N94-22812

A HIGH QUALITY VOICE CODER WITH INTEGRATED ECHO CANCELLER AND VOICE ACTIVITY DETECTOR FOR MOBILE SATELLITE APPLICATIONS

A.M.Kondoz, B.G.Evans

Centre for Satellite Engineering Research
University of Surrey, Guildford, Surrey, GU2 5XH, U.K.
Tel: (0483) 509131, Fax: (0483) 34139

ABSTRACT

In the last decade low bit rate speech coding research has received much attention resulting in newly developed good quality speech coders operating at as low as 4.8 Kb/s. Although speech quality at around 8 Kb/s is acceptable for a wide variety of applications, at 4.8 Kb/s more improvements in quality are necessary to make it acceptable to the majority of applications and users. In addition to the required low bit rate with acceptable speech quality, other facilities such as integrated digital echo cancellation and voice activity detection are now becoming necessary to provide a cost effective and compact solution. In this paper we describe a CELP speech coder with integrated echo canceller and a voice activity detector all of which have been implemented on a single DSP32C with 32 KBytes of SRAM. The quality of CELP coded speech has been improved significantly by a new codebook implementation which also simplifies the encoder/decoder complexity making room for the integration of a 64-tap echo canceller together with a voice activity detector.

1. INTRODUCTION

After the successful development of low bit rate speech coding in the last decade, we are now beginning to see its real time application in various systems ranging from simple digital speech storage to very complex cellular and satellite mobile telephony. For mobile satellite communication systems, resources such as power and bandwidth are very limited. These systems employ very small transceiver terminals requiring larger satellite power. This is particularly true for land mobile and personal communication satellite systems, for which only a few MHz of bandwidth have been allocated on a primary basis. For such services to be economical they must be very spectrally efficient. In order to be competitive and use modulation schemes that will not cause excessive distortion over the difficult satellite propagation channel, digital coding of speech at around 4.8 Kb/s or less is required.

Robustness to channel errors is an important performance measure of a speech coding algorithm designed for mobile communication systems. Unlike most line communication systems, mobile communication systems have to withstand very high channel error rates. Speech coders are not expected to operate in these severe conditions without some form of forward error correction (FEC), but should be able to cope with residual errors of up to 2×10^{-2}. Hence, when designing a speech coder, its performance under channel errors should not be compromised for higher clean channel performance.

In addition to the speech quality under various channel conditions, the echo over long delay satellite systems should also be minimised. The disruption caused by echo in a system is proportional to the delay in that system, which becomes large if satellites are used. Also important is the use of the satellite cannel for other services when the speech channel is not active. By using a voice activity detector (VAD) for about 50 % of the time each speech channel can be used for other services thus doubling the usage of the allocated channel capacity. When there is no speech activity, the transmitter may also be turned off to conserve battery power and to reduce the interference to other users.

In the following sections we present a compact robust speech coder with integrated echo canceller and a VAD, implemented on a single DSP32C, which addresses and proposes solutions to the above problems.

2. CELP SPEECH CODING

At bit rates of 4.8 Kb/s CELP is the most widely reported speech coding algorithm. Although, CELP has been around for about a decade, it is only in the last few years that working implementations became available [1][2]. The processing stages of a CELP coder can be split into three blocks. The LPC analysis, long term prediction (LTP) analysis and the codebook search. It is the type of codebook excitation, computed every subframe which makes the difference between various CELP versions.

In the CELP using a standard codebook, it is assumed that the size of the codebook is large enough to cater for both voiced and unvoiced speech excitations.

However, at low bit rates, this assumption fails at the fast voiced onsets where, the LTP cannot built up fast enough to track the built up of voiced speech. Therefore, the speech quality deteriorates significantly as the bit rate is reduced by increasing the size of the subframe (vector). In addition during voiced regions where more than one main pitch excitation is necessary for good objective matching, the LTP predictor accuracy reduces very rapidly with the increase in the subframe size hence resulting in poor performance. In some versions, mixed codebooks are used, where, part of the codebook is similar to multi-pulse excitation, catering for the above cases. However, these approaches have preselected codebook structures which limits their performance. In the following a dynamic secondary excitation codebook is described where the excitation pulse positioning is made adaptive with the LTP lag (pitch) computed for the same subframe.

In pitch adaptive mixed excitation (PAME) the static codebook is split into two parts. The first part is made adaptive with respect to the LTP lag (pitch delay) as follows. The excitation buffer is filled with unit sample amplitudes, one every D samples starting from the first position. The rest of the vector elements are set to zero. During the search of the codebook, this vector is synthesised and its phase position is determined by shifting its synthetic response one sample at a time for $D-1$ times. Each phase position is then treated as a new excitation vector. In order to guard against pitch doubling errors in the LTP search, if the lag D is greater than $2D_{min}$ the same process is applied again by placing the excitation pulses every $D/2$ samples. The total number of excitation vectors that are searched is then found by adding all the phase positions considered. This is similar to regular pulse excitation with decimation factors of D and $D/2$. After selecting the best excitation vector from the adaptive section using C_a phase positions, the search continues in the fixed part of the codebook. Here, a further $C_f = C - C_a$ vectors are searched and the best performing vector index from the overall search process is transmitted to the receiver (C is the total number of allocated codebook indices). At the receiver, after decoding the LTP lag, the corresponding excitation vector is decoded.

By forcing the secondary excitation to have pitch structure, it is possible to match voiced onsets more accurately. This is because, firstly, the LTP memory builds up faster to track the incoming periodicity more accurately and secondly, the secondary excitation provides the required periodicity where the LTP fails. This of course depends on the accurate computation of the periodicity by the LTP in the first place. Many other adaptation schemes may be used to accurately place the secondary excitation pulses every pitch period. The LTP lag adaptation is useful because it does not require extra computation or bits. If all the possible pitch and phase combinations were to be used then for a subframe size of 60 (60+1+2+...+30+29+...+20) 750 combinations would be needed, which reduces to (40+1+2+..+20) 250 possibilities for a subframe size of 40. Encoding and decoding processes of the codebook index in this algorithm can have three possibilities which are communicated to the receiver by the fixed subframe size and the decoded LTP lag D.

The fixed part of the codebook was implemented in the form of a single shift centre clipped overlapping codebook. Overlapping codebooks are useful in reducing the computation of the codebook search as well as requiring less storage. If $h_w(n)$ represents the weighted STP impulse response, the synthesised vector $\hat{s}_k(n)$ due to the k^{th} excitation vector $x_k(n)$ of a single shift codebook is,

$$\hat{s}_k(n) = \sum_{i=0}^{n} h_w(i) x_k(n-i) \qquad (1)$$

Since, in a single shift codebook the difference between two consecutive vectors is only one sample at either end of the two vectors, the synthesised vector $\hat{s}_{k+1}$ can be computed in terms of $\hat{s}_k$ as,

$$\hat{s}_{k+1}(n) = x_{k+1}(0) h_w(n) + \hat{s}_k(n-1) \qquad (2)$$

The spectrum of STP and LTP inverse filtered speech is assumed to be flat. However, this is not strictly true. Therefore, the secondary excitation should be shaped accordingly to compensate for the model inaccuracies. Although, this is applicable to all secondary excitation types, it is more important in the pitch adaptive case where, placing a single pulse in time corresponds to a DC in frequency. To further improve the speech quality, therefore, adaptive shaping using the STP parameters can be applied to the secondary excitation. This shaping should be included in the AbS loop to make it most effective. A block diagram of a CELP encoder with pitch adaptive secondary excitation including adaptive shaping is shown in Figure 1. Although, the shaping process requires extra computation, if its impulse response, is computed together with the STP filter impulse response then during the search process, no additional computation is required. The only extra computation required both at the encoder and decoder is the implementation of shaping once every subframe, to produce the final output.

3. ECHO CANCELLATION

Echo in a telecommunication system is the delayed and distorted sound which is reflected back to the source. Echo is generated at the two-to-four wire converting hybrid transformer due to imperfect impedance matching. Here, we are only concerned with the electrical

echo for which the CCITT have a set of recommendations, G.165.

The source of electrical echo can be understood by considering a simplified block diagram of a connection between a pair of subscribers S1 and S2 shown in Figure 2. It can be seen from this block diagram that each subscriber has a two wire loop over which the received signal from the far end and the transmitted signal to the far end travel. The hybrid converts from 2 to 4 wire line working. The role of the hybrid is to direct the signal energy arriving from S1 and S2 to the upper and lower paths of the four wire circuit, respectively, without allowing any leakage back to the two sources over the opposite direction line pairs. Because of the impedance mismatching in the local loops however, some of the transmitted signal returns to its original source who hears a delayed version of his speech. This is called the talker echo, the subjective effect of which depends on the round trip delay around the loop. For short delays and reasonable attenuation (6 dB or more) the talker echo cannot be distinguished from the normal side tone of the telephone and hence does not cause problems. In applications such as satellites however, as a consequence of high altitude a round trip delay of 540 ms (270 ms each way) is possible, which makes the echo very disturbing and may in fact destroy conversation. When this is the case, it is essential to control or even to remove the echo.

A block diagram of an echo canceller for one direction of transmission is shown in Figure 3 where the far end signal is denoted by $y(i)$, the unwanted echo signal $r(i)$, and the near end talker signal $x(i)$. The near end talker signal and the echo are added together at the output of the hybrid. Since the far end signal is available as a reference for the echo canceller, the replica of the echo $\hat{r}(i)$ is estimated by matching the signals on both paths of the four wire section. This echo replica is then subtracted from the total of the returned echo and the near end signal as

$$u(i) = x(i) + r(i) - \hat{r}(i) \tag{3}$$

The difference between $r(i)$ the returned echo and $\hat{r}(i)$ the estimated echo should be as small as possible for good echo cancellation performance. The echo canceller produces the echo replica by using the far end reference signal in a transversal filter. If the impulse response of the filter is the same as the echo path response then the estimated echo and the returned echo become identical resulting in a perfect echo cancellation. Since the echo path response is not known in advance and may vary slowly with time, the coefficients of the transversal filter are adapted. In order to produce no distortion on the near end talker signal, the filter coefficients are only updated when there is no near end activity.

An echo canceller should in general satisfy the following fundamental requirements: Rapid convergence of the filter coefficients when turned on, very low echo when there is no near end speech, slow divergence when there is no far or near end speech and little divergence during times that both near and far end signals are present.

An echo canceller can be split into adaptive transversal filter, near end speech detection and residual error suppression parts. In a digital echo canceller both the reference and echo signals are available in digital form. Therefore the echo path impulse response can be represented in digital form denoting it by h_k to form

$$r(i) = \sum_{k=0}^{N-1} h_k y(i-k) \tag{4}$$

Assuming the system is linear and the echo path impulse response N is of finite length, then the echo canceller form the replica of the returned echo using

$$\hat{r}(i) = \sum_{k=0}^{N-1} a_k y(i-k) \tag{5}$$

When $a_k = h_k, \quad k = 0, 1, ..., N-1$ the returned and estimated echoes are identical resulting in no residual echo. The coefficients of the transversal filter are updated to match the slowly time varying echo path impulse response by minimising the mean squared residual error. When there is no near end speech ($x(i) = 0$) the filter coefficients are updated in such a way that as a result the residual error tends to a minimum. The update of the coefficients at each iteration is controlled by a step size β,

$$a_k(i+1) = a_k(i) + 2\beta e(i) y(i-k) \tag{6}$$

The convergence of the algorithm is determined by the stepsize β and the power of the far end signal $y(i)$. In general making β large speeds up the convergence, while a smaller β reduces the asymptotic cancellation error. The convergence time constant is inversely proportional to the power of $y(i)$, and that the algorithm converges very slowly for low signal levels. To overcome this situation, the loop gain is usually normalised by an estimate of the far end signal power.

$$2\beta = 2\beta(i) = \frac{\beta_1}{P_y(i)} \tag{7}$$

where β_1 is a compromise value of the step size constant and $P_y(i)$ is an estimate of the average power in $y(i)$ at time i. The far end signal power can be estimated by

$$P_y(i) = [L_y(i)]^2 \tag{8}$$

where

$$L_y(i+1) = (1-\rho)L_y(i) + \rho|y(i)| \qquad (9)$$

where a typical value of $\rho = 2^{-7}$. The above equation is only an estimate of the average signal level which is updated for every sample using the approximation for ease of implementation in real-time.

The quality of the echo canceller can be affected significantly if the near end speech is not detected accurately. This is because the filter coefficients will be adjusted wrongly and hence will distort the near end speech. Therefore the coefficients are only updated when there is no near end speech and kept fixed during near end activity to prevent divergence. The power estimate $\hat{s}(i)$ of the near end composite signal $s(i) = x(i) + r(i)$ is usually compared with the power estimate $\hat{y}(i)$ of far end signal $y(i)$ to decide if there is near end activity. The power estimates can be computed using

$$\hat{s}(i+1) = (1-\alpha)\hat{s}(i) + \alpha|s(i)| \qquad (10)$$

and

$$\hat{y}(i+1) = (1-\alpha)\hat{y}(i) + \alpha|y(i)| \qquad (11)$$

where a typical value for α is $1/32$. Near end speech is declared when

$$\hat{s}(i) \geq \ MAX[\hat{y}(i), \hat{y}(i-1), \hat{y}(i-N)] \qquad (12)$$

In order to avoid continuous switching, every time near end speech is detected, it is assumed to last for some time (typically 600 samples).

The echo canceller performance can be improved by a residual echo suppresser. This can be done simply by comparing the returned signal power with a threshold relative to the far end signal and completely eliminating it if it falls below the threshold. Again the returned signal power is estimated using

$$L_u(i+1) = (1-\rho)L_u(i) + \rho|u(i)| \qquad (13)$$

Whenever, $L_u(i)/L_y(i) < 2^{-4}$ the residual echo suppresser is activated. In some applications however, it may be perceptually more acceptable to leave a very low level of random signal to indicate that the line is not dead. The above algorithm with 64 tap filter has been implemented with a multi rate CELP coder operating at 4.8, 6.4 and 8 Kb/s and found to satisfy the CCITT recommendations completely [3].

4. VOICE ACTIVITY DETECTION

Discontinuous transmission (DTX) may be used to allocate the channel to other uses when there is no speech to be transmitted. DTX transmission simply makes use of the fact that every speech channel is not active continuously. In a duplex line conversation each party is active only for less than 50% of the time. Even during activity there are times that sizeable gaps between words and expressions exist. Therefore, by using a voice activity detector (VAD) to indicate active times, the channel may be allocated to another call when it is not needed. The use of VAD to indicate activity on the line may also be used to transmit non-speech data during the speech pauses which is very attractive in multi-media services. Alternatively, in mobile communication systems, the transmitter may be turned off to reduce the co-channel interference and also conserve the batteries of the hand held portable mobile terminals. This has been adopted as part of the overall specification of new mobile systems such as GSM. The efficiency or the gain of such systems depends on the performance of the VAD algorithm which has to work in severe background noise environments typical in a vehicle mounted mobile. The basic assumptions that the principles of a VAD algorithm can be based on are as follows:

1. Speech is a nonstationary signal. Its spectral shape usually changes after short periods of time, e.g., 20 to 30 ms.
2. Background noise is usually stationary during much longer time periods and changes very slowly with time.
3. Speech signal level is usually higher than the background noise level (otherwise speech will be unintelligible).

Using the above assumptions a VAD algorithm can be designed to detect silence gaps as well as distinguishing background noise with and without speech. In systems where the background noise level is very low, a simple signal energy threshold can be used to detect the silence regions. However, in systems where large and varying background noise is present, a much more intelligent algorithm needs to be used. This is typical of mobile systems where the mobile terminal is placed in a moving vehicle. In these systems, the noise level is very high and changing making it impossible to distinguish speech with background noise and background noise alone by using a simple energy threshold function. Since the level of the background noise could be changing, the threshold should be made adaptive. However the threshold should be updated only when there is no speech. For this the spectral characteristics are checked to see if it is likely to be speech with frequently changing spectral shape or noise with fairly stationary frequency response. Since speech can be classified as voiced with very slowly changing strong pitch, the change in periodicity within the frame time may also be considered to reinforce the confidence about speech detection. The accuracy of the VAD decision can be improved if the CELP LPC parameters are quantised

using LSFs. The LSF parameters for stationary signals remain fairly constant and do not change very rapidly. Therefore, the change in the LSF parameters from one frame to the next may be used to indicate signal stationarity [4].

If nonspeech decision is indicated a further conditioning is applied to eliminate the possibility of cutting out speech mid-bursts. This is done by adding a hangover stage to the VAD output. Before making the final decision that speech is not present a number of nonspeech frames have to be detected consecutively. This is determined by the length of the hangover time which is in the order of 60 to 100 ms (or 3 to 5 frames). If after the hangover time the decision still indicates nonspeech then the output of the VAD is used to indicate this to the DTX controller.

5. ROBUSTNESS TO CHANNEL ERRORS

The channel error performance mainly depends on the way the parameters are quantised and error protected. The most error sensitive CELP parameters are the LPC coefficients, followed by the excitation vector gain, LTP lag and gain, and finally the codebook index. CELP is generally robust up to the error rates of 10^{-3} without any FEC. The error protection can be split into two parts, one is achieved without any redundancy which is called built-in protection and the other is implemented by using extra redundancy bits in the form of FEC. By using Line Spectrum transformation, the residual errors on the LPC parameters can be controlled. The monotonicity of the Line Spectral Frequencies (LSF) can be used for error detection and correction [5]. This scheme is very effective at bit error rates (BER) of up to 10^{-2}.

Since the optimum vector gain is quantised as an absolute value, when one or more gain values are corrupted, they produce very annoying background noise. This is especially annoying when the speech is silence and errors result in very large decoded gain values, which produce loud blasts. To reduce this problem an optimum gain control algorithm has been developed and used [5].

It is assumed that in the worst case, the ratio of any subframe gain magnitude $|g(i)|$, $i = 1, \ldots, Q$ to the mean gain magnitude $\bar{g}$ of the corresponding frame, is not greater than a factor α. The second assumption is that the residual error rate is low enough that only one gain term per frame is corrupted. Again, this is a reasonable assumption for a satellite channel, except perhaps, during periods of deep fading. In the design, it is assumed that these (rare) situations will be detected and coped with by employing lost frame reconstruction techniques. Thus, at the receiver, we wish to find any gain for which $|g(i)|/\bar{g} > \alpha$, and then adjust it to achieve the desired ratio. There are two related problems that first need to be solved.

1. Whenever gain control is invoked, we are assuming that there is at least a single gain term corrupted within the frame. Thus, the average gain $\bar{g}$ for the frame cannot simply be computed from the sum of the received gains. This problem can be overcome by computing separate sub-average gains, $\bar{g}(i)$ for Q clusters of gains. The i^{th} cluster is composed of all the other subframe gains except $|g(i)|$. Thus,

$$\bar{g}(i) = \frac{1}{Q-1} \sum_{j=1}^{Q} |g(j)|, \; j \neq i = 1, \ldots, Q; \quad (14)$$

   where $g(i)$ is the excluded gain and Q, is the number of excitation subframes per frame.

2. Since we now have Q sub-average gains, we have to determine which one to use in the corruption tests. Since we are only interested in *upwardly corrupted* gains, and taking into account the assumption that the variance of the gains is limited, we would expect the cluster of gains with the corrupted gain to have the highest variance. Therefore, the best sub-average to use is that of the cluster with the minimum variance, as this is more likely to be closest to the full average gain $\bar{g}$. The cluster variances $\sigma^2(i)$ are calculated as,

$$\sigma^2(i) = \frac{1}{Q-1} \sum_{j=1}^{Q} (\bar{g}(i) - |g(j)|)^2, \; j \neq i = 1, \ldots, Q \quad (15)$$

Let the cluster with the minimum variance be cluster I. Then if the test for *upward corruption*, $|g(I)| \leq \alpha \bar{g}(I)$ fails, $g'(I)$ the controlled gain is reset as: $g'(I) = \bar{g}(I) \times sign[g(I)]$.

The ratio α is very important in these tests. If it is set too high, then a significant proportion of corrupted gains will pass the test, resulting in a degraded performance with channel errors. On the other hand, setting α too low means some uncorrupted gains will be 'adjusted', leading to degradation in the clear channel speech quality of the coder. In practice, it is very difficult to strike a reasonable balance which is speaker independent. A more attractive alternative is to adopt an adaptive approach in which the test factor changes according to the gains being considered. Since the test is trying to determine the deviation of the suspect gain from the mean gain, the test can be changed into,

$$|g(I)| \leq \alpha' \sigma(I) + \bar{g}(I) \quad (16)$$

where α', the adaptive factor is given by,

$$\alpha' = MIN[\sigma(i)]/\sigma(I), \quad I \neq i = 1, \ldots, Q \quad (17)$$

Parameters	Number/Frame	Bits/Frame
STP(LSF)	10	37
LTP	4×1-tap	4×(7+5)=48
CB index	4	4×7=28
CB gain	4	4×5=20
Total	-	139

Table 1: Configuration of CELP at 4.43 Kb/s with 30ms frame

This produces the best performance, giving only a negligible reduction in clear channel $segSNR$, whilst detecting a high proportion (about 88%) of the corrupted gains.

Built in error control for the LSFs and excitation vector gains was found to be sufficient for a maximum residual error rate of 10^{-2}.

6. CODER PERFORMANCE

The CELP coder described in Table 1 has been implemented on a single DSP32C with 32 Kbytes of SRAM. Using the same DSP a 64-tap echo canceller together with a VAD and built in error control was integrated with the main CELP encoder/decoder. This resulted in a complete solution in one DSP enabling a very cost effective and compact implementation. Since the coder uses 133 bits of the available 144 every 30 ms frame, the remaining 11 bits were used for very robust synchronisation. Every frame a sync-pattern that differed from any possible data patterns at least in two bit positions was used, enabling very fast locking time and robust synchronisation under channel errors.

The speech quality of the coder was assessed using informal listening tests. In these tests the CELP coder defined in Table 1 scored a MOS of 3.5. Its higher bit rate version at 6.65 Kb/s again defined by the same table but with a 50 Hz frame rate scored a MOS of 3.9 and was found to be better than the full rate GSM coder. Both versions were also found to be transparent to channel errors of up to 2×10^{-3}. At 10^{-2} very slight degradation was noticed due to corrupted LTP lags.

The performance of the 64-tap echo canceller was tested against the CCITT G.165 and was found to satisfy it fully. This coder has been used in various VSAT terminals produced by different manufacturers and found to be very acceptable by the users.

7. REFERENCES

[1] M.R.Schroeder, B.S.Atal "Code-excited linear prediction (CELP): High quality speech at very low bit rates", Proc. of ICASSP-85, pp 937-940.

[2] M.R.Suddle, A.M.Kondoz, B.G.Evans "DSP Implementation of Low Bit Rate CELP Based Speech Coders", Proc. 6th Int. Conf. on Digital Processing of Signals in Communications, Loughborough, U.K., Sep-1991, pp 309-314.

[3] M.R.Suddle, A.M.Kondoz, B.G.Evans "A Single DSP Multi-Rate Voice Coder with Integrated Echo Canceller", 3rd Int. Workshop on Digital Processing Techniques Applied to Speech Applications", ESA/ESTEC, Netherland, 1992.

[4] H.G.Asjadi "Real-time Implementation of Low Bit-rate Speech Coders for Satellite and Land Mobile Communications", PhD Thesis, University of Surrey, Guildford, U.K. 1990.

[5] S.A.Atungsiri "Joint Source and Channel Coding for Low Bit-rate Communication Systems", PhD Thesis, University of Surrey, Guildford, U.K. 1991.

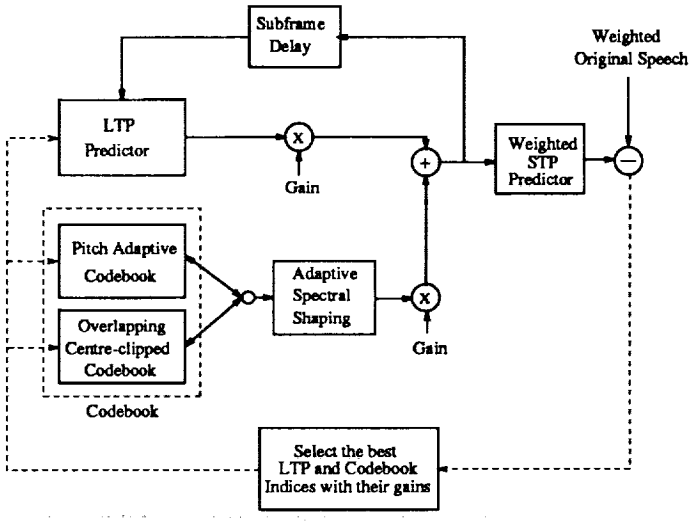

Figure 1: CELP with PAME excitation

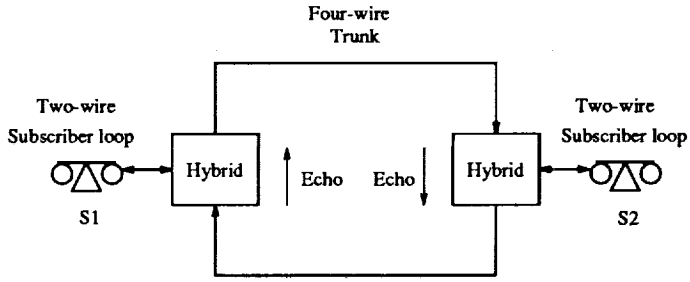

Figure 2: A duplex telphone line connection

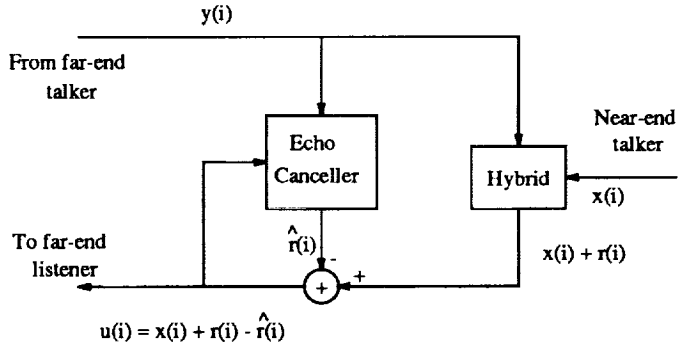

Figure 3: An Echo canceller set-up