SELECTIVE ENCRYPTION AND SCALABLE SPEECH CODING FOR VOICE COMMUNICATIONS OVER MULTI-HOP WIRELESS LINKS

J. D. Gibson, A. Servetti[1] , H. Dong, A. Gersho, T. Lookabaugh[2], J. C. De Martin[3]

Dept. of Electrical & Computer Engineering
University of California
Santa Barbara, CA 93106, USA
{gibson,huidong,gersho}@ece.ucsb.edu

[2]Dept. of Computer Science
University of Colorado
Boulder, CO 80309, USA
Tom.Lookabaugh@colorado.edu

[1]Dip. di Automatica e Informatica/[3]IEIIT-CNR
Politecnico di Torino
Corso Duca degli Abruzzi, 24 – I-10129 Torino, Italy
{servetti,demartin}@polito.it

ABSTRACT

With the rapid deployment of wireless LANs and the future development of mobile ad hoc networks, multi-hop wireless communications links are expected to become much more common. How to achieve reliable and efficient, yet secure, voice communication over such multi-hop wireless links is an important issue. This paper proposes and investigates a combination of scalable speech coding and selective encryption for secure voice communication over multi-hop wireless links that addresses both the efficient use of network and node resources and security against unwanted eavesdroppers. It is shown that when the Shannon lower bound is satisfied with equality for rate distortion optimal scalable coding, transmission of the enhancement layer in-the-clear provides no information regarding the core layer. A specific example of combining selective encryption with MPEG-4 scalable speech coding demonstrates the effectiveness and efficiency of the approach.

INTRODUCTION

Wireless local area networks (WLANs) are enjoying widespread popularity and rapid growth. As a result, efforts are underway to interconnect existing digital cellular systems with WLANs to enhance data and voice communications. Additionally, research and development on mobile ad hoc networks (MANETs) is proceeding at a rapid pace as well. Mobile ad hoc networks are a self-organizing network of mobile nodes connected through wireless links, wherein the mobile nodes operate not only as hosts but also as network routers. MANETs are attractive for a host of applications and environments, such as emergency response scenarios, law enforcement and military activities, conference and convention center communications, and "always on" connectivity for a single mobile user. One critical component of all of these applications is expected to be packet voice communications. Voice over IP (VoIP) over MANETs and other multi-hop wireless links offer particular challenges. Among these challenges are (i) reliability, because of the wireless communications channels, and in the case of MANETs, the rapid reconfigurability of the networks, (ii) security against passive eavesdroppers due to broadcast wireless communications, (ii) bandwidth efficiency due to shared communications channels, and (iv) network and node resource conservation [1, 2].

In this paper, we propose and investigate an approach for providing bandwidth efficient speech coding with security against passive eavesdropping that also conserves valuable network and node resources. The proposed method involves the combination of scalable speech coding methods with selective encryption of the core layer data stream. In the present work, we are assuming that the eavesdropper is relatively passive and will listen in on voice conversations only if it is reasonably simple to do so. This is in contrast to a dedicated attacker that might perform concentrated, off-line cryptanalysis.

An important type of scalable speech coding, called SNR scalability, consists of a minimum rate bit stream that provides acceptable coded speech quality, along with one or more enhancement bit streams, which when combined with a lower rate coded bit stream, provide improved speech quality. SNR scalable speech coding addresses both the bandwidth efficiency and the resource conservation problems by allowing the nodes to prune the enhancement layers when wireless channels become congested, or in the case of MANETs, in order to conserve mobile node battery power, by avoiding excessive transmissions of enhancement layer bit streams [3].

Selective encryption is a technique wherein only a selected subset of the transmitted data is protected and the remainder of the data stream is sent in the clear [4, 5]. By not encrypting the entire data stream, valuable node resources

A. Servetti's work was performed during a stay as visiting researcher at UCSB.

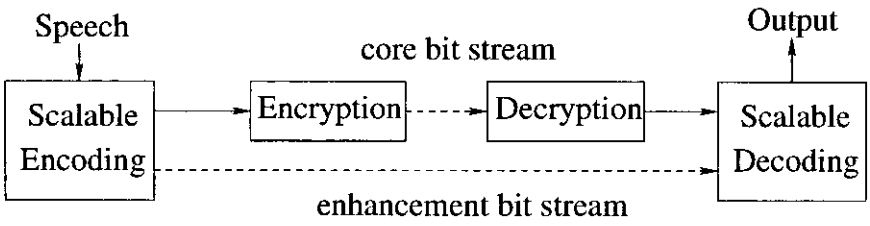


Fig. 1. Scalable coding with selective encryption

can be conserved.

In our work, we propose that the core bit stream of an SNR scalable speech coder be chosen as the data subset to be encrypted and that the enhancement layer(s) be sent in the clear, as illustrated by Fig. 1. Previous investigations of selective encryption for speech coding have not employed scalable speech coding structures, and hence, although sufficient security is provided, the efficiencies of SNR scalability are not exploited. We note here that the separate encryption of all layers of scalable video coders for video streaming has been investigated in [6].

To validate this concept, it is essential that both theoretical issues and practical speech coding structures both be addressed. We address the theoretical secrecy of the approach by using the information theoretic treatment of SNR scalability, called successive refinement of information, and the information theoretic definition of secrecy in terms of the equivocation. More practical aspects are addressed by using the MPEG-4 scalable speech coding tool to code speech in layers below 8 kilobits/sec (kbps).

The paper is outlined as follows. In Section 2, issues that are particular to voice communications over multi-hop wireless links are discussed. SNR scalability and SNR scalable speech coding techniques are examined in Section 3, followed by a treatment of selective encryption with respect to speech coding in Section 4. Section 5 presents the information theoretic essentials concerning SNR scalability and Section 6 provides an overview of theoretical secrecy from the information theory viewpoint. Specific examples of the proposed method based upon the MPEG-4 scalable speech coding tools are given in Section 7, including possible reconstruction attacks by an eavesdropper. Conclusions are presented in Section 8.

VOICE COMMUNICATIONS OVER MULTI-HOP WIRELESS LINKS

As an example of a network with multi-hop wireless links, consider the diagram of a mobile ad hoc network as shown in Fig. 2, where we wish to consider voice communications between nodes A and B. This figure shows multiple hop and multiple path connections between node A and node B. Note that even though the three paths from node A to node B in the figure are shown as nonoverlapping wired connections, these connections are over wireless links, and thus in some way, share the same channel and channel bandwidth. Therefore, this figure highlights at least two important points for voice communications over MANETs [7, 8].

One point is that the wireless links are shared resources, and therefore, bandwidth efficient coding strategies are of interest (to allow control of network loading) and the transmissions are susceptible to interception and eavesdropping. A second point, which raises two additional issues, is that the voice connections will often consist of multiple hops, wherein intervening mobile nodes serve as routers. To preserve the resources of these mobile nodes that are serving as routers, it is desirable to minimize transmissions by these nodes. Thus, when mobile node resources, such as battery power, become low, it is desirable that these mobile nodes be able to reduce their transmitted power. Furthermore, since the voice communications link flows through these nodes, the possibility of unwanted, passive eavesdropping by these nodes is raised. To provide security during the multi-hop, wireless connections, some type of encryption of the voice data stream is desirable. However, this encryption should demand as little signal processing as possible at the source and destination nodes in order to preserve the battery power resources of the nodes at the connection endpoints, i. e., nodes A and B in the figure.

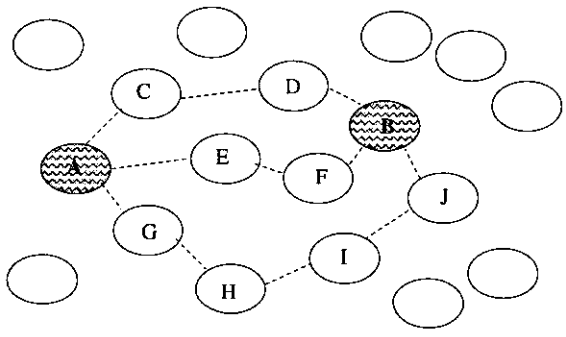


Fig. 2. Example of multi-hop and multi-path communication in an ad hoc network

SCALABLE SPEECH CODING

Scalable speech coding consists of a minimum rate bit stream that provides acceptable coded speech quality, along with one or more enhancement bit streams, which when combined with a lower rate coded bit stream, provide improved speech quality. Scalable coding has been investigated outside of the current context because it has a number of advantages[3]. However, in this paper, we focus on the key new issues for multi-hop wireless voice communications as outlined in Section 2.

The only existing standards for SNR scalable telephone bandwidth speech coding are G.727 (ADPCM)[9], which operates at rates of 16 kbps and above, and the MPEG-4 speech coding tools [10]. There is a standards activity in the digital cellular community to develop an SNR scalable speech coder, but it is not near a solution at this time.

The G.727 speech coding standard is based upon adaptive differential pulse code modulation and it operates at data rates of 16, 24, 32, and 40 kbps[9]. The core bit rate is 16 kbps, and up to three 8 kbps enhancement layers can be included. Due to the relatively high bit rates (a core layer of 16 kbps or greater), we do not consider G.727 further in this paper.

The MPEG-4 Natural Speech Coding Tool Set is a component of the MPEG-4 natural audio coding system[10]. For coding telephone bandwidth speech, it uses code-excited linear prediction (CELP) with a multipulse excitation codebook. There are 28 possible bit rates from 3.85 kbps to 12.2 kbps. SNR or bit rate scalability of any of these core layer bit rates is possible in increments of 2 kbps, with up to three enhancement layers.

The block diagram of the MPEG-4 CELP SNR scalable coder is shown in Fig. 3. SNR scalability in the MPEG-4 CELP coder is achieved by encoding the speech signal using a combination of the core coder and the bit rate scalable tool.

The enhancement layer encodes the residual signal produced at the core layer by minimizing the perceptually weighted distortion between the reconstruction error signal from the core and the output signal from the enhancement layer using additional fixed excitation codebook vectors. At the decoder, the new excitation vectors are summed with the excitation in the core layer to obtain the refined speech output.

SELECTIVE ENCRYPTION

Protection against unwanted eavesdropping is essential for the viability of wireless multimedia services. Furthermore, in many wireless applications, network resources, such as bandwidth, and node resources, such as battery power, must be conserved. Since full encryption of transmitted data streams can place a heavy signal processing burden on originating and receiving nodes, one is led to consider the concept of partial encryption of the data streams. In partial encryption only a percentage of the transmitted data stream is processed by an encryption algorithm, with the remainder of the data stream being sent in the clear.

The questions to be addressed in partial encryption are: (i) What data must be encrypted to provide the needed level of security? (ii) Can a cryptanalytic attack on the data sent in the clear be mounted that will allow important information about the transmitted data to be discerned by an eavesdropper? (iii) What is the percentage of the data stream that must be protected? Clearly, the data chosen to be protected must be the "most important" bits in terms of reconstruction of the content from the overall data stream, and this idea has lead partial encryption to sometimes be denoted as selective encryption, which is the terminology that we adopt in this paper.

While several studies of selective encryption for video and image compression have been performed and documented[11, 12, 5], very few results on selective encryption of coded speech have been presented. Servetti and De Martin [4] investigated partial encryption of G.729 at 8 kbps with respect to what bits should be encrypted to provide security with respect to several factors, including intelligibility, gender identification, plain-text identification, and speech/non-speech discrimination. They demonstrate that partial encryption of about 45% of the bitstream provides protection equivalent to full encryption, and that encryption of as little as 30% of the bitstream precludes intelligibility.

The results from [4] for G.729 could be used to address the conservation of host and destination node battery power in wireless networks. However, since G.729 is not SNR scalable and the encrypted bits alone do not allow acceptable reconstruction of the coded speech, this approach does not offer the feature of pruning transmitted data rate to conserve node transmitted power or to reduce network bandwidth utilization as needed. The combination of selective encryption with an SNR scalable coder provides an opportunity to address both issues that are important for multi-hop wireless links.

SUCCESSIVE REFINEMENT OF INFORMATION

SNR scalability has been developed from the rate distortion theory viewpoint under the label of successive refinement of information. A sequence of random variables $X_1, \cdots, X_n$ is said to be successively refined without rate loss by a two-

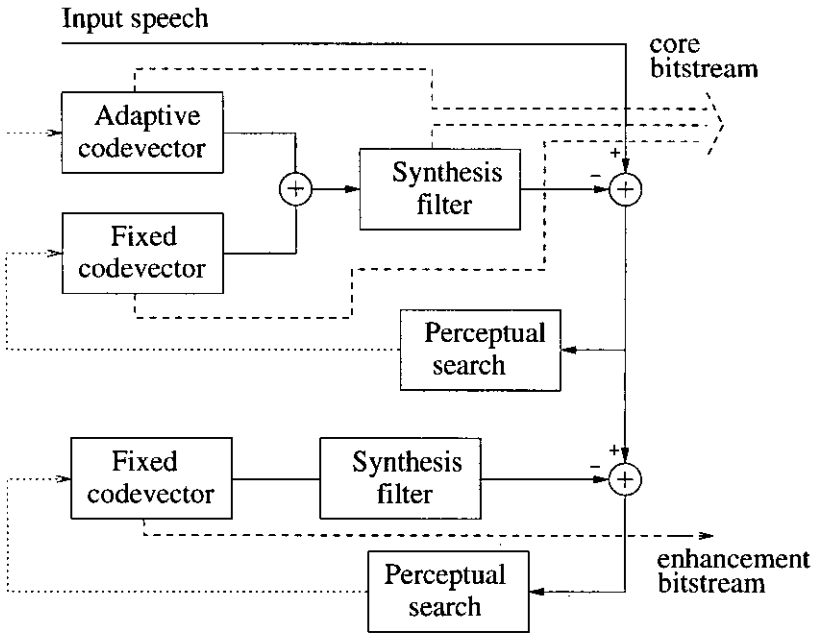


Fig. 3. The MPEG-4 CELP bit rate (SNR) scalable coder.

stage description if the description is rate distortion optimal at each stage. The X sequence is encoded as $\hat{X}$ at rate R_1 bits per symbol with average distortion D_1. Then information is added to the first message at rate R_2 bits per symbol so that the two-stage resulting reconstruction $\hat{X}_r$ now has average distortion D_2 at rate $R_2 + R_1$.

Most rate distortion theory research related to SNR scalability has been concerned with finding the conditions under which successive refinement without rate loss is achievable. The successive refinement problem was first introduced by Koshelev [13, 14] as the problem of divisibility, and he proved the sufficiency of a Markov chain relationship between the source and the refined reconstructions in 1980. Equitz and Cover [15] proved necessity in 1991. They also showed, using the Shannon backward channel formulation, that the Markov chain condition holds for Gaussian sources and squared error distortion, Laplacian sources and the absolute error distortion, and all discrete sources and Hamming distortion measures. Rimoldi[16] was the first to observe that this achievable rate region allowed different distortion measures at each refinement stage.

Theorem: [13, 14, 15, 16] Successive refinement with distortion D_1 and D_2 ($D_1 \geq D_2$) can be achieved if and only if there exists a conditional distribution $p(\hat{x}, \hat{x_r} \mid x)$ with $Ed(X, \hat{X}) \leq D_1$ and $Ed(X, \hat{X}_r) \leq D_2$, such that $R(D_1) = I(X; \hat{X})$ and $R(D_2) = I(X; \hat{X}_r)$ and $p(\hat{x}, \hat{x_r} \mid x) = p(\hat{x}_r \mid x)p(\hat{x} \mid \hat{x}_r)$.

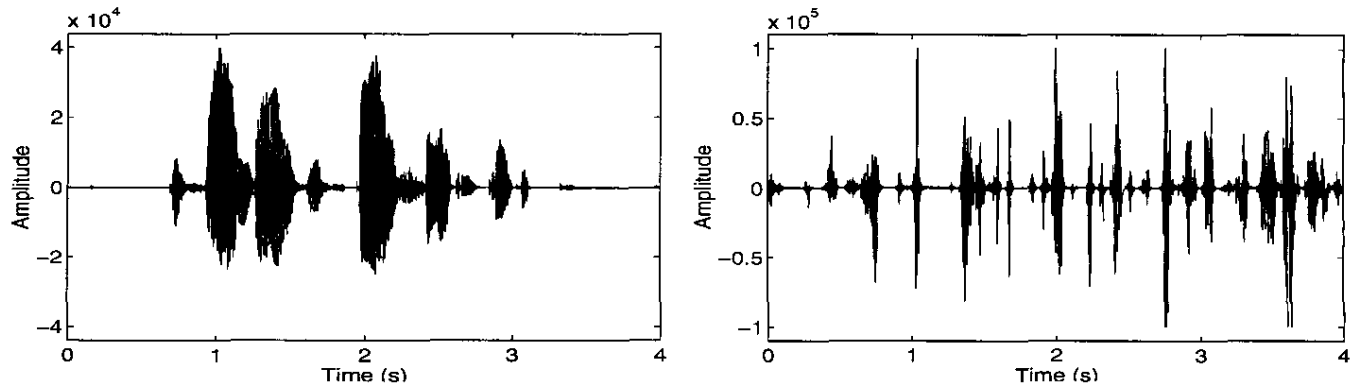
$\Diamond$

The last condition is equivalent to saying that X, $\hat{X}$, $\hat{X}_r$ can be written as the Markov chain $X \rightarrow \hat{X}_r \rightarrow \hat{X}$, or, equivalently, as $\hat{X} \rightarrow \hat{X}_r \rightarrow X$.

THEORETICAL SECRECY

Since the enhancement layer is sent in-the-clear, the eavesdropper can use the enhancement layer bitstream to glean as much information as possible about the core layer. In order to measure how much information the enhancement layer provides about the core layer, we use the equivocation of the core layer ($\hat{X}$) given the enhancement layer information ($\hat{X}_2$), denoted here as $H(\hat{X}|\hat{X}_2)$. Shannon introduced this quantity in an originally-classified 1945 report wherein he called the equivocation the theoretical secrecy since it assumes unlimited computational resources for the eavesdropper [17].

To obtain an expression for the above-defined equivocation, we start with the results concerning successive refinement of information from Section 5. In particular, we focus on the Markov chain condition that is necessary and sufficient for successive refinement without rate loss that is expressible as $X \leftrightarrow \hat{X}_r \leftrightarrow \hat{X}$. We also note that $\hat{X}$ achieves average distortion D_1 and $\hat{X}_r$ achieves average distortion D_2. Since the evaluation of rate distortion functions is often difficult, in 1959, Shannon derived a lower bound on the rate distortion function (now called the Shannon lower bound) that is valid for the important case of difference distortion measures [18, 19]. A key point is that in those situations where we can find conditions when the Shannon lower



(a) Original speech signal

(b) Selectively encrypted speech signal

Fig. 4. Selective encryption (b) of a sentence (a) (English text "The swan dive was far short of perfect.").

bound is satisfied with equality, we then have an exact expression for the rate distortion function.

When the Shannon lower bound is satisfied with equality in the Markov chain condition for successive refinement of information, we can write that $X = \hat{X}_r + E_r$, with E_r statistically independent of $\hat{X}_r$, and further that $\hat{X}_r = \hat{X} + \hat{X}_2$, where $\hat{X}_2$ is statistically independent of $\hat{X}$ and given by $\hat{X}_2 = \hat{X}_r - \hat{X}$.

Since $\hat{X}_2$ is statistically independent of $\hat{X}$, $H(\hat{X}|\hat{X}_2) = H(\hat{X})$, or equivalently, $I(\hat{X}; \hat{X}_2) = 0$, and the enhancement layer provides no information about the core layer. Tuncel and Rose [20] show that this condition is satisfied for additive successive refinement. Thus, even though the enhancement layer is sent in-the-clear, the eavesdropper gains no advantage, and the secrecy of the system is governed by the encryption method used on the core layer.

SELECTIVE ENCRYPTION IN MPEG-4 SCALABLE CODERS

We present a specific example of combining selective encryption with scalable speech coding using the MPEG-4 standard CELP scalable speech coding tool. Only the core layer of the scalable bitstream is encrypted, which is about 50% of the total bitrate.

We investigate the performance of selective encryption for the lowest MPEG-4 CELP rate, 3.85 kbps, because this rate offers the greatest possible efficiency in terms of reduced bandwidth utilization and in terms of intermediate mobile node transmitted power conservation when only the core layer is forwarded. The case of a 3.85 kbps core and two 2 kbps enhancement layers, for a total of 7.85 kbps, has been evaluated and is compared here to the result obtained by Servetti and De Martin [4] for selective encryption of G.729 encoded speech at 8.0 kbps (which is a nonscalable approach).

The input speech data is first partitioned by the MPEG-4 encoder into frames of 320 samples (40 ms) and then each frame is coded with the MPEG-4 CELP algorithm. Selective encryption of the core layer protects 154 bits out of 314 for each frame, which is about 49% of the entire compressed stream (the core layer plus two enhancement layers). Thus, the originating and destination nodes require about half the signal processing for encryption and decryption, respectively, that would be required should the entire bitstream be encrypted. Furthermore, by paring the relayed data stream back to only the core layer, the intermediate nodes drop the enhancement layer to increase the number of voice calls supported in the given bandwidth (or to improve reliability), or for MANETs, to conserve transmitted power (and hence, battery power).

The proposed encryption was applied to flat filtered clean speech taken from the NTT Multi-lingual Speech Database. The data was coded by the MPEG-4 CELP scalable coder at 7.85 kbps, and the effect of encryption of the compressed bitstreams was simulated by replacing the core layer bits with a random sequence of binary digits. Then, mimicking an eavesdropper, an MPEG-4 standard-compliant decoder generated the output material. The MPEG-4 bitstream has a header with information such as sampling frequency, frame length, operating mode, number of layers, etc. This header is not encrypted, and in our tests, we considered this information to be available to the eavesdropper.

Figure 4 shows an example sentence. Comparison of the original signal, shown in Fig. 4(a), to the signal subject to core-only selective encryption, shown in Fig. 4(b), indicates a very high degree of content destruction. Analysis of the partially encrypted signal alone does not even permit discrimination between speech and silence. Informal listening

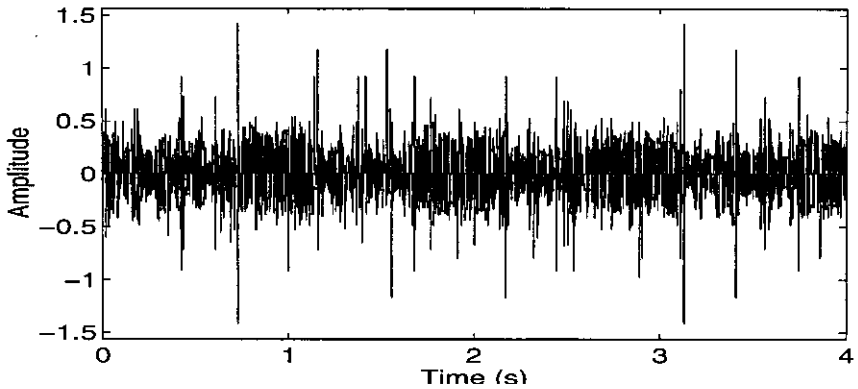


Fig. 5. Decoded excitation using only enhancement layer information for the sentence in Fig. 4(a)

tests confirm that the signal sounds noise-like, with no hints of perceptual content.

Since the common quality-oriented listening tests would not be useful for encrypted speech, subjective tests emphasizing utterance and speech feature identification have been conducted. The first experiment was related to language identification. Listeners were asked to guess if the encrypted sentence they heard was spoken in English or not. The second experiment focused on plain-text identification. Listeners were presented an unencrypted sentence followed in succession, but in a random order, by the previous sentence encrypted and a different encrypted sentence. The listeners were asked to match one of the encrypted sentences with the sentence without encryption. The listeners were free to repeat their listening experience at will. This experiment tests whether an encrypted bitstream retains information about its clear version.

For language identification, listeners were unable to identify the sentences in English, and for plain-text identification, the success rate was statistically equivalent to random guesses. Neither of the partially encrypted sentences were intelligible, and no features of the original speech could be distinguished in the encrypted streams. Therefore, selective encryption of the core layer of the MPEG-4 SNR scalable coder achieves results comparable to those obtained by Servetti and De Martin for G.729 [4].

Rather than use the received bitstream directly in an MPEG-4 standard-compliant decoder, the eavesdropper might alternatively analyze the enhancement layer information that is received in the clear to see if it reveals anything about the core layer or about the voice signal itself. To examine this possibility, we note that the two MPEG-4 enhancement layers contain codebook excitations and codebook gain information only. Knowing the MPEG-4 syntax, the eavesdropper can decode this information exactly, and we plot these decoded excitations in Fig. 5 for the sentence shown in Fig. 4(a). It appears that this waveform reveals very little about the original speech signal and listening tests confirm that it sounds noiselike with no discernible speech content. It is unlikely that the eavesdropper could use this information fruitfully to ascertain properties of the speech signal.

We also note that many other cryptanalytic attacks are possible, depending upon the scenario of interest. In the present work, we have assumed that the eavesdropper is relatively passive and will listen in on voice conversations only if it is reasonably simple to do so. We do not consider here what might be called a motivated attacker who is trying to obtain "high-value" content such as a movie or intellectual property. These situations are reserved for future work.

CONCLUSIONS

We have proposed and investigated an approach to secure, efficient voice communications in networks with multi-hop wireless links by combining SNR scalable speech coding with selective encryption of the core layer. The scalable coding method allows network load to be reduced and routing node power to be conserved by transmitting only the core layer, as network conditions and node resources indicate. It is shown that encryption of the core layer only is sufficient to ensure a high level of protection against eavesdroppers, thus significantly reducing the signal processing power needed for encryption and decryption in comparison to encryption of the full bitstream.

ACKNOWLEDGEMENTS

This research was supported, in part, by the National Science Foundation under Grant No. CCR-0243332 and by the California MICRO Program, Dolby Laboratories, Inc., Lucent Technologies, Inc., Mindspeed Technologies, Inc., and

Qualcomm, Inc.

REFERENCES

[1] C. E. Perkins, *Ad Hoc Networking*, Addison Wesley Professional, 2001.

[2] W. A. Arbaugh, "Wireless security is different," *Computer*, pp. 99–101, Aug. 2003.

[3] H. Dong, *SNR and bandwidth scalable speech coding*, Ph.D. thesis, Southern Methodist University, Dallas, Texas, December 2002.

[4] A. Servetti and J. C. De Martin, "Perception-based partial encryption of compressed speech," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 637–643, 2002.

[5] T. Lookabaugh, I. Vedula, and D. C. Sicker, "Selective encryption and MPEG-2," *ACM Multimedia*, 2003.

[6] S. J. Wee and J. G. Apostolopoulos, "Secure scalable video streaming for wireless networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2001.

[7] H. Wu, C. Hung, M. Gerla, and R. Bagrodia, "Speech support in wireless, multihop networks," in *Third International Symposium on Paprallel Architectures, Algorithms, and Networks*, Dec. 1997, pp. 282–288.

[8] V. N. Muthiah and W. C. Wong, "A speech-optimised multiple access scheme for a mobile ad hoc network," in *First Annual Workshop on Mobile and Ad Hoc Networking and Computing*, Aug. 2000, pp. 127–128.

[9] ITU-T, *5-, 4-, 3- and 2-bit/sample embedded adaptive differential pulse code modulation(ADPCM)*, Dec. 1990.

[10] ISO/IEC JTC1 SC29/WG11, ISO/IEC FDIS 14496-3, *Information Technology-Coding of Audiovisual Objects-Part 3: Audio*, May 1998.

[11] A. M. Alattar and G. I. Al-Regib, "Evaluation of selective encryption techniques for secure transmission of MPEG-compressed bit-streams," *IEEE Int. Symp. on Circuits and Systems*, pp. 340–343, 1999.

[12] H. Cheng and X. Li, "Partial encryption of compressed images and videos," *IEEE Trans. on Signal Processing*, vol. 3, pp. 2439–2451, Aug. 2000.

[13] V. Koshelev, "Multilevel source coding and data-transmission theorem," in *VII All-Union Conf. Theory of Coding and Data Trans.* 1978, pp. 85–92, USSR.

[14] V. Koshelev, "An evaluation of the average distortion for discrete scheme of sequential approximation," in *Probl. Pered. Inform.*, 1981, vol. 17, pp. 20–33.

[15] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. on Inform. Theory*, vol. 37, no. 2, pp. 269–275, March 1991.

[16] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. on Inform. Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.

[17] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, pp. 656–715, 1949.

[18] C. E. Shannon, "Coding theorem for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec., Part 4*, pp. 142–163, 1959.

[19] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[20] E. Tuncel and K. Rose, "Additive successive refinement," *IEEE Transaction on Information Theory*, vol. 49, pp. 1983–1991, Aug. 2003.