SPEECH PRIVACY FOR MODERN MOBILE COMMUNICATION SYSTEMS

José F. de Andrade Jr.[*] *Marcello L. R. de Campos*[*] *José A. Apolinário Jr.*[*]

Federal University of Rio de Janeiro
Electrical Engineering Program
P.O.Box 68504, 21941-972, Brazil
ps7jfa@urbi.com.br

Federal University of Rio de Janeiro
Electrical Engineering Program
P.O.Box 68504, 21941-972, Brazil
campos@lps.ufrj.br

Military Institute of Engineering
Department of Electrical Engineering
Praça General Tibúrcio, 80, 22290-270, Brazil
apolin@ime.eb.br

ABSTRACT

Speech privacy techniques are used to scramble clear speech into an unintelligible signal in order to avoid eavesdropping. Some analog speech-privacy equipments (scramblers) have been replaced by digital encryption devices (comsec), which have higher degree of security but require complex implementations and large bandwidth for transmission. However, if speech privacy is wanted in a mobile phone using a modern commercial codec, such as the AMR (Adaptive Multirate) Codec, digital encryption may not be an option due to the fact that it requires internal hardware and software modifications. If encryption is applied before the codec, poor voice quality may result, for the vocoder would handle digitally encrypted signal resembling noise. On the other hand, analog scramblers may be placed before the voice encoder without causing much penalty to its performance. Analog scramblers are intended in applications where the degree of security is not too critical and hardware modifications are prohibitive due to its high cost. In this article we investigate the use of different techniques of voice scramblers applied to mobile communications vocoders. We present our results in terms of LPC and cepstral distances, and PESQ values.

Index Terms— Speech privacy, scramblers, mobile communications, adaptive multirate codec, AMR.

1. INTRODUCTION

In order to make a speech signal unintelligible, analog speech scrambling algorithms use permutation of speech segments in time, frequency, or time-frequency domain. This is usually carried out in the digital domain and the signal is converted back to its analog form for transmission. Therefore, the main constraint of these techniques is the preservation of the bandwidth such that the scrambled speech can be transmitted through an ordinary speech channel. This technique was widely used in the past and comprises a class of speech privacy devices with a level of security considered *tactical*.

On the other hand, digital encryption offers a higher level of security and is widely used nowadays. The price for this so-called *strategical* level of security is usually the need of a digital transmission with a larger bandwidth.

In case speech privacy is needed in modern mobile telephony, due to the encapsulated technology of commercial codecs with no external access, digital encryption is not an option when a low cost equipment is desired. Therefore, although commercial solutions exist to avoid eavesdropping even in a higher level of security, as demanded by governmental agencies, hardware modifications may restrict its use to the general public due to the prohibitive costs involved.

This paper recasts analog scrambler techniques in an attempt to propose a low cost speech-privacy mobile phone for use with commercial off-the-shelf (COTS) equipments. Two classes of scrambling techniques are used in a modern codec and the result is assessed by means of objective measures.

The paper is organized as follows. The basic theory of analog scramblers is revisited in Section 2 while Section 3 deals with the main implementation issues concerning the use of a speech privacy in modern mobile telephony: codec, synchronization, key management, and performance evaluation. The experimental results are presented in Section 4 followed by conclusions in Section 5.

2. ANALOG SCRAMBLERS

This section briefly reviews the main concepts used in speech privacy techniques. They will be referred to as *Time-Segment Permutation* (TSP), *Frequency-Domain Scrambling* (FDS), *Time-Frequency Scrambling* (TFS), and *Transform-Domain Scrambling* (TDS).

2.1. Time-Domain Scramblers

In time-domain scrambling [1], the most used method, TSP, divides the digitized speech signal $x(n)$ into short time frames or blocks[1] (typically 20ms, i.e., 160 samples for a sampling

[*]The authors are grateful to CNPq and to the Brazilian Navy for partially supporting this work.

[1]frames and blocks will be used interchangeably in this work to denote a set of signal segments.

frequency $fs = 8000$kHz) which are divided in smaller segments that are permuted in time.

Considering M time segments, $\boldsymbol{x}_m$, $m = 1, \ldots, M$, each containing N samples, the i-th frame or block is represented as vector $\vec{\boldsymbol{x}}_i = [\boldsymbol{x}_1^T\ \boldsymbol{x}_2^T\ \cdots\ \boldsymbol{x}_M^T]^T$, with MN elements. An $M \times N$ matrix may be constructed with one segment per row, as follows:

$$\boldsymbol{X}_i = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_M]^T \tag{1}$$

The $M \times M$ permutation matrix $\boldsymbol{P}$ is defined as a matrix having only one nonzero element in each row, each nonzero element being equal to one. The scrambled speech block, vector $\vec{\boldsymbol{y}}_i$, is obtained from the concatenation of the rows of the product $\boldsymbol{Y}_i = \boldsymbol{P}\boldsymbol{X}_i$, i.e.,

$$\vec{\boldsymbol{y}}_i = [\boldsymbol{y}_1^T\ \boldsymbol{y}_2^T\ \cdots\ \boldsymbol{y}_M^T]^T \tag{2}$$

where

$$\boldsymbol{Y}_i = \boldsymbol{P}\boldsymbol{X}_i = [\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_M]^T \tag{3}$$

In the receiver, recovery of the original speech vector $\vec{\boldsymbol{x}}_i$ is obtained rearranging vector $\vec{\boldsymbol{y}}_i$ to form an $M \times N$ matrix $\boldsymbol{Y}_i$ and multiplying it by matrix $\boldsymbol{P}^{-1}$.

Time-domain scramblers have three major implementation factors that limit their application: (i) need for time synchronization, (ii) introduction of time delay, and (iii) small effective number of keys [7]. Therefore, this class of scramblers has not been considered suitable for the application of interest in this work and was included here for didactic purposes only.

2.2. Frequency-Domain Scramblers

Frequency-domain scramblers split the frequency contents of each speech signal block into M frequency bands. These bands are permuted according to some particular rule (or key), and a time sequence with scrambled frequency contents is synthesized to replace the original speech signal block. Frequency-domain scramblers are usually implemented with uniform filter banks or with wavelet transforms [2].

An M-subband multirate filter bank is a set of M filters, which span the whole frequency spectrum. The speech signal is split into M subbands after passing through the analysis filter bank, $H_i(z)$, and is critically downsampled, i.e., decimated by a factor of M. An $M \times M$ permutation matrix $\boldsymbol{P}$ is inserted after the decimators in order to scramble the signals in the subband domain. It is then fed to upsamplers (interpolators) followed by the synthesis filter bank, $F_i(z)$ (see Fig. 1). Recovery of the original speech vector is obtained with the same structure of Fig. 1 using the inverse of the permutation matrix, $\boldsymbol{P}^{-1}$.

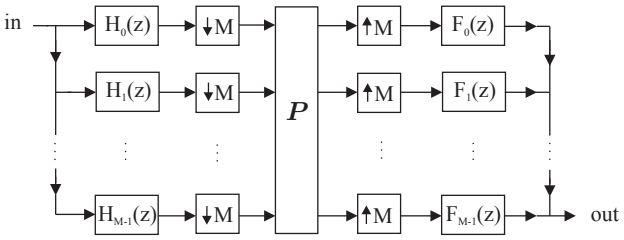

Fig. 1. Block diagram of a Frequency Domain Scrambler.

2.3. Alternative Analog Schemes

As alternative schemes, we present next the *Bi-Dimensional* (or *Time-Frequency*) Scrambler and the *Transform-Domain* Scrambler.

Time-frequency scrambling uses a combination of both time- and frequency-domain scramblers. The main idea is to split the speech signal frame into M uniform subbands and to divide the output of each subband in N segments. These time-frequency segments are then scrambled and a new scrambled speech signal frame is synthesized. Although this type of scrambler has very low residual intelligibility [3], it was also not considered suitable to be used with AMR codecs, for having equivalent or more critical implementation constraints as the TSP scheme.

Transform-domain scramblers [4] divide the speech signal in blocks and a transformation matrix is used to pre-multiply each vector, obtaining a transformed signal, which is divided in segments to be permuted. An inverse transformation is then applied, resulting in the scrambled speech signal.

Consider vector $\vec{\boldsymbol{x}}_i$, which contains MN speech samples representing the ith frame of speech signal. Let vector $\vec{\boldsymbol{x}}_i$ be pre-multiplied by an $MN \times MN$ orthogonal transformation matrix $\boldsymbol{T}$ such that $\vec{\boldsymbol{v}}_i = \boldsymbol{T}\vec{\boldsymbol{x}}_i$.

Vector $\vec{\boldsymbol{v}}_i$ can be split into M segments, which are arranged in an $M \times N$ matrix $\boldsymbol{V}_i$ with one segment per row. Then an $M \times M$ permutation matrix $\boldsymbol{P}$ is applied to $\boldsymbol{V}_i$, generating a permuted and transformed matrix $\boldsymbol{U}_i$, $\boldsymbol{U}_i = \boldsymbol{P}\boldsymbol{V}_i$.

The scrambled speech vector $\vec{\boldsymbol{y}}_i$ is obtained by applying the inverse transformation matrix $\boldsymbol{T}^{-1}$ to vector $\vec{\boldsymbol{u}}_i$, constructed via concatenation of the rows of matrix $\boldsymbol{U}_i$, i.e., $\vec{\boldsymbol{y}}_i = \boldsymbol{T}^{-1}\vec{\boldsymbol{u}}_i$.

In order to recover the original speech vector $\vec{\boldsymbol{x}}_i$, it is necessary to apply transformation, inverse permutation, and inverse transformation to each block of the scrambled signal.

In order to ensure that noise added by the channel will not be amplified in the de-scrambling process, it is necessary to constrain the scrambling and de-scrambling processes to be orthogonal transformations.

3. IMPLEMENTATION ISSUES

3.1. Adaptive Multirate Codec

The Adaptive Multirate (AMR) speech coder consists of a multirate speech data compression scheme optimized for mobile communication operation. It consists of the speech coder, a source controlled rate scheme, including a voice activity detector (and a comfort noise generation system), and an error concealment mechanism to combat the effects of transmission errors and lost packets.

AMR has been adopted as the standard speech codec by 3GPP and it is widely used in GSM (Global System for Mobile communications) and its third generation evolution (3GSM). Its most important feature is the capacity of adapting to select bit rate according to link condition. If the radio link becomes poorer, source coding is reduced and channel coding is improved. The multirate speech coder is a single integrated speech codec with eight source rates from 4.75 to 12.2 kbps, and a low rate background noise encoding mode.

The speech coder is capable of switching its bit-rate every 20ms speech frame. AMR employs different techniques such as Algebraic Code Excited Linear Prediction (ACELP), Voice Activity Detection (VAD), and Discontinuous Transmission (DTX) [5].

3.2. Synchronization

Synchronization is necessary to find boundaries in a block processing scheme. In case of speech scrambling sensitive to synchronization, this process may become mandatory in order to recover the original signal from the scrambled speech.

Whenever necessary, the following basic synchronization methods may be employed: (i) synchronization signal sent only during startup, (ii) continuously transmitted, or (iii) sent at periodic intervals.

The use of methods that do not require frame synchronization is an import choice since the synchronization design and implementation increases the final equipment cost. Synchronism is also necessary in order to offer other attractive features, such as key management, when time-varying key is used. The simulation results presented in this paper assumed that the synchronization was not previously known, i.e., no synchronization scheme was employed.

Bi-dimensional scramblers and TDS are susceptible to loss of frame synchronism, while FDS schemes (those employing sharp subband filters) do not require frame synchronization [6].

3.3. Algorithm Key

For time-domain scramblers, the effective number of algorithm keys is only $10 - 20\%$ of the key space ($M!$) [7]. In the frequency domain, there are $M!$ ways to permute the subbands, but again not all permutations produce scrambled speech signals with low residual intelligibility (possibility that listeners may fully comprehend or partially grasp the meaning of a message from a scrambled speech). From the authors personal experience, only approximately $13,000$ out of the $8! = 40,320$ keys can be considered effective.

In order to overcome the problem of key selection, two methods can be employed: key generation algorithm and time-varying keys. The second method is more suitable for reducing the residual intelligibility and increasing the robustness to cryptanalysis; however, it needs a segment synchronization scheme. In equipments with embedded GPS, their clock can be used to implement a local synchronization.

Among analog scrambler techniques, the lowest residual intelligibility is obtained with bi-dimensional or time-frequency scramblers [3].

3.4. Performance Evaluation

3.4.1. Objectives Measures

The residual intelligibility of the scrambled signal and the quality of the recovered signal were evaluated using two objective measures: LPC (*Linear Predictive Coding*) distance and Cepstral distance [8]. The distance measures represent the level of spectral similarity between the tested signals and, therefore, it is possible to quantify the residual intelligibility of the scrambled speech signal and the quality of the recovered speech signal.

3.4.2. Perceptual Evaluation of Speech Quality (PESQ)

PESQ is an objective measurement tool, defined according to [9], that predicts the results of subjective listening tests on narrow band telephony systems and speech codecs. This quality measure method uses a perceptual model to compare the original, unprocessed signal, with the degraded or processed signal. The resulting quality score, though an objective measure, is more closely related to the subjective "Mean Opinion Score" (MOS) defined according to [10].

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The selected schemes to be tested in our simulations were: *Discrete Cosine Transform Transform Domain Scrambling* (DCT TDS) and *Uniform DFT Filter Bank Frequency-Domain Scrambling* (UDFT FDS). The second scheme applies the *polyphase-component* decomposition concept [11] and uses as prototype filter an 170th-order Finite Impulse Response (FIR) filter.

All signals used in the simulations were sampled at 8kHz with 16-bit precision. The scrambling process and its inverse operation were applied to signal frames of 20ms. Time-invariant keys with $M = 8$ were used.

Table 1 shows the residual intelligibility for different coding rates for both scramblers. In this table, original and scram-

bled signals are compared using LPC and Cepstral distances. Table 2 shows the recovered signal quality also for different coding rates for both scrambling strategies. Note here the PESQ scores which were not present in the first table due to its non adequacy for evaluating ciphered speech. The DCT method (TDS) showed lower residual intelligibility than the UDFT method (FDS), but also, due to loss of synchronization, a poor quality of the recovered signal (Table 2). In order to apply this transform method on systems that employ codecs, a sophisticated synchronization scheme is required.

The UDFT filter bank method proved to be resistant to loss of synchronization and maintained an acceptable residual intelligibility, which can be improved increasing the number of subbands. Tables 1 and 2 present the results for the scrambled and the recovered signals, respectively. It can be observed that, for the test signals used[2] and the best scrambler technique for this application (FDS), there is a slightly improvement as the rate increases from 4.75 to 12.2 kbps.

Table 1. Comparison between original speech and scrambled signals with TDS (DCT) and FDS (UDFT) for AMR rates 4.75, 5.9, 7.85, and 12.2 kbps.

Method and rate	LPC Distance	Cepstral Distance (dB)
TDS 4.75 kbps	5.569	5.569
FDS 4.75 kbps	5.327	5.193
TDS 5.90 kbps	5.613	5.344
FDS 5.90 kbps	5.387	5.205
TDS 7.85 kbps	5.613	5.336
FDS 7.85 kbps	5.387	5.205
TDS 12.2 kbps	5.894	5.388
FDS 12.2 kbps	5.354	5.145

Table 2. Comparison between the recovered and the original speech signals for TDS (DCT) and FDS (UDFT) and AMR rates 4.75, 5.9, 7.85, and 12.2 kbps.

Method and rate	LPC Distance	Cepstral Distance (dB)	PESQ
TDS 4.75 kbps	0.772	0.990	1.981
FDS 4.75 kbps	0.632	0.546	2.338
TDS 5.90 kbps	0.634	0.761	2.048
FDS 5.90 kbps	0.550	0.248	2.443
TDS 7.85 kbps	0.634	0.761	2.048
FDS 7.85 kbps	0.550	0.248	2.443
TDS 12.2 kbps	0.305	−0.945	2.937
FDS 12.2 kbps	0.315	−1.286	2.938

A similar experiment was carried out using English speakers (10 speakers from the LDC corpus TIMIT) and similar results were obtained, but with a slightly higher PESQ values for both scrambled and de-scrambled signals.

[2]The corpus used for this experiment was composed of 40 males native Brazilian speakers. For each speaker, a set of four phonetically balanced Portuguese phrases were recorded, in a noiseless environment, using an electret microphone.

5. CONCLUSIONS

The analog scrambler techniques applied to AMR codecs described here provide some degree of speech privacy, graded as casual. These techniques are appropriated to systems having a non-critical degree of security and whenever hardware modifications are considered too expensive. From the basic theoretical part and the experimental results, the best analog scrambler for the proposed application is the frequency-domain scheme which does not require synchronization nor expand the signal bandwidth. The results from our experiments have shown that scrambled speech deciphered after the AMR codec presented PESQ ranging from 2.3 to 2.9 for FDS and a bit rate from 4.75 kbps to 12.2 kbps, respectively. In order to assess this score, we note that an average PESQ score for (clear speech after an AMR codec) American English is 3.71, whereas for average Spanish is 3.25. In our experiments (Brazilian Portuguese), we have obtained 3.50. Although not close to the PESQ score for clear speech at 4.75 kbps, after listening to several de-scrambled signals in all rates, we could observe that they were all intelligible, which gave us the feeling that FDS can be used for implementing a low cost voice privacy mobile phone.

6. REFERENCES

[1] N. Jayant, B. McDermott, S. Christensen, and A. Quinn, "A comparison of four methods for analog speech privacy," *IEEE Transactions on Communications*, vol. COM-29, no. 1, pp. 18–23, January 1981.

[2] F. Ma, J. Cheng, and Y. Wang, "Wavelet transform-based analogue speech scrambling scheme," *Electronics Letters*, vol. 32, no. 8, pp. 719–721, April 1996.

[3] A. S. Bopardikar, V. U. Reddy, *Speech Encryption Using Wavelet Packets*. Bangalore, Indian Institute of Science, October 2005.

[4] M. S. Ehsani, S. E. Borujeni, "Fast Fourier transform speech scrambler," *2002 First International IEEE Symposium Intelligent Systems*, pp. 248–251, September 2002.

[5] 3GPP TS 26.071 V6.0.0 (2004-12) Technical Specification 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech CODEC speech processing functions; AMR speech CODEC; General description (Release 6).

[6] L. S. Lee, G .C. Chou, and C .S. Chang, "A new frequency domain speech scrambling system which does not require frame synchronization,"*IEEE Trans. Commun.*, vol. COM-32, no. 4, pp. 444–456, April 1984.

[7] V. Senk, V. D. Delic, and V. S. Milosevic, "A new speech scrambling concept based on Hadamard matrices," *IEEE Signal Processing Letters*," vol. 4, no. 6, pp. 161–163, June 1997.

[8] J. R. Deller Jr., J. G. Proakis, J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Macmilan Publishing Company, 1993.

[9] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, International Telecommunication Union, Geneva, February 2001.

[10] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, International Telecommunication Union, Geneva, August 1996.

[11] P .P. Vaidyanathan, *Mulitrate Systems and Filter Banks*. Englewood Cliffs NJ, Prentice-Hall, 1993.

[12] ITU-T Recommendation P.862.3, *Series P: Telephone transmission quality, telephone installations, local line networks. Methods for objective and subjective assessment of quality.* International Telecommunication Union, Geneva, November 2005.