

FRAGILE SPEECH WATERMARKING FOR CONTENT INTEGRITY VERIFICATION

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

Two fragile speech watermarking schemes for content integrity, i.e. exponential-scale odd/even modulation and linear additive watermarking, are proposed and compared with respect to their ability in detecting localized malicious alterations and tolerating content preserving operations. Both algorithms embed watermarks in the DFT magnitude domain by using a simplified masking model to achieve inaudibility. Statistical analysis is performed to identify the detection failure rate and the false detection rate of these two algorithms. These two schemes require no auxiliary authentication data to perform integrity verification, and could distinguish various content preserving operations from malicious tampering. The exponential-scale odd/even modulation is better in detecting localized content alteration while linear additive watermarking is more tolerant to low bit-rate speech coders such as the CELP coders.

1. INTRODUCTION

The importance of tamper detection for audiovisual data is rapidly increasing since modern digital media editing and processing technologies allow high quality forgery to be created at a relatively low cost. 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 [8, 7], 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 preserving operations while able to detect minor content alterations.

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. Algorithms in this category usually embed watermarks in frequency domains such as the DCT domain [1, 4] or the wavelet domain [3], and a brief review of these schemes could be found in [8].

A watermarking technique called *odd/even modulation* is adopted in several image tamper-proofing schemes recently [3, 2, 6]. 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. In section 4, we will propose a speech watermarking scheme adapted from odd/even modulation.

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 an unfavorable choice.

The SNR ratios of CELP speech coders in the DFT magnitude domain are much larger than those in the time domain. Figure 1 shows the average SNR of DFT magnitude coefficients generated by the GSM-AMR codec at a bit-rate of 5.15 kbps. As expected,

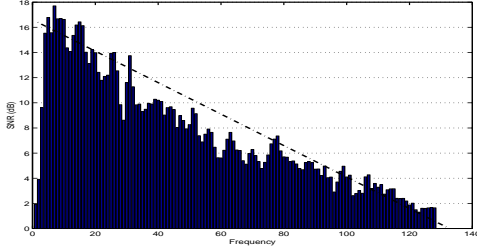


Figure 1: The average SNR of GSM-AMR in the DFT magnitude domain, where GSM-AMR operates in the 5.15kbps mode, and the DFT window size is 256. Note that the other half of the spectrum is omitted due to symmetry.

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 [2] 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.* [5] 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 of odd/even modulation is very sensitive to minor changes in the quantization step size, the detection result will be erroneous. In fact, the scheme proposed by Lu *et al.* aims at complete authentication, which does not allow content preserving operations in the channel.

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.

The embedding and detection functions of exponential-scale odd/even modulation are defined as

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

and

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

where the input sample s_i are real numbers, the watermark signal w_i could be 0 or 1, and the quantization step sizes d_i are positive real numbers. The function *round* means to be rounded toward the nearest integer. In the transmission channel, \hat{s}_i is changed into \tilde{s}_i by various noise sources.

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 (1) that the value $\hat{s}_i - s_i$ of the watermark is proportional to the value s_i of the signal. More specifically,

$$\max_{s_i} \left(\frac{|\hat{s}_i - s_i|}{s_i} \right) \approx d_i. \quad (3)$$

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 1, 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

$$d_f = 10^{-\text{SNR}_f^{sc}/20}, \quad (4)$$

where SNR_f^{sc} denotes the typical SNR value of GSM-AMR speech coding at frequency f as shown in Figure 1. 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:

$$\text{SNR}_f = 20 \log \left| \frac{s_f}{\hat{s}_f - s_f} \right| \geq 20 \log \frac{1}{d_f} = \text{SNR}_f^{sc}, \quad (5)$$

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 (3). 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.

$$d_i = \{d_f + r_i \mid f = \text{frequency of } s_i\}, \quad (6)$$

where r_i is a pseudo-random sequence of real numbers generated with the secret key.

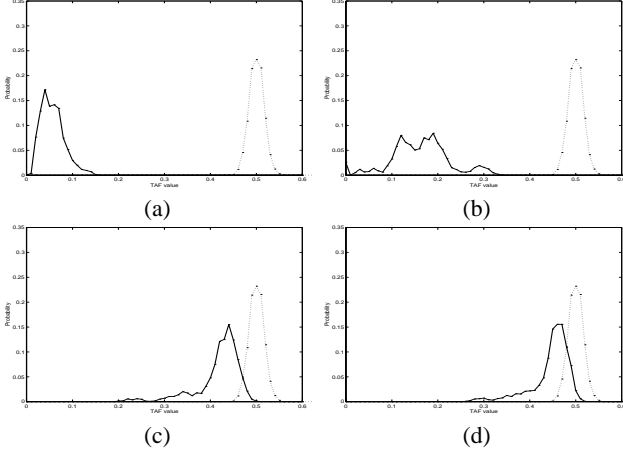


Figure 2: The TAF distribution of the malicious content replacement in comparison with that of (a) white noise pollution, (b) G.721 ADPCM speech coding, (c) G.723.1 speech coding, and (d) GSM-AMR speech coding. The probability values in each plot are generated from 10^4 trials. The horizontal and vertical 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.

5. LIMITATIONS OF QUANTIZATION-BASED WATERMARKING

The proposed exponential scale quantization algorithm distinguishes distortions due to content preserving operations and malicious content replacement by using the tamper assessment function (TAF) [3] defined as

$$TAF(\mathbf{w}, \tilde{\mathbf{w}}) = \frac{1}{N_g} \sum_{i=1}^{N_g} w_i \oplus \tilde{w}_i, \quad (7)$$

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. As shown in Figure 2, 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 and G.721 ADPCM coding are clearly separated from that of malicious content replacement.

In contrast, 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 2(c)-(d). 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 2 (c), 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%.

The other limitation of the exponential scale quantization algorithm results from the signal amplification operation in the channel. Signal amplification is a content preserving operation, but quantization-based watermarking cannot be detected from amplified signals. This problem could be solved by normalizing the power of the speech signal to a fixed value before watermark embedding, and then restoring its power to the original value. At the receiver end, watermark detection is also preceded by normalizing the power of received speech to the same fixed value. However, the receiver must receive a large segment of speech before it could normalize its power. Therefore, the content authentication process cannot start as the first fraction of speech data comes in. These two limitations can be overcome in the additive watermarking scheme proposed in the next section.

6. SPEECH WATERMARKING USING LINEAR ADDITION

The simplest form of additive watermark embedding adds a watermark sequence with a fixed magnitude to selected samples, i.e.

$$\hat{s}_i = s_i + \alpha w_i, \quad (8)$$

where the positive real number α controls the power of the watermark signal. In order to utilize the masking effect, most audio watermarking algorithms in this category adapt the watermark signal power to various masking models. Based on discussions in Sections 3, we choose DFT magnitude coefficients in the lower half of the spectrum to embed the additive watermark. The first few coefficients in the lowest frequency region are excluded because human are insensitive to them. Based on the same masking model used in Section 4, watermark embedding is defined as

$$\hat{s}_i = s_i(1 + \alpha_i w_i), \quad (9)$$

$$\alpha_i = \{\alpha_f \mid f = \text{frequency of } s_i\}, \quad (10)$$

$$\alpha_f = 10^{\frac{\text{SNR}_f^{sc}}{-20}}, \quad (11)$$

where $\{w_i\}$ is a pseudo-random sequence consisting of 1 and -1, and SNR_f^{sc} are the same values from Section 4. With this approach, noise introduced by watermarking will have a signal-to-noise ratio similar to that of GSM-AMR in each frequency range.

In the transmission channel, \hat{s}_i are changed to \tilde{s}_i by various noise sources, i.e. $\tilde{s}_i = m\hat{s}_i + n_i$, where n_i is assumed to be uncorrelated with s_i and m represents amplification (or de-amplification) by intelligent agents. Watermark detection is performed by applying the following equation to each group of samples

$$\text{correlation coefficient} = \frac{\sum_{i=1}^{N_g} \tilde{s}_i w_i}{\sigma_{\tilde{s}_i} \sigma_{w_i}}, \quad (12)$$

where $\sigma_{\tilde{s}_i}$ and σ_{w_i} are the standard deviations of \tilde{s}_i and w_i , respectively. The correlation coefficient could be approximated by

$$\begin{aligned} \text{correlation coefficient} &= \frac{\sum_{i=1}^{N_g} (ms_i(1 + \alpha_i w_i) + n_i)w_i}{\sigma_{\tilde{s}_i} \sigma_{w_i}} \\ &= \frac{\sum_{i=1}^{N_g} (ms_i \alpha_i w_i^2 + \sum_{i=1}^{N_g} s_i w_i + \frac{1}{m} \sum_{i=1}^{N_g} n_i w_i)}{\sigma_{\tilde{s}_i} \sigma_{w_i}} \\ &\approx \frac{\sum_{i=1}^{N_g} \alpha_i s_i w_i^2 + 0 + 0}{\sigma_{s_i} \sigma_{w_i}} \end{aligned} \quad (13)$$

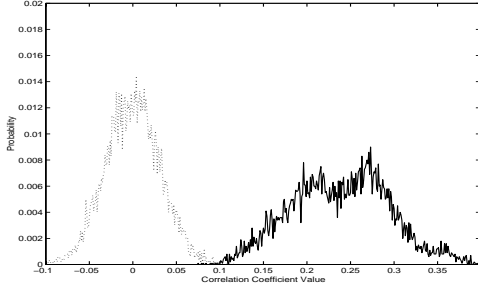


Figure 3: The distribution of correlation coefficient values with malicious content replacement and no distortion. The probability values in the figure are generated from 10^4 trials, and the group size N_g is 1000. The dotted curve represents the case of malicious content replacement.

where $\sum_{i=1}^{N_g} s_i w_i$ and $\sum_{i=1}^{N_g} n_i w_i$ are approximately zero because w_i is uncorrelated with s_i and n_i . Since α_i and s_i are all positive numbers in our scheme, the value of Equation (13) is greater than zero. In contrast, if no watermark is embedded in the signal, α_i is equal to zero, and thus (13) is approximately zero. Please note that signal amplification in the channel has no effect on the value of the correlation coefficient.

The precision of the approximation in (13) is related to the magnitude of N_g . The effect of $\sum_{i=1}^{N_g} s_i w_i$ is much larger than $\sum_{i=1}^{N_g} n_i w_i$ because the values of s_i are relatively large. In order to satisfy the design goal of individually authenticating each 0.5 sec block of speech, the number of selected coefficients (i.e. the group size N_g) is equal to 1000. As shown in Figure 3, the distributions of correlation coefficients for the cases of content replacement and no tampering are *not* clearly separated under this group size.

If the group size is doubled to 2000, which would require a block size of 1 second, the separation is significantly enhanced. As shown in Figure 4(a) with 10^4 trials, the correlation coefficient of content replacement is always below 0.075, and that of no tampering is always greater than 0.12. The horizontal axis and the vertical axis in the figure represent the correlation coefficient value and the probability, respectively. It is also shown in Figure 4(b)-(d) that content preserving operations such as white noise pollution and G.721 speech coding do not tighten the gap between these two distributions, but CELP coders such as GSM-AMR would eliminate the gap. However, since the overlap is not serious, a threshold could still be found to simultaneously achieve a high tamper detection rate and a low false detection rate. For example, in the case of GSM-AMR coding as shown in Figure 4 (d), setting the threshold to 0.040 would yield a tamper detection rate of 95.71% and a false detection rate of 10^{-4} . Although 4.29% of malicious content replacements are not detected, the hacker has no way of knowing which ones would not be detected.

However, the block size of 1 sec is much greater than the duration of a typical syllable in speech. Therefore, if the hacker only replaces one syllable instead of the whole second, the probability of failing to detect the alteration will be much higher than 4.29%. Nevertheless, for a common layman who is unaware of the details of our algorithm, it is more likely for him/her to replace or zero out a whole chunk of speech rather than precisely cut out syllables. Fragile speech watermarking in the DFT magnitude domain

by using linear addition provides an flexible approach to detecting this kind of tampering.

In summary, speech watermarking using the exponential-scale odd/even modulation is better in detecting localized content alteration, but its limitations include inability to tolerate low bit-rate speech coders such as the CELP coders and the initial latency needed to produce detection results. Linear additive speech watermarking overcomes these two limitations, but it may not detect malicious alterations by the hacker.

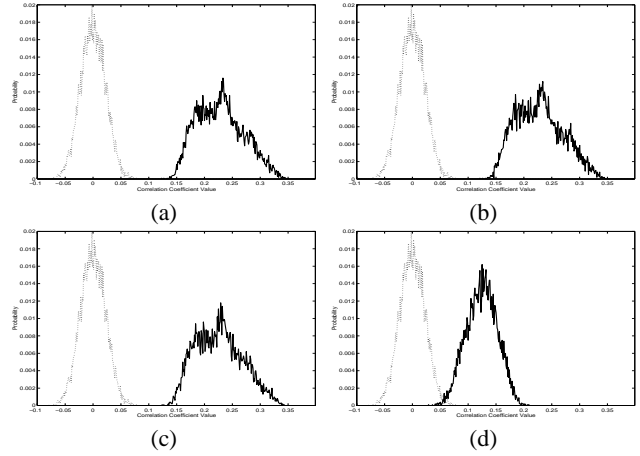


Figure 4: The distribution of correlation coefficient values of malicious content replacement in comparison with that of (a) no distortion, (b) white noise pollution, (c) G.721 ADPCM speech coding, and (d) GSM-AMR speech coding. The probability values in each plot are generated from 10^4 trials, and the dotted curve in each figure is the probability of correlation coefficient under malicious content replacement.

7. REFERENCES

- [1] J. Fridrich. Image watermarking for tamper detection. *Proc. ICIP '98*, October 1998.
- [2] H. Inoue, A. Miyazaki, and T. Katsura. Wavelet-based watermarking for tamper proofing of still images. *IEEE International Conference on Image Processing*, Sept. 2000.
- [3] 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.
- [4] 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.
- [5] C.-S. Lu, H.-Y. Liao, and L.-H. Chen. Multipurpose audio watermarking. *Proceedings of 15th IAPR International Conference on Pattern Recognition*, Sept. 2000.
- [6] 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.
- [7] C.-P. Wu and C.-C. J. Kuo. Speech content authentication integrated with celp speech coders. *ICME2001*, August 2001.
- [8] 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.