Performance Comparison on Speech Codecs for Digital Watermarking Applications

Y. Mamongkol and T. Amornraksa
Multimedia Communication Laboratory, Department of Computer Engineering
Faculty of Engineering, King Mongkut's University of Technology Thonburi
91 Pracha-Uthit Rd., Tungkru, Bangkok 10140, Thailand
Phone: +66-2-470-9083 Fax: +66-2-872-5050
Email: yothin2k@pacific.net.th, t.amornraksa@cpe.kmutt.ac.th

Abstract: Using intelligent information contained within the speech to identify the specific hidden data in the watermarked multimedia data is considered to be an efficient method to achieve the speech digital watermarking. This paper presents the performance comparison between various types of speech codec in order to determine an appropriate one to be used in digital watermarking applications. In the experiments, the speech signal encoded by four different types of speech codec, namely CELP, GSM, SBC and G.723.1codecs is embedded into a gray-scale image, and theirs performance in term of speech recognition are compared. The method for embedding the speech signal into the host data is borrowed from a watermarking method based on the zerotrees of wavelet packet coefficients. To evaluate efficiency of the speech codec used in watermarking applications, the speech signal after being extracted from the attacked watermarked image will be played back to the listeners, and then be justified whether its content is intelligible or not.

1. Introduction

The benefit of digital technology, communication and networks such as Internet not only makes various types of data to be stored, processed and transmitted in digital form, but also enable it to be easily accessed by someone else. Since digital data can be reproduced without loss in fidelity, this makes the duplicated data identical to the original copy, and difficult to differentiate from each other, leading to the problem of addressing the pirated material. There are many approaches for preventing such a circumstance not to occur. Applying cryptographic methods by encrypting the digital data before transmission is one solution that many people use nowadays. However, once the data is decrypted, an authorized user can make a copy and re-distribute it to the others. Another efficient method is based on digital watermarking. Watermarking the digital data before distributing it to the public is a way to show the possession of the copyright work, and to discourage someone to misuse the copy he owns. A small secret, which is embedded into the distributed data, can be any information related to copyright owner, authorized recipient or purchasing information.

Basically, an efficient watermarking method should fulfill the following requirements. The watermark signal should be invisible, undetectable, non-removable, unalterable and unambiguous by unauthorized person. In addition, the quality of watermarked data should not be significantly degraded.

In this paper, the performance of various types of speech codec i.e. CELP, GSM, SBC and G.723.1 codecs, which can be used in the speech based digital watermarking is compared. In the experiments, the encoded speech signal is embedded into the gray-scale image, while the method for embedding is borrowed from a watermarking method based on the zerotrees of wavelet packet coefficients. The next section provides some related work, while Section 3 describes the details of the watermarking scheme used in the experiments. The experimental setting is explained in Section 4. The results and discussions are then given in Section 5. Finally, the conclusions are drawn in Section 6.

2. Related Work

An efficient watermarking technique based on zerotrees of the discrete cosine transform (DCT) was proposed in [1]. The technique embedded the watermark signal by changing the value of coefficients derived from the DCT of the image. The watermark bit is assigned in the zerotrees by changing the number of zerotrees contained in a transformed block. This technique does not need the original image in the extraction process, but needs to carefully consider the suitable reference threshold instead.

To gain more advantage on new wavelet-based compression algorithm, a similar technique was applied by using the zerotrees derived from wavelet transform of the image [2]. In this technique, the positions of zerotrees defined in the embedded zerotree wavelet (EZW) and threshold value were used to detect the embedded data. One of the interesting techniques based on wavelet packet transform was proposed by [3]. Since the wavelet packet transform provides more energy compaction than ordinary wavelet transform, the watermark scheme applying this approach can then provide an impressive robustness against compression. To embed the watermark, the original image was first decomposed into small subbands by the wavelet packet transform. Then the watermark signal as a sequence of 0 and 1 was added to the coefficients in the chosen subbands. However, this technique still needs the original image in the extraction process.

Recently, Mukherjee *et al.* implemented a scheme for hiding 8KHz speech sampled at 16 bits/sample in a 30 frames/s QCIF video [4]. In the scheme, the successive samples of speech were vector-quantized, and the indices were embedded into the LL-HH subband coefficient. Then watermarked video was piped through a H.263 encoder. The speech and video extracted from the compressed video at high compression ratio were found to be intelligible, and acceptable for visual quality, respectively.

3. The Watermarking Scheme

The watermark embedding process is illustrated in Figure 1. The original image is segmented into 8x8 non-overlapping blocks, and each block is then transformed by using wavelet packet transform twice.

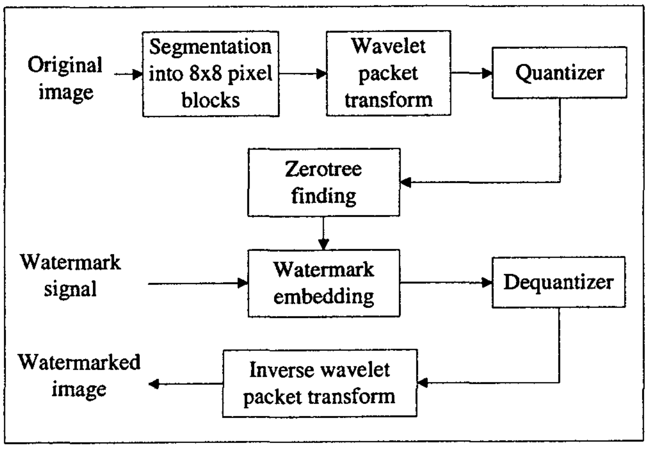


Figure 1. Block diagram of the watermark embedding process

In order to determine the zerotree in each transformed block, we use default quantization table provided by many compression standards such as JPEG, and modify it until we obtain the proper zerotree. The details of process of finding the zerotree can be found in [5]. To embed the watermark, the number of zerotrees is changed according to the watermark bit. For instance, assuming the number of zerotrees contained in the block is odd; if the watermark bit is zero, we will modify the coefficient value in that block until the resultant number of zerotrees is even. In the same way, if the watermark bit is one, the zerotrees remain untouched. Note that the value of selected coefficient to be changed should result in slightly perceptual degradation of the image, but still be robust against compression process, e.g. in the mid frequency range.

To extract the watermark signal, the same steps as used in the watermark embedding process are performed in reverse. After the zerotrees in each quantized transformed block are obtained, we count for the number of zerotrees contained within each block. If the result is odd, the embedded bit is one, and vice versa. The block diagram of the watermark extraction process is shown in Figure 5.

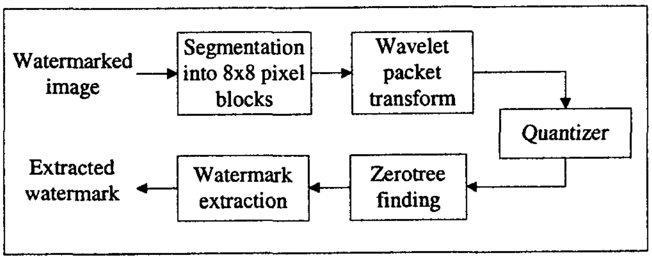


Figure 2. Block diagram of the watermark extraction process

4. Experimental Setting

Beside of embedding speech signal into an image, the intelligent information contained within the speech signal is also used as a watermark signal, instead of its characteristics alone. Our proposed method is distinct from many existing watermarking methods since, the extracted speech signal is justified by hearing perceptibility of human ear. The speech signal we considered in the experiment was obtained from a number of speech codecs i.e. the Code Excited Linear Prediction (CELP), Global System for Mobile speech coder (GSM), Sub Band Coding (SBC) and dual rate speech codec G.723.1 from TrueSpeech, which digitized and encoded the speech signal input from the microphone. The speech of the Thai word *"Yothin"*, which is encoded by the above, was chosen to be used as a watermark signal. When the coded speech signal was embedded into the image, a number of attacks was then applied to the watermarked image including brightness/contrast enhancement, high-pass filtering, Gaussian noise adding and JPEG compression scheme, at various levels of strength. After being attacked, the corrupted speech signal was extracted and then played back by a speech decoder to observe its remaining intelligent information. Ten Thai native speakers, with ages between 18-30 year-old were selected to be listening to the extracted corrupted speech. Different versions of the corrupted speech were randomly played back to the listeners. After the listeners heard the corrupted speech, they would tell whether they recognized what was said in the speech, that is, the word "*Yothin*". The format of this test is recommended in [6].

5. Experimental Results and Discussions

The coded speech signal from the codecs previously described was used as a watermark signal in the experiment. For the original images, we used two different gray-scale images *"airplane"* and *"cliff"* with the size of 512x512 pixels. Note that since the image at this size will contain 4096 8x8 pixel blocks, there will be enough space to embed around 12 × 10^3 watermark bits into one image frame i.e. three watermark bits per one transform block. The original image *"airplane"* and the watermarked images, with Peak Signal to Noise Ratio (PSNR) of 34.11 dB, are shown in Figure 3 (a) and (b), respectively, while the original raw speech waveform is shown in Figure 4.

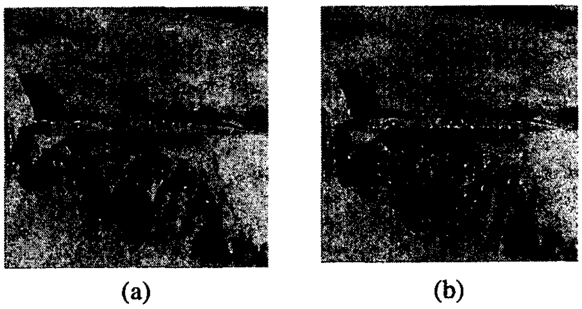


Figure 3. (a) The original image (b) its watermarked version

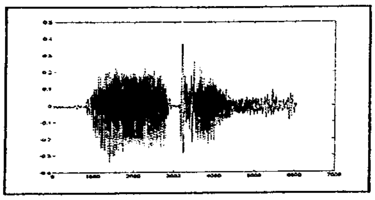
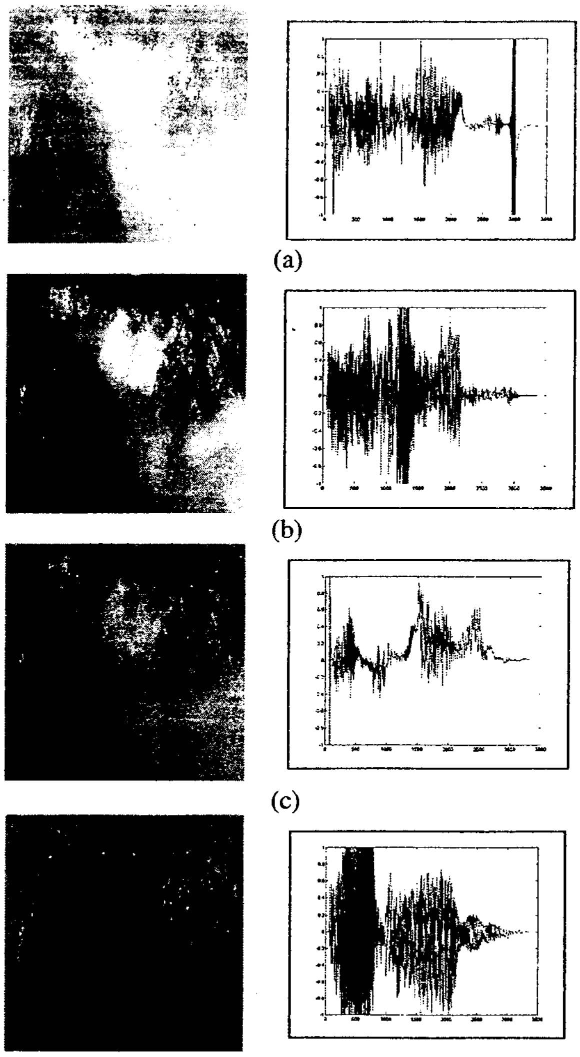


Figure 4. the original raw speech waveform "*Yothin*"

After the coded speech signal was embedded into the image using the watermarking technique as mentioned earlier, a number of attacks was then applied to the watermarked image, in such a way that the strength of each attack was raised from 0 to the level that no listeners could recognize what was said in the extracted speech signal. Some results of the watermarked image and the waveform of the extracted speech are illustrated in Figures 5 and 6.



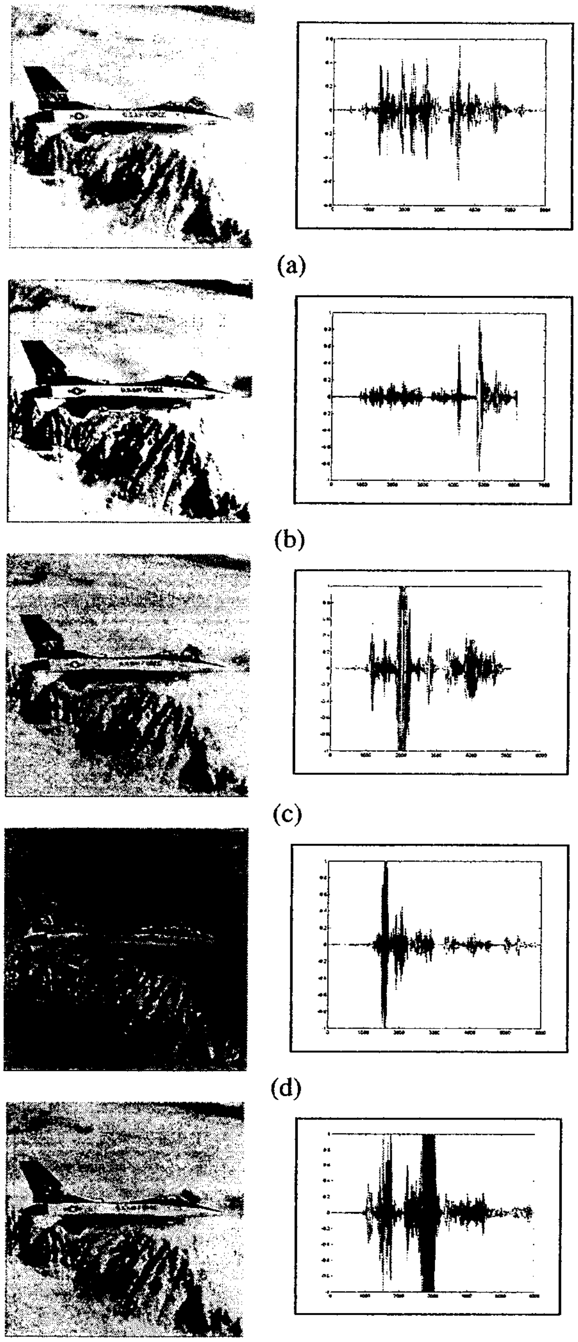
(a)

(b)

(c)

(d)

Figure 5. The watermarked image and the extracted G.723.1 coded speech waveform after applying (a) brightness enhancement by 80 % (b) contrast enhancement by 40% (c) Gaussian noise adding by 5 % (d) highpass filtering with radius of 10 pixels



(a)

(b)

(c)

(d)

(e)

Figure 6. The watermarked image and the extracted CELP coded speech waveform after applying (a) brightness enhancement by 30 % (b) contrast enhancement by 60% (c) Gaussian noise adding by 3 % (d) highpass filtering with radius of 10 pixels and (e) JPEG compression at 75 % quality

According to Figure 4 and 5, it can be clearly seen that even if the speech waveforms' shape was quite badly corrupted, most listeners could still recognize its content. In the next experiment, we compared the performances obtained from the speech watermarking scheme using various speech encoders i.e. CELP, GSM, SBC and G.723.1, and the results from the listening tests are shown in Table 1.

Table 1. Number of listeners who could recognize the word "*Yothin*"

	Types of speech coder			
Types of attack	CELP	SBC	GSM	G.723.1
Brightness enhancement by 80%	10	0	10	10
Contrast enhancement by 40%	10	0	7	8
Gaussian noise adding by 3%	10	0	7	8
JPEG encoding by 50 % quality	10	0	10	10
JPEG encoding by 95 % quality	10	10	10	10
High-pass filtering with 10 pixel radius	10	0	10	10

From Table 1, it is obvious that extracted CELP coded speech signal was robust against all attacks at all levels. Furthermore, the SBC coded speech signal could survive the JPEG compression based attack at 95 % quality only, while the other two, GSM and G.273.1 coded speech signal, gave slightly lower performance than that of extracted CELP coded speech signal.

It should be noted that we inserted the results from JPEG encoding by 95 % quality in order to show that, in that level, the content within the extracted SBC coded speech signal can still be recognized by all listeners. But after we further increased the strength of attack, the content in the speech was totally destroyed. Therefore, it is obvious that the SBC based speech coder was vulnerable against most attacks and should not be used in practice. This is because of the error propagation caused by the corrupted reference samples within the speech signal. In contrary, the CELP speech coder should be considered as it gave similar performance, or better in some attacks, to the one based on the G.273.1 coder, while occupied less bandwidth in the host data.

6. Conclusions

The performance comparison between various types of speech codec has been presented in this paper. The zerotrees-based watermarking method is just an example we used in the experiments to evaluate the performance of the speech codecs when implemented in digital watermarking applications. According to the experimental results, the speech signal encoded by CELP codec gave the best performance in both output bit-rate of the encoded signal and the robustness against attacks, at least, in some certain levels. The GSM and G.723.1 speech codecs gave slightly lower performance than the CELP, and still be worth using them. For the last one, the SBC speech codec, it gave unsatisfied results and should not be considered for the use in practice.

7. Acknowledgement

This research work is supported by the National Research Council of Thailand (NRCT).

References

[1] M. N. Xia, M. L. Zhe and H. S. Sheng, 'Digital watermarking based on zerotrees of DCT coefficients', *Proceedings of ICIP*, Vol. 2, 1996, pp. 237-240.

[2] J. M. Shapiro, 'Embedded image coding using zerotrees of wavelet coefficients', *IEEE Transactions on Signal Processing*, Vol.41. No. 12, December 1993, pp. 3445-3462.

[3] S. Onrit, Y. Attavetchakul, S. Aroonrungratsami and K. Chamnongthai, 'Digital watermarking embedding and extraction by using wavelet packet transform', *Proceedings of Electrical Engineering Conference (EECON'2000)*, Chiang Mai, Thailand, 2000, pp. 512-524.

[4] 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.

[5] Y. Mamongkol and T. Amornraksa, 'Speech watermarking using zerotress of wavelet packet transform', *Proceedings of the International Symposium on Communications and Information Technology (ISCIT2001)*, Chiang Mai, Thailand, November 14-16, 2001, pp.353-356.

[6] H. K. Dunn and S. D. White, 'Statistical measurements on conversational speech', *Journal of Acoustic Society of America*, pp. 278-288, January 1940.