PESQ-Based Selection of Efficient Partial Encryption Set for Compressed Speech

Hae-Yong Yang, Kyung-Hoon Lee, Sang-Han Lee, and Sung-Jea Ko

Adopting an encryption function in voice over Wi-Fi service incurs problems such as additional power consumption and degradation of communication quality. To overcome these problems, a partial encryption (PE) algorithm for compressed speech was recently introduced. However, from the security point of view, the partial encryption sets (PESs) of the conventional PE algorithm still have much room for improvement. This paper proposes a new selection method for finding a smaller PES while maintaining the security level of encrypted speech. The proposed PES selection method employs the perceptual evaluation of the speech quality (PESQ) algorithm to objectively measure the distortion of speech. The proposed method is applied to the ITU-T G.729 speech codec, and content protection capability is verified by a range of tests and a reconstruction attack. The experimental results show that encrypting only 20% of the compressed bitstream is sufficient to effectively hide the entire content of speech.

Keywords: Wi-Fi, partial encryption, multimedia security, ITU-T G.729 speech codec, reconstruction attack.

Manuscript received Sept. 8, 2008; revised Apr. 22, 2009; accepted Apr. 30, 2009.

This research was partially supported by Seoul Future Contents Convergence (SFCC) Cluster established by Seoul R&BD Program.

Hae-Yong Yang (phone: +82 42 870 2249, email: formant@korea.ac.kr), Kyung-Hoon Lee (email: khl@ensec.re.kr), and Sang-Han Lee (email: freewill71@ensec.re.kr) are with the Network & Communication Security Division, the Attached Institute of ETRI, Daejeon, Rep. of Korea.

Sung-Jea Ko (email: sjko@korea.ac.kr) is with the Department of Electronic Engineering, Korea University, Seoul, Rep. of Korea.

doi:10.4210/etrij.09.0108.0508

I. Introduction

Wireless fidelity (Wi-Fi) was originally developed for mobile Internet services, but it is now increasingly used for additional services such as voice over Internet protocol (VoIP), multimedia streaming, and so on. In recent years, the voice over Wi-Fi (VoWi-Fi) service using Wi-Fi phones has been regarded by many experts as a "killer application" in both public and private areas [1]. Unsurprisingly, most VoIP providers currently supply the public VoWi-Fi service within hot-spots communicating with Wi-Fi-enabled hand-held phones as shown in Fig. 1.

In terms of security, a Wi-Fi phone requires a much higher security level than a wired device because of the vulnerability of the communication media: the open air space. Therefore, the Wi-Fi standard includes security protocols known as wired equivalent privacy (WEP) and Wi-Fi protected access (WPA). These protocols are highly recommended to Wi-Fi users as a

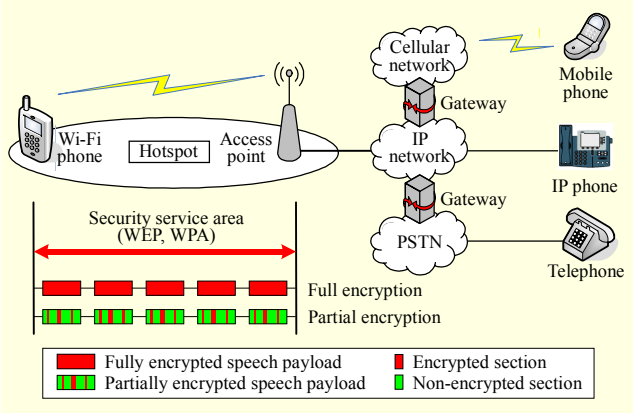

Fig. 1. Public VoWi-Fi network and the encryption methods for the compressed speech.

countermeasure against various types of attacks on the communication channel between a Wi-Fi phone and an access point.

VoWi-Fi service still faces many challenges from the security perspective. First, the encryption service is computationally demanding and inevitably requires additional power consumption [2]. Second, encryption often degrades communication quality because the encryption/decryption procedure causes additional packet loss and delay [3]. Moreover, these problems become more critical in ultra-low-power applications, such as wireless sensors and ad-hoc networks [4].

Various methods to reduce the computational burden of the encryption service have been proposed. One of these methods is the technique of partial encryption (PE) of the compressed bitstream. The PE technique involves encrypting not the entire compressed bitstream but the perceptually sensitive sections of the bitstream, namely, the partial encryption set (PES).

The first attempt to apply the PE method to compressed speech was recently made by Servetti and Martin [5], who suggested two PESs for the ITU-T G.729 speech codec. They demonstrated that, compared to full encryption (FE), the PE which uses a PES of 45% could provide the equivalent content protection capability. Moreover, they showed that the PE which uses a PES of 30% could preclude intelligibility of the restored speech. Similar works were conducted in [6] and [7] for ITU-T G.723.1 and MPEG-4 speech codecs, respectively.

The objective of this paper is twofold: first, to propose a new selection method of finding a much smaller PES, and second, to demonstrate that the proposed PES does not degrade the security level of the encrypted speech compared to FE. To achieve the first objective, we suggest some improvements to the PES selection method in the conventional PE algorithm [5]. Then, we investigate how single-bit encryption and multiple-bit encryption affect restored speech quality. Based on these evaluations and physical meanings of each speech parameter group, the PES selection criteria suitable for speech encryption are suggested. To measure the effects of encryption on speech distortion, the proposed method employs the accurate objective speech quality metric known as the ITU-T P.862 perceptual evaluation of speech quality (PESQ) algorithm. We apply the proposed method to the ITU-T G.729 speech codec. The second objective is satisfied by various objective and subjective tests which consist of time/frequency domain analyses and formal listening tests. Also, the immunity of the proposed PES against the reconstruction attack is investigated.

The rest of the paper is organized as follows. Section II briefly reviews the conventional PE algorithm and shows some inefficient aspects of existing PESs. Section III describes the proposed PES selection method. The immunity against the reconstruction attack and the experimental results are presented in sections IV and V, respectively. Finally, concluding remarks and comments on future work are given in section VI.

Table 1. Bit allocations of the G.729 codec and the existing PESs.

Group	Parameter symbol	G.729	HPS	LPS
LSP	L0	1	0	0
	L1	7	7	5
	L2	5	5	3
	L3	5	0	0
Pitch	P0	1	0	0
	P1	8	7	5
	P2	5	3	3
Gain	GA1	3	3	2
	GB1	4	4	2
	GA2	3	3	2
	GB2	4	4	2
Residual	S1	4	0	0
	C1	13	0	0
	S2	4	0	0
	C2	13	0	0
Total number of bits		80	36	24
Encryption rate			45%	30%

II. Conventional Partial Encryption Algorithm

1. ITU-T G.729 Speech Codec

The G.729 is a narrowband ITU-T speech coding standard based on the conjugate structure algebraic code-excited linear prediction (CS-ACELP) algorithm [8]. The G.729 standard is widely used in packet-switched networks including VoATM and VoIP. The length of a speech frame is 10 ms, and at every speech frame, the G.729 encoder generates a compressed 80-bit bitstream. According to the speech production model, a compressed bitstream includes information about the line spectral pair (LSP) group, the pitch group, the gain group, and the residual group. Each of these groups comprises several parameters as shown in Table 1. More detailed illustrations including physical meanings of each parameter group and the decoding procedure of the codec will be presented in subsections III.4 and IV.3.

2. Partial Encryption of the Compressed Speech

In [5], Servetti and Martin proposed two PESs, namely, a high protection set (HPS) and a low protection set (LPS), for PE of the G.729 codec. Table 1 and Fig. 2 show bit allocations and locations of the HPS and the LPS. The PE system works as follows. A bitstream is made after a frame of speech is compressed. Among bits of the bitstream, only the bits in the

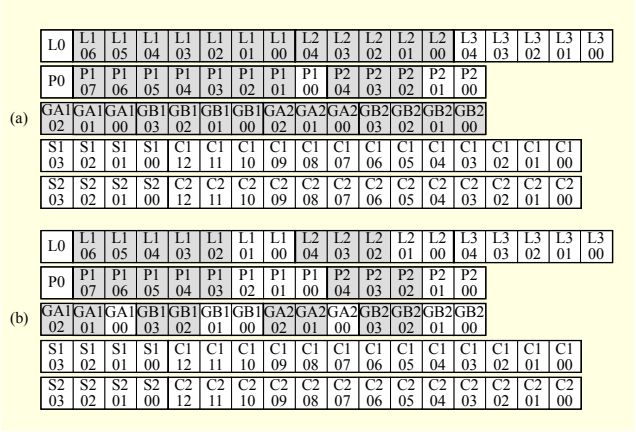

Fig. 2. Bit locations of the existing PESs: (a) HPS and (b) LPS. Most significant bit (MSB) to least significant bit (LSB) from left to right. Bits subject to encryption are shown in gray.

HPS or in the LPS, that is, the bits marked in gray in Fig. 2, are processed by an encryption function. Since the G.729 codec employs a fixed-length coding scheme, the partially encrypted bitstream maintains the standard compliance. As a result, it can be decoded with a standard decoder. However, the content of original speech cannot be acquired without the correct decryption process.

3. Problem 1: Employment of EPS as PES

The HPS originates from research on the unequal error protection (UEP) technique [9]. The UEP technique assigns different protection levels to each bit of a multimedia bitstream. The protection level is determined on the basis of the non-uniform perceptual importance of each bit of the compressed bitstream. In [9], the UEP scheme was demonstrated to enhance overall speech quality in a noisy environment. To select an efficient error protection set (EPS) for UEP, subjective listening tests were performed on every bit in parallel with objective tests.

However, it is worthwhile to note that selection criteria for an efficient PES should differ from those for an efficient EPS, since the purpose of PE is obviously different from that of UEP. The purpose of UEP is to minimize overall speech quality degradation by assigning additional protection capabilities to the bits most sensitive to errors. Conversely, the purpose of PE is to maximize speech distortion, that is, to minimize chances to extract any information from the illegally restored signal. This difference offers the following hints for the improvement of the PE algorithm.

- The EPS tends to occupy a large portion of the bitstream to maintain overall speech quality. However, removing the intelligibility can be more easily achieved by adding noise to a much smaller number of bits in the bitstream.
- The sensitive bits for the EPS were determined by the single-bit error tests under the assumption of a low bit error environment. However, to select the efficient PES, the speech distortion effect caused by encryption on multiple bits should be considered.

Table 2. Quantization methods of the G.729 speech codec parameters.

Group	Parameter symbol	Quantization method
LSP	L0, L1, L2, L3	Vector quantization
Pitch	P0	G.729 specific scheme
	P1	Scalar quantization
	P2	Differential scalar quantization
Gain	GA1, GB1, GA2, GB2	Vector quantization
Residual	S1, C1, S2, C2	G.729 specific scheme

4. Problem 2: Selection of MSBs of parameters for PES

The LPS of the study in [5] was determined by a systematic informal listening test. A tester was given a chance to change only the number of bits in each parameter (each parameter was one of L0 to L3, P0 to P2, GA1, GB1, GA2, GB2, S1, C1, S2, and C2) using a scroll bar, and was unable to select the location of each bit. After the tester's selection, the most significant bits (MSBs) of each parameter were automatically selected as shown in Fig. 2(b). This test procedure was established under an assumption that the MSBs of each parameter are always more important than the least significant bits (LSBs).

However, this heuristic assumption is unreasonable; the parameters in a bitstream do not stand for a physically meaningful value. They are vector/scalar quantized codebook indices or results of codec specific quantization schemes. In practice, the parameters of the G.729 speech codec employ the various quantization schemes shown in Table 2. As a result, the distortion sensitivity of a bit in a parameter does not depend on the associated location of the bit.

III. Proposed PES Selection Method

This section presents the PES selection method which we propose to resolve the two problems previously described. The proposed method is applied to the G.729 speech codec. It employs the PESQ algorithm to evaluate the speech distortion effect of encryption.

1. PESQ Algorithm

As substitutes for the expensive and time consuming mean

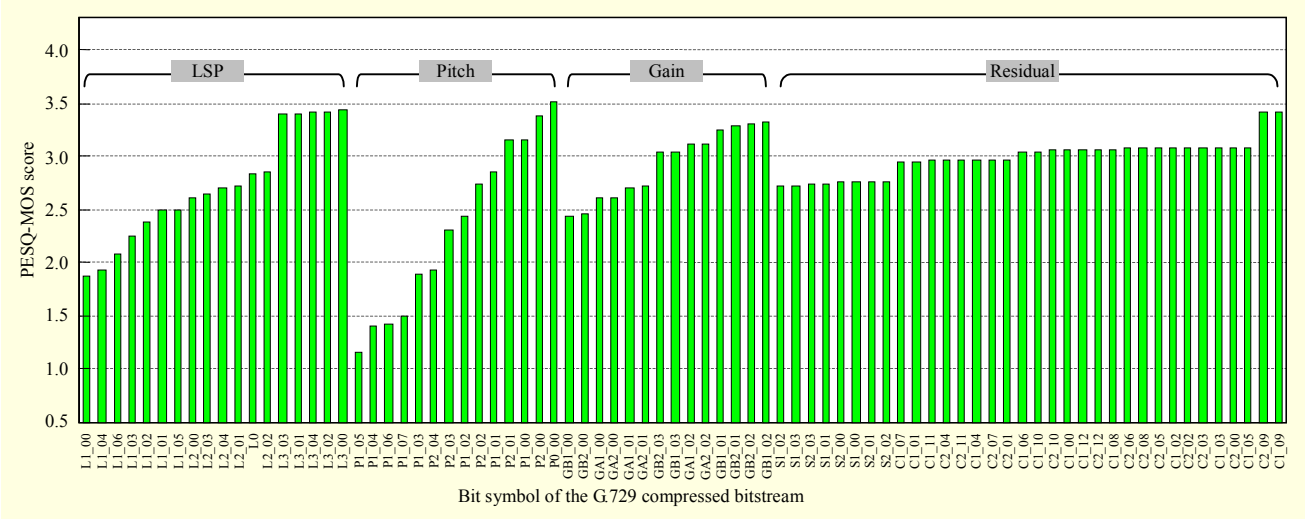

Fig. 3. PESQ-MOS scores according to single-bit encryption. Scores are arranged for each group by sensitivity order. L1_00 indicates the LSB of the L1 parameter.

opinion score (MOS) subjective test [10], several perceptual speech quality evaluation methods have been suggested. Among these alternatives, the ITU-T P.862 PESQ [11] is the most recently developed algorithm, and some benchmark tests of the PESQ have yielded an average correlation of 0.935 with the corresponding MOS values [12]. For this reason, the PESQ is found in many commercial testing and monitoring systems. The PESQ provides raw MOS scores in the range -0.5 to 4.5. The PESQ-MOS scores in our study are also within this range.

2. Distortion Caused by Single-Bit Encryption

The first step for selecting the efficient PES is to accurately evaluate how single-bit encryption on the compressed bitstream affects restored speech quality, or distorts the speech waveform. For this test, we used a speech database with 180 oral statements from 32 Korean male and 28 Korean female speakers. All statements are about six seconds in length. The evaluation steps are as follows: (1) corrupt a corresponding bit in every encoded bitstream of the database, (2) measure a PESQ-MOS score using the original and decoded speech, and (3) obtain an average value using all the scores of the speech database.

Figure 3 shows the results of the single-bit encryption test. In this figure, a lower PESQ-MOS average score means that the corresponding bit is more sensitive to encryption. Note that, in Fig. 3, the MSBs of a parameter are not always more sensitive than the LSBs. For example, L1_00 (the LSB of the L1 parameter) is more sensitive than the other bits (L1_01 to L1_06). This result confirms the former statement that the sensitivity of a bit in a parameter does not depend on the location of the bit. The other important fact observed from this result is that the pitch group is most sensitive to the waveform distortion. The LSP, the gain, and the residual groups are next in order.

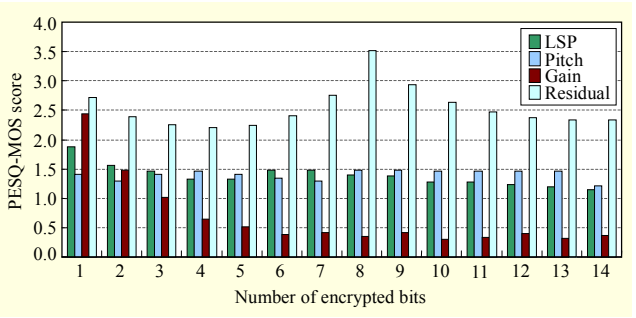

Fig. 4. PESQ-MOS scores according to multiple-bit encryption.

3. Distortion Caused by Multiple-Bit Encryption

Next, we measured the effect of the speech distortion caused by multiple-bit encryption on each parameter group with the same procedure of the single-bit encryption. The bits subject to encryption are selected in the order of the single-bit encryption sensitivity. For example, in case of 3-bit encryption in the gain group, GB1_00, GB2_00, and GA1_00 bits are corrupted.

Figure 4 presents the simulation results for multiple-bit encryption using the same speech database used in the single-bit simulation. Figure 4 shows a somewhat different tendency from the single-bit encryption test results. Increasing the number of encrypted bits in the gain group most rapidly degrades the speech quality. That is, the gain is most important in PE. The figure also indicates that the gain group is followed by the LSP group, the pitch group, and the residual group in order in terms of importance in PE.

Table 3. Bit allocations of the proposed PESs.

PES	Bit allocation (bit)				Encryption rate (%)
	LSP	Pitch	Gain	Total	
Class 1	1	1	2	4	5
Class 2	2	2	4	8	10
Class 3	5	4	7	16	20
Class 4	8	6	10	24	30
Class 5	11	9	12	32	40

4. Physical Meaning of Each Parameter Group

Since each group of the G.729 compressed bitstream individually implies a significant meaning in the speech production model [13], the physical meaning of each group helps us to select the efficient PES.

First, the LSP group models the dynamic motion of a speaker's vocal tract; this group describes distinguishable components of the speech syllables. Second, the pitch group represents the dynamic vibration of a vocal cord; the parameters of this group vary according to the speaker's gender and tone color. Third, the gain group represents the loudness (strength and weakness) of a speech signal; this group is important in PE, since random distribution of loudness has great influence on both the intelligibility and the naturalness of a speech signal. Finally, the residual group delineates the remaining signal, the LSP filter, and the pitch filter; the residual group does not require encryption because it does not reveal any information alone; physically, it displays simple pulse sequences.

5. PES Selection Criteria

Taking the results of the encryption tests and the physical meanings of each group into consideration, we can conclude that the gain group is most important in PE. Along with the gain group, encrypting the LSP group and the pitch group is also important because they could be meaningful to an attacker.

Based on the analyses and discussions described so far, the following four selection criteria for PE of the ITU-T G.729 speech codec are derived:

- The gain group, the LSP group, and the pitch group should be included together.
- A higher number of bits should be allocated to the more important groups (the gain group, the LSP group, and the pitch group in order).
- For each group, the most sensitive bits to single-bit encryption should be included (see Fig. 3).
- For small PESs, one should include as many bits of the gain group as possible.

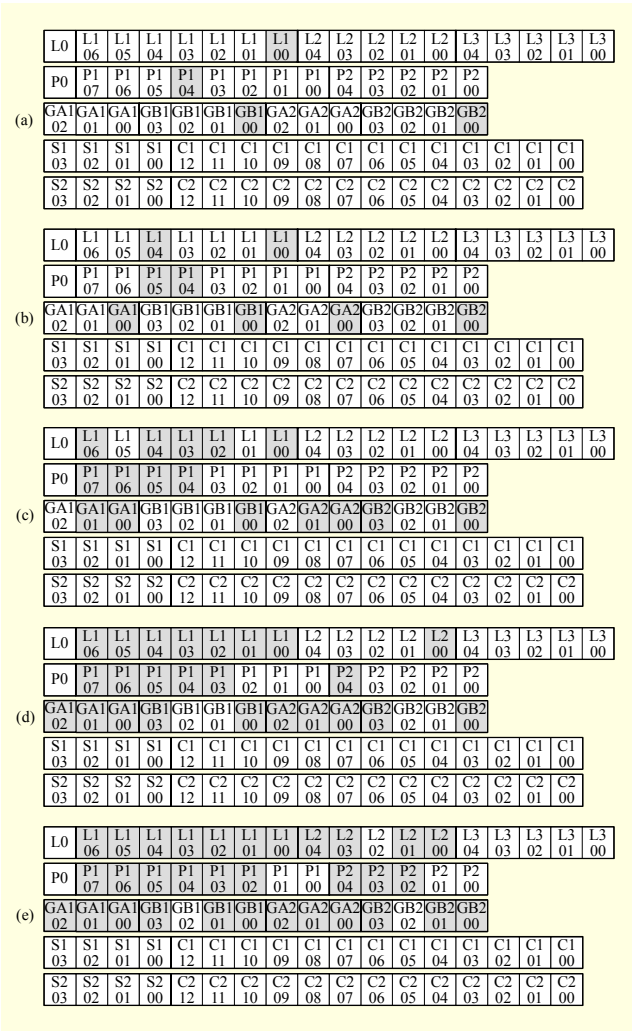

Fig. 5. Bit locations of the proposed PESs: (a) class 1, (b) class 2, (c) class 3, (d) class 4, and (e) class 5.

6. Selected PESs Based on the Proposed Criteria

Finally, according to the proposed selection criteria, five PESs for the G.729 speech codec are selected. These five PESs are named as classes 1 to 5, and they are 4-bit (5%), 8-bit (10%), 16-bit (20%), 24-bit (30%), and 32-bit (40%), respectively. Table 3 and Fig. 5 show the bit allocations and bit locations of the proposed PESs.

IV. Immunity against Reconstruction Attack

1. Reconstruction Attack on PE of Compressed Multimedia

It is commonly assumed that the strength of an encryption system is measured by the difficulty of finding the key under

various cryptanalysis attacks. However, PE of multimedia data may give an opportunity to restore the meaning of encrypted information instead of finding the key. This type of attack, namely, the reconstruction attack, is caused by the structure of the PE approach in which many of the compressed bits are left unencrypted. Practically, the reconstruction attacks succeed for partially encrypted image (JPEG) and video (MPEG) data [14], [15]. In [14], Wu and Kuo showed that the PE algorithms which encrypt some discrete cosine transform (DCT) coefficients or some parts of every DCT coefficient are not secure. They could recover a useful image by replacing the encrypted parts with typical data.

The main reason why the simple ciphertext-only attack could be successful is that not all physically meaningful values (that is, all DCT coefficients) are sufficiently distorted. An undistorted or less distorted physically meaningful value may give significant hints to an attacker. Therefore, to evaluate the confidentiality of the proposed PESs against existing reconstruction attacks, it is worthwhile to investigate whether physically meaningful values leak intelligible information.

Even though information leakage analysis is a good metric to evaluate immunity against the existing type of attacks [14], [15], it is not a sufficient condition to completely protect against an unknown reconstruction attack. This unknown approach, which may extract meaningful information, is left for further study.

2. Strengths of G.729 against Reconstruction Attack

The compression scheme of speech is essentially different from that of image or video. According to the decoding procedure whose details will be described in the following subsection, the following properties of the G.729 codec make it possible to distort all physically meaningful values by enciphering fewer parameters (for newly appearing notations, refer to Fig. 6): (a) multi-stage vector quantization (VQ): slight change of an index of a VQ codebook causes severe distortion of the output value, (b) inter-value dependency: T_2 is decoded relative to T_1, and (c) inter-frame dependency: g_c and q_i are decoded relative to the corresponding values of the previous frame. In other words, when an attacker replaces the PES with typical data, these properties make it much harder for the attacker to get useful information.

3. Physically Meaningful Values of G.729 Codec and Their Decoding Procedures

Figure 6 shows the block diagram of the G.729 decoder, where the residual signal $c(n)$ and the pitch signal $v(n)$ are generated, and these signals are scaled to the appropriate level by g_c and g_p, respectively. The excitation signal $u(n)$ is filtered by a linear prediction (LP) synthesis filter to yield the synthetic

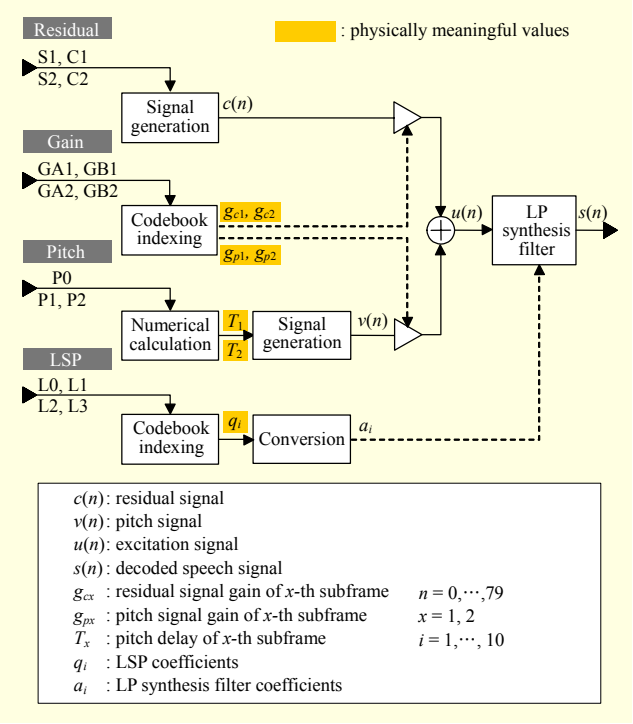

Fig. 6. Simplified block diagram and decoding procedure of the G.729.

speech $s(n)$. From the standpoint of an attacker, the physically meaningful values of the model, which correspond to the DCT coefficients of image, are the following: (a) the pitch delays T_1 and T_2, (b) the gains g_{c1}, g_{c2}, g_{p1}, and g_{p2}, and (c) the LSP coefficients q_i, $i = 1,\cdots,10$. The residual signal $c(n)$ is excluded because it only consists of pulse sequences generated from the pulse positions C1 and C2 and signs S1 and S2.

The values are decoded using the following compressed parameters [8]:

- The pitch delays T_1 and T_2: P0 is the parity bit. The pitch delay of the first frame T_1 is decoded from a numerical calculation from P1. Also, T_2 is numerically decoded relative to T_1 using P2. The dependency shows that an error at P1 results in distortion at both T_1 and T_2.
- The gains g_{c1}, g_{c2}, g_{p1}, and g_{p2}: GA1, GB1, GA2, and GB2 are indices of the VQ codebooks which comprise with the two-stage conjugate structured codebooks $\mathcal{GA}$ and $\mathcal{GB}$. The gains are decoded by

$$g_{px} = \mathcal{GA}_1(GAx) + \mathcal{GB}_1(GBx), \quad x = 1, 2, \qquad (1)$$

$$g_{cx} = \hat{g}_{cx}\,(\mathcal{GA}_2(GAx) + \mathcal{GB}_2(GBx)), \quad x = 1, 2, \qquad (2)$$

  where $\mathcal{GA}_y(z)$ denotes the y-th element of a given index z in the codebook $\mathcal{GA}$, and $\hat{g}_{cx}$ is the predicted gain based on g_{cx} of the previous speech frame.
- The LSP coefficients q_i: the current quantizer output e_i is decoded by indexing the two-stage VQ codebook. The first

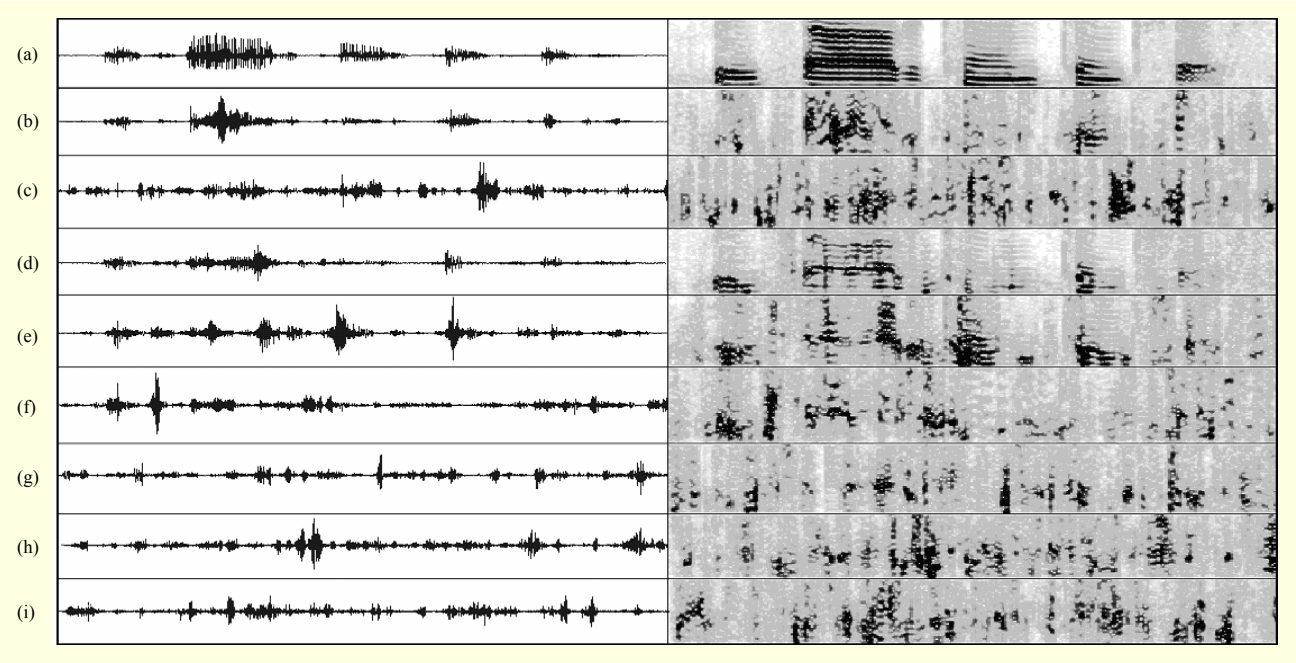

Fig. 7. Restored speech signals (left) and corresponding spectrograms (right) using the following encryption methods: (a) NE, PE using (b) LPS, (c) HPS, (d) class 1, (e) class 2, (f) class 3, (g) class 4, (h) class 5, and (i) FE.

stage is the 10-dimensional codebook $\mathcal{L}1$, and the second stage is the split VQ codebook using the two 5-dimensional codebooks, $\mathcal{L}2$ and $\mathcal{L}3$. With e_i, each q_i coefficient is decoded by

$$e_i = \begin{cases} \mathcal{L}1_i(\mathrm{L1}) + \mathcal{L}2_i(\mathrm{L2}), & i = 1,\cdots,5, \\ \mathcal{L}1_i(\mathrm{L1}) + \mathcal{L}3_{i-5}(\mathrm{L3}), & i = 6,\cdots,10, \end{cases} \quad (3)$$

$$q_i = f(e_i, \hat{e}_i, \mathrm{L0}), \qquad i = 1,\cdots,10, \quad (4)$$

where $f(x, y, z)$ denotes the numerical operation with x, y, and z, and $\hat{e}_i$ is the e_i version of the previous frame. The equations show that the L1 parameter contributes much more than the others (L2 and L3) to obtaining the LSP coefficients q_i.

V. Experimental Results

1. Signal Observation

The first evaluation method is to analyze the restored speech in both the time and frequency domains. Figure 7 shows the restored speech signals and the corresponding spectrograms. Restored speech is generated by decoding the partially (or fully) encrypted G.729 bitstream without decrypting the encrypted sections. Figures 7(a) and (i) present the restored speech signals and the corresponding spectrograms of the non-encrypted and the fully encrypted bitstreams, respectively. Figures 7(b) and (c) present the results from existing PESs, and Figs. 7(d) to (h) present the results from the proposed PESs.

Compared to the no-encryption (NE) and FE results, the signals and spectrograms in Figs. 7(c), (f), (g), and (h) show that PE using the HPS and the PESs of classes 3, 4, and 5 has characteristics similar to those of FE in terms of energy and spectrum distribution. Speech specific characteristics, such as quasi-periodicity and the formant frequency, disappear in these figures. On the other hand, the signals and the spectrograms of the LPS and the PESs of classes 1 and 2 show some resemblance to those of NE in terms of the spectrum and energy contour.

2. Objective Distortion Measurements

The second evaluation method is based on various objective metrics which quantitatively estimate the distortion effects of the encrypted speech signals in the time and frequency domain. In this measurement, the following metrics are employed: the PESQ score for a perceptual speech quality measure [11], the spectral distortion for a frequency domain distortion measure, the segmental energy difference (ED), and the segmental signal-to-noise ratio (SNR) for time domain distortion measures. Another speech database that is different from that of the single-bit encryption simulation is utilized here. The database also comprises about 180 oral statements from 60 speakers. The simulation is executed iteratively 100 times to reduce randomness in the results. Table 4 shows the simulation results. The results consistently indicate that the PESs of classes 3, 4, and 5 and the HPS have distortion effects that are

Table 4. Objective speech distortion measurement results for the existing and proposed PESs.

	Conventional PES			Proposed PES				
PES (encryption rate)	LPS (30%)	HPS (45%)	FE (100%)	Class 1 (5%)	Class 2 (10%)	Class 3 (20%)	Class 4 (30%)	Class 5 (40%)
(a) PESQ-MOS score (-0.5 to 4.5)	0.88	0.76	0.73	1.65	0.96	0.78	0.77	0.77
(b) Spectral distortion (dB)	12.56	15.72	16.07	9.82	12.38	14.84	15.45	15.75
(c) Segmental ED (dB)	9.88	13.04	12.82	7.25	9.91	12.60	13.02	13.07
(d) Segmental SNR (-dB)	1.81	3.63	3.93	1.13	2.70	3.03	3.55	3.60

Table 5. Formal listening test results for the existing and proposed PESs.

	Conventional PES						Proposed PES									
PES (encryption rate)	LPS (30%)		HPS (45%)		FE (100%)		Class 1 (5%)		Class 2 (10%)		Class 3 (20%)		Class 4 (30%)		Class 5 (40%)	
(a) Intelligibility (1 to 5)	1.06		1.00		1.00		2.58		1.04		1.00		1.00		1.00	
	No vote	False	No vote	False	No vote	False	No vote	False	No vote	False	No vote	False	No vote	False	No vote	False
(b) Plain-text identification	50.0	33.3	79.2	20.0	85.0	15.0	1.7	1.7	13.3	16.7	71.7	21.7	73.3	15.0	88.3	7.5
(c) Gender identification	67.5	15.0	86.7	8.3	85.0	10.0	1.7	21.7	30.0	2.5	70.8	12.5	83.3	7.5	84.2	10.0
(d) Speech/non-speech discrimination	28.3	11.7	46.7	22.5	37.5	37.5	10.0	5.0	16.7	13.3	40.0	26.7	38.3	32.5	42.5	30.0

comparable to those of the FE set. Note that the class 2 PES shows more improvement in content protection capability than the LPS.

3. Formal Listening Tests

Next, we performed formal subjective listening tests [5] with 20 listeners and six test sentences per test and PES. Table 5 shows the results of the listening tests.

According to the intelligibility test result (Table 5(a)), some listeners understood one or two fragments of the partially encrypted signal with the PESs of classes 1 and 2 and the LPS. In contrast, the PESs of classes 3, 4, and 5 and the HPS blocked the attack completely. All listeners gave the score of 1 (the lowest rating), which means that the listeners did not extract any information from the restored speech sample.

Next, the plain-text identification test was performed. After the test participants listened to an example sentence, they were asked to select one candidate from four PE-restored speech samples. They could select "no vote" if they were not sure enough to select one of the candidates. The plain-text identification test result (Table 5(b)) shows that PE using the PESs of classes 3, 4, and 5 and the HPS have content protection capability comparable to that of FE. The gender identification and speech/non-speech discrimination test results, as shown in Table 5(c) and 5(d), have a thread of connection to the plain-text identification test result.

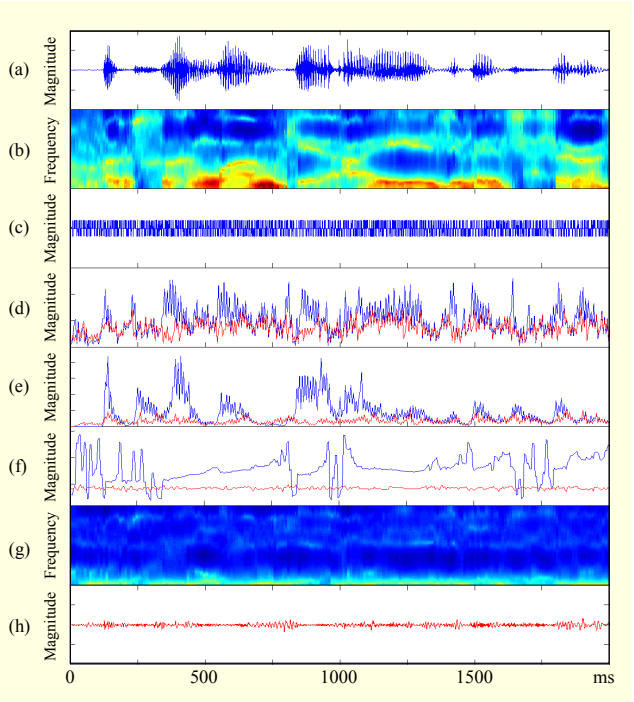

Fig. 8. Comparison of the reconstructed signals and values when all compressed bits of the class 3 PES are set to the replacement values: (a) $s(n)$, (b) frequency response contour of q_i, (c) $c(n)$, (d) g_p (blue line) and g'_p (red line), (e) g_c (blue line) and g'_c (red line), (f) T (blue line) and T' (red line), (g) frequency response contour of q'_i, and (h) $s'(n)$, where x' indicates partially encrypted and replaced version of x.

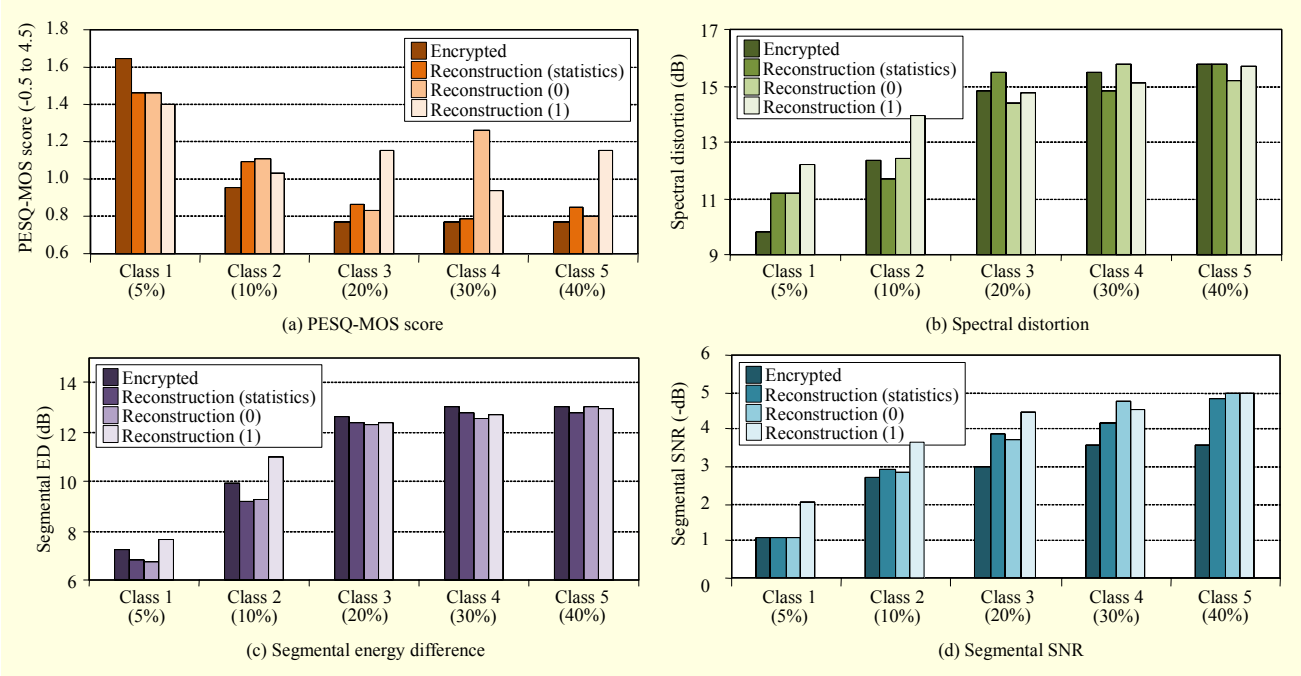

Fig. 9. Objective speech distortion measurement results for the attacked signals compared to the simply encrypted case.

4. A Reconstruction Attack

To show that the proposed class 3 PES is robust against the reconstruction attack, we attempt to practically attack with a reasonable scenario: first, collect the statistics about the probabilities that the corresponding bit value is one using a speech database; second, choose the appropriate replacement values based on the statistics; third, encrypt a speech using the proposed class 3 PES; fourth, replace the encrypted parts of the bitstream as the fixed chosen value; and fifth, decode the replaced bitstream and extract the physically meaningful values with the format compliant G.729 decoder.

The experimental results are shown in Fig. 8. Compared to the frequency response of the unencrypted LSP shown in Fig. 8(b), that of the partially encrypted and replaced LSP shown in Fig. 8(g) is sufficiently distorted. Also, Figs. 8(d), (e), and (f) show that all the other physically meaningful values are distorted enough to make it impossible to predict the original values.

To finally confirm the robustness against the reconstruction attack, we performed the same objective tests as those presented in subsection V.2 for three kinds of attacked signals: the statistics-based, the replacement of the PESs by zero, and the replacement of the PESs by one. Also, the same formal listening tests were performed with 10 listeners and four sentences per test. For fair evaluation, the overall energy level of an attacked signal was changed according to the unencrypted signal except speech/non-speech discrimination test signals. This additional processing was needed because some of the restored signals had almost zero or saturated values.

Figures 9 and 10 show the test results compared to the simply encrypted case shown in Tables 4 and 5. The results indicate that the three kinds of reconstruction attacks are not effective in the PESs of classes 3, 4, and 5. In the PESQ-MOS scores (Fig. 9(a)), the high peaks of the PESs of classes 3, 4, and 5 are caused by prediction errors of the PESQ algorithm due to the white noise-like characteristics of some of the attacked signals.

5. Computational Complexity

We evaluated the computational complexity reduction effect with the PESs in a practical Wi-Fi phone environment. Most of the recently released Wi-Fi phones have the following security features: (a) operation based on the WPA protocol, (b) adoption of the RC4 stream cipher, and (c) use of a 128-bit key and a 48-bit initial vector.

The implementation target device is the TMS320C5502 digital signal processor [16], which is a commercially popular device in low-power areas. Under the assumption that the PE algorithms are implemented on this device using the C language, the execution cycle counts of the PE algorithms with the different PESs are compared in Table 6. To quantify the computational complexity in an actual condition, the cycle counts were measured during a three-minute period, which is equivalent to the average call duration. Implementation results show that, compared to the FE algorithm, the PE algorithm

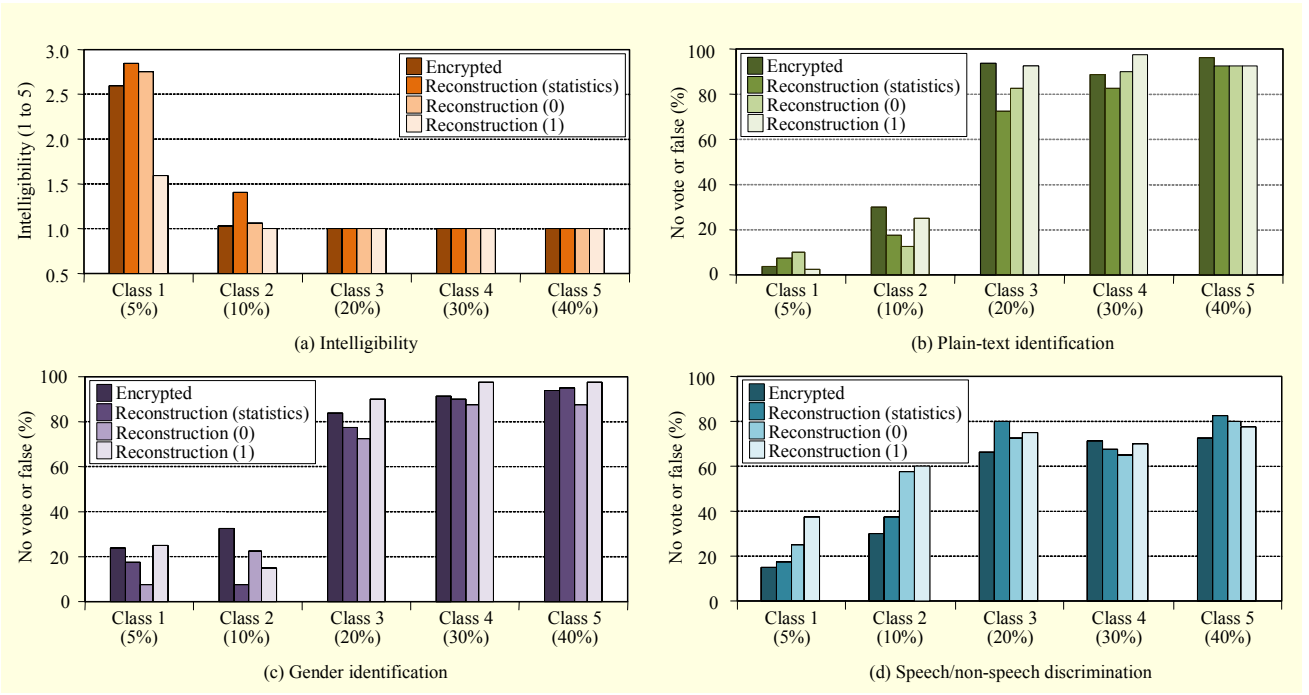

Fig. 10. Formal listening test results for the attacked signals compared to the simply encrypted case.

Table 6. Comparison of computational load of the PESs.

PES	Encryption rate (%)	Cycle counts per call (3 min.)	Cycle count ratio (%)
Class 1	5	3,528,000	26.20
Class 2	10	4,284,000	31.82
Class 3	20	5,292,000	39.30
Class 4, LPS	30	6,480,000	48.13
Class 5	40	7,704,000	57.22
HPS	45	8,388,000	62.30
FE	100	13,464,000	100.00

with the class 3 PES only requires nearly 40% cycle counts.

VI. Conclusion and Future Work

This paper proposed a novel selection method to determine an efficient PES and has shown that the proposed PES is practically secure against the various test scenarios. The experimental results showed that, with the proposed method, encrypting only 20% of the compressed speech can effectively provide information security. The proposed PESs can provide a readily applicable solution to the various problems associated with secure voice over Wi-Fi service.

Future work needs to address the issue of an unknown intelligent reconstruction attack using a more sophisticated method. In addition, we will focus on designing a new speech coding paradigm suitable for ultra-low-power applications, which have good properties for both coding and encryption efficiency.

References

[1] S. Taso, "Research Challenges and Perspectives of Voice over Wireless LAN," *Proc. IEEE EITC*, Aug. 2005.

[2] J. Goodman and A.P. Chandrakasan, "Low Power Scalable Encryption for Wireless Systems," *Wireless Networks*, no. 4, Jan. 1998, pp. 55-70.

[3] H. Xiao and P. Zarrella, "Quality Effects of Wireless VoIP Using Security Solutions," *Proc. IEEE MILCOM*, vol. 3, Oct. 2004, pp. 1352-1357.

[4] R. Chandramouli, S. Bapatla, and K.P. Subbalakshmi, "Battery Power-Aware Encryption," *ACM Trans. Information and System Security*, vol. 9, no. 2, May 2006, pp. 162-180.

[5] A. Servetti and J.C. De Martin, "Perception-Based Partial Encryption of Compressed Speech," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 8, Nov. 2002, pp. 637-643.

[6] C.-P. Wu and C.-C. J. Kuo, "Fast Encryption Methods for Audio-Visual Data Confidentiality," *Proc. SPIE Int. Symp. Information Technologies*, vol. 4209, Nov. 2000, pp. 284-295.

[7] J.D. Gibson et al., "Selective Encryption and Scalable Speech Coding for Voice Communications over Multi-hop Wireless Links," *Proc. MILCOM*, vol. 2, Oct. 2004, pp. 792-798.

[8] ITU-T Recommendation G.729, *Coding of Speech at 8 kbit/s*

Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP), ITU-T, Mar. 1996.

[9] K. Swaminathan and A.R. Hammons, "Selective Error Protection of ITU-T G.729 Codec for Digital Cellular Channels," *Proc. IEEE ICASSP*, vol. 1, May 1996, pp. 577-580.

[10] ITU-T Recommendation P.800, *Methods for Objective and Subjective Assessment of Quality*, ITU-T, Aug. 1996.

[11] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T, Jan. 2002.

[12] C. Hoene, H. Karl, and A. Wolisz, "A Perceptual Quality Model for Adaptive VoIP Applications," *Proc. SPECTS*, July 2004.

[13] W.C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, John Wiley & Sons, 2003, ch. 11, pp. 299-301.

[14] C.-P. Wu and C.-C. J. Kuo, "Design of Integrated Multimedia Compression and Encryption Systems," *IEEE Trans. Multimedia*, vol. 7, no. 5. Oct. 2005, pp. 828-839.

[15] A. Said, "Measuring the Strength of Partial Encryption Schemes," *Proc. IEEE ICIP*, vol. 2, Sept. 2005, pp. 1126-1129.

[16] Texas Instrument, *Data Manual: TMS320VC5502 Fixed-Point Digital Signal Processor*, Aug. 2006.

Hae-Yong Yang received the BS and MS degrees in electrical engineering from Pusan National University, Pusan, Korea, in 1996 and 1998, respectively. He is currently working toward the PhD in electrical engineering from Korea University. He has been working at the Attached Institute of Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, where he is currently a senior member of engineering staff. Before joining ETRI, he was with Hyundai Electronics, Korea. His research interests are A/V signal processing, multimedia security, and VoIP QoS.

Kyung-Hoon Lee received the BS degree in 1992, MS degree in 1994, and the PhD in 1998, all in electrical engineering from Korea University. Between 1998 and 2000, he was a senior member of engineering staff with Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. Between 2000 and 2001, he was a manager of the engineering department of Tellion, Daejeon, Korea. In 2001, he joined the Attached Institute of ETRI, Daejeon, Korea, where he is currently a principle member of engineering staff. His research interests are image/video signal processing, non-linear digital signal processing, and security.

Sang-Han Lee received the BS degree in 1995 and the MS degree in 1997, both in electrical engineering from Kyungpook National University, Daegu, Korea. Between 1997 and 1999, he was a researcher at the Agency for Defense Development (ADD), Daejeon, Korea. In 2000, he joined the Attached Institute of Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, where he is currently a senior member of engineering staff. His research interests are implementation of cryptographic devices and security SoCs.

Sung-Jea Ko received the BS degree in electronic engineering from Korea University in 1980, and the MS degree in 1986 and the PhD in 1988, both in electrical and computer engineering from State University of New York at Buffalo. In 1992, he joined the Department of Electronic Engineering at Korea University where he is currently a professor. From 1988 to 1992, he was an assistant professor with the Department of Electrical and Computer Engineering at the University of Michigan-Dearborn. He has published over 100 journal articles. He also holds over 30 patents on video signal processing and multimedia communications. He is currently a senior member in the IEEE, a fellow in the IET, and a Korean representative of IEEE Consumer Electronics Society. He has been the Special Sessions Chair for the IEEE Asia Pacific Conference on Circuits and Systems (1996). He served as an associate editor for *Journal of the Institute of Electronics Engineers of Korea* (IEEK) (1996), *Journal of Broadcast Engineering* (1996 to 1999), and *Journal of the Korean Institute of Communication Sciences* (KICS) (1997 to 2000). He was an editor of *Journal of Communications and Networks* (JCN) (1998 - 2000). He is the 1999 recipient of the LG Research Award given to the Outstanding Information and Communication Researcher. He received the Hae-Dong Best Paper Award from the IEEK (1997), the Best Paper Award from the IEEE Asia Pacific Conference on Circuits and Systems (1996), and the Research Excellence Award from Korea University (2004).