

Comparison of Two Speech Content Authentication Approaches

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

Speech content authentication, which is also called speech content integrity or tamper detection, protects the integrity of speech contents instead of the bitstream itself. In this work, two major approaches for flexible speech authentication are presented and compared. The first scheme is based on content feature extraction that 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 second method embeds fragile watermarks into the speech signal in the frequency domain. If the speech signal is tampered, the secret watermark sequence is also modified. The receiver detects the fragile watermark from received data and compares it to the original secret sequence. These two approaches are compared in terms of computational complexity, false detection rate, and tolerance to mis-synchronization and content preserving operations. It is shown that each approach has its own merits and shortcomings.

Keywords: content authentication, speech integrity, tamper detection, fragile watermarking

1. INTRODUCTION

Traditional cryptographic integrity schemes with hash functions, such as the one shown in Figure 1, 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 allow more implementational flexibility for audiovisual data. Integrity verification algorithms that focus on the perceptual content could allow channel noise and *content preserving* operations to exist in the transmission channel without triggering the authentication alarm.

A typical scenario of audiovisual data transmission with content authentication is shown in Figure 2(a), where the “transmission channel” could be either real-time transmission, storage devices, or a combination of both. To facilitate transmission and to maximize the playback quality, signal processing techniques are often performed on audiovisual data at various stages after their creation. The processing that occurs before the sender generates (or embeds) authentication data or after the receiver produces authentication results would have no influence on integrity verification. It is however often desirable to process the audiovisual data within the channel to accommodate the versatile multimedia transmission environment. Examples 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 background noise in the recording. Since these operations in the channel do no change the semantic meaning of the transmitted data, it is desirable that the receiver can distinguish them from malicious modifications. An authentication system that determines whether the semantic meaning of audiovisual data is tampered or not is called the *content authentication* (or *content integrity*) system.

While the simple hash function scheme as shown in Figure 1 cannot tolerate content preserving operations, a content integrity system could actually be built by using multiple cryptographic hash functions. This concept was briefly mentioned in,¹ and we will analyze it in detail below. Figure 2(b) is the system diagram of a content integrity system with hash functions. The central task of the agents in the channel is to perform content preserving operations ($A_i \rightarrow A_{i+1}$). However, in order for the hash data and the audiovisual data at the receiver end to be consistent with each other, each agent in the middle must also modify the hash data accordingly ($H_i \rightarrow H_{i+1}$). In order to access the hash data, agents should be provided with the secret key K

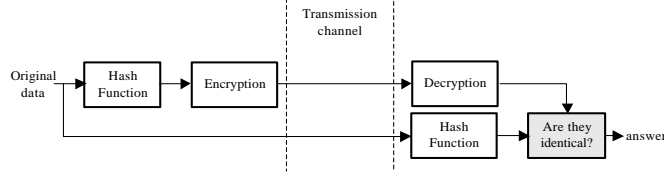


Figure 1: Blockdiagram of the traditional hash function integrity scheme

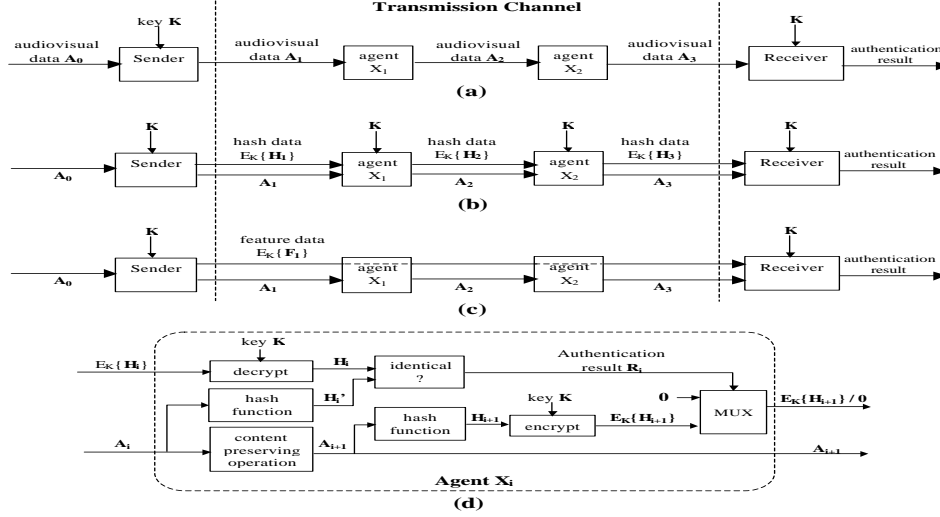


Figure 2. (a) A typical scenario of the content authentication system, (b) a content authentication system with hash functions in each agent, (c) a content authentication system with feature extraction, and (d) an internal structure of an agent in (b).

that the sender and the receiver use to encrypt and decrypt the hash data. A possible design of the internal structure of an agent is shown in Figure 2(d). The agent verifies the integrity of its input data A_i by calculating its hash and comparing it to the decrypted hash data. Its output hash data $E_k(H_{i+1})$ is conditionally zeroed out to signify that the input data A_i is not authentic.

There are several major limitations with this hash-function-based content authentication scheme. First, each agent must be equipped with proper software or hardware modules to verify the input data and generate the authentication code for the output data with hash functions. If one of the agents does not possess this capability, the whole scheme cannot work. Moreover, the receiver has to trust all agents to perform integrity verification on their input data correctly. The final verification result would be wrong if any hash verification or hash generation module in the channel fails. Finally, the secret key K is more likely to be compromised because it is shared not only between the sender and the receiver, but also among all agents. When there are many potential points of key leakage, it is more difficult to trace which party leaks the key.

A more robust approach called the *feature extraction* method is shown in Figure 2(c). As opposed to hash data, content features extracted from A_0 by the sender should be insensitive to noise and content preserving operations in the transmission channel. The agents do not process the encrypted feature data and merely pass it along. The receiver extracts content features from the received data A_3 and compares them to the decrypted content features F_1 . In this scheme, the participation of agents is reduced to the minimum. They need not perform feature extraction or encryption/decryption, and the secret key is only shared between the sender and the receiver. Thus, there are fewer points for processing errors to occur, and the secret key is more secure. In our previous work^{2,3} we presented a speech content feature extraction scheme for integrity verification, which is integrated with CELP speech codecs to minimize the overall computation cost.

Fragile watermarking for content integrity entirely eliminates the participation of the agents. 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. The similarity between them decides whether audiovisual data has gone through malicious alteration in the channel. Since each agent could be completely ignorant of the authentication process, the diagram of this scheme is actually simplified to the case as shown in Figure 2(a). In this paper, we propose a novel fragile speech watermarking algorithm and compare it to the feature extraction based scheme.

This paper is organized as follows. The speech feature extraction scheme for content integrity verification is summarized in Section 2. Previous work related to fragile speech watermarking is reviewed in Section 3. A novel speech watermarking algorithm using exponential scale quantization is presented in Section 4. Experimental results and their analysis are given in Section 5. Finally, concluding remarks are provided in Section 6.

2. SPEECH FEATURE EXTRACTION INTEGRATED WITH CELP SPEECH CODECS

A summary of our work on speech content authentication using feature extraction is presented in this section. More details could be found in our previous papers.²⁻⁴

2.1. Selection of Speech Features for Effective Tamper Detection

In image authentication, popular content features are often low-level features such as the mean intensity of image blocks⁵⁻⁸ and the edge information.^{9,10} For the speech signal, the corresponding audio features can be the short-time energy function and the short-time zero crossing rate. The computational cost of extracting low-level features is usually low, but the size of these features tends to be large. This could demand a heavy computational cost on encryption/decryption operations and a considerable amount of transmission resource.

The overall sizes of the short-time energy function and the short-time zero crossing rate depend on the window size used. For the same piece of speech data, larger window sizes result in fewer coefficients. However, larger windows also allow forgery to be done more easily. In our experiments,⁴ we are able to replace words in speech without significantly perturbing these two speech features when the window size is large enough. We conclude that, under a sampling rate of 8000 samples/sec, the window size should be set to 20 samples when either of these speech features is used alone, and set to 40 samples when these two features are used together. Both of these schemes require a feature data rate of approximately 400 bytes/sec, which is close to that of the compressed speech signal. Therefore, encrypting and transmitting the short-time energy function and/or the short-time zero crossing rate is no better than encrypting the whole coded speech bitstream.

Even though it may be possible that a cocktail of other low-level speech features could result in a lower data rate, randomly choosing combinations of features for testing is an inefficient approach. In order to achieve smaller feature data size, we need to extract speech features that are focused on the meaning of speech rather than signal characteristics. We have chosen 3 kinds of features that are relevant to speech semantic meaning.

1. *Pitch information*

The pitch information represents the intonation of a sentence and the emphasized syllable of each word. The emphasized syllable in each word has a higher pitch than others.

2. *The changing shape of the vocal tract*

The shape of the vocal tract determines the vowels and some consonants, which are important to speech semantic meaning.

3. *Energy envelope*

The energy envelope of a speech signal controls the temporal location of each syllable and the number of syllables in each time frame. Controlling the envelope can detect whether syllables has been deleted or added.

The computational cost of extracting the first two features is much higher than that of low-level features, but we could dramatically reduce this cost by integrating feature extraction with speech coding.

Resynchronization methods	Total number of math operations required
Correlation function calculated directly	400N
Correlation function calculated using FFT	128N
Salient point extraction	16N

Table 1: Comparison between the computational cost of different resynchronization methods.

2.2. Synchronization between Sent and Received Speech Signals

Before the receiver extracts speech features from the received speech signal, mis-synchronization between sent and received speech signals must be resolved. Otherwise, speech features extracted from the received signal obviously would not match the feature values of the original signal. Mis-synchronization could be caused by content preserving operations such transcoding or D/A-A/D conversion.

One trivial way for regaining synchronization is to encrypt a segment of the original speech signal and send it to the receiver. The receiver calculates the cross-correlation function between this encrypted segment and the received speech to find out the amount of mis-synchronization and realign received signals. According to our observations, the size of the encrypted segment should be at least 200 samples to avoid getting erroneous results. Assuming N samples in the received speech signal need to be checked for resynchronization, direct calculation of the cross-correlation function would require $200N$ multiplications and $200N$ summations. The total amount of basic math operations could be reduced to $128N$ if the cross-correlation function is computed via FFT and multiplication in the frequency domain. This is a standard technique for computing linear/circular convolution and correlation functions.¹¹

We propose to use an alternative approach for resynchronization, i.e. the salient point extraction method. It requires much less computation than either of the above techniques. This method is adopted from our previous work on robust audio watermarking.¹² Salient points are extracted as positions where the audio signal energy is fast climbing to a peak. These positions indicate the start of talk bursts, and they are relatively stable under content preserving operations. The location of the first 15 salient points in the original speech are encrypted and transmitted together with the authentication data. At the receiver end, salient point locations are detected from the received audio signal and compared to decrypted data to identify the amount of time-shifting during transmission. As shown in Table 1, the proposed algorithm requires less than $16N$ math operations, which is much less than the cost of the correlation function methods.

2.3. Integrating Feature Extraction with CELP Speech Coders

Since modern speech codecs are designed based on a human speech generation model, the resulting coefficients capture the semantic content of speech pretty well. The most popular speech coders are: CELP coders, which include ITU G.723.1, GSM-AMR and MPEG4-CELP. They model the changing shape of the speaker vocal tract with LSP analysis and extract the pitch from the speech signal. The cost of our speech content feature extraction process is greatly reduced by obtaining “the pitch information” and “the shape of the vocal tract” from the final or intermediate output of speech coders instead of the raw signal.

The algorithms for obtaining speech content features from these CELP coders are basically the same. Only the parameters need to be fine-tuned for each coder. The details of the algorithm are described below.

A. Feature extraction

In CELP coders, the “lag of pitch predictors” control the pitch information while the “LSP codebook indices” model the changing shape of the vocal tract. Among the 10 LSP coefficients in each frame, we only use the first 3 as content features because they contribute the most to the model. One pitch coefficient is obtained from each frame as the average of the “lag of pitch predictors” of all subframes.

The extraction of the first two features, i.e. the LSP coefficients and the pitch information, requires virtually no computation. At the sender side, calculation of the energy ratio between adjacent frames requires only one division operation per frame because the the frame energy is a byproduct of the LSP analysis already performed in the CELP coders. At the receiver side, the calculation of the frame energy requires one multiplication operation (to obtain sample energy) and one addition operation (to perform summation) for each speech sample.

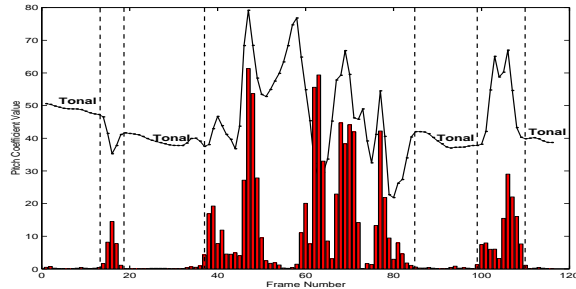


Figure 3. Pitch coefficients (the curve) extracted from a segment of the original speech signal compared with the absolute difference (the bars) between pitch coefficients extracted from the original and the received signals. Pitch information calculation is integrated with the AMR coder at both ends of the transmission channel.

B. Silent period and tonal region identification

Speech signals generally contain a considerable amount of silent periods between words and sentences, where only background noise exists. LSP coefficients in these regions do not model the shape of the speaker vocal tract, and thus could be greatly altered by content preserving operations. Consequently, it is pointless to encrypt and transmit them. Instead, the starting and ending locations of silent periods should be transmitted. While there are elaborate schemes in detecting silence periods, we use a low cost method to suit our purpose. The frames whose energy is less than 5% of the average energy of a 10-sec surrounding region are considered silent frames.

Pitch coefficients extracted from CELP speech coders in the previous step have no physical meanings in non-tonal regions such as fricative sounds and silent periods. Therefore, content preserving operations easily perturbs pitch coefficients in these regions. Similar to LSP coefficients, the location of non-tonal regions should be transmitted instead of pitch coefficients in these regions. As shown in Figure 3, tonal regions could be identified as frames whose pitch coefficients are very close to their neighboring frames.

C. Feature difference calculation

The feature difference calculation of the three features are done independently. For LSP coefficients, we take the weighted average of the 3 LSP coefficients and then compute the difference between decrypted and extracted results. For pitch information and frame average energy, we also compute the difference between decrypted and extracted features. The feature difference is only calculated for the LSP and the pitch coefficients in non-silent periods and tonal regions, respectively.

D. Threshold comparison

Before differences are compared to the threshold, a low-pass filter is applied to the difference sequence. This step ensures that random burst-type errors do not trigger the false alarm. The low-pass filter is implemented with a moving averaging window.

3. 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.

3.1. Fragile Watermarking for Image Authentication

Fragile watermarking techniques for content authentication are called semi-fragile¹³ watermarking to indicate that they are not completely fragile to all kinds of modifications. Algorithms in this category usually embed watermarks in frequency domains such as the DCT domain^{13,14} or the wavelet domain,¹⁵ and a brief review of these schemes could be found in.²

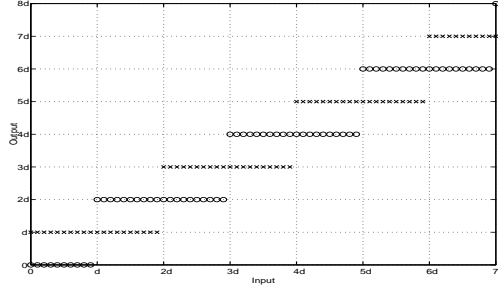


Figure 4. The plot of the embedding function for odd/even modulation, where the horizontal axis and the vertical axis represent the value of s_i and \hat{s}_i , respectively.

A watermarking technique called *odd/even modulation* is adopted in several image tamper-proofing schemes recently.^{15,16} 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:

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

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 4, 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

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

to each \tilde{s}_i of the received signal. When the noise $(\tilde{s}_i - \hat{s}_i)$ 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 suppression¹⁷ 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*)¹⁵ 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.

3.2. Audio Watermarking Techniques

A variety of audio watermarking methods were reviewed in our paper on robust audio watermarking for copyright protection.¹⁸ In this subsection, we discuss the potential of applying these methods to speech content

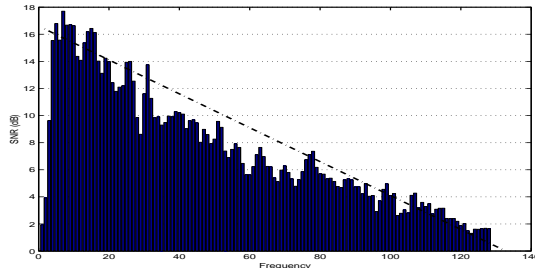


Figure 5. The average SNR of GSM-AMR in the DFT magnitude domain, where GSM-AMR operates in the 5.15kbps mode, and the DFT window size is 256. Note that the other half of the spectrum is omitted due to symmetry.

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.^{19,20} 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.²¹ 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.^{19,22} Thus, hackers are not able to detect and restore the watermark after modifying the speech signal. In following section, we will propose a fragile speech watermarking scheme that uses pseudo-random noise as the watermark.

4. FRAGILE SPEECH WATERMARKING USING EXPONENTIAL SCALE QUANTIZATION

For the feature extraction based speech content integrity scheme in Section 2 to work properly, all agents that perform content preserving operations in the transmission channel must participate in the scheme by passing the feature data along. Sometimes, it may be inconvenient or impossible to require such a level of cooperation. In this section, we investigate fragile speech watermarking for tamper detection, which requires no auxiliary data so that agents in the channel could be completely ignorant of the authentication process.

4.1. 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 unfavorable choice.

The SNR ratios of CELP speech codecs in the DFT magnitude domain are much larger than those in the time domain. Figure 5 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.

4.2. Exponential Scale Odd/Even Modulation

In this subsection, 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 approach¹⁶ 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.*²³ 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 odd/even modulation, (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 5, 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 by the dotted line in Figure 5. 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 pseudo-random sequences a_i and b_i could be optionally incorporated in watermark embedding and detection. In this case, (4) and (5) are modified to be

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

and

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

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 $\hat{s}_i - s_i$.

The computational complexity of exponential scale odd/even modulation is dominated by the exponential and logarithm functions that are performed on every watermarked DFT coefficient. The resulting computation cost is much higher compared to the cost of speech feature extraction integrated with CELP codecs.

4.3. Synchronization between Sent and Received Speech Signals

Similar to the feature extraction scheme, mis-synchronization between sent and received speech signals must also be resolved before fragile watermark detection. However, the re-synchronization scheme in Section 2.2 is not suitable for fragile watermarking because it transmits the location of the first 15 salient points as auxiliary data.

Without explicitly transmitting the salient point locations, we could utilize salient point extraction by embedding the fragile watermark beginning from the first salient point in the speech signal, and detecting the watermark starting from the first salient point at the receiver side. However, the success of this simple method depends on the stability of a single salient point, which could shift due to content preserving operations in the channel. Although minor mis-synchronization would not cause disastrous effects to DFT domain watermarking,^{12, 18} excessive shifting would reduce the success rate of watermark detection. According to our observations, the drifting of salient point locations is usually under 10 samples, but sometimes could be as high as 50 ~ 100 samples.

To enhance the synchronization precision, a synchronization signal is embedded in the time domain starting from the location where the DFT domain fragile watermark begins, i.e. the location of the first salient point

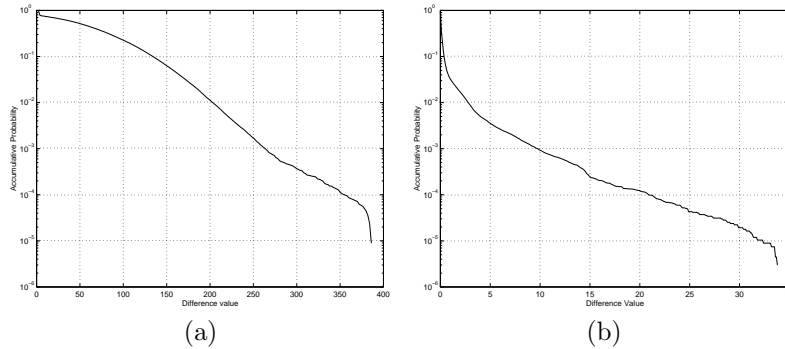


Figure 6. The effect of transcoding on false positive rates: (a) the cumulative probability function of the LSP coefficient difference between the original and the received signals and (b) the cumulative probability function of the pitch coefficient difference between the original and the received signals.

in the original speech signal. This synchronization signal is a pseudo-random sequence with its frequency spectrum shaped like that given in Figure 5. The receiver performs a localized search around the first salient point extracted from the received speech data. The position that leads to the highest correlation between the synchronization signal and the received speech is the correct location to start fragile watermark detection. Note that if the pseudo-random noise is used as the only synchronization tool instead of an enhancement phase, the search space would be significantly expanded. This does not only greatly increase the computation cost of searching, but also produces more synchronization errors.

The computation cost and implementation complexity of the hybrid re-synchronization scheme in this section is higher compared to the re-synchronization scheme in Section 2.2, which only uses salient point locations.

5. EXPERIMENTAL RESULTS

5.1. Speech Feature Extraction Integrated with GSM-AMR Speech Coder

We conducted our experiments by using CELP coders such as G.723.1 and GSM-AMR. GSM-AMR is used to demonstrate our experimental results in this paper. Experimental results using G.723.1 could be found in our previous work as given in.^{2,4} We found² that transcoding usually causes more alterations in speech features than other content preserving operations, such as re-compression, amplifying, resampling and D/A-A/D conversion. Therefore, we examine the false alarm probability of the transcoding operation in the transmission channel with statistical analysis by using 336,000 frames of speech (each frame is 20ms long). The transcoding operation transforms AMR coded data into the G.723.1 bitstream, and then back to AMR again.

Figure 6(a) shows the empirical cumulative probability function of LSP coefficient differences plotted in the semi-log scale. The resulting curve is approximately linear, which indicates that the probability for a frame to be falsely classified as maliciously altered is 10^{-5} when the threshold is set at 380. Figure 6(b) shows that the probability of false alarm in pitch difference is also 10^{-5} when its threshold is set at approximately 32. Similarly, the false positive rate of the frame average energy could be calculated in the same fashion.

5.2. DFT Domain Fragile Speech Watermarking

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 7, 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 μ -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 (with proper anti-aliasing) followed by

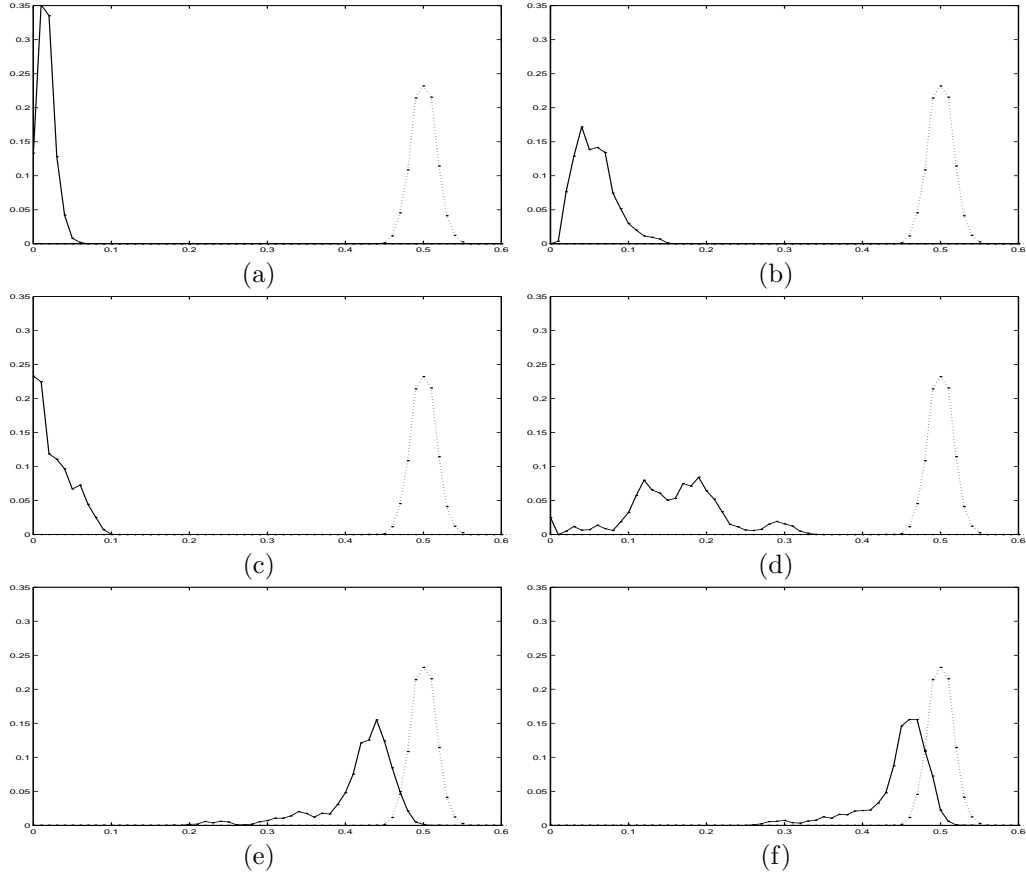


Figure 7. The TAF distribution of the malicious content replacement in comparison with that of (a) resampling, (b) white noise pollution, (c) G.711 μ -law speech coding, (d) G.721 ADPCM speech coding, (e) G.723.1 speech coding, and (f) GSM-AMR speech coding. The probability values in each plot are generated 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 1000, which corresponds to 0.5 sec of speech data.

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 7(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 7 (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.

6. CONCLUSION

Two approaches for speech content authentication, feature extraction integrated with CELP speech coders and fragile watermarking using exponential scale quantization, were presented and compared in this paper. Both

algorithms authenticate each 0.5 second of speech and achieve comparable false detection rates. The fragile speech watermarking scheme allows for more implementation flexibility because the agents in the channel are not required to pass along auxiliary data. However, the speech feature extraction scheme has a much lower computation cost and a more straightforward re-synchronization algorithm. The feature extraction scheme is fully integrated with low-bitrate CELP coders while the exponential scale odd/even modulation requires more research to improve its compatibility with the CELP coders.

REFERENCES

1. C.-Y. Lin and S.-F. Chang, "Issues and solutions for authenticating MPEG video," *SPIE Security and Watermarking of Multimedia Contents*, January 1999.
2. C.-P. Wu and C.-C. J. Kuo, "Speech content integrity verification integrated with ITU G.723.1 speech coding," *ITCC2001*, pp. 680–684, April 2001.
3. C.-P. Wu and C.-C. J. Kuo, "Speech content authentication integrated with celp speech coders," *ICME2001*, August 2001.
4. C.-P. Wu and C.-C. J. Kuo, "Robust content integrity verification of G.723.1-coded speech," *The 2001 International Conference on Imaging Science, Systems, and Technology*, June 2001.
5. D.-C. Lou and J.-L. Liu, "Fault resilient and compression tolerant digital signature for image authentication," *IEEE Transactions on Consumer Electronics* **46**, pp. 31–39, February 2000.
6. C. Rey and J.-L. Dugelay, "Blind detection of malicious alterations on still images using robust watermarks," *IEE Seminar on Secure Images and Image Authentication*, pp. 7/1–7/6, April 2000.
7. M. Schneider and S. F. Chang, "A robust content based digital signature for image authentication," *Proceedings of IEEE ICIP'96*, pp. 227–230, 1996.
8. M. Wu and B. Liu, "Watermarking for image authentication," *ICIP'98*, pp. 437–441, October 1998.
9. M. P. Queluz, "Towards robust, content based techniques for image authentication," *Proceedings of IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, December 1998.
10. M. Steinder, S. Iren, and P. D. Amer, "Progressively authenticated image transmission," *MILCOM 1999*, pp. 641–645, November 1999.
11. A. Oppenheim and R. Schaffer, *Discrete-Time Signal Processing*, Prentice-Hall, 1989.
12. C.-P. Wu, P.-C. Su, and C.-C. J. Kuo, "Robust audio watermarking for copyright protection," *SPIE 44th Annual Meeting*, July 1999.
13. J. Fridrich, "Image watermarking for tamper detection," *Proc. ICIP '98*, October 1998.
14. 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**, January 2000.
15. D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proceedings of the IEEE - Special Issue on Identification and Protection of Multimedia Information* **87**, pp. 1167–1180, July 1999.
16. H. Inoue, A. Miyazaki, and T. Katsura, "Wavelet-based watermarking for tamper proofing of still images," *IEEE International Conference on Image Processing*, Sept. 2000.
17. M. Ramkumar and A. Akansu, "Self-noise suppression schemes for multimedia steganography," *SPIE Workshop on Voice, Video and Data Communication, Multimedia Applications* **3845**, Sept. 1999.
18. C.-P. Wu, P.-C. Su, and C.-C. J. Kuo, "Robust and efficient digital audio watermarking using audio content analysis," *SPIE 12th International Symposium on Electronic Imaging* **3971**, pp. 382–392, Jan. 2000.
19. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal* **35**(3-4), p. 313, 1996.
20. J. F. Tilki and A. A. Beex, "Encoding a hidden auxiliary channel onto a digital audio signal using psychoacoustic masking," in *IEEE Southeastcon 97*, pp. 331–333, 1997.
21. J. Lacy, S. Quackenbush, A. Reibman, D. Shur, and J. Snyder, "On combining watermarking with perceptual coding," in *ICASSP*, **6**, pp. 3725–3728, 1998.
22. I. Pitas and P. Bassia, "Robust audio watermarking in the time domain," *EUSIPCO'98*, pp. 25–28, 1998.
23. C.-S. Lu, H.-Y. Liao, and L.-H. Chen, "Multipurpose audio watermarking," *Proceedings of 15th IAPR International Conference on Pattern Recognition*, Sept. 2000.