Data Hiding Robust to Mobile Communication Vocoders

Reza Kazemi, Fernando Pérez-González, *Fellow, IEEE*, Mohammad Ali Akhaee, and Fereydoon Behnia

***Abstract*—The swift growth of cellular mobile networks in recent years has made voice channels almost accessible everywhere. Besides, data hiding has recently attracted significant attention due to its ability to imperceptibly embed side information that can be used for signal enhancement, security improvement, and two-way authentication purposes. In this regard, we aim at proposing efficient schemes for hiding data in the widespread voice channel of cellular networks. To this aim, our first contribution is to model the channel accurately by considering a linear filter plus a nonlinear scaling function. This model is validated through experiments with true speech signals. Then we leverage on this model to propose two additive and multiplicative data hiding methods based on the spread spectrum techniques. In addition, inspired by the concept of M-ary biorthogonal codes, we develop novel schemes that significantly outperform the previous ones. The performance of all the methods that we present is assessed mathematically and cross-validated with simulations. These are later extended to true speech signals where the results evidence an excellent performance as predicted by the theory. Finally, we assess the imperceptibility by means of both subjective and objective benchmarks and show that the perceptual impact of our watermarks is acceptable.**

***Index Terms*—Data hiding, mobile communication vocoder, nonlinearity, spread spectrum, watermarking.**

I. Introduction

CELLULAR networks have become major means of voice communications around the world. This is to the extent that the number of mobile network users in 2014 was estimated to be 6.8 billion, showing a penetration rate of 97% among the world's population [1]. In spite of such spread, source and destination authentication is still an issue in cellular networks. There exist a number of ways such as IMSI-Catcher or VoIP termination to spoof a personal number and utilize it for malicious purposes [2]. Although some progress and solutions have been proposed to overcome this vulnerability, to the best of our knowledge, most of them rely on one of the following prerequisites.

The first category of solutions imposes modifications to the cellular network protocol [3], [4] which of course are hard to implement in already deployed cellular networks. The second category needs data transferring through mobile voice channels which is prohibitive and still not practical [5], [6]. With the previous limitations in mind, we aim at developing a data hiding method which can easily be implemented on common existing smart cellular phones and seamlessly embeds the appropriate data (e.g., authentication data) in the speech of the user while he/she is regularly talking with another person on his/her phone. Furthermore, data hiding in mobile voice channels can be exploited for improving the quality of service [7], and for enhancing the security of communications by combining watermarking and encryption algorithms [8].

Information hiding applications impose different constraints that lead to substantially different methods. In steganography, the mere existence of a hidden message is to remain unknown to the adversary, while in data hiding this requirement can be sacrificed for a larger robustness against intentional attacks and common processing operations. When the information to be hidden is simply the presence or absence of a certain secret pattern, the term one-bit data hiding is used and is sometimes equated with watermarking. Since even in the multiple-bit case data is hidden by embedding a low-power signal called *watermark*, we will use the terms data hiding and watermarking interchangeably throughout this paper. One-bit watermarking is usually utilized in integrity verification applications such as copyright protection [9]. We are instead interested in multiple-bit data hiding that must survive substantial channel distortions, so that it can be reliably decoded at the receiver side [10]. Several methods have been proposed for data embedding in audio and human voice signals using data hiding techniques. These methods can be considered as one of the following sorts. In the first one, the information is embedded in the audio files in offline applications such as copyright protection of audio files [11], while the second one concentrates on online applications such as data insertion in Voice over IP (VoIP) streams [12]. Wang *et al.* proposed a method to embed data in a G.711 vocoder by hiding the information in the LSBs of the speech signal [13]. Ditmann *et al.* proposed a general scheme for data embedding in all VoIP streams by focusing on the active frame of the speech signal [14].

Huang *et al.* introduced a new method for data hiding in G.723.1 VoIP frames. They insert the secret message into the LSB of inactive speech VoIP frames [15]. They also suggested another solution to embed data in G.729A VoIP frames based on m-sequences [16]. Huang *et al.* also proposed a novel embedding method into a low-bit rate codec. Therein, data

Manuscript received September 12, 2015; revised June 5, 2016; accepted August 5, 2016. This work was supported in part by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under Project TACTICA and Project COMPASS (TEC2013-47020-C2-1-R), and in part by the Galician Regional Government and ERDF under "Project Consolidation of Research Units" (GRC2013/009), Project REdTEIC (R2014/037), and Project AtlantTIC. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiao-Ping Zhang.

R. Kazemi and F. Behnia are with the Department of Electronic and Electrical Engineering, Sharif University of Technology, Tehran 1136511155, Iran (e-mail: rezakazemi.reza@gmail.com; behnia@sharif.edu).

F. Pérez-González is with the Signal Theory and Communications Department, EE Telecomunicacion, University of Vigo-SPAIN, Vigo 36310, Spain (e-mail: fperez@gts.uvigo.es).

M. A. Akhaee is with the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran 1417466191, Iran (e-mail: Akhaee@ut.ac.ir).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2599149

insertion is performed in the process of pitch period prediction for a G.723.1 VoIP codec [17]. Piotrowski and Gul presented methods for watermarking in VoIP applications[18], [19]. Xiao *et al.* proposed a structure for low rate vocoders based on Quantization Index Modulation (QIM) [20]. Finally, Wu *et al.* introduced another method for data hiding over high rate vocoders such as G.711 [21].

The aforementioned related art is generally focused on steganography and not watermarking. Moreover, the targeted vocoders are the common ones used in VoIP standards. As a consequence, focus is put on implementation issues of IP networks such as packet losses, jitter, etc. in order to design mechanisms to mitigate these kind of impairments. For instance, some of those works modify the standard vocoder structure and design a new one with data hiding capabilities which is robust to the targeted distortions.

Leveraging on the mentioned related art, the main contributions of this work can be summarized as follows:

1) Model the mobile communication vocoders with a combination of linear and nonlinear blocks and validate the constructed model.
2) Extend the baseline SS methods to make them robust to the constructed model, in particular, to the nonlinear scaling blocks within, even when multiple symbol constellations are used.
3) Enhance the derived methods by considering biorthogonal codes as spreading signals and carry out an accurate performance analysis of the proposed methods.

Moreover, the other novelty in this paper is related to its application. While most of the previous research in this domain has concentrated on steganography and watermarking over VoIP channels, and, consequently, on the group of vocoders which are relevant there, we aim at proposing a method for watermarking through mobile voice channels, which enjoy a tremendous penetration. The main difference between our work and the related art lies in the target vocoder, as VoIP customarily employs waveform vocoders (which do not make any prior assumption on the speech signal) for which compression rates are in the range [24-64] kbps (such as G.711, G.729) [22], whereas the vocoders utilized in mobile voice channels such as AMR are based on extracting the signal parameters and then modeling and synthesizing the speech at the receiver side, and their compression rate is in the range of [.6-13.2] kbps. It should be noted that data hiding for the second group of vocoders (such as AMR) is inherently more complicated due to the higher compression rate.

Actually, in real mobile voice communication environments, there are several effects pertinent to the vocoder systems, wireless communication channel, synchronization, etc. [23]. In this regard, and since data hiding robustness is our main motivation, we believe that the vocoders at both the transmit and receive terminals constitute the main impairment among all other effects due to the following considerations:

1) The actual over-the-air transmission is handled by the cellular network system with its (proprietary) waveform design, modulation, forward error correction (FEC), forward error detection (FED) and equalization. The digital transmission/reception subsystems are in charge of guaranteeing that the samples (now, the watermarked samples) are delivered reliably [23]. Since our work proposes a variant of spread-spectrum data hiding, which is known to be robust to sample deletion, insertion and channel errors [24], [25], the watermark will still be decoded correctly.
2) Even if several full frames may be affected by fading or other network impairments, to the point that the watermark cannot be reliably decoded, the hidden data (e.g., authentication information) can be repeated as many times as needed and with a time separation that is much larger than the coherence time of the channel, so that the critical hidden information can still be decoded. This is of course a rudimentary sort of repetition coding, but a small payload will suffice in most applications. Obviously, other more sophisticated methods for protecting the watermark are possible, but we have not pursued them and are left for future works.
3) Finally, a real implementation would need to include means for achieving synchronization, both at the watermark symbol level and at the watermark data frame level. This is also out of the scope of our paper, although some solutions have been proposed to overcome these issues, such as [24], [26], [27], which can also be integrated with our proposed methods.

Therefore, this work centers on investigating data hiding methods robust to the vocoder effects of the mobile voice channel. In doing so, we face two important constraints. On one hand, the proposed method should operate on the vocoders already deployed in cellular networks; thus, in contrast to existing solutions which modify the vocoders, for our purposes we rule out such possibility. On the other hand, it must be achieved with a low complexity that allows for implementations with low impact in CPU usage and power consumption.

In order to design an efficient data hiding method robust to mobile communication vocoders, there are two main options, namely, Quantization Index Modulation (QIM) and Spread Spectrum (SS) methods. Both have advantages and disadvantages in terms of data rate, robustness, imperceptibility and security. For applications where security is important, SS-based data hiding is arguably superior [28]. Moreover, SS-based data hiding is more robust to strong channel distortions like the ones we encounter in mobile voice networks [29], including lossy compression, nonlinear gains, analog-to-digital conversion, etc. [30]. Since high data embedding rates are not critical in the foreseen applications, we have singled out SS-based solutions.

In this paper, we first model the mobile communication vocoder as a combination of a linear filter and a non-linear block. Then, according to the constructed model, two types of robust suboptimal decoders based on the SS paradigm are designed and developed. The performance of the proposed schemes is analytically studied and the imperceptibility of the scheme is investigated and evaluated using well-known subjective and objective metrics.

Notation: Throughout this paper, we use regular lowercase letters for scalar variables and random variables, and lowercase boldface letters for vectors. Matrices are represented using regular uppercase letters, while the corresponding regular lowercase, with subscript indices, represents the entries. We use $\|\mathbf{x}\|^l$ to

denote $\sum |x_i|^l$. The probability distribution function of a random variable is denoted by $p(\cdot)$ and the probability of occurrence of a single event is written as $\Pr(\cdot)$. Besides, $\mathrm{E}[\cdot]$ is the expected value of random variables.

The rest of this paper is organized as follows. Section II focuses on modeling of the mobile communication vocoder and proposes an approximation of the nonlinear effects of the codec voice channel. Spread spectrum embedders and decoders according to the constructed model are proposed in Section III. We proceed in this section with a performance analysis of these decoders. Section IV provides the results of simulations with synthetic signals as well as with true speech signals, including an imperceptibility assessment based on both objective and subjective benchmarks. Finally, Section V contains our conclusions and discusses future research lines.

II. Modeling Mobile Communication Vocoders

The behavior of vocoder systems mainly depends on the utilized coding techniques. Generally, the codecs used in voice dedicated channels are classified into two main groups: the first group comprises waveform coders that encode the input signal without any prior assumption on the speech signal. These codecs exploit coding techniques such as PCM (G.711 international standard) and ADPCM (G.726 international standard) to achieve bit rates in the range of [24-64] kbps. Such high data-rate codecs allow transmission of most signals in the voice frequency range with minor distortion.

The second group is a set of vocoders that extract and encode some voice-specific parameters (mostly LPC-based speech modeling parameters) from the input signal and use them to synthesize the voice signal while decoding. These codecs use coding techniques such as Regular Pulse Excitation with Long-Term Prediction (RPE-LTP), Code Excited Linear Prediction (CELP), ACELP, Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP), Vector Sum Excited Linear Predictive Coding (VSELP), Mixed Excitation Linear Prediction (MELP), etc. They can achieve output data rates in the range of [0.6-13.2] kbps. Data embedding schemes that use codecs of the second type which include cellular network vocoders such as AMR are more complicated due to the higher compression rate.

Considering the nonlinear and complicated blocks of vocoders [29], [31], it seems impossible to exactly model the mentioned systems and derive the statistically optimum decoder subject to ML criteria. In order to solve this problem, we concentrate on approximating this channel for human voice inputs. We consider a linear filter plus noise and a nonlinear scaling block as the basis of our approximate model and validate it. In each stage of model validation, we will take into account the interplay between the linear and nonlinear parts. In doing so, we start constructing our model based on the linear part while constraining the input signal to be small for preventing the occurrence of nonlinear effects. Then, we model the non-linear part, and finally, we validate the entire model with actual speech signals.

For the sake of simplicity and clarity, at first we consider pure spread-spectrum (SS) watermarking (for improved SS watermarking the computations differ, see Section III). To perform the required analysis, let x_k denote the kth sample of the host signal, w_k is the corresponding sample of the watermark, and $y_k = x_k + w_k$ is then the watermarked signal.

As customary, we assume that watermarking is performed in an i.i.d fashion, with a watermark with zero mean and variance σ_w^2. As to the host, we first assume it is stationary, with zero mean, variance σ_x^2 and normalized autocorrelation function

$$\rho_i \doteq \mathrm{E}\{x_k x_{k+i}\}/\sigma_x^2. \tag{1}$$

With these definitions, the Document-to-Watermark Ratio (DWR) can be written as σ_x^2/σ_w^2.

For hypothesizing our model, first, we conjecture that the nonlinearity behaves as a linear function if the input power is small, in other words, to ensure that nonlinear effects are negligible, we input low power signals to model the linear part. Finally, after constructing both linear and nonlinear parts, we verify whether the initial conjecture holds. To proceed with constructing the linear part of our model, we are firstly interested in learning how linear time invariant (LTI) filtering affects each signal. We let h_k denote the filter impulse response, and y'_k, x'_k, w'_k the filtered versions of y_k, x_k, w_k, respectively. Notice that due to the superposition principle, we have $y'_k = x'_k + w'_k$. We assume that the decision about w_k is taken from y'_k alone; this means that even though watermarking decoding would clearly benefit from equalization of h_k, we decide not to do so. Therefore, we can write

$$y'_k = h_0 w_k + \sum_{\substack{i=-\infty \\ i\neq k}}^{\infty} h_i w_{k-i} + x_k * h_k. \tag{2}$$

The second term in the right hand side of (2) is akin to the intersymbol interference (ISI) found in communications, so we will refer to it by this name.

If we want to know the effective DWR at the output of the filter, we must compute the variance of the ISI plus host interference term, which we will denote by v_k. Noticing that w_k is white, we can write

$$\mathrm{E}\{v_k^2\} = \sigma_w^2 + \sigma_w^2 \sum_{\substack{i=-\infty \\ i\neq k}}^{\infty} h_i^2 + \sigma_x^2 \sum_{i=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_i h_m \rho_{i-m} \tag{3}$$

while for the watermark part, we have that the variance is simply $h_0^2\sigma_w^2$. Then, if the filter impulse response and the host autocorrelation are known (or can be estimated), it is possible to calculate the DWR.

We are interested in obtaining a manageable *equivalent* model for y'_k. To this end, we notice that if we scale the watermark, this affects both the useful part of the received signal and the ISI term, while if we scale the host, this affects only the host-interference term. Therefore we can write y'_k in terms of ISI term (denoted by t_k) and the host-interference term (denoted by u_k)

$$y'_k = h_0 w_k + t_k + u_k \tag{4}$$

where w_k, t_k and u_k are mutually independent, zero-mean, and

$$\mathrm{E}\{t_k^2\} = \sigma_w^2 \cdot \zeta(h);\ \mathrm{E}\{u_k^2\} = \sigma_x^2 \cdot \beta(h, \rho) \tag{5}$$

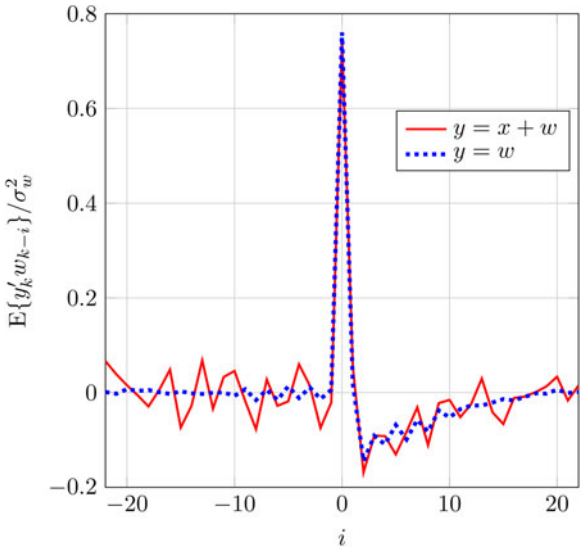


Fig. 1. Estimation of h_i based on (7), (9) for two different scenarios respectively: 1) $y = x + w$; $\sigma_x^2 = 1$, $\sigma_w^2 = 0.01$; 2) $y = w$; $\sigma_w^2 = 1$.

for some ζ and β that depend on the filter and the normalized input autocorrelation, as explicitly indicated by the notation. Finally, the effective DWR, denoted by τ, can be written as

$$\tau = \frac{\sigma_t^2 + \sigma_u^2}{h_0^2 \sigma_w^2} = \frac{\zeta(h) + \sigma_x^2/\sigma_w^2 \cdot \beta(h, \rho)}{h_0^2} \tag{6}$$

which explicitly depends on the input DWR but in an affine fashion. In this simplified model, one can use the superposition principle to estimate the scalars h_0, ζ and β.

As a first step, to validate the adequacy of the linear part of our model, we compare the estimated values for h_i, which we will denote by $\hat{h}_i$ in the sequel, obtained in two different ways. In the first, we compute the value of $\hat{h}_i$ by the following estimator:

$$\hat{\mathbf{h}} = R_{\mathrm{xx}}^{-1} \mathbf{r}_{\mathrm{yx}} \tag{7}$$

where $\hat{\mathbf{h}} \doteq [\hat{h}_{-l}, \ldots, \hat{h}_l]^T$. The autocorrelation matrix of the input signal (R_{xx}) and the input-output cross-correlation vector ($\mathbf{r}_{\mathrm{yx}}$) are defined as

$$R_{\mathrm{xx}} = \sigma_x^2 \begin{bmatrix} \rho_0 & \cdots & \rho_{-2l} \\ \vdots & \ddots & \vdots \\ \rho_{2l} & \cdots & \rho_0 \end{bmatrix}, \mathbf{r}_{\mathrm{yx}} = \begin{bmatrix} \mathrm{E}\{{y'}_k x_{k+l}\} \\ \vdots \\ \mathrm{E}\{{y'}_k x_{k-l}\} \end{bmatrix} \tag{8}$$

and l indicates the effective length of the impulse response. Alternatively, we can estimate h_i as

$$\hat{h}_i = \frac{\mathrm{E}\{y'_k w_{k-i}\}}{\sigma_w^2} \tag{9}$$

and the obtained results for both methods are shown in Fig. 1. As one can see, the achieved results here are consistent.

In addition, it should be noted that the achieved outcomes here seem noisy. In order to determine how much the estimated filter response changes with the input signal, we have conducted an experiment in which true speech samples are passed through the codec channel. In this simulation we move forward through

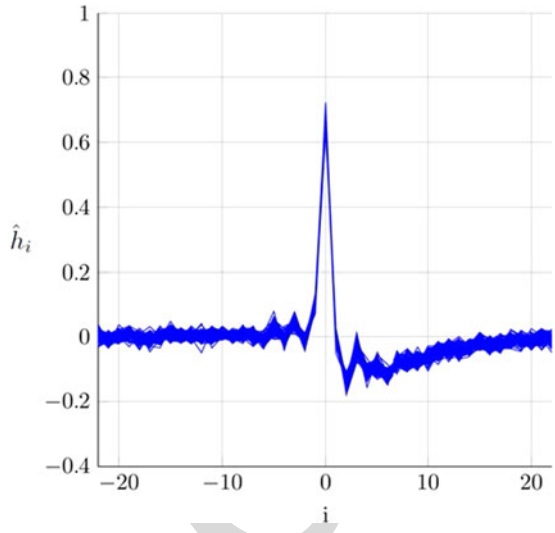


Fig. 2. Obtained values of $\hat{h}_i$ over several true speech samples.

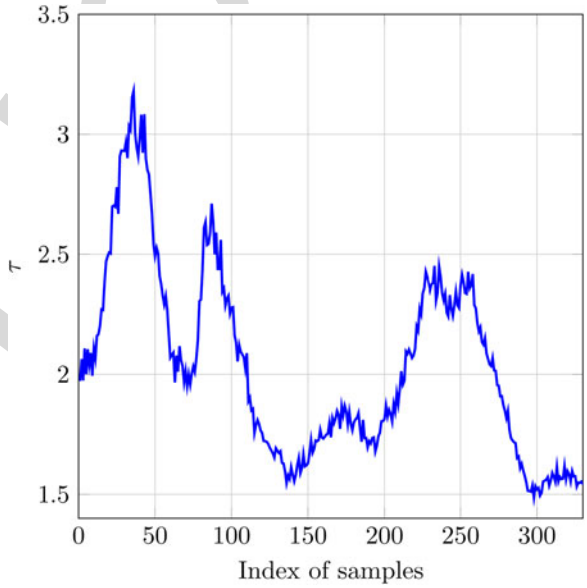


Fig. 3. Obtained value of τ over several true speech samples.

one long speech recording file with a window size of 50 samples each, while assuming a window overlap of 50%. Fig. 2 shows in one snapshot all the obtained channel responses, where we can see that the estimated channel responses have little variance; therefore, the average channel response can be taken as a good representative of the true impulse response. On the other hand, Fig. 3 represents the value of τ in (6) as a function of time for a speech signal. As we can notice, the obtained values fluctuate over time. This fluctuation could be easily justified by considering the definition of τ and its dependency on ρ_i which obviously varies in time due the non-stationary nature of true speech signals.

Moreover, to check if it is necessary to consider any noise in our modeling, we compare the theoretical output power (i.e., $\mathrm{E}\{{y'}_k^2\}$) with the average measured output power in the simulations (i.e., those conducted to plot Fig. 3). We compute the

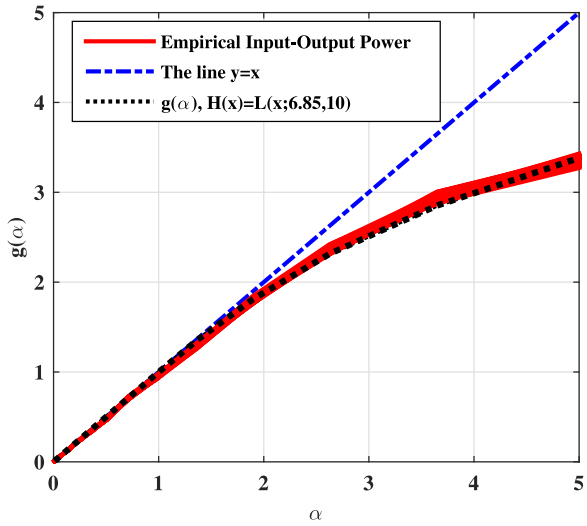


Fig. 4. Empirical input-output power curves (shown as a red band) and the approximation based on Rapp's model versus the power of input signal, denoted by α.

output energy of the signal as

$$\mathrm{E}\{y'^2_k\} = \sigma_w^2 \sum_i h_i^2 + \sigma_x^2 \sum_i \sum_m h_i h_m \rho_{i-m}\,. \tag{10}$$

The achieved results show that the difference between the measured output power and $\mathrm{E}\{y'^2_k\}$ is negligible (i.e., it was less than 0.015). Thus, we can include noise in our model with an approximate variance of 0.01 or, given its small contribution, even neglect it. For the sake of simplicity, we proceed with the latter.

As mentioned earlier, we conjectured that nonlinear effects do not arise in the case of low input power, so the constructed model up to here is entirely linear. In addition, we conducted additional simulations to decide whether it is necessary to consider any nonlinear constituent part. In other words, if the model is entirely linear, scaling the inputs should produced scaled outputs. We plot the input-output power curve in Fig. 4 to check whether such property holds. Moreover, to verify whether this nonlinear block is time-invariant, we rerun this simulation over different audio files from TIMIT dataset (in particular, the core test set of TIMIT material which contains 1920 sentences from 24 speakers) and plotted the corresponding empirical input-output curves for each file in one snapshot in Fig. 4. As illustrated in this Figure, these curves constitute a thin red band, from which it can be inferred that this block (i.e., the nonlinear scaling block) can be modeled as time-invariant and insensitive to the attributes of the input signal. This property has also been advocated for the structure of the AMR vocoder in [29].

As illustrated in Fig. 4, assuming the entire linear model is not tenable and we should consider a nonlinear block to model the full regime (including clipping and gain-saturation). To do so, we add a limiter function block to our hypothesized model as illustrated in Fig. 5. Passing $y_k = x_k + w_k$ through the limiter function denoted by $H(\cdot)$, we have $y''_k = H(y_k)$ at the output of the limiter. Recalling that the watermark magnitude must

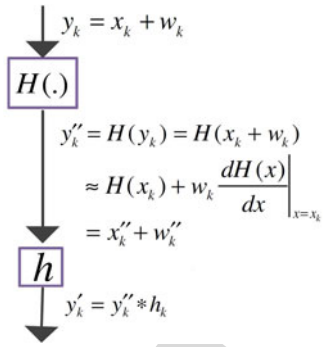


Fig. 5. Block-diagram of the complete model.

be small (i.e., $|w| \ll 1$) for perceptual reasons, we can deduce that $|w|^n \ll |w|$ for any $n > 1$. Thus, we approximate $H(\cdot)$ by applying a first-order Taylor expansion around $y_k = x_k$ as

$$H(y_k) = H(x_k + w_k) \approx H(x_k) + w_k \left.\frac{dH(x)}{dx}\right|_{x=x_k}. \tag{11}$$

Lets denote $H(x_k)$ and $w_k \left.\frac{dH(x)}{dx}\right|_{x=x_k}$ by x''_k and w''_k respectively. Considering the nonlinear part and according to Fig. 5, we can update (6) by substituting σ_x and σ_w by $\sigma_{x''}$ and $\sigma_{w''}$ respectively. The variance of x'' and w'' now must be calculated as

$$\sigma^2_{x''} = \int_{-\infty}^{\infty} H^2(x) p_x(x) dx \tag{12}$$

$$\sigma^2_{w''} = \sigma_w^2 \int_{-\infty}^{\infty} \left(\frac{d}{dx} H(x)\right)^2 p_x(x) dx. \tag{13}$$

It is noteworthy to say that, since $H(\cdot)$ is an odd function and $p_x(x)$ assumed to be an even function, the mean of x'' is zero. Moreover, the power of the output signal (i.e., $\mathrm{E}\{(y''_k)^2\}$) can be computed as

$$\begin{aligned}\mathrm{E}\{(y'')^2\} &= \int_{-\infty}^{\infty} H^2(y) p_y(y) dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H^2(x+w) p_x(x) p_w(w) dx dw. \end{aligned} \tag{14}$$

Let us denote the input-output power relationship by function $g(\cdot)$. Invoking (11) and doing some algebraic simplifications

$$\begin{aligned} g(\alpha) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H^2\left(\frac{(x+w)\sqrt{\alpha}}{\sqrt{\sigma_x^2+\sigma_w^2}}\right) p_x(x) p_w(w) dx dw \\ &\approx \int_{-\infty}^{\infty} H^2\left(\frac{x\sqrt{\alpha}}{\sqrt{\sigma_x^2+\sigma_w^2}}\right) p_x(x) dx \\ &\quad + \frac{\alpha\sigma_w^2}{\sigma_x^2+\sigma_w^2} \int_{-\infty}^{\infty} \left(\frac{d}{dx} H\left(\frac{x\sqrt{\alpha}}{\sqrt{\sigma_x^2+\sigma_w^2}}\right)\right)^2 p_x(x) dx. \end{aligned} \tag{15}$$

It should be noticed that in practice we encounter the inverse problem, that is, we know $g(\cdot)$ and we want to find $H(\cdot)$. To solve this inverse problem, since an explicit expression for $H(x)$

cannot be obtained, we approximate $H(x)$ by a function in the following set, parameterized by $x_{\max}$

$$L(x; x_{\max}, q) = \frac{x}{\left(1 + \left(\frac{|x|}{x_{\max}}\right)^{2q}\right)^{\frac{1}{2q}}}. \quad (16)$$

The family given by (16) corresponds to the limiter functions encompassed by Rapp's model [32]. To find the best match, we numerically solve the following optimization problem using true speech samples

$$x^*_{\max}, q^* = \arg \min_{x_{\max}, q} \int_{0.1}^{5} \|g(\alpha) - \hat{g}(\alpha)\| d\alpha \quad (17)$$

where $\hat{g}(\alpha)$ is defined in the same fashion as $g(\alpha)$, replacing $H(x)$ by $L(x; x_{\max}, q)$ in (15). The optimization in (17) yields $x^*_{\max} = 6.85$ and $q^* = 10$ which, as can be seen in Fig. 4, results in a good approximation of $H(x)$.

Our hypotheses for constructing the model in Fig. 5 include the assumption that if the input power signal were low enough, the linearity would hold and nonlinear effects would not be significant. Fig. 4 validates such assumption: the utilized input signal for modeling the linear part was small enough to avoid noticeable nonlinear effects. Thus, since our conclusions regarding the linear part have been drawn for such operating point, the proposed nonlinear block does not alter them. As a consequence, our global model that consists of a linear and a nonlinear part will be exploited in the subsequent analyses performed in this paper. Notice that with this model, even if x'_k followed a known distribution (e.g., a generalized Gaussian), y'_k would not, so we will focus instead on decoders that do not make assumptions on $p_x(x)$. This is pursued in the next section. Last, to give more insights on the comfortability of proposed model with vocoder structure let us say that, we have used the standard version of AMR 12.2 codec which is officially released by ETSI and written in ANSI-C in all simulations to ensure the conformance of our results with the AMR 12.2 codec utilized in cellular networks [33]. Recalling that the main purpose of our model is to capture the impairments that the codec causes on the watermark signal, it is worth pinpointing the roots of the model in the constituent blocks of the codec.

The nonlinear block in Section II, namely $H(\cdot)$, straightforwardly corresponds to the (A-law, μ-law) compander. As to the LTI filter $h[n]$, it can be explained by the lossy encoding/decoding of the LPC filter and by the low-pass filter that performs subframe interpolation and long-term synthesis. To elaborate, let $\hat{A}(z)$ denote the LPC synthesis filter obtained by quantizing the LSP (line spectral pairs) coefficients of the analysis filter $A(z)$, and let $G(z)$ denote the low-pass filter. Then, the input-output transfer function can be modeled by $L(z) \doteq \frac{A(z)G(z)}{\hat{A}(z)}$. Experiments conducted on real speech signals confirm that the impulse response $l[n]$ so obtained is remarkably close to $h[n]$ and despite the fact that both $A(z)$ and $\hat{A}(z)$ are time-varying, their ratio, and thus $l[n]$, are quite stationary, in accordance with our observations for $h[n]$.

Leveraging on the modeling methodology here presented, one of the important advantages is that it can be extended to other AMR codecs such as AMR 10.2, AMR-WB. Moreover, as our watermarking methods have been designed to achieve a large degree of robustness, they can be expected to perform well with other vocoders having a similar underlying structure. We have checked this for the AMR 10.2 and GSM FR codecs, with promising results.

III. Spread Spectrum Data Embedding

Spread Spectrum (SS) methods are arguably the most popular for data hiding. The SS scheme was first presented by Cox *et al.* [34] in 1997. The authors proposed a method by which the information could be embedded into the host signal with a shared key. There are both additive [35] and multiplicative [36] versions of SS. At the receiver side, the information is decoded and extracted by using the same key as in embedding.

A. Improved Additive Spread Spectrum

In the case of additive spread spectrum, we insert one data bit into one block of the host signal, i.e., N consecutive samples of the host signal. The samples of the watermarked signal $\mathbf{y}$ for each block are computed as

$$\mathbf{y} = \mathbf{x} + b\mathbf{w}. \quad (18)$$

Where the data bit $b \in \{-1, 1\}$ is modulated and added to N host coefficients $\mathbf{x}$. The watermark signal $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ is a key-dependent pseudorandom sequence. The imposed distortion to the host signal can be written as

$$D = \frac{1}{N}\mathrm{E}\{\|b\mathbf{w}\|^2\} = \sigma_w^2.$$

Having introduced the distortion parameter, the document to watermark ratio (DWR) is DWR $\doteq \sigma_{\mathbf{x}}^2/D$. Since the presented embedding procedure does not compensate the interference from the host contents, the resulting performance is generally not acceptable. In order to improve the performance and have a host-rejection approach at the receiver side which decreases the error probability, the improved additive spread spectrum (IASS) method can be used as follows [37]:

$$\mathbf{y} = \mathbf{x} + b\mathbf{w} - \gamma\mathbf{w}\mathbf{w}^T\mathbf{x} = \mathbf{x} + b\mathbf{w} + \mathbf{u} \quad (19)$$

where γ is set to $1/N\sigma_w^2$ to minimize the probability of error and $\mathbf{u}$ denotes the host-rejection term. It is worth noticing that the embedding distortion can be computed as

$$D = \mathrm{E}\{\|\mathbf{y} - \mathbf{x}\|^2\} = \sigma_w^2 + \frac{\sigma_x^2}{N}. \quad (20)$$

Next, according to the constructed model in Section II, the output y' is derived as

$$y'_k = h_k * H(y_k) \quad (21)$$

which clearly illustrates the linear and nonlinear operations on the watermarked signal y_k. In the following sections, we modify the mentioned watermarking structure to tackle the issues of both nonlinear scaling (i.e., the $H(\cdot)$ function) and linear filtering (i.e., convolution with h_k). For the sake of simplicity, we assume that $H(\cdot)$ operates pointwise on its input arguments

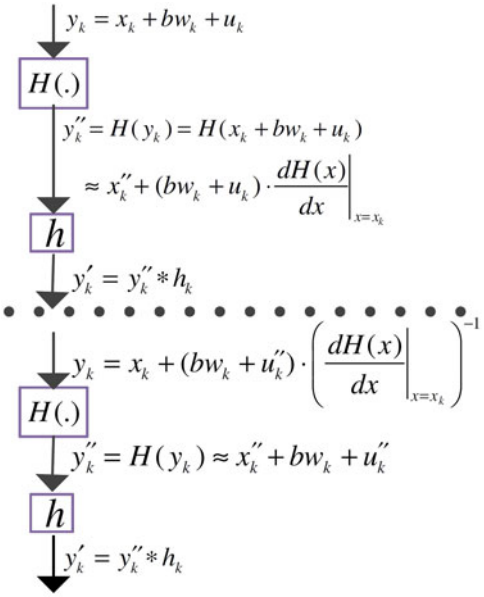


Fig. 6. Block-diagram of the complete model for IASS.

(whether vectors or scalars), e.g., for an arbitrary vector $\mathbf{a}$ with length N, $H(\mathbf{a}) = [H(a_1), \ldots, H(a_N)]^T$.

1) Nonlinear Scaling: In the case of IASS we have

$$H(y_k) = H(x_k + bw_k + u_k). \tag{22}$$

Recalling (11) and considering that for large N, we have $\sigma_u^2 = \frac{\sigma_x^2}{N} \ll \sigma_x^2$, the mentioned equation can be approximated as

$$H(y_k) \approx H(x_k) + (bw_k + u_k) \cdot \frac{dH(x)}{dx}\Big|_{x=x_k}. \tag{23}$$

To counterbalance the effect of nonlinear scaling, we first replace the host-rejection part by $\mathbf{u}'' \doteq -\gamma \mathbf{w}\mathbf{w}^T H(\mathbf{x})$ and then modulate both the watermark and new host-rejection terms by $(dH(x)/dx)^{-1}$. In other words, we reformulate the encoder for IASS as follows:

$$y_k = x_k + (bw_k + u_k'') \cdot \left(\frac{dH(x)}{dx}\Big|_{x=x_k}\right)^{-1}. \tag{24}$$

According to the parameters of Rapp's model (i.e., $x_{\max} = 6.85$), we can assume that almost all of the true speech samples are less than $x_{\max}$. So, since even in the extreme case of $x_k = x_{\max}$, the value of $(dH(x_k)/dx)^{-1}$ is less than two, it is still reasonable to hold the small signal assumption for this part (i.e., $(bw_k + u_k'') \cdot (dH(x_k)/dx)^{-1}$). Therefore, recalling (11), after passing the mentioned signal through the nonlinear scaling part of our model we have

$$H(y_k) \approx H(x_k) + bw_k + u_k''. \tag{25}$$

Consequently, as seen in Fig. 6 we can compensate the effect of nonlinear scaling by applying the proposed method and just considering $H(x_k)$ instead of x_k as the host signal. The cost of this compensation is the degradation of the error probability, i.e., since $H(x_k)$ is always smaller than x_k, then the watermark power in (25) must be smaller to guarantee the same DWR; this in turn produces an increase in the error probability. This performance degradation was already expected due to the nonlinear scaling of the vocoder. Finally, one might think of applying the inverse of $H(\cdot)$ to y_k to completely remove the nonlinearity. However, this would increase the dynamic range of the input signal to the voice channel, which would be unacceptable in practice. In addition, denoting $\mathrm{E}\{\frac{dH(x)}{dx}|_{x=x_k}\}$ by β, the embedding distortion in this new structure can be shown to be

$$D = \frac{\sigma_w^2 + \frac{\sigma_{x''}^2}{N}}{\beta^2}. \tag{26}$$

2) Linear Filtering: To mathematically discuss the effect of linear filtering, let us periodically repeat the watermark signal to make an infinite sequence, i.e., let us assume that $w_i = w_{i-mN}$ for all integer values of m. Now, considering (25), at the output of model we have

$$y_k' = \sum_{i=-\infty}^{\infty} h_i H(y_{k-i}) \approx \sum_{i=-\infty}^{\infty} h_i \left(x_{k-i}'' + bw_{k-i} + u_{k-i}''\right). \tag{27}$$

By applying the correlator decoder, i.e., the inner product of $\mathbf{y}'$ and $\mathbf{w}$ at the decisor, we have

$$z_{\mathrm{A}} = \mathbf{w}^T \mathbf{y}' = \sum_{k=1}^{N} \sum_{i=-\infty}^{\infty} w_k h_i \left(x_{k-i}'' + bw_{k-i} + u_{k-i}''\right) \tag{28}$$

where z_{A} indicates the test statistic. After some algebraic manipulations, we can compute the mean and variance of z_{A}, denoted by m_{A} and σ_{A}^2, respectively, as

$$\begin{aligned} m_{\mathrm{A}} &= Nbh_0\sigma_w^2 \\ \sigma_{\mathrm{A}}^2 &= N\sigma_w^2\sigma_{x''}^2 \sum_{i\neq 0}\sum_{m\neq 0} h_i h_m \rho_{i-m}'' \end{aligned} \tag{29}$$

where ρ_{i-m}'' indicates the normalized autocorrelation function of x'', i.e., $\rho_i'' = \mathrm{E}\{x_k'' x_{k+i}''\}/\sigma_{x''}^2$. Recalling the central limit theorem (CLT), for large N we can assume that test statistic z_A in (28) approximately follows a Gaussian distribution with mean m_{A} and variance σ_{A}^2. Assuming an equal prior probability for the information bit, i.e., $\Pr(b=+1) = \Pr(b=-1) = 1/2$, we can approximate the error probability as follows[1]:

$$\begin{aligned} \Pr(e) &= \Pr(e|b=1) = \Pr(e|b=-1) \\ &\approx Q\left(\frac{m_{\mathrm{A}}}{\sigma_{\mathrm{A}}}\right) = Q\left(\frac{\sqrt{N}h_0\sigma_w}{\sigma_{x''}\sqrt{\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m \rho_{i-m}''}}\right). \end{aligned} \tag{30}$$

considering (26) and defining $\kappa \doteq \beta^2 \frac{\sigma_x^2}{\sigma_{x''}^2}$, the approximate error probability can be rewritten as

$$\Pr(e) \approx Q\left(\frac{h_0\sqrt{\frac{\kappa N}{\mathrm{DWR}} - 1}}{\sqrt{\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m \rho_{i-m}''}}\right). \tag{31}$$

[1] $Q(x) = (2\pi)^{-1/2}\int_x^{\infty} \exp(-v^2/2)dv$.

One way to increase the data rate of the introduced structure is to use multi bi-orthogonal codes as spreading signals. To this end, we propose a new structure (embedding/decoding structure) which aims at achieving host rejection with multiple simultaneous watermark carriers. Let $\mathbf{w}^i, i \in \{1, \ldots, M\}$ with length N_m denote a set of orthogonal codes, M indicates the number of orthogonal watermark carriers and $\mathbf{w}^* \in \{\mathbf{w}^1, \ldots, \mathbf{w}^M\}$ represents the embedded symbol. Denoting the new host-rejection term by $\mathbf{r}$, i.e., $\mathbf{r} = (\sum_{i=1}^{M} \mathbf{w}^i \mathbf{w}^{i^T}) H(\mathbf{x})$, the embedder and decoder for the M-ary bi-orthogonal Additive (MA) structure is proposed as follows:

$$y_k = x_k + \left(b w_k^* - \frac{r_k}{N_m \sigma_w^2} \right) \cdot \left(\left. \frac{dH(x)}{dx} \right|_{x=x_k} \right)^{-1} \tag{32}$$

where the decision is made by the decoder as

$$\hat{d}_{\mathrm{MA}} = \hat{\jmath} \cdot \operatorname{sgn}\left(\mathbf{y}'^T \mathbf{w}^{\hat{\jmath}} \right) \tag{33}$$

with

$$\hat{\jmath} = \underset{j \in \{1 \ldots M\}}{\arg\max} \, |\mathbf{y}'^T \mathbf{w}^j|. \tag{34}$$

The embedded distortion for this structure is

$$D = \frac{1}{N} \mathrm{E} \left\{ \left\| b\mathbf{w}^* - \frac{(\sum_{i=1}^{M} \mathbf{w}^i \mathbf{w}^{i^T}) \mathbf{x}}{N_m \sigma_w^2} \right\|^2 \right\} = \frac{\sigma_w^2 + \frac{M \sigma_{x''}^2}{N_m}}{\beta^2}. \tag{35}$$

Since we now choose one among $2M$ spreading sequences, there are $\log_2 2M$ bits encoded in the decision, so we can increase the length of the spreading sequences by this amount for the same effective rate. Therefore $N_m = N \log_2 2M$. Carrying out some algebraic manipulations, the error probability for the mentioned structure can be approximated as [38]

$$\begin{aligned} \Pr(e) &= 1 - \Pr(c|b=1) = 1 - \Pr(\mathbf{w}^* = \mathbf{w}^{\hat{\jmath}} | b = 1) \\ &\overset{(33)}{=} \frac{2^{M-1}}{2^M - 1} \left[1 - \prod_{\substack{j=1 \\ j \neq \hat{\jmath}}}^{M} \Pr\left(\mathbf{y}'^T \mathbf{w}^* > |\mathbf{y}'^T \mathbf{w}^j| \right) \right] \\ \Pr(e) &\approx \frac{2^{M-1}}{2^M - 1} \left[1 - \frac{1}{\sqrt{2\pi\sigma_{\mathrm{A}}^2}} \right. \\ &\quad \left. \times \int_0^\infty \left(1 - 2Q\left(\frac{x}{\sigma_{\mathrm{A}}} \right) \right)^{M-1} e^{-\frac{(x - m_{\mathrm{A}})^2}{2\sigma_{\mathrm{A}}^2}} dx \right]. \end{aligned} \tag{36}$$

B. Improved Multiplicative Spread Spectrum Data Embedding

In the improved multiplicative spread spectrum (IMSS), the watermarked signal is generated as [30]

$$y_k = x_k + b x_k^2 w_k + u_k. \tag{37}$$

Now, similarly to Section III-A1, to take into account the effect of nonlinear scaling, we modify (37) to

$$y_k = x_k + \left(b H^2(x_k) w_k + u_k'' \right) \cdot \left(\left. \frac{dH(x)}{dx} \right|_{x=x_k} \right)^{-1}. \tag{38}$$

In this scheme, after some straightforward computations, D can be shown to be

$$D = \frac{\sigma_w^2 \mathrm{E}\{(x'')^4\} + \frac{\sigma_{x''}^2}{N}}{\beta^2}. \tag{39}$$

After passing y_k through the nonlinear scaling function of our model, we have

$$H(y_k) \approx x_k'' + b(x_k'')^2 w_k + u_k'' \tag{40}$$

recalling the linear part of our model and akin to (27), the output of linear block is

$$\begin{aligned} y_k' &= \sum_{i=-\infty}^{\infty} h_i H(y_{k-i}) \\ &\approx \sum_{i=-\infty}^{\infty} h_i \left(x_{k-i}'' + b(x_{k-i}'')^2 w_k + u_{k-i}'' \right). \end{aligned} \tag{41}$$

By applying the correlator decoder (i.e., the inner product of y', w as the test statistic) we have

$$z_{\mathrm{M}} = \mathbf{w}^T \mathbf{y}' \approx \sum_{k=1}^{N} \sum_{i=-\infty}^{\infty} h_i \left(x_{k-i}'' + b(x_{k-i}'')^2 w_k + u_{k-i}'' \right). \tag{42}$$

Next, to apply the CLT and compute the error probability, we need to find the values of the mean and variance of z_{M} denoted by m_{M}, σ_{M}^2 respectively. Noticing that the variables in the second sum of (42) are zero-mean and uncorrelated with the watermark, we can write

$$m_{\mathrm{M}} \approx N h_0 b \sigma_w^2 \sigma_{x''}^2 \tag{43}$$

whereas σ_{M}^2 can be computed as

$$\begin{aligned} \sigma_{\mathrm{M}}^2 &= \mathrm{E}\{z_{\mathrm{M}}^2\} - m_{\mathrm{M}}^2 \\ &\approx N \left(\sigma_w^2 \sigma_{x''}^2 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}'' + \sigma_w^4 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \varphi_{i-m} \right) \\ &\quad + N h_0^2 \mathrm{E}\{w_i^4\} \varphi_0 + h_0^2 \sigma_w^4 \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varphi_{i-j} - N^2 h_0^2 \sigma_w^4 \sigma_{x''}^4 \end{aligned} \tag{44}$$

in which $\varphi_i \doteq \mathrm{E}\{(x_{k-i}'')^2 (x_k'')^2\}$. Assuming that $\mathrm{E}\{w_i^4\} \ll \mathrm{E}\{w_i^2\}$, σ_{M}^2 can be approximated as

$$\sigma_{\mathrm{M}}^2 \approx N \sigma_w^2 \sigma_{x''}^2 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}''. \tag{45}$$

Consequently, similarly to the additive case, the error probability is

$$\Pr(e) \approx Q\left(\frac{m_{\mathrm{M}}}{\sigma_{\mathrm{M}}} \right) = Q\left(\frac{\sqrt{N} h_0 \sigma_{x''} \sigma_w}{\sqrt{\sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}''}} \right). \tag{46}$$

Denoting $\frac{\sqrt{\mathrm{E}\{(x'')^4\}}}{\sigma^2_{x''}}$ by η and considering (39), the error probability can be reformulated in terms of the DWR as

$$\Pr(e) \approx Q\left(\frac{h_0\sqrt{\frac{\kappa N}{\mathrm{DWR}}-1}}{\eta\sqrt{\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m \rho''_{i-m}}}\right). \tag{47}$$

Moreover, inspired by the M-ary bi-orthogonal watermarking in Section III-A, a M-ary Multiplicative (MM) structure can be proposed in the case of Improved multiplicative SS as follows:

$$y_k = x_k + \left(b(x''_k)^2 w^*_k - \frac{r_k}{N_m\sigma^2_w}\right)\cdot\left(\frac{dH(x)}{dx}\Big|_{x=x_k}\right)^{-1} \tag{48}$$

in which $w^*_k, w^j_k r_k$ are defined in the same fashion as in Section III-A and, consequently, the decoder makes the decision as

$$\hat{d}_{\mathrm{MM}} = \hat{\jmath}\cdot\mathrm{sgn}\left(\mathbf{y}'^T\mathbf{w}^{\hat{\jmath}}\right) \tag{49}$$

where

$$\hat{\jmath} = \underset{j\in\{1\ldots M\}}{\arg\max}\,|\mathbf{y}'^T\mathbf{w}^j|. \tag{50}$$

Furthermore, the embedding distortion can be written as

$$D = \frac{\sigma^2_w\mathrm{E}\{(x'')^4\} + \frac{M\sigma^2_{x''}}{N_m}}{\beta^2}. \tag{51}$$

Let σ^2_{MM} denote the variance of the interference that results after multiplying by other spreading signal (i.e., $\mathrm{E}\left\{\mathbf{y}^T\mathbf{w}^j \,\middle|\, \mathbf{w}^j\neq\mathbf{w}^*\right\}$). Next, recalling (36) and after some algebraic manipulations the error probability becomes

$$\begin{aligned}\Pr(e) &= 1-\Pr(c|b=1) = 1-\Pr(\mathbf{w}^*=\mathbf{w}^{\hat{\jmath}}|b=1)\\ &\overset{(49)}{=} \frac{2^{M-1}}{2^M-1}\left[1-\prod_{\substack{j=1\\ j\neq\hat{\jmath}}}^{M}\Pr\left(\mathbf{y}'^T\mathbf{w}^* > \left|\mathbf{y}'^T\mathbf{w}^j\right|\right)\right]\\ &\approx \frac{2^{M-1}}{2^M-1}\left[1-\frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{M}}}}\right.\\ &\quad\left.\times\int_0^\infty\left(1-2Q\left(\frac{x}{\sigma_{\mathrm{MM}}}\right)\right)^{M-1} e^{-\frac{(x-m_{\mathrm{M}})^2}{2\sigma^2_{\mathrm{M}}}}dx\right]\end{aligned} \tag{52}$$

where σ^2_{MM} can be computed as

$$\begin{aligned}\sigma^2_{\mathrm{MM}} &= N\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m\left(\sigma^2_w\sigma^2_{x''}\rho''_{i-m} + \sigma^4_w\varphi_{i-m}\right)\\ &\quad + Nh^2_0\sigma^4_w\varphi_0 \approx N\sigma^2_w\sigma^2_{x''}\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m\rho''_{i-m}.\end{aligned} \tag{53}$$

IV. Simulations and Results

In this section, we validate our analysis (in particular; error probability formulas) with several experiments. The good conformance between experimental results and theory serves as an additional supporting validation for the vocoder modeling from Section II. Afterwards, we assess the imperceptibility of the proposed methods according to subjective and objective benchmarks.

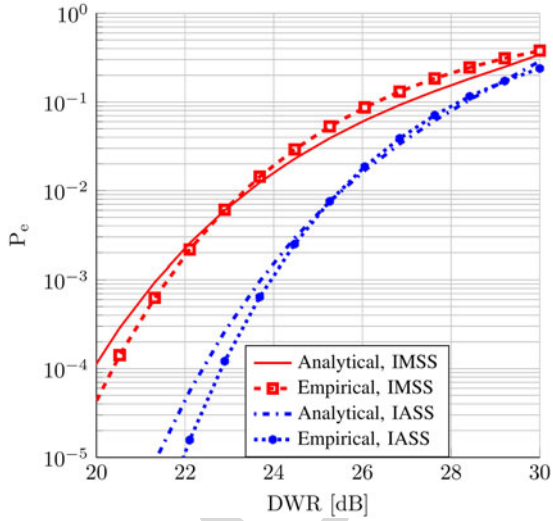


Fig. 7. Empirical and analytical results over true speech samples for $N = 2500$, AMR12.2.

A. Performance Analysis

According to (6), the analytical error probability is highly dependent on the autocorrelation of the input signals. We consider two different scenarios to measure the goodness of our proposed model. In the first one, we perform the simulations over true speech samples and consider the average autocorrelation of the input signal in our analytical formula. As shown in Fig. 7, the empirical results are close to the analytical ones but do not accurately match.

In the second scenario, we synthesize a signal with i.i.d Generalized Gaussian Distribution (GGD) samples [39], [40]. This ensures a time-invariant autocorrelation. As illustrated in Fig. 8, now the analytical and empirical results match more closely. We can conclude that for stationary signals the analytical prediction of the error probability is quite accurate, and that some accuracy is lost for real speech signals, as stationarity does not hold in this case. It is worth noticing that all the human voice samples used in our simulations have been selected from the TIMIT database [41].

It is worthy to note that the most common vocoders in cellular networks are GSM FR, GSM HR in 2.5G and 2.75G, and AMR in 2.5G, 2.75G, 3G. To the best of our knowledge, the most prevalent one among the mentioned vocoders is AMR 12.2 [42], [43]. So much so, we have focused on AMR 12.2 rather than other compression rates. Additionally, as illustrated in Fig. 9, we have conducted another simulation for AMR 10.2, AMR 7.95, to show that the insights given by the proposed model can be applied to other compression rates of AMR as well.

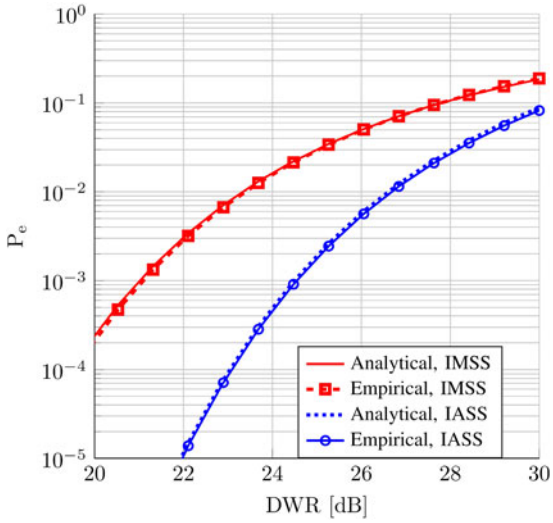


Fig. 8. Empirical and analytical results over synthetic signals for $N = 2500$, AMR12.2.

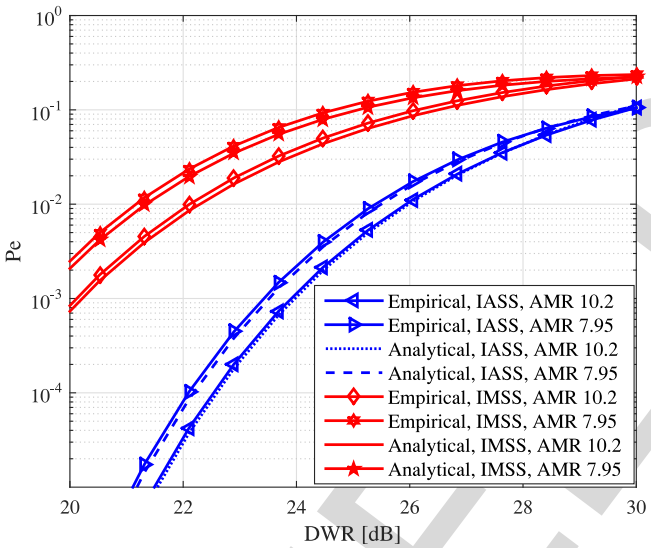


Fig. 9. Empirical and analytical results over synthetic signals for AMR 10.2 and AMR 7.95.

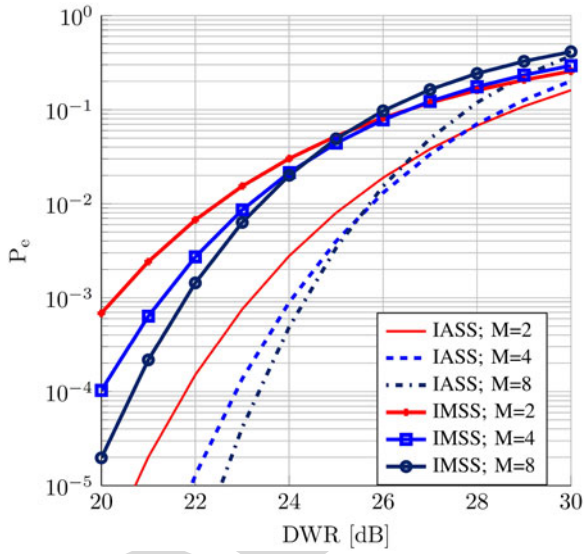


Fig. 10. Empirical results over synthetic signals for $N = 2500$, AMR12.2.

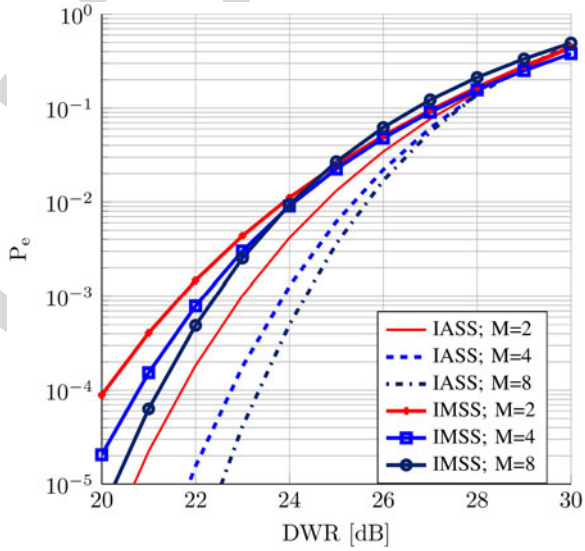


Fig. 11. Empirical results over true speech samples for $N = 2500$, AMR12.2.

In addition, we conducted simulations for the case of M-ary symbols. As illustrated in Figs. 10 and 11, the proposed M-ary structures outperform the results shown in Figs. 7 and 8, especially for low DWRs. It should be noted that for the sake of clarity we do not plot the analytical results in Figs. 10 and 11, but the match is similar to that observed in Figs. 7 and 8.

We also compared the proposed methods with the baseline additive and multiplicative SS schemes [35] [36], and with the scheme proposed by Cheng known as Generalized Embedding of Multiplicative (GEM) watermarking [44]. We plot the results in Figs. 12 and 13 . As illustrated in these figures, the proposed methods outperform the mentioned prior art.

The results shown in Figs. 7 and 11 have been obtained by running the simulations over samples of audio files from the TIMIT dataset [41], and plotting the average results. Experiments were conducted over the core test set of the TIMIT material which contains 1920 sentences from 24 speakers. Moreover, all simulations in this section have been conducted with the following setup:

1) Watermark frame length: 2500, i.e., $N = 2500$.
2) Codec: AMR 12.2.
3) Synthetic signals follow a GGD distribution with zero mean and unit variance, with a shape parameter of 1.5 which is matched with human voice properties [39].
4) The total length of human speech utilized in the simulations was around 737 million samples which corresponds to around 300,000 watermark frames.

Moreover, our proposed methods are based on spread-spectrum. Although we have customized and particularized them for our specific problem, they still inherit the main properties of spread-spectrum, discussed at length in [24]. This

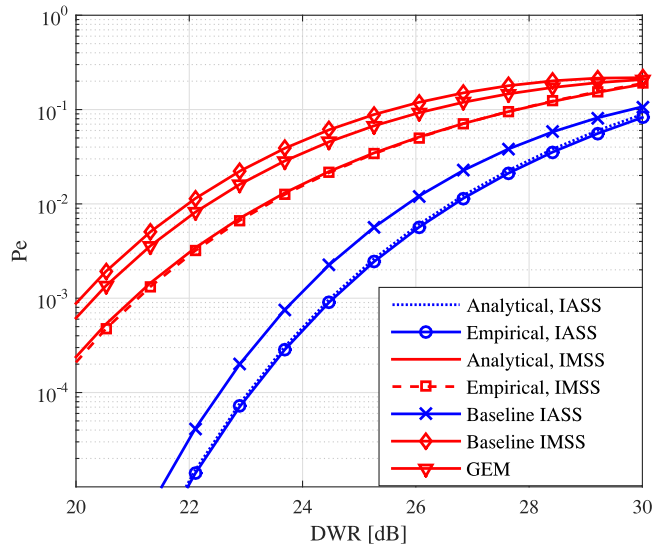


Fig. 12. Empirical and analytical results over synthetic signals based on the proposed methods, baseline spread spectrum methods, and GEM.

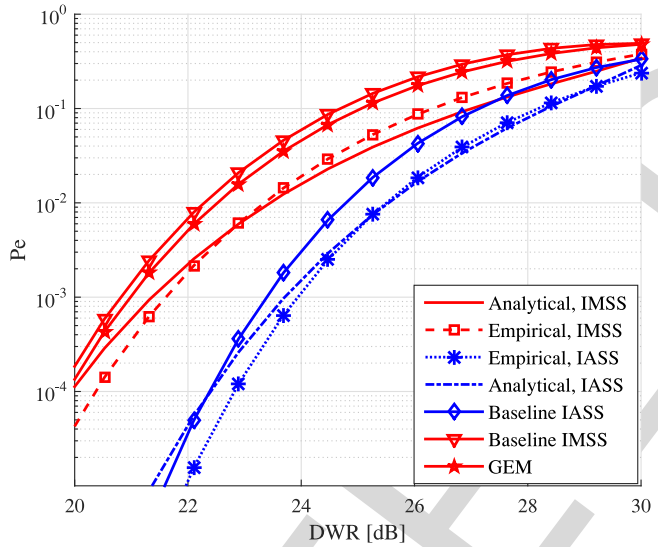


Fig. 13. Empirical and analytical results over true speech samples based on the proposed methods, baseline spread spectrum methods, and GEM.

means that they can be made robust against desynchronization attacks, spectrum filtering, chess watermarking, cut-sampling, zero-padding, resampling, noise addition, sample clipping, etc.

B. Imperceptibility Analysis

In this section we evaluate the proposed structures using objective and subjective benchmarks to assess their imperceptibility. ITU P.862 and ITU P.860 recommendations are two well-known standards which are widely used to evaluate the subjective quality of speech and the imperceptibility of embedded watermarks. In these tests, we considered $N = 2500$ and a sampling rate of 8 KHz, so the embedding bit rate is $8000/2500 = 3.2$ bps. Moreover, we considered $M = 8$ (i.e., 8-ary symbols), $\Pr(e) = 0.001$, and in order to achieve this target $\Pr(e)$, we set the operating DWR at 22.5 dB and 24.5 dB for IMSS and IASS, respectively. To check the imperceptibility of the proposed methods in accordance with the ITU P.860 standard, we prepared four sound file groups, each consisting of five audio files with a length of 10 seconds. We asked five persons to execute the $A/B/X$ test [17]. In the $A/B/X$ test, A indicates the watermarked signal, B stands for plain signal and X is assigned randomly to be A or B. In each stage of this experiment, whenever we played the X file for each listener, we asked him to decide between A and B. We summarized the results in Table I, where $\mathbf{M1}, \mathbf{M2}$ indicate 8-ary IMSS and 8-ary IASS, respectively. It should be noted that in the case of ideal watermarking in this sort of test, the expected percentage of failure would be 50%.

TABLE I
TEST RESULTS FOR ITU P.860 (PERCENTAGE OF FAILURE)

	Group I		Group II		Group III		Group IV	
	M1	M2	M1	M2	M1	M2	M1	M2
Tester 1	45%	55%	45%	35%	50%	55%	60%	55%
Tester 2	55%	50%	40%	40%	55%	45%	55%	50%
Tester 3	55%	35%	40%	45%	60%	55%	60%	45%
Tester 4	45%	45%	50%	50%	50%	60%	50%	55%
Tester 5	60%	55%	45%	40%	55%	55%	55%	55%
Average	52%	48%	44%	42%	54%	54%	56%	52%

TABLE II
TEST RESULTS FOR ITU P.862

	File Num	MOSLQO		Average	
		M1	M2	M1	M2
Group I	F1.1	4.320	4.275	**4.3352**	**4.2537**
	F1.2	4.311	4.206		
	F1.3	4.373	4.241		
	F1.4	4.377	4.268		
	F1.5	4.295	4.279		
Group II	F2.1	4.328	4.304	**4.3474**	**4.2274**
	F2.2	4.321	4.246		
	F2.3	4.353	4.189		
	F2.4	4.363	4.191		
	F2.5	4.371	4.206		
Group III	F3.1	4.294	4.287	**4.3167**	**4.2563**
	F3.2	4.359	4.205		
	F3.3	4.355	4.284		
	F3.4	4.276	4.204		
	F3.5	4.300	4.269		
Group IV	F4.1	4.330	4.219	**4.3334**	**4.2228**
	F4.2	4.403	4.197		
	F4.3	4.304	4.205		
	F4.4	4.344	4.256		
	F4.5	4.286	4.236		

ITU P.862 recommendation describes an objective method for predicting the subjective quality of narrowband speech codecs. In this method, perceptual evaluation speech quality (PESQ) and mean opinion score listening quality objective (MOSLQO) are used to compare the proposed schemes. The results are shown in Table II. It should be noticed that the PESQ value of the host

TABLE III
TEST RESULTS OF COMPARING THE COMPUTATIONAL COMPLEXITY OF VOICE ENCODING/DECODING AND THE PROPOSED DATA HIDING ENCODING/DECODING IN TERMS OF MIPS

Methods	Required MIPS
	($N=2500$)
Voice Enc/Dec based on AMR 12.2	225-300
Enc/Dec based on IASS	20-40
Enc/Dec based on IMSS	20-40
Enc/Dec based on MA (M=4)	30-50
Enc/Dec based on MM (M=4)	30-50

signal before embedding the data was equal to 4.5. As expected and shown in Tables I and II, although both methods exhibit an acceptable level of imperceptibility, the multiplicative structure behaves better than the additive one. In this work and according to its main application (i.e., hiding data in regular voice calls between two persons), we aimed at proposing a watermarking method which would be imperceptible to the human hearing system, as opposed to undetectability by steganalytic methods such as [45], [46]. Therefore, since our primary application is data hiding and not steganography, we have put emphasis on transparency as measured by ITU P.860 and ITU P.860 which are subjective tests of imperceptibility.

C. Complexity Analysis

To assess the complexity of the proposed schemes, we measured the required million instructions per second (MIPS) for processing one time frame of data hiding (in particular, in the case $N = 2500$) which encompasses both encoding and decoding processes. We report the achieved results in Table III together with the required MIPS for voice encoding/decoding (in particular, considering AMR 12.2 as the decoder/encoder) the same time frame. As shown in Table III, the complexity of the proposed schemes is a fraction of that of the vocoder in cellular phones. We must remark that our implementation has not been fully optimized, and even more so, no adaptation to the specific architecture of cellular phone CPUs has been intended. Considering this fact, we believe that a further reduction in the results reported in Table III is feasible.

V. CONCLUSION

In this paper we have shown that for watermarking purposes a mobile communication vocoder can be accurately modeled by considering a non-linear scaling function plus a linear filter. Adhering to such model, we have proposed two Spread-Spectrum-based data hiding methods, termed IASS and IMSS. Moreover, their performance can be largely improved by considering multi bi-orthogonal codes as spreading signals. Finally, the experimental assessment using both subjective and objective measures has revealed that our proposed schemes exhibit an acceptable level of imperceptibility.

Although we have concentrated on detectors which do not rely on the probability distribution of received signals, the approximation of such distribution and the derivation of the corresponding maximum likelihood detectors are topics for further research. In addition, in order to increase the technological readiness of the proposed methods for practical implementation, a technique to assure integrity of the hidden messages in the case of packet losses, and a method for synchronization considering the limited bandwidth constraints and nonlinearities deserve further attention.

REFERENCES

[1] "List of countries by number of mobile phones in use," 2014. [Online]. Available: http://en.wikipedia.org/wiki/List of countries by number of mobile phones in use

[2] D. Strobel, *IMSI Catcher: Chair for Communication Security*., Bochum, Germany: Ruhr-Universität Bochum, 2007, p. 14.

[3] M. Toorani and A. Beheshti, "Solutions to the GSM security weaknesses," in *Proc. 2008 2nd Int. Conf. Next Generation Mobile Appl., Services Technol.*, 2008, pp. 576–581.

[4] L. Buttyan, C. Gbaguidi, S. Staamann, and U. Wilhelm, "Extensions to an authentication technique proposed for the global mobility network," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 373–376, Mar. 2000.

[5] M. Boloursaz, R. Kazemi, B. Barazandeh, and F. Behnia, "Bounds on compressed voice channel capacity," in *Proc. 2014 Iran Workshop Commun. Inform. Theory*, 2014, pp. 1–6.

[6] M. Boloursaz, R. Kazemi, D. Nashtaali, M. Nasiri, and F. Behnia, "Secure data over GSM based on algebraic codebooks," in *Proc. 2013 East-West Des. Test Symp.*, 2013, pp. 1–4.

[7] W. Mazurczyk and Z. Kotulski, "Adaptive VoIP with audio watermarking for improved call quality and security," *J. Inf. Assurance Security*, vol. 2, no. 3, pp. 226–234, 2007.

[8] J. Singh, P. Garg, and A. Nath De, "A combined watermarking and encryption algorithm for secure VoIP," *Inf. Security J., Global Perspective*, vol. 18, no. 2, pp. 99–105, 2009.

[9] N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Trans., Inform. Theory*, vol. 54, no. 1, pp. 255–274, Jan. 2008.

[10] F. Pérez-González, F. Balado, and J. R. H. Martin, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 960–980, Apr. 2003.

[11] Y. Diqun, W. Rangding, and Z. Liguang, "Quantization step parity-based steganography for mp3 audio," *Fundamenta Informaticae*, vol. 97, no. 1, pp. 1–14, 2009.

[12] A. Naofumi, "A technique of lossless steganography for G.711 telephony speech," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, 2008, pp. 608–611.

[13] C. Wang and Q. Wu, "Information hiding in real-time VoIP streams," in *Proc. 9th IEEE Int. Symp. Multimedia*, Dec. 2007, pp. 255–262.

[14] J. Dittmann, D. Hesse, and R. Hillert, "Steganography and steganalysis in voice-over IP scenarios: Operational aspects and first experiences with a new steganalysis tool set," in *Proc. Electron. Imag. 2005*, 2005, pp. 607–618.

[15] Y. F. Huang, S. Tang, and J. Yuan, "Steganography in inactive frames of voip streams encoded by source codec," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 296–306, Jun. 2011.

[16] Y.-m. Su, Y.-f. Huang, and X. Li, "Steganography-oriented noisy resistance model of G.729a," in *Proc. IMACS Multiconf. Comput. Eng. Syst. Appl.*, 2006, vol. 1, pp. 11–15.

[17] Y. Huang, C. Liu, S. Tang, and S. Bai, "Steganography integration into a low-bit rate speech codec," *IEEE Trans. Inform. Forensics Security*, vol. 7, no. 6, pp. 1865–1875, Dec. 2012.

[18] Z. Piotrowski, J. Wojtuń, and J. Ośka, "Hardware watermark token for voip telephony," *Przeglad Elektrotechniczny*, vol. 89, pp. 196–198, 2013.

[19] F. C. Er and E. Gul, "Comparison of digital audio watermarking techniques for the security of voip communications," in *Proc. 2011 7th Int. Conf. Inform. Assurance Security*, 2011, pp. 13–18.

[20] B. Xiao, Y. Huang, and S. Tang, "An approach to information hiding in low bit-rate speech stream," in *Proc. IEEE GLOBECOM Global Telecommun. Conf.*, Nov.–Dec. 2008, pp. 1–5.

[21] Z. Wu and W. Yang, "G. 711-based adaptive speech information hiding approach," in *Proc. Intell. Comput.*, 2006, pp. 1139–1144.

[22] W. Mazurczyk, "VoIP steganography and its detectiona survey," *ACM Comput. Surveys*, vol. 46, no. 2, 2013, Art. no. 20.
[23] C. K. LaDue, V. V. Sapozhnykov, and K. S. Fienberg, "A data modem for GSM voice channel," *IEEE Trans., Veh. Technol.*, vol. 57, no. 4, pp. 2205–2218, Jul. 2008.
[24] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1020–1033, Apr. 2003.
[25] A. Nadeau, and G. Sharma, "Insertion, deletion robust audio watermarking: a set theoretic, dynamic programming approach," in *Proc. IS&T/SPIE Electron. Imag.*, 2013, pp. 866–503–866–503.
[26] C. Baras, N. Moreau, and P. Dymarski, "Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1772–1782, Sep. 2006.
[27] L. de CT Gomes, E. Gómez, and N. Moreau, "Resynchronization methods for audio watermarking," presented at the 111th AES Conv., New York, NY, USA, 2001.
[28] L. Pérez-Freire and F. Pérez-González, "Spread-spectrum watermarking security," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 2–24, Mar. 2009.
[29] "Digital cellular telecommunications system (phase 2+); adaptive multirate (amr) speech transcoding (GSM 06.90 version 7.2.1)," ETSI, Tech. Rep. ETSI EN 301 704, 1998.
[30] A. Valizadeh, and Z. J. Wang, "An improved multiplicative spread spectrum embedding scheme for data hiding," *IEEE Trans. Inform. Forensics Security*, vol. 7, no. 4, pp. 1127–1143, Aug. 2012.
[31] M. Boloursaz, A. Hadavi, R. Kazemi, and F. Behnia, "Secure data communication through GSM adaptive multi rate voice channel," in *Proc. 2012 6th Int. Symp., Telecommun.*, 2012, pp. 1021–1026.
[32] A. A. Eltholth, A. R. Mekhail, A. Elshirbini, M. Dessouki, and A. Abdelfattah, "Modeling the effect of clipping and power amplifier non-linearities on OFDM systems," *Ubiquitous Comput. Commun. J.*, vol. 3, no. 1, pp. 54–59, 2009.
[33] "Universal mobile telecommunication systaem (UMTS); AMR speech codec general description (3gpp ts 26.071 version 5.0.0 released 5)," ETSI, Tech. Rep. ETSI EN 126 071, 2002.
[34] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.
[35] Q. Cheng and T. S. Huang, "An additive approach to transform-domain information hiding and optimum detection structure," *IEEE Trans. Multimedia*, vol. 3, no. 3, pp. 273–284, Sep. 2001.
[36] A. Valizadeh and J. Wang, "A framework of multiplicative spread spectrum embedding for data hiding: Performance, decoder and signature design," in *Proc. GLOBECOM IEEE Global Telecommun. Conf.*, Nov.–Dec. 2009, pp. 1–6.
[37] A. Valizadeh and Z. J. Wang, "Efficient blind decoders for additive spread spectrum embedding based data hiding," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–21, 2012.
[38] J. Proakis, *Digital Communications* (ser. Communications and signal processing). New York, NY, USA: McGraw-Hill, 1995. [Online]. Available: http://books.google.com/books?id=cIqYQgAACAAJ
[39] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett*, vol. 10, no. 7, pp. 204–207, Jul. 2003.
[40] K. Kokkinakis and A. K. Nandi, "Speech modelling based on generalized Gaussian probability density functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 1, pp. 381–384.
[41] J. S. Garofolo *et al.*, *TIMIT: Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[42] "The world's premier supplier of speech and audio codec-voiceage," 2015. [Online]. Available: http://www.voiceage.com/AMR-NB.AMR.html
[43] "Mobile telecommunications-radio-electronics," 2015. [Online]. Available: http://www.radio-electronics.com/info/cellulartelecomms.php
[44] Q. Cheng, "Generalized embedding of multiplicative watermarks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 978–988, Jul. 2009.
[45] S. T. Yong feng Huang, and Y. J. Y. Chunlai Bao, "Steganalysis of compressed speech to detect covert VoIP channels," *IET Inf. Security*, vol. 5, no. 1, pp. 1–7, Mar. 2011.
[46] Y. Huang, S. Tang, and Y. Zhang, "Detection of covert voice-over internet protocol communications using sliding window-based steganalysis," *Commun. IET*, vol. 5, no. 7, pp. 929–936, 2011.

Reza Kazemi was born in Ilam, Iran, in 1986. He received the B.S., M.S., and Ph.D. degrees in communication systems from the Sharif University of Technology, Tehran, Iran, in 2008, 2010, and 2015, respectively.

His research interests include watermarking, steganography, information forensics, DoV, and M2M communication.

Fernando Pérez-González (S'98–M'94–SM'09–F'16) received the Telecommunication Engineer degree from the University of Santiago, Santiago, Spain in 1990, and the Ph.D. degree in telecommunications engineering from the University of Vigo, Vigo, Spain, in 1993.

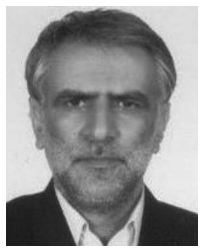
In 1990, he became an Assistant Professor with the School of Telecommunication Engineering, University of Vigo. From 2007 to 2010, he was Program Manager of the Spanish National R&D Plan on Electronic and Communication Technologies, Ministry of Science and Innovation. From 2009 to 2011, he was the Prince of Asturias Endowed Chair of Information Science and Technology, University of New Mexico, Albuquerque, NM, USA. From 2007 to 2014, he was the Executive Director of the Galician Research and Development Center in Advanced Telecommunications. He has been the Principal Investigator of the University of Vigo Group, which participated in several European projects, including CERTIMARK, ECRYPT, REWIND, NIFTY, and WITDOM. He is currently a Professor in the School of Telecommunication Engineering, University of Vigo, Vigo, Spain, and a Research Professor in Information Science and Technology, University of New Mexico, Albuquerque, NM, USA. He has coauthored more than 50 papers in leading international journals and 160 peer-reviewed conference papers. He has coauthored several international patents related to watermarking for video surveillance, integrity protection of printed documents, fingerprinting of audio signals, and digital terrestrial broadcasting systems. His research interests include the areas of digital communications, adaptive algorithms, privacy enhancing technologies, and information forensics and security.

Prof. Pérez-González was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (2005–2009) and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (2006–2010). He is currently is an Associate Editor of the LNCS *Transactions on Data Hiding and Multimedia Security*, and the EURASIP *International Journal on Information Forensics and Security*.

Mohammad Ali Akhaee (S'07–M'07) received the B.Sc. degree in electronics and communications engineering from the Amirkabir University of Technology, Tehran, Iran, and the M.Sc. and Ph.D. degrees from the Sharif University of Technology, Tehran, Iran, in 2005 and 2009, respectively.

He is currently an Assistant Professor with the College of Engineering and the Director of the Secure Communication Laboratory, University of Tehran, Tehran, Iran. He has authored or coauthored more than 50 papers, and holds one Iranian patent. His research interests include the area of signal processing, in particular multimedia security, watermarking, and statistical signal processing.

Prof. Akhaee was the Technical Program Chair of EUSIPCO '11 and the Executive Chair of ISCISC '14. He received the Governmental Endeavour Research Fellowship from Australia in 2010.



Fereydoon Behnia was born in Tarom, Iran, in 1958. He received the B.Sc., M.Sc., and Ph.D. degrees from the Sharif University of Technology, Tehran, Iran, in 1985, 1987, and 1997, respectively.

Since 1988, he has been with the Electrical Engineering Department, Sharif University of Technology.

Data Hiding Robust to Mobile Communication Vocoders

Reza Kazemi, Fernando Pérez-González, *Fellow, IEEE*, Mohammad Ali Akhaee, and Fereydoon Behnia

***Abstract*—The swift growth of cellular mobile networks in recent years has made voice channels almost accessible everywhere. Besides, data hiding has recently attracted significant attention due to its ability to imperceptibly embed side information that can be used for signal enhancement, security improvement, and two-way authentication purposes. In this regard, we aim at proposing efficient schemes for hiding data in the widespread voice channel of cellular networks. To this aim, our first contribution is to model the channel accurately by considering a linear filter plus a nonlinear scaling function. This model is validated through experiments with true speech signals. Then we leverage on this model to propose two additive and multiplicative data hiding methods based on the spread spectrum techniques. In addition, inspired by the concept of M-ary biorthogonal codes, we develop novel schemes that significantly outperform the previous ones. The performance of all the methods that we present is assessed mathematically and cross-validated with simulations. These are later extended to true speech signals where the results evidence an excellent performance as predicted by the theory. Finally, we assess the imperceptibility by means of both subjective and objective benchmarks and show that the perceptual impact of our watermarks is acceptable.**

***Index Terms*—Data hiding, mobile communication vocoder, nonlinearity, spread spectrum, watermarking.**

I. Introduction

CELLULAR networks have become major means of voice communications around the world. This is to the extent that the number of mobile network users in 2014 was estimated to be 6.8 billion, showing a penetration rate of 97% among the world's population [1]. In spite of such spread, source and destination authentication is still an issue in cellular networks. There exist a number of ways such as IMSI-Catcher or VoIP termination to spoof a personal number and utilize it for malicious purposes [2]. Although some progress and solutions have been proposed to overcome this vulnerability, to the best of our knowledge, most of them rely on one of the following prerequisites.

The first category of solutions imposes modifications to the cellular network protocol [3], [4] which of course are hard to implement in already deployed cellular networks. The second category needs data transferring through mobile voice channels which is prohibitive and still not practical [5], [6]. With the previous limitations in mind, we aim at developing a data hiding method which can easily be implemented on common existing smart cellular phones and seamlessly embeds the appropriate data (e.g., authentication data) in the speech of the user while he/she is regularly talking with another person on his/her phone. Furthermore, data hiding in mobile voice channels can be exploited for improving the quality of service [7], and for enhancing the security of communications by combining watermarking and encryption algorithms [8].

Information hiding applications impose different constraints that lead to substantially different methods. In steganography, the mere existence of a hidden message is to remain unknown to the adversary, while in data hiding this requirement can be sacrificed for a larger robustness against intentional attacks and common processing operations. When the information to be hidden is simply the presence or absence of a certain secret pattern, the term one-bit data hiding is used and is sometimes equated with watermarking. Since even in the multiple-bit case data is hidden by embedding a low-power signal called *watermark*, we will use the terms data hiding and watermarking interchangeably throughout this paper. One-bit watermarking is usually utilized in integrity verification applications such as copyright protection [9]. We are instead interested in multiple-bit data hiding that must survive substantial channel distortions, so that it can be reliably decoded at the receiver side [10]. Several methods have been proposed for data embedding in audio and human voice signals using data hiding techniques. These methods can be considered as one of the following sorts. In the first one, the information is embedded in the audio files in offline applications such as copyright protection of audio files [11], while the second one concentrates on online applications such as data insertion in Voice over IP (VoIP) streams [12]. Wang *et al.* proposed a method to embed data in a G.711 vocoder by hiding the information in the LSBs of the speech signal [13]. Ditmann *et al.* proposed a general scheme for data embedding in all VoIP streams by focusing on the active frame of the speech signal [14].

Huang *et al.* introduced a new method for data hiding in G.723.1 VoIP frames. They insert the secret message into the LSB of inactive speech VoIP frames [15]. They also suggested another solution to embed data in G.729A VoIP frames based on m-sequences [16]. Huang *et al.* also proposed a novel embedding method into a low-bit rate codec. Therein, data

Manuscript received September 12, 2015; revised June 5, 2016; accepted August 5, 2016. This work was supported in part by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under Project TACTICA and Project COMPASS (TEC2013-47020-C2-1-R), and in part by the Galician Regional Government and ERDF under "Project Consolidation of Research Units" (GRC2013/009), Project REdTEIC (R2014/037), and Project AtlantTIC. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiao-Ping Zhang.

R. Kazemi and F. Behnia are with the Department of Electronic and Electrical Engineering, Sharif University of Technology, Tehran 1136511155, Iran (e-mail: rezakazemi.reza@gmail.com; behnia@sharif.edu).

F. Pérez-González is with the Signal Theory and Communications Department, EE Telecomunicacion, University of Vigo-SPAIN, Vigo 36310, Spain (e-mail: fperez@gts.uvigo.es).

M. A. Akhaee is with the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran 1417466191, Iran (e-mail: Akhaee@ut.ac.ir).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2599149

insertion is performed in the process of pitch period prediction for a G.723.1 VoIP codec [17]. Piotrowski and Gul presented methods for watermarking in VoIP applications[18], [19]. Xiao *et al.* proposed a structure for low rate vocoders based on Quantization Index Modulation (QIM) [20]. Finally, Wu *et al.* introduced another method for data hiding over high rate vocoders such as G.711 [21].

The aforementioned related art is generally focused on steganography and not watermarking. Moreover, the targeted vocoders are the common ones used in VoIP standards. As a consequence, focus is put on implementation issues of IP networks such as packet losses, jitter, etc. in order to design mechanisms to mitigate these kind of impairments. For instance, some of those works modify the standard vocoder structure and design a new one with data hiding capabilities which is robust to the targeted distortions.

Leveraging on the mentioned related art, the main contributions of this work can be summarized as follows:

1) Model the mobile communication vocoders with a combination of linear and nonlinear blocks and validate the constructed model.
2) Extend the baseline SS methods to make them robust to the constructed model, in particular, to the nonlinear scaling blocks within, even when multiple symbol constellations are used.
3) Enhance the derived methods by considering biorthogonal codes as spreading signals and carry out an accurate performance analysis of the proposed methods.

Moreover, the other novelty in this paper is related to its application. While most of the previous research in this domain has concentrated on steganography and watermarking over VoIP channels, and, consequently, on the group of vocoders which are relevant there, we aim at proposing a method for watermarking through mobile voice channels, which enjoy a tremendous penetration. The main difference between our work and the related art lies in the target vocoder, as VoIP customarily employs waveform vocoders (which do not make any prior assumption on the speech signal) for which compression rates are in the range [24-64] kbps (such as G.711, G.729) [22], whereas the vocoders utilized in mobile voice channels such as AMR are based on extracting the signal parameters and then modeling and synthesizing the speech at the receiver side, and their compression rate is in the range of [.6-13.2] kbps. It should be noted that data hiding for the second group of vocoders (such as AMR) is inherently more complicated due to the higher compression rate.

Actually, in real mobile voice communication environments, there are several effects pertinent to the vocoder systems, wireless communication channel, synchronization, etc. [23]. In this regard, and since data hiding robustness is our main motivation, we believe that the vocoders at both the transmit and receive terminals constitute the main impairment among all other effects due to the following considerations:

1) The actual over-the-air transmission is handled by the cellular network system with its (proprietary) waveform design, modulation, forward error correction (FEC), forward error detection (FED) and equalization. The digital transmission/reception subsystems are in charge of guaranteeing that the samples (now, the watermarked samples) are delivered reliably [23]. Since our work proposes a variant of spread-spectrum data hiding, which is known to be robust to sample deletion, insertion and channel errors [24], [25], the watermark will still be decoded correctly.
2) Even if several full frames may be affected by fading or other network impairments, to the point that the watermark cannot be reliably decoded, the hidden data (e.g., authentication information) can be repeated as many times as needed and with a time separation that is much larger than the coherence time of the channel, so that the critical hidden information can still be decoded. This is of course a rudimentary sort of repetition coding, but a small payload will suffice in most applications. Obviously, other more sophisticated methods for protecting the watermark are possible, but we have not pursued them and are left for future works.
3) Finally, a real implementation would need to include means for achieving synchronization, both at the watermark symbol level and at the watermark data frame level. This is also out of the scope of our paper, although some solutions have been proposed to overcome these issues, such as [24], [26], [27], which can also be integrated with our proposed methods.

Therefore, this work centers on investigating data hiding methods robust to the vocoder effects of the mobile voice channel. In doing so, we face two important constraints. On one hand, the proposed method should operate on the vocoders already deployed in cellular networks; thus, in contrast to existing solutions which modify the vocoders, for our purposes we rule out such possibility. On the other hand, it must be achieved with a low complexity that allows for implementations with low impact in CPU usage and power consumption.

In order to design an efficient data hiding method robust to mobile communication vocoders, there are two main options, namely, Quantization Index Modulation (QIM) and Spread Spectrum (SS) methods. Both have advantages and disadvantages in terms of data rate, robustness, imperceptibility and security. For applications where security is important, SS-based data hiding is arguably superior [28]. Moreover, SS-based data hiding is more robust to strong channel distortions like the ones we encounter in mobile voice networks [29], including lossy compression, nonlinear gains, analog-to-digital conversion, etc. [30]. Since high data embedding rates are not critical in the foreseen applications, we have singled out SS-based solutions.

In this paper, we first model the mobile communication vocoder as a combination of a linear filter and a non-linear block. Then, according to the constructed model, two types of robust suboptimal decoders based on the SS paradigm are designed and developed. The performance of the proposed schemes is analytically studied and the imperceptibility of the scheme is investigated and evaluated using well-known subjective and objective metrics.

Notation: Throughout this paper, we use regular lowercase letters for scalar variables and random variables, and lowercase boldface letters for vectors. Matrices are represented using regular uppercase letters, while the corresponding regular lowercase, with subscript indices, represents the entries. We use $\|\mathbf{x}\|^l$ to

denote $\sum |x_i|^l$. The probability distribution function of a random variable is denoted by $p(\cdot)$ and the probability of occurrence of a single event is written as $\Pr(\cdot)$. Besides, $\mathrm{E}[\cdot]$ is the expected value of random variables.

The rest of this paper is organized as follows. Section II focuses on modeling of the mobile communication vocoder and proposes an approximation of the nonlinear effects of the codec voice channel. Spread spectrum embedders and decoders according to the constructed model are proposed in Section III. We proceed in this section with a performance analysis of these decoders. Section IV provides the results of simulations with synthetic signals as well as with true speech signals, including an imperceptibility assessment based on both objective and subjective benchmarks. Finally, Section V contains our conclusions and discusses future research lines.

II. Modeling Mobile Communication Vocoders

The behavior of vocoder systems mainly depends on the utilized coding techniques. Generally, the codecs used in voice dedicated channels are classified into two main groups: the first group comprises waveform coders that encode the input signal without any prior assumption on the speech signal. These codecs exploit coding techniques such as PCM (G.711 international standard) and ADPCM (G.726 international standard) to achieve bit rates in the range of [24-64] kbps. Such high data-rate codecs allow transmission of most signals in the voice frequency range with minor distortion.

The second group is a set of vocoders that extract and encode some voice-specific parameters (mostly LPC-based speech modeling parameters) from the input signal and use them to synthesize the voice signal while decoding. These codecs use coding techniques such as Regular Pulse Excitation with Long-Term Prediction (RPE-LTP), Code Excited Linear Prediction (CELP), ACELP, Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP), Vector Sum Excited Linear Predictive Coding (VSELP), Mixed Excitation Linear Prediction (MELP), etc. They can achieve output data rates in the range of [0.6-13.2] kbps. Data embedding schemes that use codecs of the second type which include cellular network vocoders such as AMR are more complicated due to the higher compression rate.

Considering the nonlinear and complicated blocks of vocoders [29], [31], it seems impossible to exactly model the mentioned systems and derive the statistically optimum decoder subject to ML criteria. In order to solve this problem, we concentrate on approximating this channel for human voice inputs. We consider a linear filter plus noise and a nonlinear scaling block as the basis of our approximate model and validate it. In each stage of model validation, we will take into account the interplay between the linear and nonlinear parts. In doing so, we start constructing our model based on the linear part while constraining the input signal to be small for preventing the occurrence of nonlinear effects. Then, we model the non-linear part, and finally, we validate the entire model with actual speech signals.

For the sake of simplicity and clarity, at first we consider pure spread-spectrum (SS) watermarking (for improved SS watermarking the computations differ, see Section III). To perform the required analysis, let x_k denote the kth sample of the host signal, w_k is the corresponding sample of the watermark, and $y_k = x_k + w_k$ is then the watermarked signal.

As customary, we assume that watermarking is performed in an i.i.d fashion, with a watermark with zero mean and variance σ_w^2. As to the host, we first assume it is stationary, with zero mean, variance σ_x^2 and normalized autocorrelation function

$$\rho_i \doteq \mathrm{E}\{x_k x_{k+i}\}/\sigma_x^2. \tag{1}$$

With these definitions, the Document-to-Watermark Ratio (DWR) can be written as σ_x^2/σ_w^2.

For hypothesizing our model, first, we conjecture that the nonlinearity behaves as a linear function if the input power is small, in other words, to ensure that nonlinear effects are negligible, we input low power signals to model the linear part. Finally, after constructing both linear and nonlinear parts, we verify whether the initial conjecture holds. To proceed with constructing the linear part of our model, we are firstly interested in learning how linear time invariant (LTI) filtering affects each signal. We let h_k denote the filter impulse response, and y'_k, x'_k, w'_k the filtered versions of y_k, x_k, w_k, respectively. Notice that due to the superposition principle, we have $y'_k = x'_k + w'_k$. We assume that the decision about w_k is taken from y'_k alone; this means that even though watermarking decoding would clearly benefit from equalization of h_k, we decide not to do so. Therefore, we can write

$$y'_k = h_0 w_k + \sum_{\substack{i=-\infty \\ i\neq k}}^{\infty} h_i w_{k-i} + x_k * h_k. \tag{2}$$

The second term in the right hand side of (2) is akin to the intersymbol interference (ISI) found in communications, so we will refer to it by this name.

If we want to know the effective DWR at the output of the filter, we must compute the variance of the ISI plus host interference term, which we will denote by v_k. Noticing that w_k is white, we can write

$$\mathrm{E}\{v_k^2\} = \sigma_w^2 + \sigma_w^2 \sum_{\substack{i=-\infty \\ i\neq k}}^{\infty} h_i^2 + \sigma_x^2 \sum_{i=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_i h_m \rho_{i-m} \tag{3}$$

while for the watermark part, we have that the variance is simply $h_0^2\sigma_w^2$. Then, if the filter impulse response and the host autocorrelation are known (or can be estimated), it is possible to calculate the DWR.

We are interested in obtaining a manageable $equivalent$ model for y'_k. To this end, we notice that if we scale the watermark, this affects both the useful part of the received signal and the ISI term, while if we scale the host, this affects only the host-interference term. Therefore we can write y'_k in terms of ISI term (denoted by t_k) and the host-interference term (denoted by u_k)

$$y'_k = h_0 w_k + t_k + u_k \tag{4}$$

where w_k, t_k and u_k are mutually independent, zero-mean, and

$$\mathrm{E}\{t_k^2\} = \sigma_w^2 \cdot \zeta(h);\ \mathrm{E}\{u_k^2\} = \sigma_x^2 \cdot \beta(h, \rho) \tag{5}$$

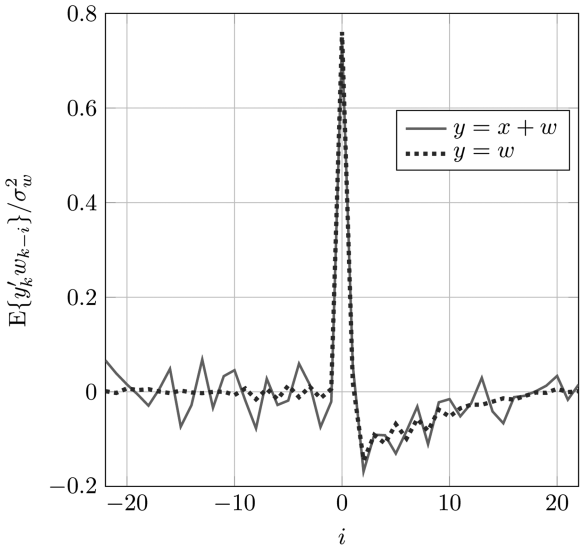


Fig. 1. Estimation of h_i based on (7), (9) for two different scenarios respectively: 1) $y = x + w$; $\sigma_x^2 = 1$, $\sigma_w^2 = 0.01$; 2) $y = w$; $\sigma_w^2 = 1$.

for some ζ and β that depend on the filter and the normalized input autocorrelation, as explicitly indicated by the notation. Finally, the effective DWR, denoted by τ, can be written as

$$\tau = \frac{\sigma_t^2 + \sigma_u^2}{h_0^2 \sigma_w^2} = \frac{\zeta(h) + \sigma_x^2/\sigma_w^2 \cdot \beta(h, \rho)}{h_0^2} \tag{6}$$

which explicitly depends on the input DWR but in an affine fashion. In this simplified model, one can use the superposition principle to estimate the scalars h_0, ζ and β.

As a first step, to validate the adequacy of the linear part of our model, we compare the estimated values for h_i, which we will denote by $\hat{h}_i$ in the sequel, obtained in two different ways. In the first, we compute the value of $\hat{h}_i$ by the following estimator:

$$\hat{\mathbf{h}} = R_{\mathrm{xx}}^{-1} \mathbf{r}_{\mathrm{yx}} \tag{7}$$

where $\hat{\mathbf{h}} \doteq [\hat{h}_{-l}, \ldots, \hat{h}_l]^T$. The autocorrelation matrix of the input signal (R_{xx}) and the input-output cross-correlation vector ($\mathbf{r}_{\mathrm{yx}}$) are defined as

$$R_{\mathrm{xx}} = \sigma_x^2 \begin{bmatrix} \rho_0 & \cdots & \rho_{-2l} \\ \vdots & \ddots & \vdots \\ \rho_{2l} & \cdots & \rho_0 \end{bmatrix}, \mathbf{r}_{\mathrm{yx}} = \begin{bmatrix} \mathrm{E}\{{y'}_k x_{k+l}\} \\ \vdots \\ \mathrm{E}\{{y'}_k x_{k-l}\} \end{bmatrix} \tag{8}$$

and l indicates the effective length of the impulse response. Alternatively, we can estimate h_i as

$$\hat{h}_i = \frac{\mathrm{E}\{y'_k w_{k-i}\}}{\sigma_w^2} \tag{9}$$

and the obtained results for both methods are shown in Fig. 1. As one can see, the achieved results here are consistent.

In addition, it should be noted that the achieved outcomes here seem noisy. In order to determine how much the estimated filter response changes with the input signal, we have conducted an experiment in which true speech samples are passed through the codec channel. In this simulation we move forward through one long speech recording file with a window size of 50 samples each, while assuming a window overlap of 50%. Fig. 2 shows in one snapshot all the obtained channel responses, where we can see that the estimated channel responses have little variance; therefore, the average channel response can be taken as a good representative of the true impulse response. On the other hand, Fig. 3 represents the value of τ in (6) as a function of time for a speech signal. As we can notice, the obtained values fluctuate over time. This fluctuation could be easily justified by considering the definition of τ and its dependency on ρ_i which obviously varies in time due the non-stationary nature of true speech signals.

Moreover, to check if it is necessary to consider any noise in our modeling, we compare the theoretical output power (i.e., $\mathrm{E}\{{y'}_k^2\}$) with the average measured output power in the simulations (i.e., those conducted to plot Fig. 3). We compute the

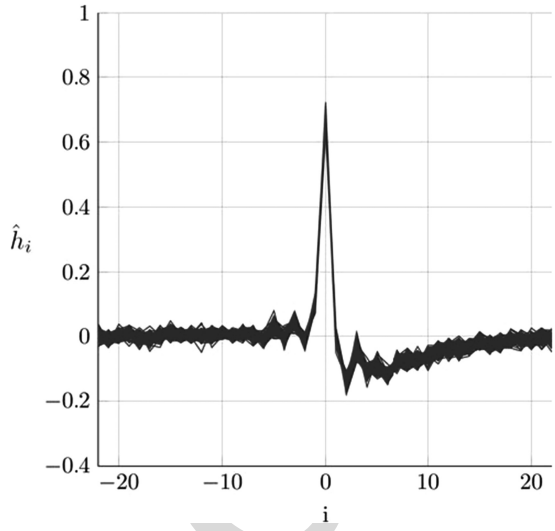


Fig. 2. Obtained values of $\hat{h}_i$ over several true speech samples.

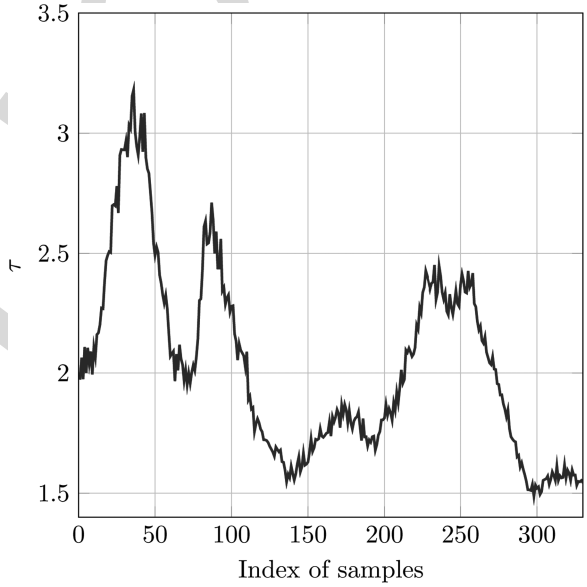


Fig. 3. Obtained value of τ over several true speech samples.

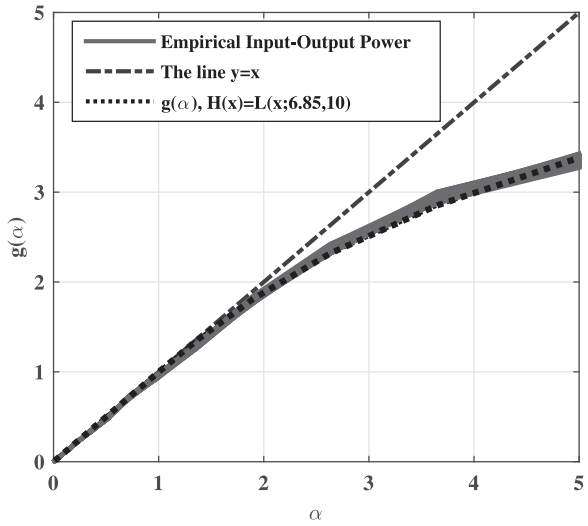


Fig. 4. Empirical input-output power curves (shown as a red band) and the approximation based on Rapp's model versus the power of input signal, denoted by α.

output energy of the signal as

$$\mathrm{E}\{y'^2_k\} = \sigma_w^2 \sum_i h_i^2 + \sigma_x^2 \sum_i \sum_m h_i h_m \rho_{i-m}\,. \tag{10}$$

The achieved results show that the difference between the measured output power and $\mathrm{E}\{y'^2_k\}$ is negligible (i.e., it was less than 0.015). Thus, we can include noise in our model with an approximate variance of 0.01 or, given its small contribution, even neglect it. For the sake of simplicity, we proceed with the latter.

As mentioned earlier, we conjectured that nonlinear effects do not arise in the case of low input power, so the constructed model up to here is entirely linear. In addition, we conducted additional simulations to decide whether it is necessary to consider any nonlinear constituent part. In other words, if the model is entirely linear, scaling the inputs should produced scaled outputs. We plot the input-output power curve in Fig. 4 to check whether such property holds. Moreover, to verify whether this nonlinear block is time-invariant, we rerun this simulation over different audio files from TIMIT dataset (in particular, the core test set of TIMIT material which contains 1920 sentences from 24 speakers) and plotted the corresponding empirical input-output curves for each file in one snapshot in Fig. 4. As illustrated in this Figure, these curves constitute a thin red band, from which it can be inferred that this block (i.e., the nonlinear scaling block) can be modeled as time-invariant and insensitive to the attributes of the input signal. This property has also been advocated for the structure of the AMR vocoder in [29].

As illustrated in Fig. 4, assuming the entire linear model is not tenable and we should consider a nonlinear block to model the full regime (including clipping and gain-saturation). To do so, we add a limiter function block to our hypothesized model as illustrated in Fig. 5. Passing $y_k = x_k + w_k$ through the limiter function denoted by $H(\cdot)$, we have $y''_k = H(y_k)$ at the output of the limiter. Recalling that the watermark magnitude must

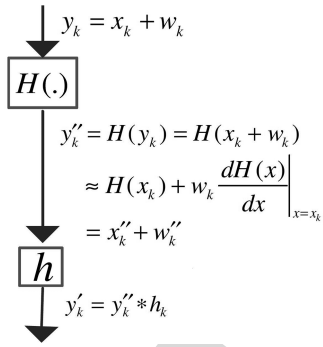


Fig. 5. Block-diagram of the complete model.

be small (i.e., $|w| \ll 1$) for perceptual reasons, we can deduce that $|w|^n \ll |w|$ for any $n > 1$. Thus, we approximate $H(\cdot)$ by applying a first-order Taylor expansion around $y_k = x_k$ as

$$H(y_k) = H(x_k + w_k) \approx H(x_k) + w_k \left.\frac{dH(x)}{dx}\right|_{x=x_k}. \tag{11}$$

Lets denote $H(x_k)$ and $w_k \left.\frac{dH(x)}{dx}\right|_{x=x_k}$ by x''_k and w''_k respectively. Considering the nonlinear part and according to Fig. 5, we can update (6) by substituting σ_x and σ_w by $\sigma_{x''}$ and $\sigma_{w''}$ respectively. The variance of x'' and w'' now must be calculated as

$$\sigma^2_{x''} = \int_{-\infty}^{\infty} H^2(x) p_x(x) dx \tag{12}$$

$$\sigma^2_{w''} = \sigma_w^2 \int_{-\infty}^{\infty} \left(\frac{d}{dx} H(x)\right)^2 p_x(x) dx. \tag{13}$$

It is noteworthy to say that, since $H(\cdot)$ is an odd function and $p_x(x)$ assumed to be an even function, the mean of x'' is zero. Moreover, the power of the output signal (i.e., $\mathrm{E}\{(y''_k)^2\}$) can be computed as

$$\begin{aligned}\mathrm{E}\{(y'')^2\} &= \int_{-\infty}^{\infty} H^2(y) p_y(y) dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H^2(x+w) p_x(x) p_w(w) dx dw.\end{aligned} \tag{14}$$

Let us denote the input-output power relationship by function $g(\cdot)$. Invoking (11) and doing some algebraic simplifications

$$\begin{aligned} g(\alpha) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H^2\left(\frac{(x+w)\sqrt{\alpha}}{\sqrt{\sigma_x^2+\sigma_w^2}}\right) p_x(x) p_w(w) dx dw \\ &\approx \int_{-\infty}^{\infty} H^2\left(\frac{x\sqrt{\alpha}}{\sqrt{\sigma_x^2+\sigma_w^2}}\right) p_x(x) dx \\ &\quad + \frac{\alpha\sigma_w^2}{\sigma_x^2+\sigma_w^2} \int_{-\infty}^{\infty} \left(\frac{d}{dx} H\left(\frac{x\sqrt{\alpha}}{\sqrt{\sigma_x^2+\sigma_w^2}}\right)\right)^2 p_x(x) dx. \end{aligned} \tag{15}$$

It should be noticed that in practice we encounter the inverse problem, that is, we know $g(\cdot)$ and we want to find $H(\cdot)$. To solve this inverse problem, since an explicit expression for $H(x)$

cannot be obtained, we approximate $H(x)$ by a function in the following set, parameterized by $x_{\max}$

$$L(x; x_{\max}, q) = \frac{x}{\left(1 + \left(\frac{|x|}{x_{\max}}\right)^{2q}\right)^{\frac{1}{2q}}}. \quad (16)$$

The family given by (16) corresponds to the limiter functions encompassed by Rapp's model [32]. To find the best match, we numerically solve the following optimization problem using true speech samples

$$x^*_{\max}, q^* = \arg \min_{x_{\max}, q} \int_{0.1}^{5} \|g(\alpha) - \hat{g}(\alpha)\| d\alpha \quad (17)$$

where $\hat{g}(\alpha)$ is defined in the same fashion as $g(\alpha)$, replacing $H(x)$ by $L(x; x_{\max}, q)$ in (15). The optimization in (17) yields $x^*_{\max} = 6.85$ and $q^* = 10$ which, as can be seen in Fig. 4, results in a good approximation of $H(x)$.

Our hypotheses for constructing the model in Fig. 5 include the assumption that if the input power signal were low enough, the linearity would hold and nonlinear effects would not be significant. Fig. 4 validates such assumption: the utilized input signal for modeling the linear part was small enough to avoid noticeable nonlinear effects. Thus, since our conclusions regarding the linear part have been drawn for such operating point, the proposed nonlinear block does not alter them. As a consequence, our global model that consists of a linear and a nonlinear part will be exploited in the subsequent analyses performed in this paper. Notice that with this model, even if x'_k followed a known distribution (e.g., a generalized Gaussian), y'_k would not, so we will focus instead on decoders that do not make assumptions on $p_x(x)$. This is pursued in the next section. Last, to give more insights on the comfortability of proposed model with vocoder structure let us say that, we have used the standard version of AMR 12.2 codec which is officially released by ETSI and written in ANSI-C in all simulations to ensure the conformance of our results with the AMR 12.2 codec utilized in cellular networks [33]. Recalling that the main purpose of our model is to capture the impairments that the codec causes on the watermark signal, it is worth pinpointing the roots of the model in the constituent blocks of the codec.

The nonlinear block in Section II, namely $H(\cdot)$, straightforwardly corresponds to the (A-law, μ-law) compander. As to the LTI filter $h[n]$, it can be explained by the lossy encoding/decoding of the LPC filter and by the low-pass filter that performs subframe interpolation and long-term synthesis. To elaborate, let $\hat{A}(z)$ denote the LPC synthesis filter obtained by quantizing the LSP (line spectral pairs) coefficients of the analysis filter $A(z)$, and let $G(z)$ denote the low-pass filter. Then, the input-output transfer function can be modeled by $L(z) \doteq \frac{A(z)G(z)}{\hat{A}(z)}$. Experiments conducted on real speech signals confirm that the impulse response $l[n]$ so obtained is remarkably close to $h[n]$ and despite the fact that both $A(z)$ and $\hat{A}(z)$ are time-varying, their ratio, and thus $l[n]$, are quite stationary, in accordance with our observations for $h[n]$.

Leveraging on the modeling methodology here presented, one of the important advantages is that it can be extended to other AMR codecs such as AMR 10.2, AMR-WB. Moreover, as our watermarking methods have been designed to achieve a large degree of robustness, they can be expected to perform well with other vocoders having a similar underlying structure. We have checked this for the AMR 10.2 and GSM FR codecs, with promising results.

III. Spread Spectrum Data Embedding

Spread Spectrum (SS) methods are arguably the most popular for data hiding. The SS scheme was first presented by Cox *et al.* [34] in 1997. The authors proposed a method by which the information could be embedded into the host signal with a shared key. There are both additive [35] and multiplicative [36] versions of SS. At the receiver side, the information is decoded and extracted by using the same key as in embedding.

A. Improved Additive Spread Spectrum

In the case of additive spread spectrum, we insert one data bit into one block of the host signal, i.e., N consecutive samples of the host signal. The samples of the watermarked signal $\mathbf{y}$ for each block are computed as

$$\mathbf{y} = \mathbf{x} + b\mathbf{w}. \quad (18)$$

Where the data bit $b \in \{-1, 1\}$ is modulated and added to N host coefficients $\mathbf{x}$. The watermark signal $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ is a key-dependent pseudorandom sequence. The imposed distortion to the host signal can be written as

$$D = \frac{1}{N}\mathrm{E}\{\|b\mathbf{w}\|^2\} = \sigma_w^2.$$

Having introduced the distortion parameter, the document to watermark ratio (DWR) is DWR $\doteq \sigma_{\mathbf{x}}^2/D$. Since the presented embedding procedure does not compensate the interference from the host contents, the resulting performance is generally not acceptable. In order to improve the performance and have a host-rejection approach at the receiver side which decreases the error probability, the improved additive spread spectrum (IASS) method can be used as follows [37]:

$$\mathbf{y} = \mathbf{x} + b\mathbf{w} - \gamma\mathbf{w}\mathbf{w}^T\mathbf{x} = \mathbf{x} + b\mathbf{w} + \mathbf{u} \quad (19)$$

where γ is set to $1/N\sigma_w^2$ to minimize the probability of error and $\mathbf{u}$ denotes the host-rejection term. It is worth noticing that the embedding distortion can be computed as

$$D = \mathrm{E}\{\|\mathbf{y} - \mathbf{x}\|^2\} = \sigma_w^2 + \frac{\sigma_x^2}{N}. \quad (20)$$

Next, according to the constructed model in Section II, the output y' is derived as

$$y'_k = h_k * H(y_k) \quad (21)$$

which clearly illustrates the linear and nonlinear operations on the watermarked signal y_k. In the following sections, we modify the mentioned watermarking structure to tackle the issues of both nonlinear scaling (i.e., the $H(\cdot)$ function) and linear filtering (i.e., convolution with h_k). For the sake of simplicity, we assume that $H(\cdot)$ operates pointwise on its input arguments

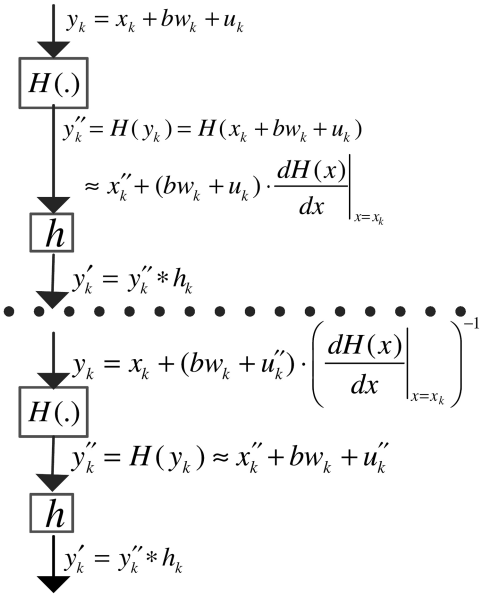


Fig. 6. Block-diagram of the complete model for IASS.

(whether vectors or scalars), e.g., for an arbitrary vector $\mathbf{a}$ with length N, $H(\mathbf{a}) = [H(a_1), \ldots, H(a_N)]^T$.

1) Nonlinear Scaling: In the case of IASS we have

$$H(y_k) = H(x_k + bw_k + u_k). \tag{22}$$

Recalling (11) and considering that for large N, we have $\sigma_u^2 = \frac{\sigma_x^2}{N} \ll \sigma_x^2$, the mentioned equation can be approximated as

$$H(y_k) \approx H(x_k) + (bw_k + u_k) \cdot \left.\frac{dH(x)}{dx}\right|_{x=x_k}. \tag{23}$$

To counterbalance the effect of nonlinear scaling, we first replace the host-rejection part by $\mathbf{u}'' \doteq -\gamma \mathbf{w}\mathbf{w}^T H(\mathbf{x})$ and then modulate both the watermark and new host-rejection terms by $(dH(x)/dx)^{-1}$. In other words, we reformulate the encoder for IASS as follows:

$$y_k = x_k + (bw_k + u_k'') \cdot \left(\left.\frac{dH(x)}{dx}\right|_{x=x_k} \right)^{-1}. \tag{24}$$

According to the parameters of Rapp's model (i.e., $x_{\max} = 6.85$), we can assume that almost all of the true speech samples are less than $x_{\max}$. So, since even in the extreme case of $x_k = x_{\max}$, the value of $(dH(x_k)/dx)^{-1}$ is less than two, it is still reasonable to hold the small signal assumption for this part (i.e., $(bw_k + u_k'') \cdot (dH(x_k)/dx)^{-1}$). Therefore, recalling (11), after passing the mentioned signal through the nonlinear scaling part of our model we have

$$H(y_k) \approx H(x_k) + bw_k + u_k''. \tag{25}$$

Consequently, as seen in Fig. 6 we can compensate the effect of nonlinear scaling by applying the proposed method and just considering $H(x_k)$ instead of x_k as the host signal. The cost of this compensation is the degradation of the error probability, i.e., since $H(x_k)$ is always smaller than x_k, then the watermark power in (25) must be smaller to guarantee the same DWR; this in turn produces an increase in the error probability. This performance degradation was already expected due to the nonlinear scaling of the vocoder. Finally, one might think of applying the inverse of $H(\cdot)$ to y_k to completely remove the nonlinearity. However, this would increase the dynamic range of the input signal to the voice channel, which would be unacceptable in practice. In addition, denoting $\mathrm{E}\{\left.\frac{dH(x)}{dx}\right|_{x=x_k}\}$ by β, the embedding distortion in this new structure can be shown to be

$$D = \frac{\sigma_w^2 + \frac{\sigma_{x''}^2}{N}}{\beta^2}. \tag{26}$$

2) Linear Filtering: To mathematically discuss the effect of linear filtering, let us periodically repeat the watermark signal to make an infinite sequence, i.e., let us assume that $w_i = w_{i-mN}$ for all integer values of m. Now, considering (25), at the output of model we have

$$y_k' = \sum_{i=-\infty}^{\infty} h_i H(y_{k-i}) \approx \sum_{i=-\infty}^{\infty} h_i \left(x_{k-i}'' + bw_{k-i} + u_{k-i}'' \right). \tag{27}$$

By applying the correlator decoder, i.e., the inner product of $\mathbf{y}'$ and $\mathbf{w}$ at the decisor, we have

$$z_{\mathrm{A}} = \mathbf{w}^T \mathbf{y}' = \sum_{k=1}^{N} \sum_{i=-\infty}^{\infty} w_k h_i \left(x_{k-i}'' + bw_{k-i} + u_{k-i}'' \right) \tag{28}$$

where z_{A} indicates the test statistic. After some algebraic manipulations, we can compute the mean and variance of z_{A}, denoted by m_{A} and σ_{A}^2, respectively, as

$$\begin{aligned} m_{\mathrm{A}} &= N b h_0 \sigma_w^2 \\ \sigma_{\mathrm{A}}^2 &= N \sigma_w^2 \sigma_{x''}^2 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}'' \end{aligned} \tag{29}$$

where ρ_{i-m}'' indicates the normalized autocorrelation function of x'', i.e., $\rho_i'' = \mathrm{E}\{x_k'' x_{k+i}''\}/\sigma_{x''}^2$. Recalling the central limit theorem (CLT), for large N we can assume that test statistic z_A in (28) approximately follows a Gaussian distribution with mean m_{A} and variance σ_{A}^2. Assuming an equal prior probability for the information bit, i.e., $\Pr(b = +1) = \Pr(b = -1) = 1/2$, we can approximate the error probability as follows[1]:

$$\begin{aligned} \Pr(e) &= \Pr(e|b = 1) = \Pr(e|b = -1) \\ &\approx Q\left(\frac{m_{\mathrm{A}}}{\sigma_{\mathrm{A}}} \right) = Q\left(\frac{\sqrt{N} h_0 \sigma_w}{\sigma_{x''} \sqrt{\sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}''}} \right). \end{aligned} \tag{30}$$

considering (26) and defining $\kappa \doteq \beta^2 \frac{\sigma_x^2}{\sigma_{x''}^2}$, the approximate error probability can be rewritten as

$$\Pr(e) \approx Q\left(\frac{h_0 \sqrt{\frac{\kappa N}{\mathrm{DWR}} - 1}}{\sqrt{\sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}''}} \right). \tag{31}$$

[1] $Q(x) = (2\pi)^{-1/2} \int_x^{\infty} \exp(-v^2/2) dv$.

One way to increase the data rate of the introduced structure is to use multi bi-orthogonal codes as spreading signals. To this end, we propose a new structure (embedding/decoding structure) which aims at achieving host rejection with multiple simultaneous watermark carriers. Let $\mathbf{w}^i, i \in \{1, \ldots, M\}$ with length N_m denote a set of orthogonal codes, M indicates the number of orthogonal watermark carriers and $\mathbf{w}^* \in \{\mathbf{w}^1, \ldots, \mathbf{w}^M\}$ represents the embedded symbol. Denoting the new host-rejection term by $\mathbf{r}$, i.e., $\mathbf{r} = (\sum_{i=1}^{M} \mathbf{w}^i \mathbf{w}^{i^T}) H(\mathbf{x})$, the embedder and decoder for the M-ary bi-orthogonal Additive (MA) structure is proposed as follows:

$$y_k = x_k + \left(b w_k^* - \frac{r_k}{N_m \sigma_w^2} \right) \cdot \left(\left. \frac{dH(x)}{dx} \right|_{x=x_k} \right)^{-1} \tag{32}$$

where the decision is made by the decoder as

$$\hat{d}_{\mathrm{MA}} = \hat{\jmath} \cdot \operatorname{sgn}\left(\mathbf{y}'^T \mathbf{w}^{\hat{\jmath}} \right) \tag{33}$$

with

$$\hat{\jmath} = \underset{j \in \{1 \ldots M\}}{\arg\max} \, |\mathbf{y}'^T \mathbf{w}^j|. \tag{34}$$

The embedded distortion for this structure is

$$D = \frac{1}{N} \mathrm{E} \left\{ \left\| b\mathbf{w}^* - \frac{(\sum_{i=1}^{M} \mathbf{w}^i \mathbf{w}^{i^T}) \mathbf{x}}{N_m \sigma_w^2} \right\|^2 \right\} = \frac{\sigma_w^2 + \frac{M \sigma_{x''}^2}{N_m}}{\beta^2}. \tag{35}$$

Since we now choose one among $2M$ spreading sequences, there are $\log_2 2M$ bits encoded in the decision, so we can increase the length of the spreading sequences by this amount for the same effective rate. Therefore $N_m = N \log_2 2M$. Carrying out some algebraic manipulations, the error probability for the mentioned structure can be approximated as [38]

$$\begin{aligned} \Pr(e) &= 1 - \Pr(c|b=1) = 1 - \Pr(\mathbf{w}^* = \mathbf{w}^{\hat{\jmath}} | b=1) \\ &\overset{(33)}{=} \frac{2^{M-1}}{2^M - 1} \left[1 - \prod_{\substack{j=1 \\ j \neq \hat{\jmath}}}^{M} \Pr\left(\mathbf{y}'^T \mathbf{w}^* > |\mathbf{y}'^T \mathbf{w}^j| \right) \right] \\ \Pr(e) &\approx \frac{2^{M-1}}{2^M - 1} \left[1 - \frac{1}{\sqrt{2\pi\sigma_{\mathrm{A}}^2}} \right. \\ &\quad \left. \times \int_0^\infty \left(1 - 2Q\left(\frac{x}{\sigma_{\mathrm{A}}} \right) \right)^{M-1} e^{-\frac{(x - m_{\mathrm{A}})^2}{2\sigma_{\mathrm{A}}^2}} dx \right]. \end{aligned} \tag{36}$$

B. Improved Multiplicative Spread Spectrum Data Embedding

In the improved multiplicative spread spectrum (IMSS), the watermarked signal is generated as [30]

$$y_k = x_k + b x_k^2 w_k + u_k. \tag{37}$$

Now, similarly to Section III-A1, to take into account the effect of nonlinear scaling, we modify (37) to

$$y_k = x_k + \left(b H^2(x_k) w_k + u_k'' \right) \cdot \left(\left. \frac{dH(x)}{dx} \right|_{x=x_k} \right)^{-1}. \tag{38}$$

In this scheme, after some straightforward computations, D can be shown to be

$$D = \frac{\sigma_w^2 \mathrm{E}\{(x'')^4\} + \frac{\sigma_{x''}^2}{N}}{\beta^2}. \tag{39}$$

After passing y_k through the nonlinear scaling function of our model, we have

$$H(y_k) \approx x_k'' + b(x_k'')^2 w_k + u_k'' \tag{40}$$

recalling the linear part of our model and akin to (27), the output of linear block is

$$\begin{aligned} y_k' &= \sum_{i=-\infty}^{\infty} h_i H(y_{k-i}) \\ &\approx \sum_{i=-\infty}^{\infty} h_i \left(x_{k-i}'' + b(x_{k-i}'')^2 w_k + u_{k-i}'' \right). \end{aligned} \tag{41}$$

By applying the correlator decoder (i.e., the inner product of y', w as the test statistic) we have

$$z_{\mathrm{M}} = \mathbf{w}^T \mathbf{y}' \approx \sum_{k=1}^{N} \sum_{i=-\infty}^{\infty} h_i \left(x_{k-i}'' + b(x_{k-i}'')^2 w_k + u_{k-i}'' \right). \tag{42}$$

Next, to apply the CLT and compute the error probability, we need to find the values of the mean and variance of z_{M} denoted by m_{M}, σ_{M}^2 respectively. Noticing that the variables in the second sum of (42) are zero-mean and uncorrelated with the watermark, we can write

$$m_{\mathrm{M}} \approx N h_0 b \sigma_w^2 \sigma_{x''}^2 \tag{43}$$

whereas σ_{M}^2 can be computed as

$$\begin{aligned} \sigma_{\mathrm{M}}^2 &= \mathrm{E}\{z_{\mathrm{M}}^2\} - m_{\mathrm{M}}^2 \\ &\approx N \left(\sigma_w^2 \sigma_{x''}^2 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}'' + \sigma_w^4 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \varphi_{i-m} \right) \\ &\quad + N h_0^2 \mathrm{E}\{w_i^4\} \varphi_0 + h_0^2 \sigma_w^4 \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varphi_{i-j} - N^2 h_0^2 \sigma_w^4 \sigma_{x''}^4 \end{aligned} \tag{44}$$

in which $\varphi_i \doteq \mathrm{E}\{(x_{k-i}'')^2 (x_k'')^2\}$. Assuming that $\mathrm{E}\{w_i^4\} \ll \mathrm{E}\{w_i^2\}$, σ_{M}^2 can be approximated as

$$\sigma_{\mathrm{M}}^2 \approx N \sigma_w^2 \sigma_{x''}^2 \sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}''. \tag{45}$$

Consequently, similarly to the additive case, the error probability is

$$\Pr(e) \approx Q\left(\frac{m_{\mathrm{M}}}{\sigma_{\mathrm{M}}} \right) = Q\left(\frac{\sqrt{N} h_0 \sigma_{x''} \sigma_w}{\sqrt{\sum_{i \neq 0} \sum_{m \neq 0} h_i h_m \rho_{i-m}''}} \right). \tag{46}$$

Denoting $\frac{\sqrt{\mathrm{E}\{(x'')^4\}}}{\sigma^2_{x''}}$ by η and considering (39), the error probability can be reformulated in terms of the DWR as

$$\Pr(e) \approx Q\left(\frac{h_0\sqrt{\frac{\kappa N}{\mathrm{DWR}}-1}}{\eta\sqrt{\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m \rho''_{i-m}}}\right). \tag{47}$$

Moreover, inspired by the M-ary bi-orthogonal watermarking in Section III-A, a M-ary Multiplicative (MM) structure can be proposed in the case of Improved multiplicative SS as follows:

$$y_k = x_k + \left(b(x''_k)^2 w^*_k - \frac{r_k}{N_m\sigma^2_w}\right)\cdot\left(\left.\frac{dH(x)}{dx}\right|_{x=x_k}\right)^{-1} \tag{48}$$

in which $w^*_k, w^j_k r_k$ are defined in the same fashion as in Section III-A and, consequently, the decoder makes the decision as

$$\hat{d}_{\mathrm{MM}} = \hat{\jmath}\cdot\mathrm{sgn}\left(\mathbf{y}'^T\mathbf{w}^{\hat{\jmath}}\right) \tag{49}$$

where

$$\hat{\jmath} = \underset{j\in\{1\ldots M\}}{\arg\max}\,|\mathbf{y}'^T\mathbf{w}^j|. \tag{50}$$

Furthermore, the embedding distortion can be written as

$$D = \frac{\sigma^2_w\mathrm{E}\{(x'')^4\} + \frac{M\sigma^2_{x''}}{N_m}}{\beta^2}. \tag{51}$$

Let σ^2_{MM} denote the variance of the interference that results after multiplying by other spreading signal (i.e., $\mathrm{E}\left\{\mathbf{y}^T\mathbf{w}^j \,\middle|\, \mathbf{w}^j\neq\mathbf{w}^*\right\}$). Next, recalling (36) and after some algebraic manipulations the error probability becomes

$$\begin{aligned}\Pr(e) &= 1-\Pr(c|b=1) = 1-\Pr(\mathbf{w}^*=\mathbf{w}^{\hat{\jmath}}|b=1)\\ &\overset{(49)}{=} \frac{2^{M-1}}{2^M-1}\left[1-\prod_{\substack{j=1\\ j\neq\hat{\jmath}}}^{M}\Pr\left(\mathbf{y}'^T\mathbf{w}^* > \left|\mathbf{y}'^T\mathbf{w}^j\right|\right)\right]\\ &\approx \frac{2^{M-1}}{2^M-1}\left[1-\frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{M}}}}\right.\\ &\quad\left.\times\int_0^\infty\left(1-2Q\left(\frac{x}{\sigma_{\mathrm{MM}}}\right)\right)^{M-1} e^{-\frac{(x-m_{\mathrm{M}})^2}{2\sigma^2_{\mathrm{M}}}}dx\right]\end{aligned} \tag{52}$$

where σ^2_{MM} can be computed as

$$\begin{aligned}\sigma^2_{\mathrm{MM}} &= N\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m\left(\sigma^2_w\sigma^2_{x''}\rho''_{i-m} + \sigma^4_w\varphi_{i-m}\right)\\ &\quad + Nh_0^2\sigma^4_w\varphi_0 \approx N\sigma^2_w\sigma^2_{x''}\sum_{i\neq 0}\sum_{m\neq 0} h_i h_m\rho''_{i-m}.\end{aligned} \tag{53}$$

IV. Simulations and Results

In this section, we validate our analysis (in particular; error probability formulas) with several experiments. The good conformance between experimental results and theory serves as an additional supporting validation for the vocoder modeling from Section II. Afterwards, we assess the imperceptibility of the proposed methods according to subjective and objective benchmarks.

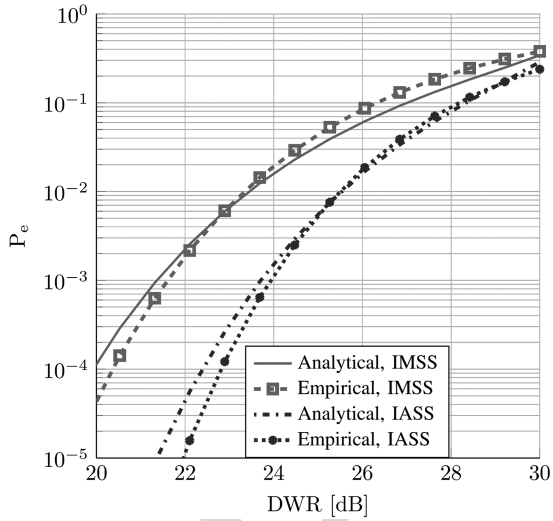


Fig. 7. Empirical and analytical results over true speech samples for $N = 2500$, AMR12.2.

A. Performance Analysis

According to (6), the analytical error probability is highly dependent on the autocorrelation of the input signals. We consider two different scenarios to measure the goodness of our proposed model. In the first one, we perform the simulations over true speech samples and consider the average autocorrelation of the input signal in our analytical formula. As shown in Fig. 7, the empirical results are close to the analytical ones but do not accurately match.

In the second scenario, we synthesize a signal with i.i.d Generalized Gaussian Distribution (GGD) samples [39], [40]. This ensures a time-invariant autocorrelation. As illustrated in Fig. 8, now the analytical and empirical results match more closely. We can conclude that for stationary signals the analytical prediction of the error probability is quite accurate, and that some accuracy is lost for real speech signals, as stationarity does not hold in this case. It is worth noticing that all the human voice samples used in our simulations have been selected from the TIMIT database [41].

It is worthy to note that the most common vocoders in cellular networks are GSM FR, GSM HR in 2.5G and 2.75G, and AMR in 2.5G, 2.75G, 3G. To the best of our knowledge, the most prevalent one among the mentioned vocoders is AMR 12.2 [42], [43]. So much so, we have focused on AMR 12.2 rather than other compression rates. Additionally, as illustrated in Fig. 9, we have conducted another simulation for AMR 10.2, AMR 7.95, to show that the insights given by the proposed model can be applied to other compression rates of AMR as well.

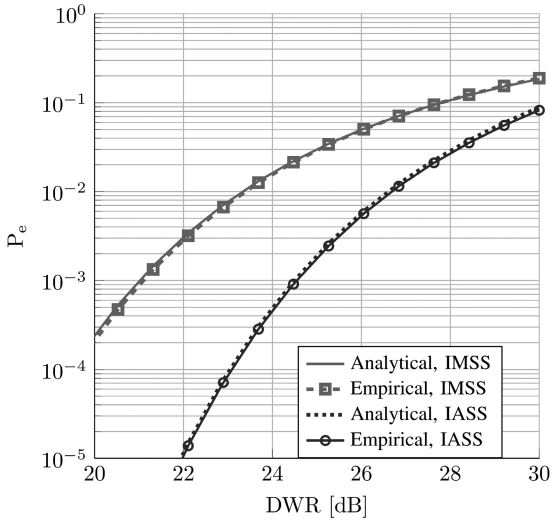


Fig. 8. Empirical and analytical results over synthetic signals for $N = 2500$, AMR12.2.

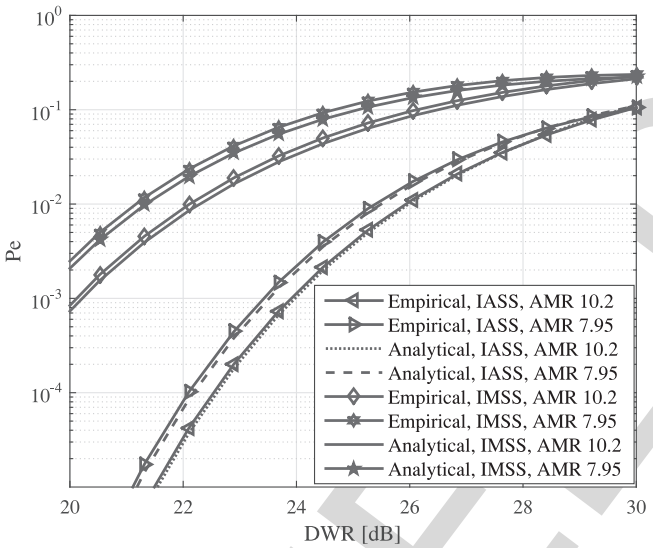


Fig. 9. Empirical and analytical results over synthetic signals for AMR 10.2 and AMR 7.95.

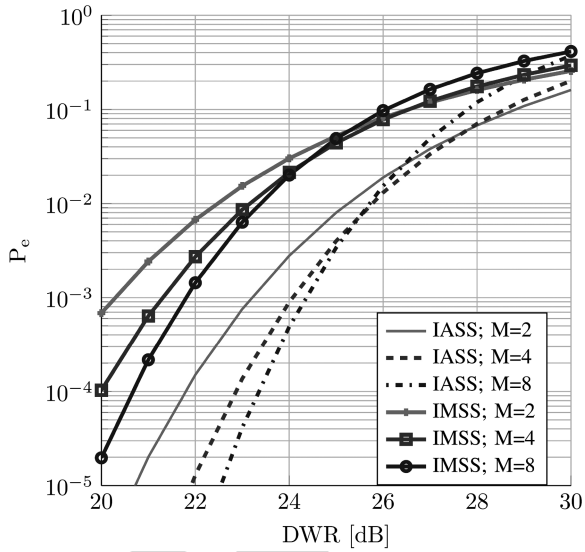


Fig. 10. Empirical results over synthetic signals for $N = 2500$, AMR12.2.

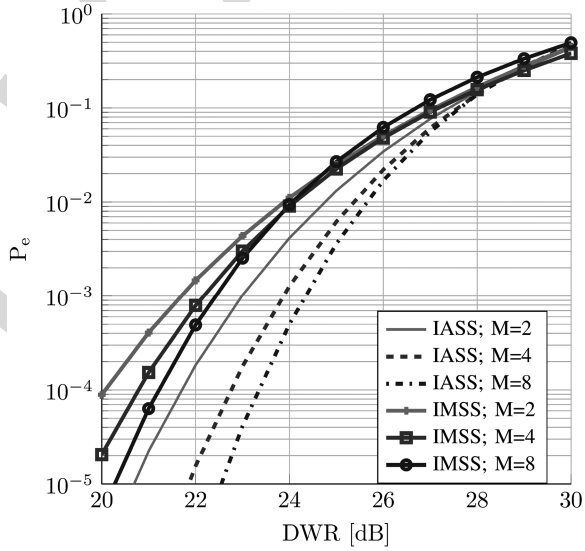


Fig. 11. Empirical results over true speech samples for $N = 2500$, AMR12.2.

In addition, we conducted simulations for the case of M-ary symbols. As illustrated in Figs. 10 and 11, the proposed M-ary structures outperform the results shown in Figs. 7 and 8, especially for low DWRs. It should be noted that for the sake of clarity we do not plot the analytical results in Figs. 10 and 11, but the match is similar to that observed in Figs. 7 and 8.

We also compared the proposed methods with the baseline additive and multiplicative SS schemes [35] [36], and with the scheme proposed by Cheng known as Generalized Embedding of Multiplicative (GEM) watermarking [44]. We plot the results in Figs. 12 and 13 . As illustrated in these figures, the proposed methods outperform the mentioned prior art.

The results shown in Figs. 7 and 11 have been obtained by running the simulations over samples of audio files from the TIMIT dataset [41], and plotting the average results. Experiments were conducted over the core test set of the TIMIT material which contains 1920 sentences from 24 speakers. Moreover, all simulations in this section have been conducted with the following setup:

1) Watermark frame length: 2500, i.e., $N = 2500$.
2) Codec: AMR 12.2.
3) Synthetic signals follow a GGD distribution with zero mean and unit variance, with a shape parameter of 1.5 which is matched with human voice properties [39].
4) The total length of human speech utilized in the simulations was around 737 million samples which corresponds to around 300,000 watermark frames.

Moreover, our proposed methods are based on spread-spectrum. Although we have customized and particularized them for our specific problem, they still inherit the main properties of spread-spectrum, discussed at length in [24]. This

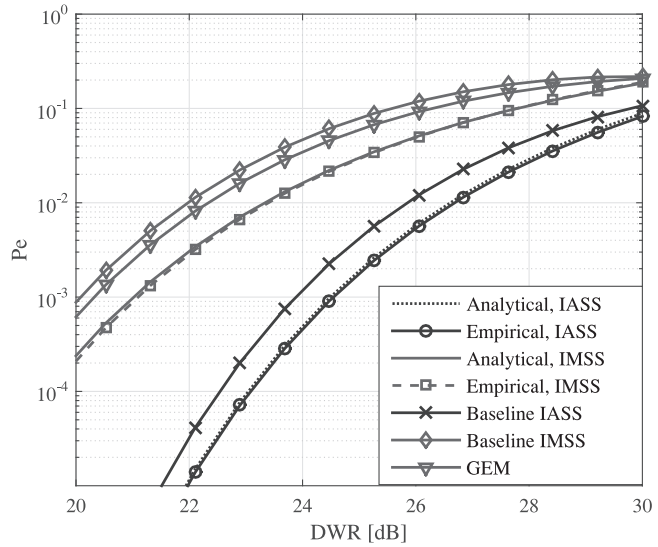


Fig. 12. Empirical and analytical results over synthetic signals based on the proposed methods, baseline spread spectrum methods, and GEM.

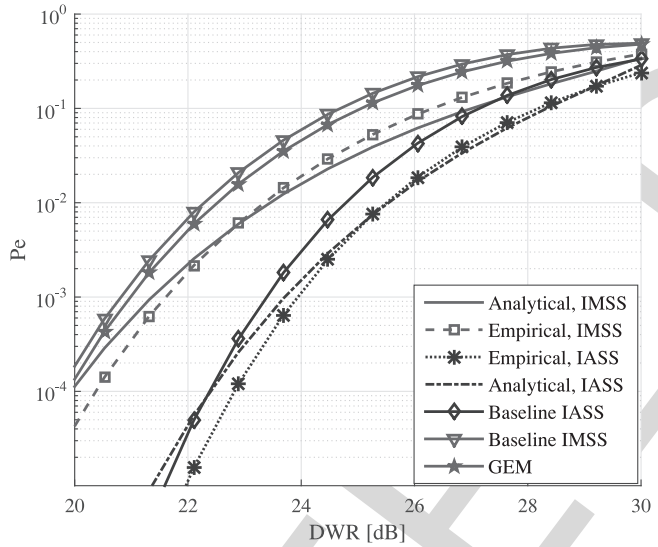


Fig. 13. Empirical and analytical results over true speech samples based on the proposed methods, baseline spread spectrum methods, and GEM.

means that they can be made robust against desynchronization attacks, spectrum filtering, chess watermarking, cut-sampling, zero-padding, resampling, noise addition, sample clipping, etc.

B. Imperceptibility Analysis

In this section we evaluate the proposed structures using objective and subjective benchmarks to assess their imperceptibility. ITU P.862 and ITU P.860 recommendations are two well-known standards which are widely used to evaluate the subjective quality of speech and the imperceptibility of embedded watermarks. In these tests, we considered $N = 2500$ and a sampling rate of 8 KHz, so the embedding bit rate is $8000/2500 = 3.2$ bps. Moreover, we considered $M = 8$ (i.e., 8-ary symbols), $\Pr(e) = 0.001$, and in order to achieve this target $\Pr(e)$, we set the operating DWR at 22.5 dB and 24.5 dB for IMSS and IASS, respectively. To check the imperceptibility of the proposed methods in accordance with the ITU P.860 standard, we prepared four sound file groups, each consisting of five audio files with a length of 10 seconds. We asked five persons to execute the $A/B/X$ test [17]. In the $A/B/X$ test, A indicates the watermarked signal, B stands for plain signal and X is assigned randomly to be A or B. In each stage of this experiment, whenever we played the X file for each listener, we asked him to decide between A and B. We summarized the results in Table I, where $\mathbf{M1}, \mathbf{M2}$ indicate 8-ary IMSS and 8-ary IASS, respectively. It should be noted that in the case of ideal watermarking in this sort of test, the expected percentage of failure would be 50%.

TABLE I
TEST RESULTS FOR ITU P.860 (PERCENTAGE OF FAILURE)

	Group I		Group II		Group III		Group IV	
	M1	M2	M1	M2	M1	M2	M1	M2
Tester 1	45%	55%	45%	35%	50%	55%	60%	55%
Tester 2	55%	50%	40%	40%	55%	45%	55%	50%
Tester 3	55%	35%	40%	45%	60%	55%	60%	45%
Tester 4	45%	45%	50%	50%	50%	60%	50%	55%
Tester 5	60%	55%	45%	40%	55%	55%	55%	55%
Average	52%	48%	44%	42%	54%	54%	56%	52%

TABLE II
TEST RESULTS FOR ITU P.862

	File Num	MOSLQO		Average	
		M1	M2	M1	M2
Group I	F1.1	4.320	4.275	**4.3352**	**4.2537**
	F1.2	4.311	4.206		
	F1.3	4.373	4.241		
	F1.4	4.377	4.268		
	F1.5	4.295	4.279		
Group II	F2.1	4.328	4.304	**4.3474**	**4.2274**
	F2.2	4.321	4.246		
	F2.3	4.353	4.189		
	F2.4	4.363	4.191		
	F2.5	4.371	4.206		
Group III	F3.1	4.294	4.287	**4.3167**	**4.2563**
	F3.2	4.359	4.205		
	F3.3	4.355	4.284		
	F3.4	4.276	4.204		
	F3.5	4.300	4.269		
Group IV	F4.1	4.330	4.219	**4.3334**	**4.2228**
	F4.2	4.403	4.197		
	F4.3	4.304	4.205		
	F4.4	4.344	4.256		
	F4.5	4.286	4.236		

ITU P.862 recommendation describes an objective method for predicting the subjective quality of narrowband speech codecs. In this method, perceptual evaluation speech quality (PESQ) and mean opinion score listening quality objective (MOSLQO) are used to compare the proposed schemes. The results are shown in Table II. It should be noticed that the PESQ value of the host

TABLE III
TEST RESULTS OF COMPARING THE COMPUTATIONAL COMPLEXITY OF VOICE ENCODING/DECODING AND THE PROPOSED DATA HIDING ENCODING/DECODING IN TERMS OF MIPS

Methods	Required MIPS
	($N=2500$)
Voice Enc/Dec based on AMR 12.2	225-300
Enc/Dec based on IASS	20-40
Enc/Dec based on IMSS	20-40
Enc/Dec based on MA (M=4)	30-50
Enc/Dec based on MM (M=4)	30-50

signal before embedding the data was equal to 4.5. As expected and shown in Tables I and II, although both methods exhibit an acceptable level of imperceptibility, the multiplicative structure behaves better than the additive one. In this work and according to its main application (i.e., hiding data in regular voice calls between two persons), we aimed at proposing a watermarking method which would be imperceptible to the human hearing system, as opposed to undetectability by steganalytic methods such as [45], [46]. Therefore, since our primary application is data hiding and not steganography, we have put emphasis on transparency as measured by ITU P.860 and ITU P.860 which are subjective tests of imperceptibility.

C. Complexity Analysis

To assess the complexity of the proposed schemes, we measured the required million instructions per second (MIPS) for processing one time frame of data hiding (in particular, in the case $N = 2500$) which encompasses both encoding and decoding processes. We report the achieved results in Table III together with the required MIPS for voice encoding/decoding (in particular, considering AMR 12.2 as the decoder/encoder) the same time frame. As shown in Table III, the complexity of the proposed schemes is a fraction of that of the vocoder in cellular phones. We must remark that our implementation has not been fully optimized, and even more so, no adaptation to the specific architecture of cellular phone CPUs has been intended. Considering this fact, we believe that a further reduction in the results reported in Table III is feasible.

V. CONCLUSION

In this paper we have shown that for watermarking purposes a mobile communication vocoder can be accurately modeled by considering a non-linear scaling function plus a linear filter. Adhering to such model, we have proposed two Spread-Spectrum-based data hiding methods, termed IASS and IMSS. Moreover, their performance can be largely improved by considering multi bi-orthogonal codes as spreading signals. Finally, the experimental assessment using both subjective and objective measures has revealed that our proposed schemes exhibit an acceptable level of imperceptibility.

Although we have concentrated on detectors which do not rely on the probability distribution of received signals, the approximation of such distribution and the derivation of the corresponding maximum likelihood detectors are topics for further research. In addition, in order to increase the technological readiness of the proposed methods for practical implementation, a technique to assure integrity of the hidden messages in the case of packet losses, and a method for synchronization considering the limited bandwidth constraints and nonlinearities deserve further attention.

REFERENCES

[1] "List of countries by number of mobile phones in use," 2014. [Online]. Available: http://en.wikipedia.org/wiki/List of countries by number of mobile phones in use

[2] D. Strobel, *IMSI Catcher: Chair for Communication Security*., Bochum, Germany: Ruhr-Universität Bochum, 2007, p. 14.

[3] M. Toorani and A. Beheshti, "Solutions to the GSM security weaknesses," in *Proc. 2008 2nd Int. Conf. Next Generation Mobile Appl., Services Technol.*, 2008, pp. 576–581.

[4] L. Buttyan, C. Gbaguidi, S. Staamann, and U. Wilhelm, "Extensions to an authentication technique proposed for the global mobility network," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 373–376, Mar. 2000.

[5] M. Boloursaz, R. Kazemi, B. Barazandeh, and F. Behnia, "Bounds on compressed voice channel capacity," in *Proc. 2014 Iran Workshop Commun. Inform. Theory*, 2014, pp. 1–6.

[6] M. Boloursaz, R. Kazemi, D. Nashtaali, M. Nasiri, and F. Behnia, "Secure data over GSM based on algebraic codebooks," in *Proc. 2013 East-West Des. Test Symp.*, 2013, pp. 1–4.

[7] W. Mazurczyk and Z. Kotulski, "Adaptive VoIP with audio watermarking for improved call quality and security," *J. Inf. Assurance Security*, vol. 2, no. 3, pp. 226–234, 2007.

[8] J. Singh, P. Garg, and A. Nath De, "A combined watermarking and encryption algorithm for secure VoIP," *Inf. Security J., Global Perspective*, vol. 18, no. 2, pp. 99–105, 2009.

[9] N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Trans., Inform. Theory*, vol. 54, no. 1, pp. 255–274, Jan. 2008.

[10] F. Pérez-González, F. Balado, and J. R. H. Martin, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 960–980, Apr. 2003.

[11] Y. Diqun, W. Rangding, and Z. Liguang, "Quantization step parity-based steganography for mp3 audio," *Fundamenta Informaticae*, vol. 97, no. 1, pp. 1–14, 2009.

[12] A. Naofumi, "A technique of lossless steganography for G.711 telephony speech," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, 2008, pp. 608–611.

[13] C. Wang and Q. Wu, "Information hiding in real-time VoIP streams," in *Proc. 9th IEEE Int. Symp. Multimedia*, Dec. 2007, pp. 255–262.

[14] J. Dittmann, D. Hesse, and R. Hillert, "Steganography and steganalysis in voice-over IP scenarios: Operational aspects and first experiences with a new steganalysis tool set," in *Proc. Electron. Imag. 2005*, 2005, pp. 607–618.

[15] Y. F. Huang, S. Tang, and J. Yuan, "Steganography in inactive frames of voip streams encoded by source codec," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 296–306, Jun. 2011.

[16] Y.-m. Su, Y.-f. Huang, and X. Li, "Steganography-oriented noisy resistance model of G.729a," in *Proc. IMACS Multiconf. Comput. Eng. Syst. Appl.*, 2006, vol. 1, pp. 11–15.

[17] Y. Huang, C. Liu, S. Tang, and S. Bai, "Steganography integration into a low-bit rate speech codec," *IEEE Trans. Inform. Forensics Security*, vol. 7, no. 6, pp. 1865–1875, Dec. 2012.

[18] Z. Piotrowski, J. Wojtuń, and J. Ośka, "Hardware watermark token for voip telephony," *Przeglad Elektrotechniczny*, vol. 89, pp. 196–198, 2013.

[19] F. C. Er and E. Gul, "Comparison of digital audio watermarking techniques for the security of voip communications," in *Proc. 2011 7th Int. Conf. Inform. Assurance Security*, 2011, pp. 13–18.

[20] B. Xiao, Y. Huang, and S. Tang, "An approach to information hiding in low bit-rate speech stream," in *Proc. IEEE GLOBECOM Global Telecommun. Conf.*, Nov.–Dec. 2008, pp. 1–5.

[21] Z. Wu and W. Yang, "G. 711-based adaptive speech information hiding approach," in *Proc. Intell. Comput.*, 2006, pp. 1139–1144.

[22] W. Mazurczyk, "VoIP steganography and its detectiona survey," *ACM Comput. Surveys*, vol. 46, no. 2, 2013, Art. no. 20.
[23] C. K. LaDue, V. V. Sapozhnykov, and K. S. Fienberg, "A data modem for GSM voice channel," *IEEE Trans., Veh. Technol.*, vol. 57, no. 4, pp. 2205–2218, Jul. 2008.
[24] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1020–1033, Apr. 2003.
[25] A. Nadeau, and G. Sharma, "Insertion, deletion robust audio watermarking: a set theoretic, dynamic programming approach," in *Proc. IS&T/SPIE Electron. Imag.*, 2013, pp. 866–503–866–503.
[26] C. Baras, N. Moreau, and P. Dymarski, "Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1772–1782, Sep. 2006.
[27] L. de CT Gomes, E. Gómez, and N. Moreau, "Resynchronization methods for audio watermarking," presented at the 111th AES Conv., New York, NY, USA, 2001.
[28] L. Pérez-Freire and F. Pérez-González, "Spread-spectrum watermarking security," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 2–24, Mar. 2009.
[29] "Digital cellular telecommunications system (phase 2+); adaptive multirate (amr) speech transcoding (GSM 06.90 version 7.2.1)," ETSI, Tech. Rep. ETSI EN 301 704, 1998.
[30] A. Valizadeh, and Z. J. Wang, "An improved multiplicative spread spectrum embedding scheme for data hiding," *IEEE Trans. Inform. Forensics Security*, vol. 7, no. 4, pp. 1127–1143, Aug. 2012.
[31] M. Boloursaz, A. Hadavi, R. Kazemi, and F. Behnia, "Secure data communication through GSM adaptive multi rate voice channel," in *Proc. 2012 6th Int. Symp., Telecommun.*, 2012, pp. 1021–1026.
[32] A. A. Eltholth, A. R. Mekhail, A. Elshirbini, M. Dessouki, and A. Abdelfattah, "Modeling the effect of clipping and power amplifier non-linearities on OFDM systems," *Ubiquitous Comput. Commun. J.*, vol. 3, no. 1, pp. 54–59, 2009.
[33] "Universal mobile telecommunication systaem (UMTS); AMR speech codec general description (3gpp ts 26.071 version 5.0.0 released 5)," ETSI, Tech. Rep. ETSI EN 126 071, 2002.
[34] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.
[35] Q. Cheng and T. S. Huang, "An additive approach to transform-domain information hiding and optimum detection structure," *IEEE Trans. Multimedia*, vol. 3, no. 3, pp. 273–284, Sep. 2001.
[36] A. Valizadeh and J. Wang, "A framework of multiplicative spread spectrum embedding for data hiding: Performance, decoder and signature design," in *Proc. GLOBECOM IEEE Global Telecommun. Conf.*, Nov.–Dec. 2009, pp. 1–6.
[37] A. Valizadeh and Z. J. Wang, "Efficient blind decoders for additive spread spectrum embedding based data hiding," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–21, 2012.
[38] J. Proakis, *Digital Communications* (ser. Communications and signal processing). New York, NY, USA: McGraw-Hill, 1995. [Online]. Available: http://books.google.com/books?id=cIqYQgAACAAJ
[39] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett*, vol. 10, no. 7, pp. 204–207, Jul. 2003.
[40] K. Kokkinakis and A. K. Nandi, "Speech modelling based on generalized Gaussian probability density functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 1, pp. 381–384.
[41] J. S. Garofolo *et al.*, *TIMIT: Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[42] "The world's premier supplier of speech and audio codec-voiceage," 2015. [Online]. Available: http://www.voiceage.com/AMR-NB.AMR.html
[43] "Mobile telecommunications-radio-electronics," 2015. [Online]. Available: http://www.radio-electronics.com/info/cellulartelecomms.php
[44] Q. Cheng, "Generalized embedding of multiplicative watermarks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 978–988, Jul. 2009.
[45] S. T. Yong feng Huang, and Y. J. Y. Chunlai Bao, "Steganalysis of compressed speech to detect covert VoIP channels," *IET Inf. Security*, vol. 5, no. 1, pp. 1–7, Mar. 2011.
[46] Y. Huang, S. Tang, and Y. Zhang, "Detection of covert voice-over internet protocol communications using sliding window-based steganalysis," *Commun. IET*, vol. 5, no. 7, pp. 929–936, 2011.

Reza Kazemi was born in Ilam, Iran, in 1986. He received the B.S., M.S., and Ph.D. degrees in communication systems from the Sharif University of Technology, Tehran, Iran, in 2008, 2010, and 2015, respectively.

His research interests include watermarking, steganography, information forensics, DoV, and M2M communication.

Fernando Pérez-González (S'98–M'94–SM'09–F'16) received the Telecommunication Engineer degree from the University of Santiago, Santiago, Spain in 1990, and the Ph.D. degree in telecommunications engineering from the University of Vigo, Vigo, Spain, in 1993.

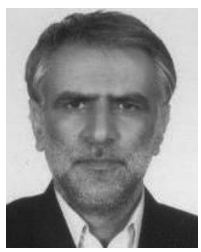
In 1990, he became an Assistant Professor with the School of Telecommunication Engineering, University of Vigo. From 2007 to 2010, he was Program Manager of the Spanish National R&D Plan on Electronic and Communication Technologies, Ministry of Science and Innovation. From 2009 to 2011, he was the Prince of Asturias Endowed Chair of Information Science and Technology, University of New Mexico, Albuquerque, NM, USA. From 2007 to 2014, he was the Executive Director of the Galician Research and Development Center in Advanced Telecommunications. He has been the Principal Investigator of the University of Vigo Group, which participated in several European projects, including CERTIMARK, ECRYPT, REWIND, NIFTY, and WITDOM. He is currently a Professor in the School of Telecommunication Engineering, University of Vigo, Vigo, Spain, and a Research Professor in Information Science and Technology, University of New Mexico, Albuquerque, NM, USA. He has coauthored more than 50 papers in leading international journals and 160 peer-reviewed conference papers. He has coauthored several international patents related to watermarking for video surveillance, integrity protection of printed documents, fingerprinting of audio signals, and digital terrestrial broadcasting systems. His research interests include the areas of digital communications, adaptive algorithms, privacy enhancing technologies, and information forensics and security.

Prof. Pérez-González was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (2005–2009) and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (2006–2010). He is currently is an Associate Editor of the LNCS *Transactions on Data Hiding and Multimedia Security*, and the EURASIP *International Journal on Information Forensics and Security*.

Mohammad Ali Akhaee (S'07–M'07) received the B.Sc. degree in electronics and communications engineering from the Amirkabir University of Technology, Tehran, Iran, and the M.Sc. and Ph.D. degrees from the Sharif University of Technology, Tehran, Iran, in 2005 and 2009, respectively.

He is currently an Assistant Professor with the College of Engineering and the Director of the Secure Communication Laboratory, University of Tehran, Tehran, Iran. He has authored or coauthored more than 50 papers, and holds one Iranian patent. His research interests include the area of signal processing, in particular multimedia security, watermarking, and statistical signal processing.

Prof. Akhaee was the Technical Program Chair of EUSIPCO '11 and the Executive Chair of ISCISC '14. He received the Governmental Endeavour Research Fellowship from Australia in 2010.



Fereydoon Behnia was born in Tarom, Iran, in 1958. He received the B.Sc., M.Sc., and Ph.D. degrees from the Sharif University of Technology, Tehran, Iran, in 1985, 1987, and 1997, respectively.

Since 1988, he has been with the Electrical Engineering Department, Sharif University of Technology.