Using Raw Speech as a Watermark, Does it work?

P. Nintanavongsa and T. Amornraksa
Multimedia Communications Laboratory, Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

Key words: Digital Watermarking, Speech Coding

Abstract: Data piracy has become a great concern in today's multimedia applications since digital data can be reproduced without any quality loss. Digital watermarking is one technique that can be used against piracy. In this paper, we propose an idea of using raw speech data as a watermark signal to enhance the robustness of the watermark. To recover the information in the watermark, the extracted watermark will be played back as speech to a listener, and through the intelligent audio perception of the human listener, the contents of the speech may be recognized. Our approach is based on the fact that raw speech contains a considerable amount of redundancy, therefore, its contents can still be recognized after the extraction process, even if the watermarked data is badly attacked. As long as the raw speech, extracted from the attacked watermarked data, contains enough important information, its contents can be intelligible. Furthermore, the impressive intelligence of the human perceptual system, as it tends to adjust and learn quickly to determine the repeated speech, enhances the probability of recognizing the contents of the extracted raw speech. A set of experiments was carried out to show that the proposed method not only successfully survives the common attacks, but also yields high intelligibility.

1. INTRODUCTION

The rapid growth of digital technology allows data to be readily stored in a digital form. Moreover, the unlimited reproduction with no loss in quality makes the digital data preferable to its analog counterpart. However, reproduction of the copyrighted data without permission from the owner results in data piracy. There are many approaches to protecting this kind of

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35413-2_36
R. Steinmetz et al. (eds.), *Communications and Multimedia Security Issues of the New Century*

piracy, including the encryption of the data prior to. However, the problem is often not completely solved since an authorized user can copy and re-distribute the data after the decryption process. Thus, for digital images, digital watermarking techniques serve as the last line of defense by embedding secret information into the original data in such a way that it is invisible to the Human Visual System (HVS). The embedded watermark data can be any information related to the copyright owner, authorized recipient or purchasing information.

In this paper, we use the intelligibility contained in speech as a watermark signal. When a copyright violation occurs, the speech embedded as a watermark is extracted to determine the related information. The next section provides an overview of related work done by other researchers. Section 3 describes the details of the algorithm used and related principles. The experimental results and analysis are illustrated in sections 4 and 5, respectively. Finally, the conclusions are drawn in section 6.

2. PREVIOUS WORKS

Sakaguchi *et al.* [1] investigated how the polarity inversion of speech signals affects human perception and applied this technique for data hiding. The experimental results showed that the data was successfully hidden and could then be restored automatically. A scheme for hiding a high bit-rate supplementary data was proposed in [2]. In this scheme, the high bit-rate supplementary data such as secondary video, was hidden into a digital video stream by directly modifying the pixels in the video frames. The experimental results showed that the instructional video can be transmitted with four language options to cover a wide range of interested parties.

Recently, Mukherjee *et al.* [3] implemented a scheme for hiding 8KHz speech sampled at 16 bits/sample in a 30 frames/s QCIF video. The speech and video extracted from the compressed video were found to be intelligible, and acceptable for visual quality, respectively, even at high compression ratio. Rhoads [4] described a method that added or subtracted small random quantities from each pixel. The watermark was subtracted by first computing the difference between the original and watermarked images and then by examining the sign of the difference, pixel by pixel, to determine if it corresponded to the original sequence of additions and subtractions.

In [5], two techniques that modified the least significant bits (LSB) of an image, based on the assumption that the LSB data were insignificant, were proposed. However, these two methods were highly sensitive to noise and easily destroyed since only the LSB data are modified. A similar result was obtained in [6], by the method called Patchwork. However, based on the

experimental results, these two techniques showed a lack of robustness to attacks. Kutter *et al.* [7] proposed the method for digitally signing the image using amplitude modulation. In this method, the signature bits were multiply embedded by modifying pixel values in the blue channel. The experimental results showed that this proposed method was immune to a variety of attacks. The main improvement brought by this proposed method is that the watermark can be retrieved regardless of the original, unmarked image.

3. BACKGROUND

Digital watermarking is a technique, which secretly embeds robust and hidden marks into the material to designate its copyright-related information such as the origin, ownership, rights, and destinations. The basic requirements of the digital watermark, as stated in [8] and [9], can be concisely described as following: invisible, undetectable, irremovable, unalterable, unambiguous, and acceptable quality of the original data.

Based on the watermark insertion methods, the watermark can be categorized into 2 types:

- *Spatial watermark*: In this method, the watermark is inserted in the spatial domain; thus, no time-frequency domain transformation is required.
- *Spectral watermark*: When the watermark is inserted in the frequency domain. In this case, some type of time domain-frequency domain is required.

3.1 Characteristics of speech signals

Speech waveforms have a number of useful properties that can be exploited when designing an efficient encoder. Some of the properties that are most often utilized in encoder design include the non-uniform probability of distribution of speech amplitude, the nonzero auto-correlation between successive speech samples, the non-flat nature of the speech spectra, the existence of voiced and unvoiced segment in speech, and the quasi-periodicity of voiced speech signals. The most basic property of the speech waveforms that is exploited by all speech coding is that they are band limited. A finite bandwidth means that it can be sampled at a finite time interval and reconstructed completely from these samples, provided that the sampling frequency is greater than twice the highest frequency component in the low pass signal [10]. While the band limited property of the speech signal makes sampling possible, the aforementioned properties allow

quantization, another important process in speech coding, to be performed with greater efficiency.

3.2 Watermarking algorithm

A watermarking algorithm based on [7] is used in this paper. The algorithm embeds the information into the blue channel, which is the one the human eye is least sensitive to. Furthermore, changes in regions of high frequency and high luminance are less perceptible, and thus favored. However, we make some modifications to the original scheme. The algorithm is described in this section.

Let (x, y) be the coordinate corresponding to the position within image $I = (R, G, B)$. The watermark bit is embedded by modifying the blue channel at the position p by a fraction of the luminance $L = 0.299R + 0.587G + 0.114B$ as follow.

$$B_{xy} = \begin{cases} B_{xy} + L_{xy}\alpha & ; \ \textit{when watermark bit is } 1 \\ B_{xy} - L_{xy}\alpha & ; \ \textit{when watermark bit is } 0 \end{cases} \tag{1}$$

where α is a constant determining the signature strength.

The value of α is selected to offer the best trade-off between robustness and invisibility. In order to recover the embedded bit, a prediction of the original value of the pixel containing the information is needed. This prediction is based on a linear combination of pixel values in a neighborhood around considered positions. Empirical results show that taking a cross-shaped neighborhood gives the best performance. The prediction B'_{xy} is thus computed as follows:

$$B'_{xy} = \frac{1}{4c}\left(\sum_{k=-c}^{c} B_{x+k,y} + \sum_{k=-c}^{c} B_{x,y+k} - 2B_{xy}\right) \tag{2}$$

where c is the size of the cross-shaped neighborhood.

To retrieve the embedded bit, the difference β between the prediction and the actual value of the pixel at position (x, y) is taken:

$$\beta = B_{xy} - B'_{xy} \tag{3}$$

The sign of the difference β determines the value of the embedded bit. That is, if β has a negative value then the embedded bit is zero and vice

versa. The embedding and the retrieval functions are not symmetric; that is, the retrieval function is not the inverse of thee embedding functions. Although correct retrieval is very likely, it is not guaranteed. To enhance the probability of correct retrieval, the bit is embedded several times, as described below.

To improve the retrieval performance, the bit can be embedded n times at different locations. The embedded position is such that the bit is sequentially embedded bit-by-bit into the original image pixel-by-pixel. The information is embedded into the pixel sequentially until the last bit of the information is embedded, then the first bit of information is re-embedded into the pixel position next to where the information bit was previously embedded.

The bit retrieval can be improved by computing the difference between the prediction and the value of the pixel for each embedded position k similar to (3). These differences are then averaged as shown in (4).

$$\beta = \frac{1}{n}\sum_{k} \beta_k \tag{4}$$

where is a number of positions that each information bit is embedded.

3.3 Mode of Simulation

Our proposed method exploits the fact that ordinary speech has considerable redundancy [11]. That is, even though parts of the raw speech are corrupted, the intelligibility of the remaining uncorrupted speech can still be perceptively recognized. The speech encoders reflect this fact as a wide range of speech quality can be provided. Note that less bit-rate speech normally gives lower quality of speech. In this paper, the content of speech is used as a watermark signal instead of its signal's characteristics, to increase the robustness of watermarking scheme. The proposed method is distinct from many existing watermarking methods since, in our method, the watermark signal is the intelligent information contained in the speech, which most people can easily recognize through auditory perception.

In the experiments, we used the Pulse Code Modulation (PCM) [12] technique for digitizing the speech input from a microphone. PCM is a widely used technique since it encodes each sample of the input waveform independently from all other samples. Therefore, the technique is inherently capable of encoding an arbitrarily random waveform. However, since it encodes every sample independently without regard to the correlation between each sample, a considerable amount of redundancy will be contained in each successive sample. Flanagan *et al.* [13] stated that the

correlation coefficient between adjacent 8KHz samples is generally 0.85 or higher, and hence the high redundancy in the PCM codes can be used to provide more strength in the watermarking scheme. In fact, at 8KHz sampling bit-rates, the significant correlations exist for upto two to three samples distances. In addition, the samples become even more correlated if the sampling rate is increased.

The 11KHz raw speech of the word *"Thank you"*, which is encoded by the PCM is chosen to be used as a watermark signal. The model of the watermarking process will be that the raw speech is embedded into the original color image with the amplitude modulation in the blue channel to obtain the watermarked image. The algorithm to be used to embed the watermark signal is similar to [7] while the attacks to be applied to the watermarked image are brightness and contrast enhancements, blurring, Gaussian noise, JPEG encoding/decoding, high-pass filtering and equally distributed random noise. After attacking, the corrupted raw speech as an embedded watermark signal is extracted and then played back by a speech decoder to observe its corrupted contents. The 10 English native speakers, 5 male and 5 female, with ages between 23-54 year-old are chosen to listen to the extracted corrupted speech. Different versions of the corrupted speech will be randomly played back to a listener. After the listener hears the corrupted speech, he/she will tell what is recognized in the speech, that is, to indicate that they hear the phrase *"Thank you"*. The format of the tests is recommended in [14] and [15]. The model of the watermarking scheme is illustrated in the figure below.

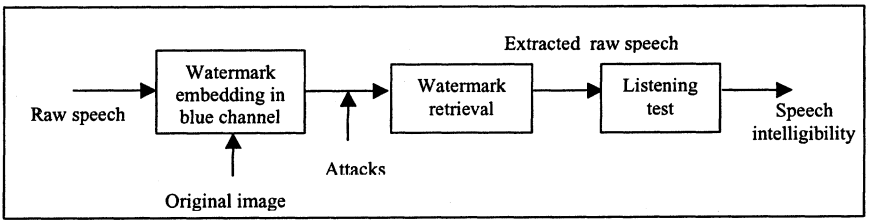
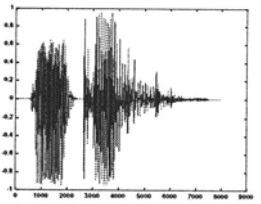


Figure 1. Model of the watermarking scheme

4. EXPERIMENTAL RESULTS

In the experiment, the color image, 512 by 512 pixels and 24-bit color, of Lena taken from [16] is used as the original image. Then the image undergoes the watermark embedding process previously described with the trade-off factor of $\alpha = 0.2$. The size of the cross-shape window, c, is set to 3

to embed the raw speech into the blue channel. The raw speech is embedded 4 times in different locations in the blue channel to enhance the probability of efficient retrieval. The simulation of watermark attack is then performed on the watermarked image. The raw speech to be embedded is 11KHz sampling frequency, 8-bit per sample. The original image and the original waveform of the speech are shown in figure 2.

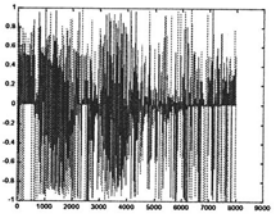


(a) *(b)*

Figure 2. a) The original image 'Lena' b) the original waveform of the speech

4.1 Brightness and contrast enhancement

Figure 3 illustrates an example of different versions of watermarked image after brightness enhancement is performed. In a), the watermarked image is brightness enhanced by 40% from the watermarked image and the extracted speech is shown in b).

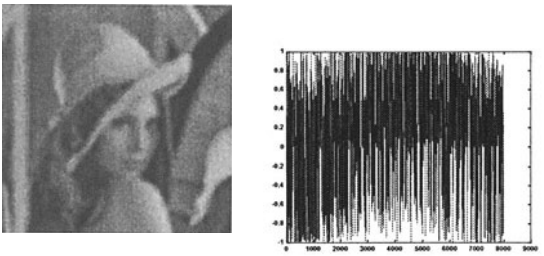


(a) *(b)*

Figure 3. a) The watermarked image of 'Lena' with brightness enhancement b) the speech waveform with 40% brightness enhancement

4.2 Blurring

Figure 4 shows another example of the watermarked images after blurring. The watermarked image is blurred by the factor of 5, which correspond to the number of neighboring pixels involved i.e. neighborhood of averaging filter = 5. The waveform of the extracted speech after blurring is also shown below.



(a) (b)

Figure 4. a) The watermarked image of 'Lena' after blurring by the factor of 5 b) the speech waveform after blurring by the factor of 5

In the experiments, we also performed three common attacks to the watermarked image, namely Gaussian additive noise adding, JPEG encoding and high-pass filtering. That is, the Gaussian additive noise is added into the watermarked image by different amount, 15% and 30% of the watermarked image. In JPEG encoding, the JPEG encoding cycle is performed on the watermarked image. Two different degrees of image quality, 70% quality and 40 % quality, are individually applied to the watermarked image. The final attack that we tested is by applying the high-pass filter with the radius of 10 pixels to the watermarked image and then observe whether the embedded watermark can still survive this type of attack. The experimental results are shown in the table below.

Table 1. Number of people who can recognize the word "Thank you"

Type of attack applied to watermarked image	Number of people
40% of Brightness enhancement	10
40% of Contrast enhancement	10
Blurring by factor of 1	10
Blurring by factor of 5	0
Gaussian additive noise adding by 15%	10
Gaussian additive noise adding by 30%	0
JPEG cycle by 70% quality	10
JPEG cycle by 40% quality	8
High-pass filtering	10

4.3 Equally distributed random noise adding

To further illustrate the effectiveness of our proposed technique, the equally distributed random noise is added to the image at various levels to simulate the attack. The different versions of extracted corrupted speech were played back to the listeners, and the listening tests'results are shown in table 2.

Table 2. Number of people who can recognize the word *"Thank you"*

Noise adding level (%)	10	20	30	40	50	60	70	80	90	100
Number of people	10	10	10	10	10	10	8	7	5	3

5. EXPERIMENTAL ANALYSIS AND DISCUSSIONS

From the table 1, each person could recognize what was saying in the corrupted speech when the watermarked image passed through brightness enhancement. The same result was also achieved when contrast enhancement was applied to the watermarked image since all the participants precisely determined the content of the extracted speech. Moreover, when the watermarked image was blurred by a factor of 1, the embedded content can be correctly determined. However, when the watermarked image was severely attacked after blurring by a factor of 5, none of participants could determine the content of the extracted speech. The Gaussian additive noise attack also yielded similar results since the extracted speech could be precisely determined for the Gaussian additive noise added by 15% of image area. On the contrary, the content of the extracted speech could not be determined from the watermarked image with 30% added Gaussian additive noise. The experimental results also showed that it could impressively survive the brightness and contrast enhancement. Furthermore, the extracted speech after applying high-pass filter was highly perceptible since all participants recognized the content of the extracted speech. The JPEG encoding/decoding has the effect on the extracted speech, as the perceptibility of the extracted speech is proportional to the encoded image quality. That is, the lower the quality of the encoded image, the less perceptibility of the extracted speech is observed. However, a trade-off should be made between quality of the encoded image and the perceptibility of the extracted speech. From the table 2, at 10% of noise, everyone could recognize what was said in the corrupted speech. However, the number decreased when the level of noise got higher, for instance, at 60% of additive noise. Nevertheless, at 100% of noise adding, there were 3 out of 10 could

recognize the contents of the corrupted speech. Likewise, a majority of the extracted speech waveform reveals that even if the watermarked image was attacked so that the waveform of the extracted speech was almost totally different, some people recognized the contents inside the extracted speech. Furthermore, we believe that the perceptibility will be greatly improved if the corrupted speech is examined by the copyright owner, since he/she will know exactly what he/she has placed in the embedded speech, and then make the identification of the content in the speech considerably easier.

6. CONCLUSIONS

This paper has presented the idea of using raw speech as a watermark to enhance the robustness of the watermarking schemes. The major advantage is that it can easily be incorporated into the existing schemes for watermarking. In the proposed method, the embedded speech also carries a unique message in its content, thus serving as a direct means of ownership identification. The experimental results show that the proposed method is immune to various forms of attacks, since in most cases, speech intelligibility remained in the corrupted image. Although the method has not been fully explored, we have demonstrated in this paper that it meets all the requirements of an effective watermarking application.

Acknowledgments: This work was supported by project NT-B-06-4C-20-319 funded by the National Electronics and Computer Technology Center (NECTEC).

References

[1] S. Sakaguchi, T. Arai and Y. Murahara, "The effect of polarity inversion of speech on human perception and data hiding as an application", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 917-920, 2000.

[2] M. D. Swanson, B. Zhu and A. H. Tewfik, "Data hiding for video-in-video", Proceedings International Conference on Image Processing, Vol. 2, pp. 676-679, 1997.

[3] D. Mukherjee, J. J. Chae, S. K. Mitra and B. S. Manjunath, "A source and channel-coding framework for vector-based data hiding in video", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 4, June 2000.

[4] G. B. Rhoads, "Indentification/authentication coding method and apparatus", Rep. WIPO WO 95/14289, World Intellectual Property Organization, 1995.

[5] R.G. Van Schyndel, A. Z. Tirkel and C. F. Osborne, "A digital watermark", International Conference on Image Processing, Vol. 2, pp. 86-90, 1994.

[6] W. Bender, D. Gruhl and N. Morimoto, "Techniques for data hiding", Proceedings SPIE, Vol. 2420, pp. 40, February 1995.

[7] M. Kutter, F. Jordan and F. Bossen, "Digital Signature of color images using amplitude modulation", Journal of Electronic Imaging, Vol.7, pp.326-332, 1998.

[8] M. George, J. Y. Chouinard and N. Georganas, "Digital watermarking of images and video using direct sequence spread spectrum techniques", Proceedings IEEE Canadian Conference on Electrical and Computer Engineer, Shaw Conference Center, Edmonton, Alberta, Canada, May 1999.
[9] F. Hartung and M. Kutter, "Multimedia watermarking techniques", Proceedings of IEEE, Vol. 87, No. 7, pp. 1079-1107, July 1999.
[10] T. S. Rappaport, "Wireless communications principles and practice", Prentice Hall, 1996.
[11] J. Bellamy, "Digital telephony", John Wiley & Sons, pp. 123-130, 1991.
[12] S. Haykin, "Communication systems", John Willy & Sons, INC, 1994.
[13] J. Flanagan, M. Schroeder, B. Atal, R. Crochiere, N. Jayant and J. Tribolet, "Speech coding", IEEE Transactions Communications, pp. 710-737, April 1979.
[14] M. H. Segal, "Speech intelligibility in the space shuttle mid-deck noise Environment: The Effect of Active Noise Reduction Technology", http://ergo.human.cornell.edu/, November 1999.
[15] H. K. Dunn and S. D. White, "Statistical measurements on conversational speech", Journal of Acoustic Society of America, pp. 278-288, January 1940.
[16] CityU Image Processing Lab, http://www.image.cityu.edu.hk/imagedb/