SECURE SPEECH CODING FOR VOICE MESSAGING APPLICATIONS

J. Leis*, S. Sridharan† and W. Millan‡

*Faculty of Engineering & Surveying
University of Southern Queensland

†Signal Processing Research Center
Queensland University of Technology

‡Information Security Research Center
Queensland University of Technology

ABSTRACT – Voice messaging provides to the user the advantages of both electronic mail (providing message store and forward) and conventional voice telephony. It has the potential to take advantage of both aforementioned technologies, transcending both time and space in providing a more convenient and natural form of communication. A number of issues present themselves in the context of a store-and-forward voice mail system, in particular the ability to compress an entire message at once, and the problem of error control coding for a message which may not be decoded immediately. When combined with the need to secure the message via some cryptographic algorithm, the delayed-decode characteristics of voice mail create a problematic situation in the event of channel or storage errors, as conventional cryptographic algorithms propagate bit errors for the remainder of the message. Although the compression and message encryption may be viewed as independent, sequential stages in the processing of a voice message, this paper proposes a joint coding and encryption approach based on tolerable levels of distortion in the received voice message.

INTRODUCTION

The transmission of digitally encoded voice allows a convenient communications mode for humans, whilst providing the possibility of enhanced services in addition to the store-and-forward mode of textual email. Paramount in importance from the user's perspective is the provision of enhanced security provided via algorithmic encryption of the digitized voice. A less obvious, but arguably more important advantage of digital voice messaging systems is the ability to reduce both the transmission bandwidth (or, alternatively, the transmission time) and the storage capacity required for messages, thanks to sophisticated voice-compression algorithms.

SECURE VOICE MESSAGING SYSTEM

It is desirable for a voice messaging system to provide secure, reliable transmission of highly compressed voice data with high residual inteligibility, even over a noisy channel. The system shown in Figure 1 provides this using a variable rate speech coding algorithm, a self- synchronising cipher (SSC), and appropriate use of error-control techniques.

One of the major problems in speech encryption is bit-slip errors in the transmission channel which can cause catastrophic loss of synchronisation at the receiver. This results in a shifted keystream being used for decryption, so that the rest of the decrypted message is lost (half of the bits are in error) after the bit slip. Without a method for automatic re-synchronisation, the entire message will need to be resent, thus reducing throughput and potentially damaging

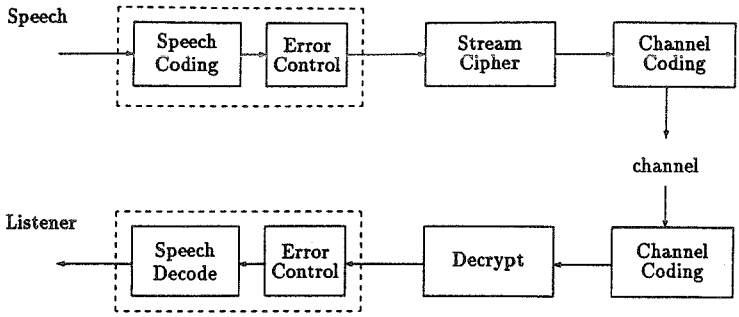

Figure 1: Block Diagram of Proposed System

security. Note that the need for a retransmission may not be discovered immediately, resulting in long delays. Message retransmission is a major problem in some applications, particularly in real time interactive situations, or when very large messages are involved. The overhead of retransmission can be quite significant when the channel is particularly error prone.

Consider also that the sender has no way of knowing that retransmission is required, unless the receiver requests they resend. In the case where the bit slip occurs early in the message, corrupting header information, the identity of the sender may be unknown to the receiver, so that the request for retransmission cannot be sent. In this case the message is totally lost, with the sender ignorant of the receivers' problem.

One technique for overcoming this problem is to use self-synchronising ciphers, which use the ciphertext itself as the synchronising information. Self-synchronous ciphers use ciphertext feedback to provide automatic re-synchronisation after a channel error. However, this is achieved at the cost of some error propagation. Each received bit will affect the decryption of $N+1$ message bits, where N is the amount of ciphertext feedback. A single error will produce a burst of errors $N+1$ bits in length. The value of N canot be too small without seriously reducing the security of the cipher. Maurer (Maurer 1991) calculated that a value of about $N = 100$ will be safe from an electronic codebook attack.

For some encryption applications, this error propagation may be intolerable. However, in speech transmission applications of 8000 bits per second (bps), a burst of 128 error bits will be perceived as an audible burst lasting only 16ms. Occurring infrequently, this will not significantly reduce the intelligibility of the received speech. In combination with effective error control on the channel, self-synchronous ciphers provide fast, secure encryption with automatic re-synchronisation.

THE VOICE SIGNAL

The voice signal may be viewed as a quasi-stationary signal from the analysis viewpoint. Figure 2 shows a segment of length 128 samples of a voiced speech waveform. In the encoding process, in all but the simplest one-sample quantization coding systems, a number of samples are taken in a block or frame of (usually) fixed length (some variable-length schemes have been reported in the literature, in order to provide a more 'natural' frame alignment boundary instead of arbitrary segmentation, for example (Shiraki & Honda 1988)). This is in order to exploit the quasi-stationarity referred to above, to achieve effectively non-integer bit rates. The frame length

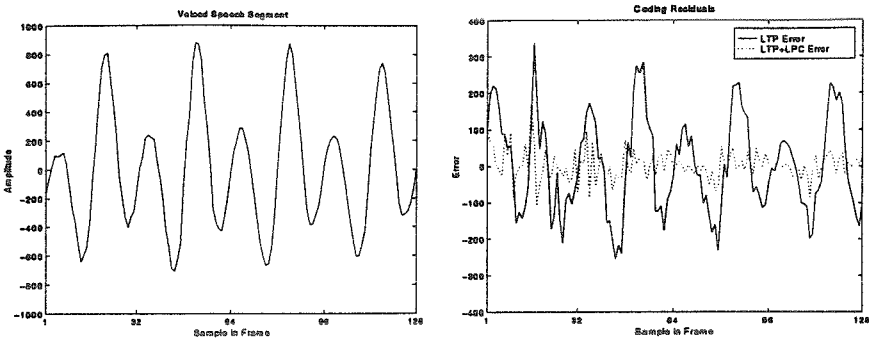

Figure 2: Voiced Speech *left* and Coding Residual

is normally of the order of 20 to 30 milliseconds in length (Campbell, Tremain & Welch 1992), (Yong 1994).In the present study, a frame length of 128 samples was chosen, together with a sample rate of 8 kHz. This gives an effective frame rate of 16 ms. The TIMIT voice database was chosen for the simulation experiments. This database consists of a number of sentences from differing dialect regions, sampled at a rate of 16 kHz and quantized linearly to 16 bit precision. This was then decimated to the required 8 kHz rate (to simulate telephone-bandwidth speech) using an 8^{th}-order Chebychev filter.

A number of compression algorithms specifically for speech have evolved over the years. We have used Code-Excited Linear Prediction algorithm for the current investigation. A pitch prediction algorithm is applied, using a one-tap pitch predictor. The autocorrelation method is used to produce a pitch lag estimate α and a pitch gain factor β. The pitch-prediction filter is then

$$H_l(z) = \frac{1}{1 - P_l(z)} = \frac{1}{1 - \beta z^{-\alpha}} \tag{1}$$

The residual from the pitch prediction is then modelled using an autoregressive model to form a 10^{th}-order all-pole short-term predictor of the form

$$H_s(z) = \frac{1}{1 - P_s(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

A 1024-entry Vector Quantizaion (VQ) stochastically-populated codebook is then generated from the resulting residual. The CELP algorithm is described further in a number of sources, for example (Salami, Hanzo, Steel, Wong & Wassell 1992).

SPEECH ENCRYPTION

Speech encryption can be carried out in voice messaging applications to provide privacy. Only authorised receivers (who have the secret key) can recover the original message. Conventional private key data encryption recognises two major classes of ciphers: block ciphers and stream ciphers. We have found that these broad classes are inadequate for voice mail applications in

Type of Cipher	Loss from Substitution Error	Loss from Bit Slip Error	Relative Speed
Block	Burst	all after	slow
Stream	1 bit	all after	fast
Self-synch	Burst	Burst	fast

Table 1: Comparison of Cipher Classes

situations where the channel is prone to certain types of transmission errors. We now discuss the effects of errors on the performance of these ciphers and explain the need for a self-synchronising cipher. Self-synchronising ciphers have some similarity to both block and stream ciphers, but display very different behavior in the presence of transmission errors. Self-synchronising ciphers have been classified as a sub-class of the stream ciphers (despite being closer to block ciphers in their internal working), and have been grossly understudied from a cryptographic perspective, since the theoretical framework they require differs markedly from that used in the analysis of stream ciphers, which have been thoroughly investigated. Self-synchronous ciphers must be considered as a third form of cipher. A comparison of the practical features of these three broad classes of ciphers is presented in Table 1. Given that for a voice messaging system, the loss of the an entire message is unacceptable, it is evident that a self-synchronising cipher must be used. Errors in a transmission channel occur in three forms: single substitution errors, insertion/deletion errors and burst errors. A substitution error occurs in a binary channel when noise causes the receiver to interpret one symbol as the other. Thus an incorrect substitution has occured. These errors generally occur independantly with low probability. Burst errors are said to occur in a transmission when many substitution errors occur close together in time, so that for the duration of the burst, the probability that any single bit is in error is much higher than normal. A bit slip error occurs when a bit is inserted into or deleted from the received transmission. This causes the information stream following the bit slip to be shifted. In conventional ciphers, these errors cause loss of synchronisation between sender and receiver and cause the remaining message to be decrypted incorrectly. The subsequent decrypted message appears random until re-synchronisation occurs. Self-synchronising ciphers automatically regain synchronisation a small time after a bit slip error, without the use of additional protocols.

SELF-SYNCHRONISATION

The essential structure of a self-synchronising cipher was discussed in (Maurer 1991).The canonical form of an SSC is shown in Figure 3. The keystream bit is the output of a keyed Boolean function. The function is selected by the secret key from a suitable set. The input to the function is the last N bits of ciphertext. This structure provides self-synchronisation by using the ciphertext itself as the synchronization information. The encryption of any bit uses the last N bits of ciphertext. Proper decryption requires the last N bits also, so that whenever $N+1$ bits of ciphertext are recieved without error, then a single bit will be decrypted correctly. When $N+n$ ciphertext bits are received without error, then n consecutive message bits will be correctly decrypted. When a single substitution error occurs in the recieved ciphertext, it effects the decryption of $N+1$ message bits. For a function that is cryptographically good about one-half of these decrypted bits will be in error. Thus the advantage of self-synchronisation cannot be separated from the propagation of errors. Once the parameter N is chosen, both the extent of error propagation and the level of basic security are set. An SSC output will not possess individual random errors: burst errors appear with the same probability that single errors appear at the input. It is recommended that SSC's be used with a burst-error correction scheme (to be applied after the decryption and before the decompression coding), and suitable error control on the channel to minimise the rate of single errors into the decryption stage. This minimises

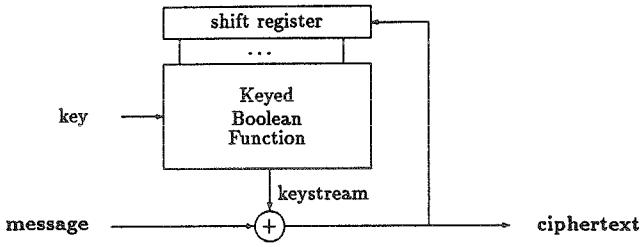

Figure 3: Essential Form of Self Synchronous Cipher

the rate of errors occuring at the input to the decoding stage.

SECURITY REQUIREMENTS FOR AN SSC

The problem of designing an SSC reduces to the problem of designing a keyed Boolean function. That is, finding a large set of functions that can be efficiently selected from by a pseudo-random key, such that each individual function satisfies a set of essential cryptographic criteria. A function which is poor with respect to any one of these criteria can be attacked easily. The basic criteria that a function must satisfy are:

- Balance in the number of 1s and 0s in the truth table;
- Output changes half the time when any input is changed, and
- Mimimal correlation to linear functions.

We approach the design of the SSC function from the perspective of attempting to find a set of functions which satisfy all the essential criteria discussed above, and then seeking ways to implement them, rather than the existing approach of guessing an implementation structure and then hoping that the resulting functions are secure. Fortunately, a set of functions have been identified which are almost balanced, have optimal avalanche characteristics with regard to changing any subset of function inputs, are the most non-linear functions (maximum Hamming distance from linear functions), and minimise the maximum correlation to any linear function. Rothaus (1976) introduced these, calling them "bent" functions. We recommend the use of functions that are have a small Hamming distance to bent functions for use in SSCs.

RESULTS AND CONCLUSIONS

Informal listening tests were conducted to simulate the effect of an entire corrupted frame. On the assumption that corrupted frames occur in isolation, and that the corrupted frame could be detected, the samples for individual frames were set to zero. This was found to have negligible effect on the intelligibility of the message. Increasing the number of samples set to zero indicates that, for the sections of the waveform with the greatest amplitude, a zero-run of 500 samples is perceptible but has little effect on the intelligibility of the speech. A zero-run of 1000 samples has a relatively minor effect on the recovered speech. We therefore conclude that up to approximately eight frames (1024 samples) may be discarded with only minor problems resulting for the message recipient. The pitch $P_l(z)$ and short-term prediction $P_s(z)$ parameters

must, however, be checked at the receiver for instability if another error detection strategy is not applied, as this may result in large-amplitude noise bursts.

REFERENCES

Campbell, J. P., Tremain, T. E. & Welch, V. C. (1992), The DOD 4.8kbps Standard, *in* 'Advances in Speech Coding', Marcel Dekker, Inc.

Corporation, L. D. (1990), *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology.

Kraus, T. P., Shure, L. & Little, J. N. (1994), *MATLAB Signal Processing Toolbox*, The Math Works, Inc.

Maurer, U. M. (1991), New Approaches to the Design of Self-Synchronisng Stream Cyphers, *in* 'Proc. Eurocrypt-91', pp. 458–471.

Salami, R. A., Hanzo, L., Steele, R., Wong, K. H. J. & Wassell, I. (1992), *Mobile Radio Communications*, Pentech, chapter 3 - Speech Coding.

Shiraki, Y. & Honda, M. (1988), 'LPC Speech Coding Based on Variable-Length Segment Quantization', *IEEE Transactions on Acoustics, Speech and Signal Processing* **36**(9), 1437–1439.

Yong, M. (1994), 'A New LPC Interpolation Technique for CELP Coders', *IEEE Transactions on Communications* **42**(1), 34–38.