A DISCRETE COSINE TRANSFORM BASED SPEECH ENCRYPTION SYSTEM

B.Goldburg*,S.Sridharan* and E. Dawson**

*School of Electrical and Electronic Systems Engineering
Queensland University of Technology

**School of Mathematics
Queensland University of Technology

ABSTRACT - A speech encryption system suitable for use on bandlimited transmission channels is described. Scrambling is achieved using permutation of discrete cosine coefficients. A method for removing energy variation in the scrambled speech has been incorperated into the scheme which significantly enhances its performance. Simulation results presented in the paper indicate that the scheme provides scrambled speech of low residual intelligibility, and recovered speech of good quality.

INTRODUCTION

Recent work in the field of analog encryption has focused on the use of transform domain scramblers (Matsunaga et. al., 1989) (Sridharan et.al, 1990). Such schemes are frame based and convert time domain vectors into the transform domain chosen. It is in this domain that the encryption is performed. The encrypted transform samples are converted back to the time domain and then transmitted. The main attraction of this method is that it should result in significantly lower residual intelligibility than the once common time domain schemes (Beker and Piper, 1982), provided the transformation removes speech redundancy.

The transform considered here is the Discrete Cosine Transform (DCT). The DCT is of interest since it is known to approach the optimal performance of the Karhunen Loeve Transform in terms of decorrelation of information. The paper aims to present the findings of research that has been conducted on a speech encryption scheme in which the DCT coefficients are permuted. The following sections introduce the process by which bandlimited scrambling is achieved, describe a method used to give a constant energy scrambled speech signal, address the need for synchronization and equalization, and outline the factors effecting the recovered speech quality. Finally the performance of the system, in relation to the residual intelligibility of the scrambled speech and the quality of the recovered speech, is evaluated using results obtained by simulation.

FORMULATION OF THE TRANSFORM DOMAIN SCRAMBLING PROCESS

Consider the DCT of a vector x of length N representing one frame of speech samples given by

$$u = Fx$$

where F denotes the transform matrix given by

$$F_{ij} = c(j)\cos[(2n+1)\frac{\pi j}{2N}] \qquad i,j = 0,1,\ldots,N-1$$

$$\begin{aligned} c(j) &= 1 && j = 0 \\ &= 2^{\frac{1}{2}} && j = 1,2,\ldots,N-1 \end{aligned}$$

The scrambling is performed by an NxN matrix P applied to the DCT vector u to produce a vector v given by

$$v = Pu$$

The scrambled speech y in the time domain is obtained by applying the inverse transformation F^{-1} on v given by

$$y = F^{-1}v$$

Restrictions have to be imposed on P due to bandwidth limitations of speech. Speech for telephony is restricted in bandwidth to 300-3400 Hz. The components outside the desired bandwidth are set to zero and the remainder are permuted. In the case for N = 256 samples per analysis frame, M = 197 components are permuted. This is in contrast to the 87 components available for permutation when the discrete Fourier transform is used under the same conditions (Sridharan et. al., 1990). Methods for generating such matrices so that all M! permutations are available have been addressed in (Sridharan et. al., 1990). This would imply for the proposed system, that 197! permutations are possible. Clearly, an exhaustive key search is infeasible.

The proposed DCT based implementation uses a fast algorithm to transform coefficients and requires only $(3N/2)(\log_2 N-1)+2$ real additions and $N\log_2 N - 3N/2+4$ real multiplications operations for each Fast Cosine Transform (FCT) (Chen et. al., 1977). This represents a factor of six improvement over the conventional double sized FFT approach.

ENERGY MODIFICATION USING INSERTION OF DUMMY COMPONENTS

The permutation of DCT coefficients preserves the signal energy within a given frame. Talk spurt and intonation information is still recoverable from the scrambled speech. To overcome this problem (Hasui et. al.,1984) proposed the substitution of dummy spectral components for a predefined block of components from the original speech spectrum. The magnitude of the components are chosen such that for any given frame the energy will be close to an established upper limit. This produces constant energy scrambled speech which resembles white noise. (Matsunaga et. al. 1989) suggested that dummy components should be adaptively positioned so that significant transform components would not be disgarded by the process. This requires that the components be recognizable as dummy components and hence run the risk of being detected and removed by a cryptanalyst.

In the proposed scheme the location of the dummy components is fixed, but is chosen carefully to ensure that the insertion process has little effect on the recovered speech quality. Five components with random amplitudes are used. They are scaled in order to maintain the desired constant energy limit. This operation is performed before permutation of the DCT coefficients. Following the scrambling operation the dummy components will be distributed throughout the spectrum. In this fashion these components should be undetectable due to their random nature.

Observe that for silent frames the five dummy components will be very easily detected since they will be the only components in the spectrum with a significant amplitude. To overcome this problem silent frames are treated as a special case. When the energy of an input frame of speech falls below a predefined threshold it is said to be silent. In this case the entire spectrum is replaced by a dummy spectrum whose components have been selected randomly such that their magnitudes match the amplitude distribution of non-silent frames. One component in the spectrum is used to indicate that such a substitution has been made. The scrambling process will move this component to a position unknown to a cryptanalyst.

Following the descrambling process the receiver must determine whether the current frame was originally a silent frame. It does this by interrogating the component used to signal such an event. If the frame was originally silent then the entire spectrum is replaced with a silent frame. If not then only the five dummy components are zeroed. If a sampling rate of 8 KHz is used, with a selection of N = 256 samples in each analysis frame, 197 spectral components lie within the usable bandwidth from 300 to 3400 Hz. Thus only three percent, (five dummy components and one signalling component) of the usable spectral components are lost as a result of this process.

FACTORS AFFECTING THE QUALITY OF THE RECEIVED SIGNAL

Permutation must be restricted to DCT components lying between 300 Hz and 3400 Hz so that the scrambling process does not increase the bandwidth. The components lying below 300 Hz and above 3400 Hz may carry a significant amount of information. They must therefore be set to zero to avoid passing this information on to unauthorized listeners. Thus there is degradation due to the bandlimitation of the signal. It should be noted however that this would occur regardless of the presence of the scrambling device.

Discontinuities are introduced at frame boundaries in the recovered speech due to the frame by frame analysis-synthesis process. This is evidenced by a low level flutter at the frame rate. A third and most important reason for the degradation is due to the channel impairments such as group delay distortion. This can be improved to a large extent using a channel equalizer. Adaptive update of the equalizer taps may be carried out using Widrow's algorithm (Widrow and Stearns, 1984). The scrambled voice was passed through a simulated telephone line during testing. It was found that an adaptively optimized 50 tap FIR filter was able to inverse model the channel effectively up to about 4kHz.

MEASURES FOR RESIDUAL INTELLIGIBILITY AND RECOVERED VOICE QUALITY

Voice quality of the recovered speech and the residual intelligibility of the encrypted speech are usually judged by subjective quality tests. Unfortunately these tests take much time and labour and require a large number of trained listeners. Even though intelligibility is a substantially subjective matter it is possible to use objective tests which are useful (if not ideal) indicators of intelligibility. Four objective measures were found useful in indicating the residual intelligibility of encrypted speech and the corresponding subjective quality of recovered speech. These were the LPC distance measure, cepstral distance measure, the segmental spectral signal to noise ratio and the frequency variant spectral distance measure. A description of these measures can be found in (Sridharan et. al., 1990).

SIMULATION RESULTS AND DISCUSSION

The scrambling schemes performance in terms of residual intelligibility and recovered voice quality under impaired channel conditions, was evaluated using the four objective measures mentioned in Section 5.

Tables 1 and 2 show the averaged LPC distance, cepstral distance, spectral segmental signal to noise ratio and frequency variant distance measure for 1000 frames (32 secs) of speech. Table 1 contains objective measures for the encrypted and recovered speech using the proposed system without the use of energy modification. Table 2 shows the corresponding measures obtained for the same speech sample but with the introduction of the dummy component insertion technique. The results for the scrambled speech show a dramatic improvement over the first case. However there is a slight degradation in recovered voice quality due to the loss of the components used in the energy modification process.

In general the objective measures suggest scrambled speech with extremely low residual intelligibility can be obtained using an energy modification technique. The recovered speech quality following passage through the transmission channel and equalization is quite acceptable.

Informal listening tests conducted by the authors confirm the findings of the objective tests.

CONCLUSIONS

An encryption system is described in which the DCT coefficients of speech segments are permuted to destroy speech intelligibility. Objective measures known to give good correlation to subjective testing indicate that this technique provides scrambled speech with very low residual intelligibility. The quality of the recovered speech is also reflected by the same measures. The proposed system uses adaptive equalization to remove channel distortion. It was shown that good quality speech can be recovered after adaptive equalization to compensate for channel distortion.

The system offers a significantly higher level of security than a similar DFT system. The number of transform components available for permutation is significantly more in the case of the DCT.

A method for the reduction of energy variations in the scrambled speech was described. It was noted that such a process significantly improves the scramblers performance.

REFERENCES

Beker, H. and Piper, F.(1982),"Cipher Systems : The Protection of Communications", (John Wiley and Sons).

Chen, W.,Harrison-Smith, C. and Fralick, S.C.(1977),"A Fast Computational Algorithm for the Discrete Cosine Transform", IEEE Transactions on Communications, p. 1004-1009.

Hasui, K. (1984),"A New Voice Band Encryption Method Using a Constant Envelope Scrambler", IEE Communications 84, p. 142-146.

Matsunaga, A., Koga, K. and Ohkawa, M.(1989),"An Analog Speech Scrambling System Using the FFT Technique with High Level Security",IEEE Journal on Selected Areas in Communications,VOL. 7,p. 540-547.

Sridharan, S.,Dawson, E. and Goldburg, B.(1990),"Speech Encryption using Discrete Orthogonal Transforms", IEEE Proceeding of ICASSP, p. 1646-1650.

Sridharan, S.,Dawson, E. and Goldburg, B.(1990),"A Fast Fourier Transform Based Speech Encryption System", IEE Communications Speech and Vision.

Widrow,B., and Stearns, S.D.(1984),"Adaptive signal processing", Prentice-Hall Englewood Cliffs.

Speech under Investigation	LPC Distance	Cepstral Distance	Spectral SSNR	F.V.S.D
Scrambled Speech	1.8193	3.4063	-1.0037	47.2679
Recovered Speech	0.2543	-2.5792	30.9711	16.8315

Table 1 - Comparison of Objective Measures for DCT scrambler without energy modification

Speech under Investigation	LPC Distance	Cepstral Distance	Spectral SSNR	F.V.S.D
Scrambled Speech	3.1068	4.6536	-37.7103	61.6520
Recovered Speech	0.3215	-1.6918	29.9156	23.6659

Table 2 - Comparison of Objective Measures for DCT scrambler with energy modification