

# FRAGILE SPEECH WATERMARKING BASED ON EXPONENTIAL SCALE QUANTIZATION FOR TAMPER DETECTION

Chung-Ping Wu and C.-C. Jay Kuo

Department of Electrical Engineering  
University of Southern California, Los Angeles, CA 90089-2564  
E-mail: {chungpin, cckuo}@sipi.usc.edu

## ABSTRACT

A fragile speech watermarking scheme of low complexity is proposed in this research as an effective way to detect malicious content alterations while tolerating content preserving operations. Since no auxiliary data are required, the authentication process is transparent to agents that perform content preserving operations such as transcoding during the transmission process. The proposed scheme is based on the modified odd/even modulation scheme with exponential scale quantization and a localized frequency masking model while assuring no mismatch between quantization steps used in watermark embedding and detection. The noise introduced by watermarking is shown to be smaller than that of speech coders. It is shown by experimental results that the proposed system is able to distinguish malicious alterations from resampling, white noise pollution, G.711 speech coding and G.721 speech coding with very low error probabilities.

## 1. INTRODUCTION

Since modern digital media editing and processing technologies allow high quality forgery to be created at a low cost, tamper detection is increasingly essential to secure speech applications. The semantic meaning of speech could be altered by simply reordering sentences or dropping out a few words. Thus, judging the authenticity of speech data by human perception alone is not enough anymore. Traditional cryptographic integrity schemes with hash functions are designed to ensure that every bit in the data stream is unmodified. Even though it is possible to apply these schemes directly to audiovisual data by treating them as binary data streams, it is desirable to achieve more implementational flexibility by allowing content preserving operations to exist in the transmission channel without triggering the authentication alarm. Examples of content preserving operations include transcoding, re-sampling and D/A-A/D conversion that are needed to transform signals from one form to another to match the requirements of a segment of the channel. Another example is the intelligent server that automatically adjusts the audio volume or filters out noise in the recording.

An authentication system that distinguishes content preserving operations from malicious modifications is called the *content integrity* system, and fragile watermarking and feature extraction are two major approaches studied in this field. In our previous work [12, 11], we proposed a speech feature extraction scheme, which is integrated with CELP speech coders to minimize the total computational cost. Speech features relevant to semantic meaning are extracted, encrypted and attached as the header information. The system was shown to be tolerant to many kinds of content pre-

serving operations while able to detect minor content alterations. Seven requirements for effective content integrity systems were also discussed and achieved.

However, for the feature extraction system to work properly, all agents that perform content preserving operations in the transmission channel must participate in the scheme by passing feature data along. Sometimes, it may be inconvenient or impossible to require such a level of cooperation. In this paper, we investigate fragile speech watermarking for tamper detection, which entirely eliminates the participation of agents. A fragile watermark is a secret pseudo-random sequence embedded into the host audiovisual data. If host data are tampered, the secret sequence is also modified. The receiver detects the fragile watermark from received data and compares it to the original secret sequence. Since no auxiliary data are needed, agents in the channel could be completely ignorant of the authentication process.

## 2. PREVIOUS WORK RELATED TO FRAGILE SPEECH WATERMARKING

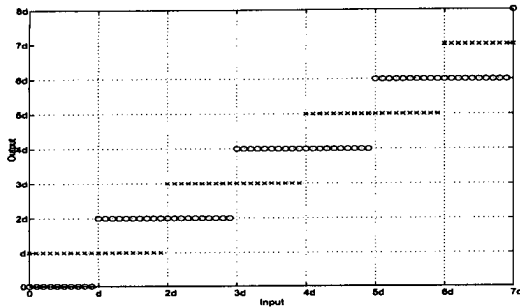
We are not aware of any previous work in the literature directly addressing speech integrity protection with fragile watermarking. Most papers on fragile watermarking have been published in the field of image/video authentication. Recently, there are some audio watermarking techniques proposed for copyright protection and auxiliary data embedding. Although they are not designed tailored to narrow-band speech signals and integrity verification, we will examine their potential applicability to speech authentication.

### 2.1. Fragile Watermarking for Image Authentication

Fragile watermarking techniques for content authentication are called semi-fragile [2] watermarking to indicate that they are not completely fragile to all kinds of modifications. Algorithms in this category usually embeds watermarks in frequency domains such as the DCT domain [2, 6] or the wavelet domain [4], and a brief review of these schemes could be found in [12].

A watermarking technique called *odd/even modulation* is adopted in several image tamper-proofing schemes recently [4, 3]. It is based on uniform scalar quantization and proven to be simple yet effective for correctly detecting a watermark from a small amount of carrying data without the need of the original data. One bit of watermark data is embedded into each selected sample in the spatial or the frequency domains by using the following equation:

$$\hat{s}_i = Q(s_i, w_i, d_i) = \text{round}\left(\frac{s_i + w_i d_i}{2d_i}\right) \times 2d_i - w_i d_i, \quad (1)$$



**Fig. 1.** The plot of the embedding function for odd/even modulation, where the horizontal axis and the vertical axis represent the value of  $s_i$  and  $\hat{s}_i$ , respectively.

where the input sample  $s_i$  is a real number, the watermark signal  $w_i$  could be 0 or 1, and the quantization step size  $d_i$  is a positive real number. The function *round* means to be rounded toward the nearest integer.

The plot of (1) is shown in Figure 1, where the circle and the cross marks represent the values of  $Q(s_i, w_i, d_i)$  for  $w_i = 0$  and 1, respectively. As shown in the figure, depending on the value of  $w_i$ , the output is equal to an even integer or an odd integer times the quantization step size  $d_i$  (hence the name odd/even modulation). The noise  $\hat{s}_i - s_i$  introduced by watermark embedding is uniformly distributed in the range  $(-d_i, d_i)$ .

In the transmission channel,  $\hat{s}_i$  is changed into  $\tilde{s}_i$  by various noise sources. Watermark detection is performed by applying

$$\tilde{w}_i = D(\tilde{s}_i, d_i) = \text{round}\left(\frac{\tilde{s}_i}{d_i}\right) \pmod{2} \quad (2)$$

to each  $\tilde{s}_i$  of the received signal. When the noise  $(\tilde{s}_i - s_i)$  introduced in the channel is smaller than  $d_i/2$ , the detection result  $\tilde{w}_i$  will be identical to the embedded watermark  $w_i$ . The purpose of self-noise suppression [9] in blind watermark detection is achieved because the signal that carries the watermark does not interfere with watermark detection.

In order to evaluate whether the transmitted signal has been maliciously altered, detection results from a group of samples are usually averaged, and one example is the tamper assessment function (*TAF*) [4] defined as

$$TAF(\mathbf{w}, \tilde{\mathbf{w}}) = \frac{1}{N_g} \sum_{i=1}^{N_g} w_i \oplus \tilde{w}_i, \quad (3)$$

where  $N_g$  is the size of the group, and the output of *TAF* ranges from 0 to 1. The value zero means that  $\tilde{\mathbf{w}}$  is identical to  $\mathbf{w}$ , and the value 0.5 means that  $\tilde{\mathbf{w}}$  and  $\mathbf{w}$  are completely uncorrelated. If the signal goes through content preserving operations, *TAF* would be smaller than 0.5. On the other hand, if the hacker replaces most of the samples in the block, *TAF* would be close to 0.5. Apparently, the group size should be small enough so that meaningful content alteration cannot be done by only replacing a fraction of the group.

## 2.2. Audio Watermarking Techniques

A variety of audio watermarking methods were reviewed in our paper on robust audio watermarking for copyright protection[13]. In

Media Type	Compression Codec	Typical Signal to Noise Ratio
Image	JPEG	30 ~ 40 dB
Wideband Audio	MPEG audio layer 3	15 ~ 20 dB
Speech	G.723.1	8 ~ 12 dB
Speech	GSM-AMR	1 ~ 2 dB

**Table 1.** Comparison between typical signal-to-noise ratio (SNR) values of several popular audiovisual compression algorithms.

this subsection, we discuss the potential of applying these methods to speech content authentication. Early work in the area of audio watermarking embedded watermarks in human insensitive regions such as the high-frequency components or the DFT phase coefficients in order to achieve inaudibility [1, 10]. This method could be insecure for integrity verification since the attacker may avoid disturbing the watermark by replacing the watermark embedded regions of maliciously modified speech signal with that of the original signal. The artifacts produced by this replacement may not be audible by human ears. Therefore, fragile watermarks should be embedded in the areas to which the human auditory system is highly sensitive.

Another class of algorithms embed watermarks in selected compression coefficients to prevent the watermark from being interfered by the coding process before transmission [5]. However, this benefit would be lost if content preserving operations in the channel remove the watermark by significantly altering those coefficients. Therefore, only those compression coefficients that are insensitive to content preserving operations and re-compressions are suitable for fragile watermark embedding. In popular speech codecs such as the CELP coders, only a small portion of coefficients meet this requirement. For example, we observe that only the LSP coefficients and the lag of pitch predictors are stable enough among all coefficients of G.723.1. In order to individually authenticate a speech segment of 0.5-second long, the amount of data suitable for carrying the watermark would be less than 80 bytes. It is difficult to properly embed and detect a watermark signal into such a small amount of host data.

Watermark signals embedded as pseudo-random noise in the time or the frequency domain are statistically undetectable without prior knowledge of the watermark sequence [1, 8]. Thus, hackers are not able to detect and restore the watermark after modifying the speech signal. In following sections, we will propose a fragile speech watermarking scheme that uses pseudo-random noise as the watermark.

## 3. SELECTION OF SPEECH SIGNAL DOMAIN TO EMBED FRAGILE WATERMARKS

In terms of the signal-to-noise ratio, the distortion introduced by low bit-rate speech codecs is significantly larger than that in image and audio coding. Image compression algorithms usually provides a signal-to-noise ratio as high as 30 dB or above. A wide-band audio codec generally aims at limiting the noise power in each sub-band. Based on our observations, the overall SNR of an MP3 codec under a bit rate of 128 kbps is in the range of 15 to 20 dB. The model-based speech coders could generate an SNR value as low as 1 or 2 dB, such as the case of GSM-AMR. Consequently, a fragile watermark directly embedded in the time domain can be greatly distorted by speech compression. This problem makes watermarking in the time domain via pseudo-random noise an unfa-

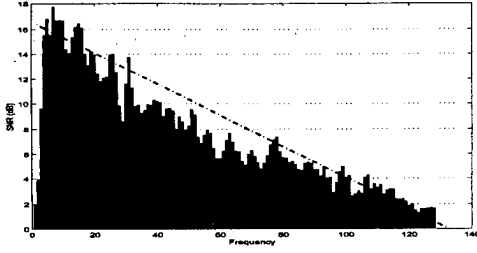


Fig. 2. The average SNR of GSM-AMR in the DFT magnitude domain, where GSM-AMR operates in the 5.15kbps mode, and the DFT window size is 256. Note that the other half of the spectrum is omitted due to symmetry.

variable choice. A comparison among the SNR values of codecs for several media types is shown in Table 1.

The SNR ratios of CELP speech codecs in the DFT magnitude domain are much larger than those in the time domain. Figure 2 shows the average SNR of DFT magnitude coefficients generated by the GSM-AMR codec at a bit-rate of 5.15 kbps. As expected, the signal-to-noise ratios of a few lowest-frequency coefficients are very low since CELP coders perform high-pass filtering to eliminate low frequency signals that do not exist in human voice. Except these coefficients, the SNR values in the low frequency region are in the range of 15 to 20 dB, and gradually decrease to zero as the frequency increases to the maximum value. To conclude, the lower half of the spectrum, excluding the region below 70Hz, is a good candidate for fragile watermark embedding.

#### 4. SPEECH WATERMARKING USING EXPONENTIAL SCALE QUANTIZATION

In this section, we propose a fragile watermarking scheme by using modified odd/even modulation with exponential scale quantization.

In speech/audio watermarking, the amount of watermark energy allowed without introducing audible noise is dependent on the masking model. If the quantization step size in odd/even modulation is fixed or only dependent on a secret key, the watermarking scheme disregards the localized masking model. Some image watermarking systems take this approach [3] because adaptive visual masking customized to the local environment may not be essential in some image domains. Nevertheless, the auditory masking model is necessary in most audio applications since the absolute inaudibility threshold is very low. Lu *et al.* [7] proposed an audio watermarking scheme by using the FFT domain uniform quantization scheme with an adaptive quantization step size. The step size is set to the masking threshold computed by the MPEG audio psychoacoustic model. Thus, the noise energy of watermarking would be no larger than that of MPEG audio compression in each frequency range. However, if the signal go through content preserving operations in the channel, the masking thresholds computed from the received audio are almost guaranteed to be non-identical to those calculated based on the original audio. Given the fact that the detection scheme as specified by (2) is very sensitive to minor changes in  $d_i$ , the output of *TAF* will be close to 0.5. In fact, the scheme proposed by Lu *et al.* aims at complete authentication rather than content authentication.

Our scheme is designed to utilize the frequency masking effect

and, at the same time, guarantee that watermark embedding and detection will use the same set of quantization step sizes. In the MPEG audio psychoacoustic model, the masking threshold of a DFT coefficient is a weighted sum of its surrounding coefficients, and the coefficient in question has the largest weight value. We roughly approximate this model by using only the DFT coefficient itself to calculate the masking threshold. In other words, the maximum watermark magnitude allowed to be embedded on a DFT coefficients is directly proportional to the magnitude of the coefficient. In order to meet this constraint, the size of quantization intervals should grow exponentially instead of being uniform. This is also equivalent to performing uniform quantization on the logarithm of the coefficient. In order to incorporate exponential scale quantization into the odd/even modulation scheme, (1) and (2) are modified to be

$$\hat{s}_i = Q(s_i, w_i, d_i) = \exp(\text{round}(\frac{\ln(s_i) + w_i d_i}{2d_i}) \times 2d_i - w_i d_i), \quad (4)$$

and

$$\tilde{w}_i = D(\tilde{s}_i, d_i) = \text{round}(\frac{\ln(\tilde{s}_i)}{d_i}) \pmod{2}. \quad (5)$$

It is worthwhile to point out that there is no potential disaster of mismatching quantization step sizes because the same set of  $\{d_i\}$  is used in both embedding and detection. It can also be mathematically proved from (4) that the value  $\hat{s}_i - s_i$  of the watermark is proportional to the value  $s_i$  of the signal. More specifically,

$$\max_{s_i} \left( \frac{|\hat{s}_i - s_i|}{s_i} \right) \approx d_i. \quad (6)$$

In order for speech watermarking to be inaudible, its noise level should not exceed that of speech coding. The typical coding SNR values of the GSM-AMR coder in each frequency bin are shown in Figure 2, and we will adopt this curve as the target level of our watermarking noise. With the SNR values in this figure, the quantization step size at each frequency is calculated as

$$d_f = 10^{-SNR_f^{sc}/20}, \quad (7)$$

where  $SNR_f^{sc}$  denotes the SNR value of GSM-AMR speech coding at frequency  $f$  as shown in Figure 2. By using  $d_f$  computed from this equation, it can be proved that the watermarking noise will be no greater than the speech coding noise:

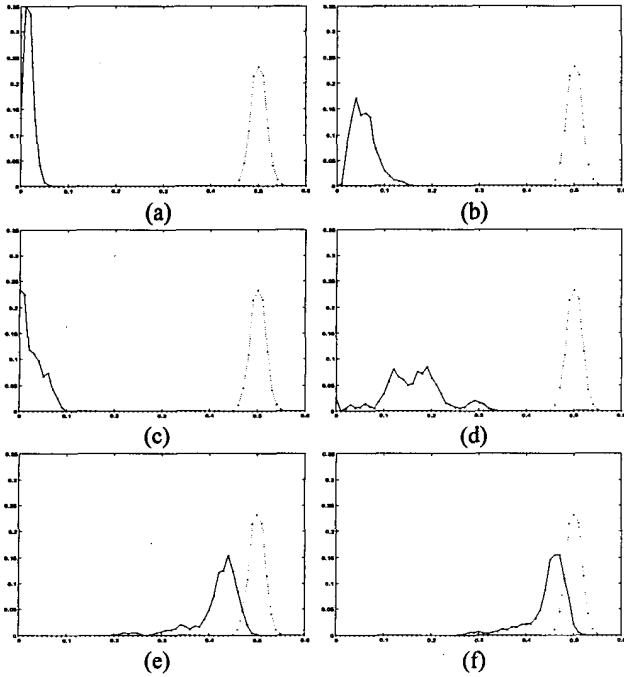
$$SNR_f = 20 \log \left| \frac{s_f}{\hat{s}_f - s_f} \right| \geq 20 \log \frac{1}{d_f} = SNR_f^{sc}, \quad (8)$$

where  $SNR_f$  is the SNR of watermarking at frequency  $f$ ,  $s_f$  denotes the value of any DFT magnitude coefficient at frequency  $f$ , and the inequality comes from (6). The same table of predetermined  $d_f$  values are used at both watermark embedding and detection processes. Please note that this table is not adaptive to each individual speech piece and, therefore, there is no risk of quantization step mismatch due to content preserving processes.

Finally, since the value  $d_f$  is fixed and will be eventually discovered by hackers, the security of the scheme cannot rely on the secrecy of  $d_f$ . Therefore, the actual quantization step sizes used in watermark embedding and detection should deviate a little from  $d_f$ , i.e.

$$d_i = \{d_f + r_i \mid f = \text{frequency of } s_i\}, \quad (9)$$

where  $r_i$  is a pseudo-random sequence of real numbers generated with the secret key. If a higher security level is desired, two other



**Fig. 3.** The TAF distribution of the malicious content replacement in comparison with that of (a) resampling, (b) white noise pollution, (c) G.711  $\mu$ -law speech coding, (d) G.721 ADPCM speech coding, (e) G.723.1 speech coding, and (f) GSM-AMR speech coding. The probability values in each plot are generated from  $10^4$  trials. Resampling consists of down-sampling to 4000kHz followed by up-sampling back to 8000kHz. The horizontal and vertical axis represent the TAF value and its probability, respectively. The group size  $N_g$  is set to 1000, which corresponds to 0.5 sec of speech data in each group.

pseudo-random sequences could be optionally incorporated in watermark embedding and detection. They are

$$\begin{aligned} \hat{s}_i &= Q(s_i, w_i, r_i, a_i, b_i) \\ &= \exp(\text{round}(\frac{\ln(s_i + a_i) + b_i + w_i d_i}{2d_i})2d_i - b_i - w_i d_i) - a_i, \end{aligned} \quad (10)$$

and

$$\tilde{w}_i = D(\hat{s}_i, r_i, a_i, b_i) = \text{round}(\frac{\ln(\hat{s}_i + a_i) + b_i}{d_i}) \pmod{2}. \quad (11)$$

It can be mathematically proved that the introduction of  $a_i$  and  $b_i$  in (10) will not increase the average energy of the watermark noise  $\hat{s}_i - s_i$ .

## 5. EXPERIMENTAL RESULTS

The proposed exponential scale quantization algorithm distinguishes distortions due to content preserving operations and malicious content replacement by using the tamper assessment function. As shown in Figure 3, the distribution of TAF for content replacement is centered at 0.5, and ranges from 0.45 to 0.55 in the  $10^4$  trials performed. It is represented by dotted curves and compared to that of various content preserving operations. It is shown that

the TAF distributions of operations such as white noise pollution, resampling, G.711  $\mu$ -law coding and G.721 ADPCM coding are clearly separated from that of malicious content replacement. The TAF values of resampling are the smallest because the effect of downsampling followed by up-sampling is similar to low-pass filtering, which does not affect our proposed watermarking scheme performed in the lower-half spectrum of the DFT domain.

The TAF distribution of CELP speech coders such as G.723.1 and GSM-AMR are not well separated from that of malicious content alteration as shown in Figure 3(e)-(f). Therefore, there is no suitable threshold value that could tolerate CELP speech coding while detecting all malicious replacement acts at the same time. For example, as shown in Figure 3 (e), in order to limited the probability of false detection (which falsely detects the presence of malicious alteration when the speech is only processed with content preserving operations) for G.723.1 coding to  $10^{-3}$ , the threshold should be set at 0.49. However, this would make the failure rate for detecting malicious content replacements about 38%.

Therefore, tamper detection using odd/even modulation with exponential scale quantization is an efficient approach for tolerating mild content preserving operations. The detection block size could be set at 0.5 sec or even smaller, and the probability of false detection is very small. However, its compatibility with CELP speech coders needs to be improved in our future work.

## 6. REFERENCES

- [1] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3-4):313, 1996.
- [2] J. Fridrich. Image watermarking for tamper detection. *Proc. ICIP '98*, October 1998.
- [3] H. Inoue *et al.*. Wavelet-based watermarking for tamper proofing of still images. *IEEE ICIP 2000*, Sept. 2000.
- [4] D. Kundur and D. Hatzinakos. Digital watermarking for tell-tale tamper proofing and authentication. *Proceedings of the IEEE - Special Issue on Identification and Protection of Multimedia Information*, 87(7):1167-1180, July 1999.
- [5] J. Lacy, S. Quackenbush, A. Reibman, D. Shur, and J. Snyder. On combining watermarking with perceptual coding. In *ICASSP*, volume 6, pages 3725-3728, 1998.
- [6] E. T. Lin, C. I. Podilchuk, and E. J. Delp. Detection of image alterations using semi-fragile watermarks. *SPIE International Conf. on Security and Watermarking of Multimedia Contents II*, 3971(14), January 2000.
- [7] C.-S. Lu, H.-Y. Liao, and L.-H. Chen. Multipurpose audio watermarking. *Proceedings of 15th IAPR International Conference on Pattern Recognition*, Sept. 2000.
- [8] I. Pitas and P. Bassia. Robust audio watermarking in the time domain. In *EUSIPCO'98*, pages 25-28, 1998.
- [9] M. Ramkumar and A. Akansu. Self-noise suppression schemes for multimedia steganography. *SPIE International Workshop on Voice, Video and Data Communication, Multimedia Applications*, 3845, Sept. 1999.
- [10] J. F. Tilki and A. A. Beex. Encoding a hidden auxiliary channel onto a digital audio signal using psychoacoustic masking. In *IEEE Southeastcon 97*, pages 331-333, 1997.
- [11] C.-P. Wu and C.-C. J. Kuo. Speech content authentication integrated with celp speech coders. *ICME2001*, August 2001.
- [12] C.-P. Wu and C.-C. J. Kuo. Speech content integrity verification integrated with itu g.723.1 speech coding. *ITCC2001*, pages 680-684, April 2001.
- [13] C.-P. Wu, P.-C. Su, and C.-C. J. Kuo. Robust and efficient digital audio watermarking using audio content analysis. *SPIE EI2000*, 3971:382-392, Jan. 2000.