new approach based on complex adaptive filtering in separate frequency bands is proposed.

III. Adaptive Filtering in Separate Frequency Bands

Fig. 5 describes the block diagram of the proposed dereverberation system which is based on adaptive filtering in separate frequency bands.

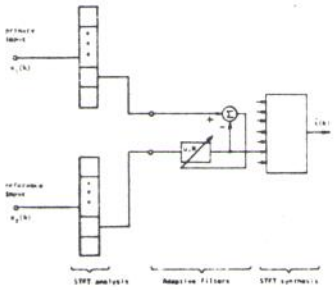

Fig. 5 Block diagram of the dereverberation system based on adaptive filtering in separate frequency bands.

Here, the spectrum of each recorded signal is divided into frequency bands of equal width and spacing according to the lowest possible pitch frequency expected in the source signal. In this way, each subband will contain only one source harmonic component. The signal in each subband is then shifted to baseband by complex modulation and decimated. Every two baseband signals corresponding to the same frequency band in the recorded signals, are passed through a complex adaptive filter. The prediction outputs baseband signals are interpolated, reallocated by complex demodulation and summed up to produce the dereverberated signal. This system enables one to independently choose the bandwidth of each of the BPF which are located at the separate subbands.

Fig. 6 describes the block diagram of the m-th channel.

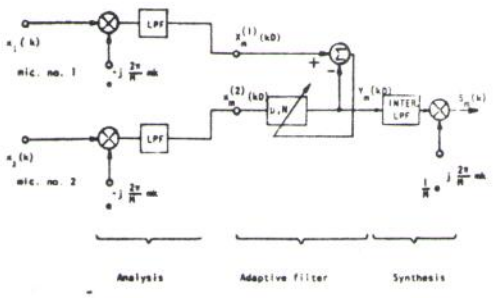

Fig. 6 Block diagram of the m-th channel of the system depicted in Fig. 5.

The analized signals are the short time Fourier transform (STFT) [11] of the two input signals.

$$X_m^{(i)}(kD) = \sum_{r=-\infty}^{\infty} x_{(i)}(r)h(kD-r)e^{-j\frac{2\pi}{M}rm} \qquad \begin{matrix} i = 1,2 \\ m = 0,1,\ldots,M-1 \end{matrix} \quad (9)$$

where h(k) is the unit sample response of a prototype LPF, D is the decimation ratio, M is the number of subbands, and the bandwidth of each subband is given by $\Delta\omega = 2\pi/MT$ rad/sec. (T is the sampling period of the input signals). The STFT in (9) can be implemented efficiently by using the FFT algorithm (if h(k) is of finite duration (FIR) and M is an integral power of 2) as described in [12,13]. The complex adaptive filtering is based on the complex LMS algorithm [14]:

$$\underline{W}_{k+1} = \underline{W}_k + 2\mu e(k)\underline{X}_k^* \quad (10)$$

where $\underline{X}_k^*$ is the complex conjugate of $\underline{X}_k$ in (3). The principles of the complex adaptive filtering of complex sinusoidals signals [9], are similar to those of the real case. Here, the bandwidth of each BPF is:

$$BW_\ell = \frac{4\mu NC_\ell^2}{T} \text{ [rad/sec]}; \quad \ell = 1,\ldots,L \quad (11)$$

and the τ_{mse_ℓ} for each component in the mean square error is:

$$\tau_{mse_\ell} \tilde{=} 1/BW_\ell \text{ [sec]}; \quad \ell = 1,\ldots,L \quad (12)$$

The synthesis is done by the FBS (Filter Bank Summation) method [11,13]. Its rule is:

$$\hat{s}(k) = \frac{1}{M}\sum_{m=0}^{M-1} Y_m(k)e^{j\frac{2\pi}{M}km} \qquad \forall k \quad (13)$$

$Y_m(k)$ is computed from $Y_m(kD)$ by interpolation. The interpolation and the weighted summation in (13) can be done efficiently by means of the FFT algorithm as discussed in [13].

The analysis-synthesis system alone is a unity system, iff h(k) has the following property [13]:

$$h(k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm M, \pm 2M, \ldots \end{cases} \quad (14)$$

One possible approach for designing the prototype LPF h(k), is by windowing [15]. Specifically, the unit sample response $(\sin(k\pi/M))/(k\pi/M)$ of an ideal LPF with cutoff frequencies $\pm\pi/M$ is multiplied by a smooth finite duration window. Here we used a Hamming window. h(k) designed in this manner will satisfy condition (14), and will result in nonoverlapping frequency subbands (except of the overlap due to the finite impulse response of h(k) which sum up to unity). The length and the shape of the window will determine the precise specifications of h(k). As the window length is increased, a better LPF is achieved, but then, the STFT signals do not reflect rapid changes in the input speech signals. We selected this length to be of 256 samples, for a sampling rate of 10 KHz. (This means that the STFT is done over a segment of 25.6 msec in which the speech signal is assumed to be quasi-stationary).

M is determined so that each subband will contain at most one source harmonic component. If we assume 100 Hz to be a typical fundamental frequency, than M = 128 will yield frequency subbands with bandwidths of 78.15 Hz (1/MT).

From Fig. 6 it can be seen that the STFT signal is the output of a LPF with bandwidth of $2\pi/MT$ rad/sec. Thus, from the sampling theorem it follows that the STFT signal can be computed every D-th value of k, where $D \leq M$. Since speech signals are quasi-stationary, as D increase, the adaptive filter input signals become less stationary, and a smaller τ_{mse} value is needed, giving a wider BPF which is built around each source and echoes harmonic components. This will cause an increase in the power of the reverberation at the prediction outputs. On the other hand, increases D has the advantage of the proportionally reducing the number of computations [9]. The exact value of D was determined by simulations as will be discussed in the sequel.

The interpolation filter can be designed in a similar way as the analysis LPF. Here, the unit sample response $(\sin(k\pi/D))/(k\pi/D)$ of an ideal LPF with cutoff frequencies $\pm\pi/D$ is multiplied by a finite duration Hamming window. The resulting impulse response fulfills the conditions required in [16]. Since we used a low value for D (see later), a relative short filter can be used. We have chosen a Hamming window of a length of 256 samples for 10 KHz sampling rate.

For efficient operation of the system, the values of (μ,N) for every adaptive filter should be properly determined. We have chosen to operate all adaptive filters with the same values for τ_{mse} and for N. Since we know only the expression for τ_{mse_ℓ} (12), we assumed τ_{mse} to be an average of $\{\tau_{mse_\ell}\}_{\ell=1}^{L}$ in the following sense:

$$\frac{1}{\tau_{mse}} = \sum_{\ell=1}^{L} \frac{1}{\tau_{mse_\ell}}$$

$$= 4\,\mu N \sum_{\ell=1}^{L} C_\ell^2 \qquad (15)$$

$$= 4\,\mu NP$$

where P is the power of the reference input signal of the adaptive filter. Thus, once the values of N and τ_{mse} are determined, the value of μ is computed from (15). The exact values for D,N, and τ_{mse} were determined by simulations and according to the spectral matching criterion. The values were: D = 16, N = 4, τ_{mse} = 33 iterations [9]. An asymmetrical behaviour of the system relative to its two inputs was observed. In all the simulations we have done, better results were achieved when the more reverberant signal (among the two inputs) feeds the primary input of the system, since this way, fewer sinusoidal components are present at the reference input, resulting in fewer BPFs in each subband.

IV Summary and Discussion

In this paper we introduced two methods based on adaptive filtering for removing long term reverberation from speech recorded in a room. The direct adaptive filtering approach was first analyzed and examined by simulations. To overcome the disadvantages of the DAF, a more flexible system, based on complex adaptive filtering in separate frequency bands was proposed and tested. These methods were compared among themselves and with a method for speech dereverberation, recently developed in Bell Laboratories (by Allen [5]), by informal listening tests and by a spectral distance measure.

All three methods were used to dereverberate speech recorded by two microphones in a closed room. The recording conditions were: room dimensions - (5.6, 6.0, 4.0) meters, radius vector to the source - (4.8, 4.0, 1.5) meters, radius vector to each microphone was (3.6, 3.9, 1.8) and (3.6, 4.1, 1.8) meters. Informal listening tests have shown that the signal processed by the DAF method, contains less reverberation than each recorded signals, but there is a degradation in speech quality. This method requires about 60 multiplications and 60 additions per input sample. The results obtained by the second method and by Allen's method seems to be similar. That is, the reverberation were significantly reduced without severely damaging speech quality. However, the distance measure used, favored our proposed system. Both methods are quite complex and require about 300 multiplications and 300 additions per input sample. For low reverberant situations, the less complex direct adaptive filtering approach could be applied.

Fig. 7 demonstrates the results obtained by the above three methods by means of spectrograms. The superiority of the adaptive filtering in separate frequency bands on the DAF is clearly seen. The final test for the proposed approach is in the score obtained in speech recognition tests in a reverberant environment. This is a subject for further research.

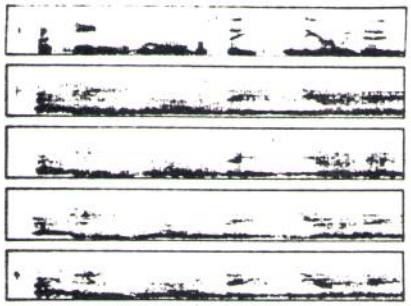

Fig. 7 Speech spectrograms: (a) source signal, (b) reverberant signal, (c) signal processed by direct adaptive filtering, (d) signal processed by complex adaptive filtering in separate frequency bands, (e) signal processed by Allen's processor.

References

[1] "Adaptive Filter", in Aspects of Network and System Theory, R. Kalman and N. Declaris, Eds. New York: Holt, Rinehart and Winston, 1971, pp. 563-587.

[2] B. Widrow et al., "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter", Proceedings of the IEEE, Vol. 64, No. 8, Aug. 1976.

[3] B. Widrow, "A Comparison of Adaptive Algorithms Based on the Methods of Steepest Descent and Random Search", IEEE Trans. on Antennas and Propagation, Vol. AP-24, No. 5, Sept. 1976.

[4] J.L. Flanagan, "Signal Processing to Reduce Multipath Distortion in Small Rooms", The Journal of the Acoustical Society of America, Vol. 47, No. 6, Feb. 1970, pp. 1475-1481.

[5] J.B. Allen et al., "Multimicrophone Signal Processing Technique to Remove Room Reverberation from Speech Signals", J. Acoust. Soc. Am., Vol. 62, No. 4, October 1977.

[6] A.H. Koenig et al., "Determination of Making-Level Differences in a Reverberant Environment", J. Acoust. Soc. Am., Vol. 61, No. 5, May, 1977.

[7] B. Widrow et al., "Adaptive Noise Cancelling: Principles and Applications", Proceedings of IEEE, Vol. 63, No. 12, December 1975.

[8] John R. Glover, "Adaptive Noise Cancelling Applied to Sinusoidal Interferences", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, No. 6, Dec. 1977.

[9] Y. Ephraim, "Adaptive Speech Signal Dereverberation". MS.c Thesis. Faculty of Electrical Engineering, Technion, I.I.T., Haifa, Israel, February 1979.

[10] Dennis H. Klatt, "A Digital Filter Bank for Spectral Matching", IEEE International Conference on Acoustics, Speech and Signal Processing, 1976.

[11] Jont B. Allen and Lawrence R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis", Proceedings of the IEEE, Vol. 65, No. 11, Nov. 1977.

[12] R.W. Schafer and L.R. Rabiner, "Design and Simulation of a Speech Analysis-Synthesis System Based on Short Time Fourier Analysis," IEEE Trans. Audio Electroacoust. Vol. AU-21, No. 3, pp. 165-174, June 1973.

[13] Michael R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transfor", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 3, June 1976.

[14] B. Widrow et al., "The Complex LMS Algorithm", Proceedings of the IEEE, April 1975, pp.719-720.

[15] Alan V. Oppenheim, Ronald W. Schafer, "Digital Signal Processing", Prentice-Hall, INC., Englewood Cliffs, New Jersey.

[16] Ronald W. Schafer and Lawrence R. Rabiner, "A Digital Signal Processing Approach to Interpolation", Proceedings of the IEEE, Vol. 61, No. 6, June 1973.