This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2009 proceedings

Robust Multiplicative Audio and Speech Watermarking Using Statistical Modeling

Mohammad Ali Akhaee
Department of Electrical Engineering
Sharif University of Technology

Nima Khademi Kalantari
Department of Electrical Engineering
Amirkabir University of Technology

Farokh Marvasti
Department of Electrical Engineering
Sharif University of Technology

***Abstract*—In this paper, a semi-blind multiplicative watermarking approach for audio and speech signals has been presented. At the receiver end, the optimal Maximum Likelihood (ML) detector aided by the channel side information for Gaussian and Laplacian signals in noisy environment is designed and implemented. The performance of the proposed scheme is analytically calculated and verified by simulation. Then, we adapt the proposed scheme to speech and audio signals. To improve robustness, the algorithm is applied to low frequency components of the host signal. Besides, the power of the watermark is controlled elegantly to have inaudibility using Perceptual Evaluation of Audio Quality (PEAQ) and Perceptual Evaluation of Speech Quality (PESQ) algorithms. Experimental results over several audio and speech signals show the higher robustness of the proposed technique in comparison with a recent watermarking scheme.**

***Index Terms*—Audio and speech signals, Maximum likelihood detector, Multiplicative watermarking, Inaudibility control.**

I. Introduction

Among several approaches presented so far, the multiplicative watermarking technique has been widely studied in recent years [2]–[4]. The main advantage of the multiplicative watermarking is its adaptation to the Human Visual System (HVS) and Human Auditory System (HAS). It is known that a disturbance in proportion to the host signal features is very difficult to perceive [1]. The multiplicative method and its extension to multi bit version have been proposed in [2] and [3], respectively. Cheng and Huang, presented optimum decoders for multiplicative watermarking in Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT) transform domains [4]. The distribution of the DWT coefficients was modeled by Generalized Gaussian Distribution (GGD) while the DFT coefficients were considered as Weibull distribution.

Nearly all of the aforementioned multiplicative watermarking schemes embed a logo within the original signal. The decoder should make a binary decision whether the received signal contains the logo or not. In fact, these methods are useful in signature verification applications. In some applications, the problem is not restricted to the logo verification. Thus, the decoder should be able to extract the watermark data from the received signal. In the latter case, the watermark serves as a transmission code and the decoder's task is more complex than the former [5]. The proposed method belongs to the second category.

In this paper, a multiplicative embedding method, proposed in [6], is used to embed watermark data into the speech and audio signals. To obtain more transparency and robustness, the watermark bits are embedded by scaling the amplitude of the low frequency parts of the host signal. In order to smooth the watermarked signal in each frame boundary, a specific window is used to modify the coefficients. Besides, in order to control the inaudibility of the watermark insertion, an iterative scheme, aided by PEAQ [7] and PESQ [8], is used to achieve the desired quality. At the receiver end, for maximizing the robustness, the detector is optimized under the Additive White Gaussian Noise (AWGN) by considering the speech and audio signals to follow Laplacian and Gaussian distribution [9] and [10]. The requirements of the proposed decoder are the variance of each frame of the host signal as well as the strength factor. By using this side information, the noise variance can be estimated during the detection process. Furthermore, the probability of error is analyzed exactly for the Gaussian host signal; while for the Laplacian one we calculate it approximately for low and large Signal to Noise Ratios (SNR). Note that proposing a method which can resist against desynchronization attacks is not the scope of this paper. To this aim, algorithms such as [12] should be combined with the proposed method.

II. System Modeling

According to [9] the speech signal is modeled with Laplacian function defined as:

$$f(x) = \frac{b}{2} \exp\left(-\frac{|x-\mu|}{b}\right) \tag{1}$$

where b is calculated as $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{|x_i|}{n}$ and μ is the mean of the random variable. As the mean value in the speech signals is approximately zero, we use the zero mean Laplacian function as the probability density function.

The noisy watermarked speech signal received at the decoder can be denoted as $y = x + n$, where x is the watermarked speech signal with Laplacian distribution and n is the channel noise. The noise is assumed to be WGN. In order to to optimize our decoder for the received signal, the distribution of y must be calculated. Since the received signal is the summation of the watermarked speech and the channel noise, regarding the probability theory, the distribution of the received signal is achieved by the convolution of two

978-1-4244-3435-0/09/$25.00 ©2009 IEEE

distribution functions. Thus, the distribution function of y is as follows:

$$f_y(y) = \frac{b}{2\sqrt{2\pi\sigma_n^2}} \int_{-\infty}^{\infty} e^{-\frac{-n^2}{2\sigma_n^2}} e^{-b|y-n|} dn \tag{2}$$

Using the error function $\mathcal{Q}(.)$, defined in communication literature, the probability density function of the received signal can be obtained as:

$$f_y(y) = \frac{b}{2} e^{\frac{b^2\sigma_n^2}{2}} \left[e^{by} \mathcal{Q}(\frac{y + b\sigma_n^2}{\sigma_n}) + e^{-by}\Big(1 - \mathcal{Q}(\frac{y - b\sigma_n^2}{\sigma_n})\Big)\right] \tag{3}$$

III. Proposed Method

A. watermark embedding

Here, the watermark is embedded by adjusting the amplitude of the host signal samples depending on the message bit similar to [6]. In this manner, the host signal is segmented into non-overlapping frames with the length of N. Then, the watermark data is simply embedded by scaling the amplitude of each frame as follows:

$$\begin{aligned} x_i' &= \alpha \times x \qquad \text{For embedding 0} \\ x_i' &= \frac{1}{\alpha} \times x \qquad \text{For embedding 1} \end{aligned} \tag{4}$$

where α is the watermark strength factor that should be slightly greater than one. Larger value for α results in more robustness while reduces the perceptual quality of the watermarked signal.

B. watermark extraction

Decoder extracts the watermark data using the host signal variance, noise variance and watermark strength factor for each frame. Thus, our decoder acts in semi-blind way. Suppose that $x_{i,k}$ represents the ith host signal sample in the kth frame and $y_{i,k}$ is the corresponding received signal. The watermarked signal is multiplied or divided by the strength factor depending on the watermark data. Therefore, for zero bit embedding, we have:

$$y_{i,k}|0 = \alpha^{-1}x_{i,k} + n_i \tag{5}$$

$$\begin{aligned} f(y_{i,k}|0) = \frac{\alpha^{-1}b_k}{2} e^{\frac{\alpha^{-2}b_k^2\sigma_n^2}{2}} \Bigg[& e^{\alpha^{-1}b_k y} \mathcal{Q}(\frac{y + \alpha^{-1}b_k\sigma_n^2}{\sigma_n}) \\ & + e^{-\alpha^{-1}b_k y} \mathcal{Q}(\frac{y - \alpha^{-1}b_k\sigma_n^2}{\sigma_n}) \Bigg] \end{aligned} \tag{6}$$

where n_i is WGN. $y_{i,k}|1$ and $f(y_{i,k}|1)$ are obtained easily by replacing α^{-1} with α in (6). By assuming that y is an iid sequence, we have

$$f(y_{1,k}, ..., y_{N,k}|0) = \prod_{i=1}^{N} f(y_{i,k}|0)$$

Similar equation is correct for $f(y_{1,k}, ..., y_{N,k}|1)$. Then, the ML decoder can extract the watermark bits '0' or '1' using the following hypothesis test

$$\prod_{i=1}^{N} f(y_{i,k}|0) \gtrless_1^0 \prod_{i=1}^{N} f(y_{i,k}|1) \tag{7}$$

The value of the $\mathcal{Q}$ function is obtained quickly from a lookup table and hence the detector can be implemented in real time. The ML detector for the Gaussian host signal in the Gaussian noisy enviornment can be described similarly as

$$\sum_{i=1}^{N} y_i^2 \gtrless_0^1 \frac{2\sigma_{y|0}^2\sigma_{y|1}^2 N \ln(\frac{\sigma_{y|0}}{\sigma_{y|1}})}{\sigma_{y|1}^2 - \sigma_{y|0}^2} \tag{8}$$

where $\sigma_{y|0}^2 = \alpha^{-2}\sigma_x^2 + \sigma_n^2$ and $\sigma_{y|1}^2 = \alpha^2\sigma_x^2 + \sigma_n^2$.

IV. Performance Analysis

A. Laplacian host

Since there is no closed-form for the probability of error in general, we solve the problem in two cases where a closed-form statistical model can be found for it.

1) Low noise variance: The received signal distribution for low noise variance is analogous to the Laplacian distribution, since the host signal dominates in this case. In the low noise variance the ML decoder is as follow:

$$\sum_{i=1}^{N} |y_i| \gtrless_0^1 \frac{2N\alpha \cdot \ln(\alpha)}{b_k(\alpha^2 - 1)} \tag{9}$$

In order to calculate the probability of error, we have to find the distribution of two conditional random variables defined as

$$z|0 = \sum_{i=1}^{N} |\alpha^{-1}x_i + n_i|, \quad z|1 = \sum_{i=1}^{N} |\alpha x_i + n_i|$$

Assuming equal probability for zero and one, the probability of error can be calculated as follows:

$$P_e = \frac{1}{2}\left[1 - \mathcal{Q}(\frac{T - N\alpha b}{\sqrt{N\sigma_n^2}}) + \mathcal{Q}(\frac{T - Nb/\alpha}{\sqrt{N\sigma_n^2}})\right] \tag{10}$$

where

$$T = \frac{2N\alpha\ln(\alpha)}{b(\alpha^2 - 1)}$$

2) Large noise variance: In the case of large noise variance, since noise dominates, the distribution of the recieved signal, is similar to the Gaussian distribution. Thus, we should obtain the ML decoder by considering Gaussian distribution for the received signal. This will result in equation (8).

Using Central Limit Theorem (CLT) and by defining $\mu_\alpha = 2N\alpha^2b^2 + N\sigma_n^2$, $\sigma_\alpha^2 = 8N\alpha^2\sigma_n^2b^2 + 2N\sigma_n^4$, $\mu_{1/\alpha} = 2N\alpha^{-2}b^2 + N\sigma_n^2$, and $\sigma_{1/\alpha}^2 = 8N\alpha^{-2}\sigma_n^2b^2 + 2N\sigma_n^4$, the probability of error can be calculated as

$$P_e = \frac{1}{2}\left[1 - \mathcal{Q}(\frac{T' - \mu_\alpha}{\sigma_\alpha}) + \mathcal{Q}(\frac{T' - \mu_{\frac{1}{\alpha}}}{\sigma_{\frac{1}{\alpha}}})\right] \tag{11}$$

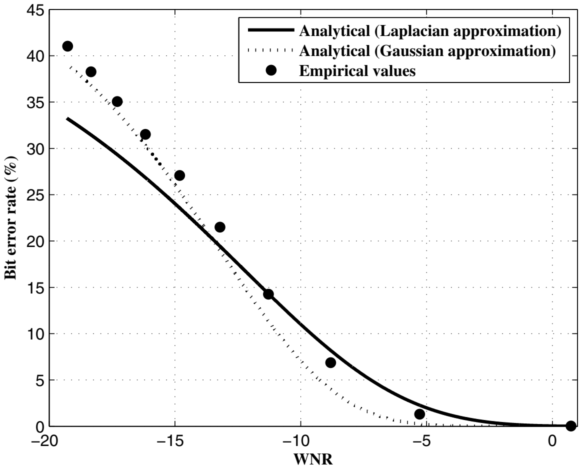

Fig. 1. Emperical and analytical values of Bit Error Rate (BER) for Laplacian host under AWGN attack (DWR=10dB)

where

$$T' = \frac{2\sigma_{y|0}^2\sigma_{y|1}^2 N ln(\frac{\sigma_{y|0}}{\sigma_{y|1}})}{\sigma_{y|1}^2 - \sigma_{y|0}^2}$$

B. Gaussian host

The same procedure as the case of Laplacian host with large noise variance can be used to derive the probability of error. In this case, the error probability can be calculated with the same approach like (11) by replacing $2b^2$ with σ_x^2.

V. Performance Evaluation

In order to evaluate the system performance, the proposed method for both kinds of host signals (Laplacian and Gaussian) is simulated. We generate a Gaussian host signal combined with AWGN. The Laplacian host is also generated using the Laplacian random variables artificially. We embedded one bit in each frame with the length of 32 samples. The results are obtained in different Watermark to Noise Ratios (WNR), defined as:

$$\text{WNR} = 10 \log \frac{E\|X' - X\|^2}{E\|Y - X'\|^2} \tag{12}$$

where X is the host signal vector, X' is the watermarked one, and Y is the received signal vector, respectively. The Document to Watermark Ratio (DWR) defined below is fixed for each plot.

$$\text{DWR} = 10 \log \frac{E\|X\|^2}{E\|X' - X\|^2} \tag{13}$$

The empirical value for each WNR is obtained after averaging the results over 100 simulations with 32000 bits. Moreover, all the simulations have been performed under the assumption that the decoder is aware of the power of the attack.

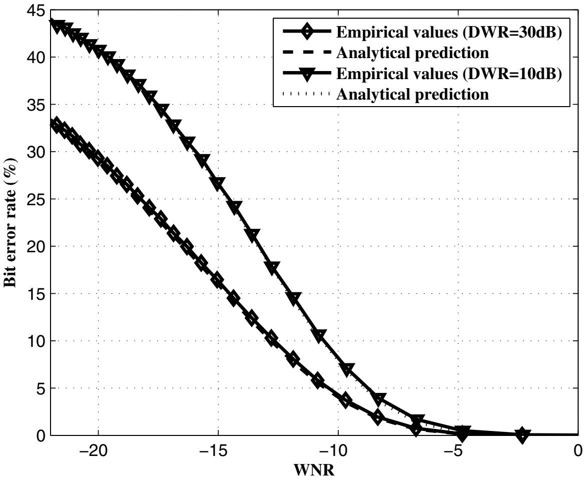

Fig. 2. Empirical and analytical values of BER for Gaussian host under AWGN attack.

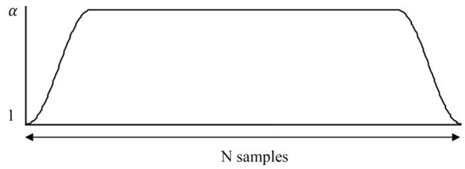

Fig. 3. The scaling window used in the watermarking process.

Fig.1 represents the analytical results and empirical results for the Laplacian host. In Fig.1 the DWR is fixed at 10dB. The results are shown in linear scale for better illustration of the differences between Gaussian and Laplacian approximations. Since DWR=10 dB , SNR which is equal to WNR+DWR falls into the range (10.5dB,-9.5dB). As we can see in Fig.1, the analytical prediction using Laplacian approximation is good in low WNRs. In high WNRs the Gaussian approximation is close to the empirical results whereas the Laplacian approximation fails.

Fig.2 shows the empirical results and analytical prediction for the Gaussian host signal. Simulations have been performed with two different DWRs. Since there is no approximation in calculation except for the usage of CLT, which is a precise approximation due to the large frame length, the error probability and the analytical prediction for both DWRs are well agreed.

VI. Requirements of Practical Implementations

Since the proposed algorithm scales the amplitude of each frame, this will cause discontinuity in the frame boundaries which results in an audible distortion. Thus, we use a specific window, shown in Fig.3, to smooth the discontinuities near the frame boundaries. Furthermore, in order to have both inaudibility of watermark insertion and robustness to high pass attacks, we use a bandpass filter to extract the middle-low frequency of the host signal and embed the watermark data into this part. Then, the remainder will be added to this part for constructing the watermarked signal.

In order to achieve maximum strength of watermark embedding and also controlling the inaudibility of the watermark data, the value of the strength factor should be changed over the time regarding the host signal specifications. To this aim, we use the PESQ [8] and PEAQ [7] for the speech and audio host signals, respectively. Both of these algorithms use some features of both reference and test audio signals and represents the quality in the form of Objective Difference Grade (ODG). ODG values are between -0.5 and 4.5 for PESQ and between 0 and -4 for PEAQ. Higher ODG values show more perceptual similarity of the reference and the test signals. In order to achieve maximum watermark strength, the strength factor is updated regarding the ODG value every I frames. The watermarked signal is obtained by considering an initial value (near to 1) for α. Then, the ODG value is attained by PEAQ algorithm. If the difference of desired ODG and obtained ODG is smaller than a specific threshold, the algorithm is stopped. Otherwise, α is increased and the above process is repeated until the stop condition is satisfied.

The the noise variance is also reqiured for the proposed ML decoder as well as the variance of the host signal. Fortunately, since the variance of the host signal is available at the decoder, the noise variance can be cumulatively estimated by the following procedure.

$$\mathcal{C}_1 = \sum_{i=1}^{N} y_i^2(k) - \alpha^2(k)\sigma_x^2(k) - \hat{\sigma}_n^2(k-1)$$

$$\mathcal{C}_0 = \sum_{i=1}^{N} y_i^2(k) - \alpha^{-2}(k)\sigma_x^2(k) - \hat{\sigma}_n^2(k-1)$$

$$\hat{\sigma}_n^2(k) = \hat{\sigma}_n^2(k-1) + \frac{\min(\mathcal{C}_1, \mathcal{C}_0)}{k} \qquad (14)$$

where k is the frame index and $\hat{\sigma}_n^2$ represents the estimated noise variance. In order to perform this process on the first frame, $\mathcal{C}_1$ and $\mathcal{C}_0$ are computed for the first and second frames assuming $\hat{\sigma}_n^2(0)$ and $\hat{\sigma}_n^2(-1)$ to be zero. Then, among the two values found for each frame, similar ones across the frames are considered as the noise variance of the first frame.

For Laplacian host, we have b instead of σ_x^2 at the receiver side. We can not use the above equation in this form. Thus, the noise variance is estimated for Laplacian host by substituting $2b^2$ instead of σ_x^2 and replacing $\sum_{i=1}^{N} y_i^2(k)$ with $2(\sum_{i=1}^{N} |y_i(k)|)^2$ in (14), respectively.

Note that in the case of some other attacks such as MP3 compression or filtering, the estimated noise variance is a negative value which for the best performance we set it to zero.

VII. EXPERIMENTAL RESULTS ON REAL SIGNALS

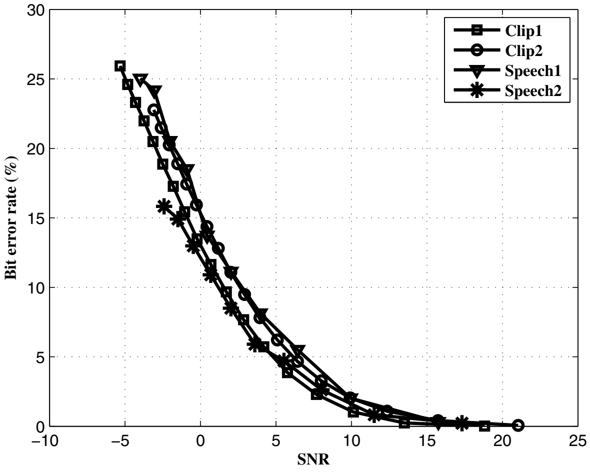

Fig. 4. BER(%) vs. SNR(dB) of the audio and speech signals under AWGN attack.

In order to evaluate the performance of the proposed technique in real condition, simulations are performed on different types of audio signals including pop, jazz, classic, folk and also different types of speech signals (male and female). In addition, to illustrate the results we selected two audio clips denoted by clip1 and clip2 and also two speech signals denoted by speech1 and speech2. Audio clips are mono, sampled at 44.1 KHz, quantized with 16 bits and both of them with the length of two minutes. Moreover, speech signals are mono with the sampling rate equal to 16 KHz, quantized with 16 bits and each of which with the length of approximately 30 seconds. The length of each frame is set to 256 for both of the speech and audio signals. Therefore, a watermark data rate of about 172 bps for audio signal and 40 bps for speech is achieved. To simultaneously have better resistance against attacks and imperceptibility, we embed the watermark in the low frequency components of the host signal. To this aim, a sharp bandpass filter with cut off frequencies at 150 Hz and 1.50 KHz has been used for filtering the original audio signal, while the passband of the respective filter for speech signals is set to [125 Hz, 625 Hz]. The desired ODG was selected as -0.3 for audio signals and 4.2 for speech signals to ensure inaudibility of watermark data insertion. In this way the strength factor gets updated every 34 (200ms) and 62 (one second) frames for audio and speech signals, respectively. As side information, the Laplacian parameter (b) or the variance of the filtered speech or audio signal is estimated. Since the decoder needs these values with high precision, these values are assigned 12 bits. Thus, for a 1*sec* block of speech or audio signal $40 \times 12 = 480$ or $172 \times 12 = 2064$ bits are allocated, respectively. On the contrary, only four bits are adequate for storing each strength factor which sums up to 20 or 4 bits for 1*sec* of speech or audio signals. The total side information after compression and scrambling is sent as a header along with the watermarked signal. This header occupies less than 0.2% of the size of the watermarked signal for speech and 0.3% for audio signals. It is noteworthy to mention that in the data extraction process, to satisfy the i.i.d. assumption used in (7), the watermarked samples are decimated by a factor of 4. This process also reduces the complexity of the proposed decoder. Fig.4 illustrate the BER of the detection

TABLE I
BER(%) PERFORMANCE AFTER SOME COMMON ATTACKS FOR THE AUDIO SIGNALS.

Attacks	Clip1	Clip2
Lowpass (2kHz)	0	0
Highpass (50Hz)	0	0
Requantization (16 to 8)	0	0
Resampling (44/11/44)	0	0
Resampling (44/6/44)	0	0
Mp3 96kbps	0	0.02
Mp3 64kbps	0.09	0.12
Mp3 32kbps	0.62	1.03

TABLE II
BER(%) PERFORMANCE AFTER SOME COMMON ATTACKS FOR SPEECH SIGNALS.

Attacks	Speech1	Speech2
Lowpass (2kHz)	0	0
Highpass (50Hz)	0	0
Requantization (16 to 8)	0	0
Resampling (16/8/16)	0	0
Resampling (16/6/16)	0.15	0

TABLE III
COMPARISON OF OUR METHOD WITH [11] FOR SOME COMMON ATTACKS

Attacks	Proposed method	Chen *et.al* [11]
Closed loop	0	0.9
Noise (20dB)	0	22.8
Requant. (16 to 8)	0	11.9
Downsampling (2)	0	6.4
Mp3 64kbps	0.03	6.5
BPF (100Hz-8000Hz)	0	4.2
Echo (40% and 100 ms)	4.35	1.6

process after adding white Gaussian noise to the watermarked signal. The results are averaged over 50 runs for each SNR. As seen in this figure, the proposed scheme has outstanding robustness against noise attack for both of the speech and audio signals. This good robustness is due to optimizing the decoders for noisy environment and also choosing appropriate strength factor using PEAQ and PESQ.

Besides, the proposed scheme is tested for audio and speech signals after applying common attacks such as MP3 compression, resampling, requantization, filtering, etc. The results are summarized in Tables I and II. Table I represents the BER for common audio attacks and Table II shows the results after the speech attacks. As it can be seen, the suggested method is highly robust in both cases.

Furthermore, we compared the proposed method with Chen and Wu method which is a low payload one [11]. They used 15 audio and 3 speech signals to obtain the final results. In order to have a fair comparison, we obtained the result after averaging all the results for 15 audio clips and 6 pieces of speech signals. Moreover we increased our frame length to reach their data rate. Table III demonstrates the BER of the proposed technique with Chen and Wu method . The ODG of our algorithm is 4.2 for speeches and -0.3 for audio signals while in [11] it is about -0.6. Nevertheless, the proposed method has outstanding robustness in comparison with their algorithm.

VIII. CONCLUSION

We have presented a new multiplicative watermarking method which is suitable for both speech and audio signals. The embedding process is performed on the low frequency components of the host signal. Thus, the proposed technique is inherently robust to compression and filtering attacks. In order to insert the maximum watermark power while keeping the imperceptibility, we have used PESQ and PEAQ to optimize the strength factor. In detection process, the maximum likelihood detector is optimized utilizing side information. The side information consists the statistical parameter of each frame and the strength factor while occupies only 0.2% of the host signal size. Since the the noise variance is also required for our detector, we estimate it cumulatively using the side information. The probability of error is analytically calculated and verified by artificially simulation on Laplacian and Gaussian signals. Extensive simulations over speech and audio signals indicate that the proposed algorithm has greater robustness against common attacks than the recently proposed algorithm.

REFERENCES

[1] M. Barni, C.I. Podilchuk, F. Bartolini, and E.J. Delp, “Watermark embedding: Hiding a signal within a cover image,” *IEEE Commun. Mag.*, vol.39, no.8, pp. 102-108, 2001.

[2] M. Barni, F. Bartolini, A. D. Rosa, and A. Piva, “A new decoder for the optimum recovery of nonadditive watermarks,” *IEEE Trans. Image Processing*, vol. 10, pp. 755-766, May 2001.

[3] M. Barni, F. Bartolini, A. D. Rosa, and A. Piva, “Optimum decoding and detection of multiplicative watermarks,” *IEEE Trans. on Signal Process.*, vol. 51, no. 4, pp. 1118-1123, 2003.

[4] Q. Cheng and T. S. Huang, “Robust optimum detection of transform domain multiplicative watermarks,” *IEEE Trans. signal Processing*, vol. 51, no. 4, pp. 906-924, 2003.

[5] P. Moulin and R. Koetter, “Data-hiding codes,” *Proceedings IEEE*, vol. 93, No. 12, pp. 2083-2127, Dec. 2005.

[6] S. M. E. Sahraeian, M. A. Akhaee, B. Sankur, and F. Marvasti, “Robust multiplicative watermarking technique with maximum likelihood detector,” *to be published in the European Signal processing Conference (EUSIPCO)*, Lausanne, Switzerland, 2008.

[7] *Recommendation B.S. 1387: Method for Objective Measurements of Perceived Audio Quality*, Int. Telecommunication Union, Geneva, Switzerland, 2001.

[8] *Perceptual Evaluation of Speech Quality (PESQ), An objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs*, ITU-T Recommendation P.862, Feb. 2001.

[9] S. Gazor and W. Zhang, “Speech probability distribution,” *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204-207, July 2003.

[10] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete- Time Processing of Speech Signals*, 2nd edition, IEEE Press, 2000.

[11] O.T.-C. Chen, W. -C. Wu, “Highly robust, secure, and perceptual-quality echo hiding scheme,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 629-638, March 2008.

[12] K. Solanki, N. Jacobsen, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, “Robust image-adaptive data hiding based on erasure and error correction,” *IEEE Trans. Image Process.*, vol. 13, no. 12, pp. 1627–1639, Dec. 2004.