PERCEPTION-BASED SELECTIVE ENCRYPTION OF G.729 SPEECH

Antonio Servetti[1], *Juan Carlos De Martin*[2]

[1] Dipartimento di Automatica e Informatica/[2]IRITI-CNR
Politecnico di Torino
Corso Duca degli Abruzzi, 24 — I-10129 Torino, Italy
E-mail: [servetti|demartin]@polito.it

ABSTRACT

Mobile multimedia applications, the focus of many forthcoming wireless services, increasingly demand low-power techniques implementing content protection and customer privacy. In this paper a low-complexity, perception-based partial encryption scheme for telephone-bandwidth speech is presented. Speech compressed by a widely-used speech coding algorithm, ITU-T G.729 CS-ACELP at 8 kb/s, is partitioned in two classes, one, the most perceptually relevant, to be encrypted, the other, to be left unprotected. Encryption of about 45% of the bitstream achieves content protection equivalent to full encryption of the bitstream, as verified by both objective measures and formal listening tests. Low-power, portable devices can, therefore, implement very high levels of speech-content protection at a fraction of the computational load of current techniques, freeing resources for other tasks and enabling longer battery life.

1. INTRODUCTION

Content protection and customer privacy are becoming increasingly important as multimedia applications become pervasive. Communications can be intercepted, especially over wireless links. Encryption can effectively prevent unauthorized access and its use is widely advocated. Unfortunately, encryption and decryption are computationally demanding, a severe problem in mobile, portable devices, where power consumption needs to be minimized. The need for encryption in wireless systems has led to intense activity aimed at reducing the complexity of encryption algorithms [1]. Partial encryption is an effective way to introduce encryption into power-constrained, real-time multimedia applications. Instead of encrypting multimedia signals in their entirety, only a perceptually relevant fraction of the bitstream is subject to encryption, while the remaining part is transmitted unencrypted. The bits subject to encryption are selected to cause the desired degree of degradation after decoding.

Encryption of only a fraction of the bitstream could significantly improve the battery life of wireless devices, especially in the case of short-range wireless systems, e.g., Bluetooth, where power consumed by encryption can represent a considerable portion of the overall energy budget. Partial encryption (sometimes referred to also as *selective encryption*) techniques have already been proposed for compressed image and video data. Efficient encryption of MPEG compressed video was proposed, for instance, in [2]. More recently, the approach was extended to image compression [3]. We propose to extend the selective encryption approach to audio and speech signals as well.

Specifically, we present a partial encryption scheme for speech compressed by the ITU-T G.729 8 kb/s speech coding standard. The proposed scheme delivers extremely effective content protection, and can be straightforwardly adapted to several other international speech coding standards based on the ACELP algorithm, e.g., ETSI GSM-AMR or TIA IS-641.

The paper is organized as follows. Section 2 describes partial encryption of G.729 compressed speech and proposes a methodology to evaluate its performance. Section 3 presents and discusses the experimental results, which consist of both objective distortion measures and the output of formal listening tests.

2. PARTIAL ENCRYPTION OF COMPRESSED SPEECH

The bits of a compressed speech bitstream are not perceptually equally important. Bit errors can have vastly different perceptual impact, depending on specifically which bit is corrupted. Non-uniform bit sensitivity has been exploited to create channel coding schemes that deliver different levels of protection to different classes of bits. Such Unequal Error Protection approach achieves, for a given capacity, significantly better perceptual performance than uniform protection of all bits.

Perception–based partial encryption, too, depends on perceptual classification of the bitstream. The objective, however, is not to preserve perceptual quality above given levels after transmission and decoding, but, on the contrary,

to cause the desired degree of content degradation. More formally, a pre-requisite of partial encryption is to identify *the smallest subset of the compressed bitstream that, if made unavailable due to encryption, causes the desired amount of degradation at the decoder*. A smaller subset would not offer enough content protection, while a larger one would wastefully increase system complexity, with adverse effects on the battery life of portable devices.

For speech, degradation may mean a decrease of *naturalness*, a decrease of *intelligibility*, or both. While in speech transmission the focus is mostly on the former, for speech protection the attention is essentially on the latter.

2.1. Partial encryption of G.729 speech

We chose to investigate partial encryption solutions for the ITU-T G.729 CS-ACELP speech coding standard [4]. G.729 is widely used, most conspicuously in Voice over IP applications, and it provides toll quality at 8 kb/s with low algorithmic delay and moderate complexity. Our goal was to find a partial encryption scheme that performs as well as full encryption, that is, no perceptual information of any sort should be noticeable by human ears.

To identify the desired subset of bits to be encrypted, we considered UEP schemes proposed for G.729 (see, e.g., [5] and [6]). The minimum set of class 1 bits reported in [6] was taken as the starting point of our investigation. Different schemes were tested to classify G.729 parameters and bits in order of perceptual importance. This analysis was carried out by systematically corrupting a given bit and then measuring the corresponding drop in performance over a speech database. Both objective distortion measures and informal listening tests were employed. Knowledge of the perceptual significance of each parameter guided the process. In the end, the proposed partial encryption scheme coincides with the minimum set of class 1 bits in [6], i.e., the 36 bits (out of 80) absolutely requiring error protection, shown in Tab. 1. The parameters subject to encryption are such that three essential features of speech, spectral envelope, pitch contour and gain contour, are severely, if not completely, degraded. The stochastic codebook contributions are left integrally unprotected.

It is straightforward to adapt the proposed partial encryption scheme to other ACELP speech coders, such as the ETSI GSM EFR and AMR standards, or the TIA IS-641 EFR.

2.2. Performance Evaluation

Absolute performance of the proposed partial encryption scheme was evaluated: 1) by signal inspection, in both the time and the frequency domains; 2) by means of objective distortion measures; 3) by means of formal listening tests.

Table 1. Proposed partial encryption scheme for G.729. In the rightmost column the number of bits (starting from the MSB) subject to encryption is shown for each protected parameter.

Symbol	Description	Bits	Encry.
L1	LPC 1st codebook idx	7	7
L2	LPC 2nd codebook low portion idx	5	5
P1	Pitch period 1st subframe	8	7
GA1	Codebook gain (stage 1) 1st subframe	3	3
GB1	Codebook gain (stage 2) 1st subframe	4	4
P2	Pitch period 2nd subframe	5	3
GA2	Codebook gain (stage 1) 2nd subframe	3	3
GB2	Codebook gain (stage 2) 2nd subframe	4	4

The first approach consisted in analyzing specific features of speech signals subjected to partial encryption. Both spectral and temporal features, in fact, carry well-studied perceptual significance: the spectral envelope, and in particular formant frequencies and bandwidths, is closely related to phoneme identification. Quasi-periodicity of the time-domain signal, or, equivalently, harmonic structure of the signal spectrum, are connected to voicing and, through the absolute value of the fundamental frequency, to gender identification.

The second evaluation approach was based on objective quality measures. In particular, segmental signal-to-noise ratio (segSNR) and Spectral Distortion (SD) were employed to evaluate the effectiveness of the proposed partial encryption scheme.

The third approach consisted of formal listening tests concerning the absolute performance of the proposed scheme. Since standard quality-oriented listening tests, such as Mean Opinion Scores (MOS) or A-B comparison tests, would not provide direct information about intelligibility or plain-text identification, new subjective tests needed to be designed. The attention was focused on the following tasks: 1) intelligibility; 2) plain-text identification; 3) speech/non speech discrimination. A fundamental objective of the proposed encryption scheme is to prevent comprehension. Listeners were, therefore, asked to listen to a sentence, in their own mother language, but unknown to them, and then give an intelligibility score based on a 5-point scale: “5–Full,” “4–Good,” “3–Fair,” “2–Poor,” “1–Nothing.” If the score was greater than one, that is, if the listener deemed to have understood one or more words, he/she was asked to write them down and then to listen to the clear-text sentence, to verify his/her judgment. In case of match between perceived and original words, the score was confirmed, otherwise it was downgraded to “1–Nothing.” Matches were rather loosely evaluated: acoustical similarity between estimates and original words was enough to confirm the first score. Listeners were free to listen to the sen-

tences as many time as desired, thus approaching the condition of an offline analyst rather than a real-time eavesdropper.

The second experiment modeled the case of speech communications based on a limited vocabulary, e.g., commands, or prompts of an automated response system. In such cases, the eavesdropper is assumed to know the vocabulary and the objective of the protection scheme is to obtain rates of successful identification equivalent to random choice. In our case, listeners heard four clear-text sentences. Successively, they heard one of the four sentences, randomly chosen, encrypted and they were asked to match it with one of the four original sentences, or to vote "don't know." Repeated listenings, in any order, were allowed.

Finally, the third experiment evaluated the ability to discriminate between speech and non-speech encrypted signals. This ability could be of interest to an eavesdropper scanning for speech vs. silence in a database of speech recordings. In this case, listeners were asked to classify the signals as "speech," "non-speech" or "don't know."

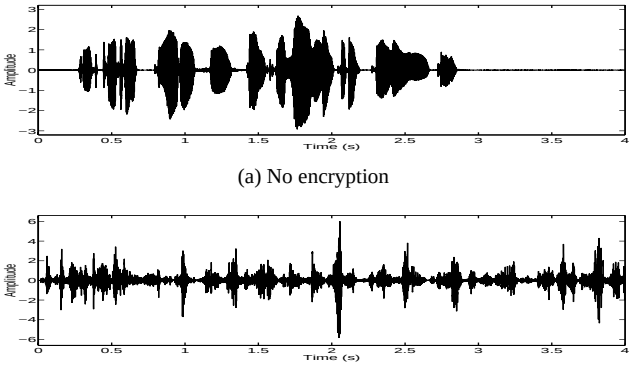
3. RESULTS AND DISCUSSION

The encryption schema was applied to flat filtered clean speech taken from the NTT Multi-lingual Speech Database. The material was encoded using the ITU-T G.729 floating-point reference software. The resulting bitstreams were partitioned according to the encryption scheme shown in Tab. 1, with the encrypted bits replaced with a random sequence of binary digits. Finally, a standard-compliant G.729 decoder generated the output material.

3.1. Signal Inspection

Fig. 1 shows an example sentence. Comparison of the original signal, shown in Fig. 1(a), to the signal subject to partial encryption, shown in Fig. 1(b), indicates a very high degree of content destruction. Analysis of the signal subject to partial encryption does not even permit to discriminate between speech and silence. Informal listening confirms that the signal sounds noise-like, with no hints of perceptual content.

Since most of the analysis performed by the hearing system is in the frequency domain, the spectrum of the penultimate vowel of the word fragment, /tiere/, has been analyzed. Fig. 2(a) shows the magnitude spectrum of the original vowel. A clear harmonic structure with fundamental frequency of approximately 250 Hz characterizes most of the spectrum. A spectral envelope with at least two formants is distinctly discernible. The partial encryption scheme completely eliminate the harmonic structure, as shown in Fig. 2(b), and leaves almost no detectable spectral envelope, confirming the total absence of perceptual content suggested by informal listening.



(a) No encryption

(b) Partial encryption

Fig. 1. Partial encryption of a sentence (Italian text "Il carrettiere frusta il cavallo troppo lento"): (a) original signal, (b) partially-encrypted signal.

3.2. Objective performance measures

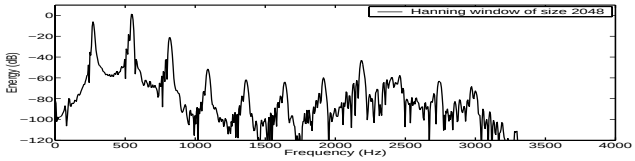
Segmental signal-to-noise ratio (segSNR) and spectral distortion (SD) were used to objectively assess the performance of the partial encryption scheme under investigation. Speech material consisting of 192 sentences, each four seconds long, spoken by both male and female speakers, were encoded using the ITU-T G.729 reference software and then partially encrypted with the proposed scheme. Although absolute segSNR and SD values are not descriptive per se, they are quite useful to compare the performance of partial encryption schemes and to supplement other performance evaluation data. Segmental SNR was used to compare partially encrypted signals to corresponding clear-text sentences. The segment size was 20 ms, no overlapping, no threshold. The proposed partial encryption scheme performs very similarly to full encryption: its segSNR value, in fact, is -18.41 dB, only 0.3 dB below full encryption. Also the Spectral Distortion measure of partial encryption, 7.60 dB, is close to that of full encryption, 8.14 dB.

In the following Section, we will see how the insight gained by signal inspection and objective performance measures correlate with the responses of listeners in formal listening tests.

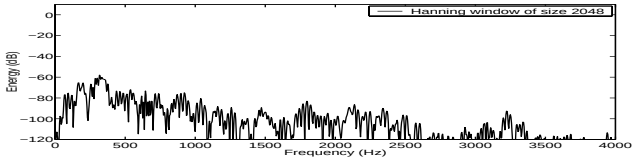
3.3. Formal Listening Tests

The test material was presented to 13 different listeners, all using headphones in a controlled environment. The three experiments described in Section 2.2 were carried out to assess the subjective performance of partial encryption. For the first (intelligibility) and third (speech/non-speech discrimination) experiment, eight sentences each, four seconds long, were used. For the second experiment (plain-text identification), three sets of four sentences each were used.

Table 2 shows the overall results. For the proposed partial encryption scheme, the intelligibility average score



(a) No encryption



(b) Partial encryption

Fig. 2. Magnitude spectra of partially encrypted speech signal corresponding to a vowel (the penultimate /e/ sound of the Italian word "carrettiere"): (a) original spectrum, (b) partially-encrypted spectrum.

was precisely 1.0, corresponding to "Nothing" in the five point scale described in Section 2.2: nothing intelligible was ever discerned in the 104 stimuli presented to the listeners. Moreover, no score downgrading ever took place, i.e., no listener ever had the perception of understanding a word.

For plain-text identification success rate is statistically equivalent to random guesses; one-third of the responses were "don't know."

Also for speech/non-speech discrimination, the success rate is again statistically equivalent to random guesses. Although only part of the compressed bitstream is ciphered, the resulting speech signal can not be distinguished from the encryption of random noise or silence.

The results of the formal listening tests confirm that the proposed partial encryption scheme achieves content protection virtually equivalent to full encryption of the compressed bitstream.

4. CONCLUSIONS

Mobile multimedia applications, the focus of many forthcoming wireless services, increasingly demand low-power techniques implementing content protection and customer privacy. In this paper low complexity perception-based partial encryption scheme for telephone bandwidth speech was presented. Speech compressed by a widely-used speech coding algorithm, ITU-T G.729 at 8 kb/s, was partitioned in two classes, one, the most perceptually relevant, to be encrypted, the other, to be left unprotected. The proposed scheme covers about 45% of the bitstream and achieves content protection equivalent to that obtained by full encryption, as verified by both objective measures and formal listening tests.

Less destructive partial encryption of speech signals is also possible. Informal results show that selective encryption of as little as 20% of the bitstream still leads to extremely low levels of intellegibility. When equivalency to full encryption is not needed, secure speech communications can thus be obtained with even lower complexity than the proposed scheme.

Low-power, portable devices can, therefore, achieve very high levels of speech-content protection at only a fraction of the computational load of current techniques, freeing resources for other tasks and enabling longer battery life.

Table 2. Listening test results.

	Score		
Intelligibility (1 to 5)	1.00		
	No vote	Correct Responses	Erroneous Responses
Plain-Text Identification	30.8 %	17.9 %	51.3 %
Speech/Non-Speech Discrimination	14.4 %	44.2 %	41.4 %

5. ACKNOWLEDGEMENTS

The authors would like to thank prof. Michele Elia of the Politecnico di Torino for his insightful comments and suggestions.

6. REFERENCES

[1] J. Goodman and A.P. Chandrakasan, "Low Power Scalable Encryption For Wireless Systems," *Wireless Networks*, vol. 4, no. 1, pp. 55–70, 1998.

[2] G.A. Spanos and T.B. Maples, "Security for Real-Time MPEG Compressed Video in Distributed Multimedia Applications," in *Conference on Computers and Communications*, March 1996, pp. 72–78.

[3] H. Cheng and X. Li, "Partial Encryption of Compressed Images and Videos," *IEEE Transactions on Signal Processing*, vol. 48, no. 8, pp. 2439–2451, Aug. 2000.

[4] R. Salami et al., "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 116–130, March 1998.

[5] P. Kroon and Y. Shoham, "Performance of the Proposed ITU-T 8 kb/s Speech Coding Standard for a Rayleigh Fading Channel," in *Proceedings IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, Maryland, September 1995, pp. 11–12.

[6] K. Swaminathan, A.R. Hammons Jr., and M. Austin, "Selective Error Protection of ITU-T G.729 CODEC for Digital Cellular Channels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 577–580.