

# SPEECH CONTENT AUTHENTICATION INTEGRATED WITH CELP SPEECH CODERS

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

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

## ABSTRACT

A speech content authentication scheme, which is integrated with CELP speech coders to minimize the total computational cost is proposed in this research. Speech features relevant to semantic meaning are extracted, encrypted and attached as the header information. A low cost synchronization algorithm is used to resolve mis-synchronization caused by content preserving operations. Silent and tonal regions in the speech are identified with algorithms of low complexity to enhance the precision of integrity verification. This scheme is not only much faster than traditional cryptographic bitstream integrity algorithms, but also more compatible for a variety of applications. Experimental results are collected by using the GSM-AMR speech coder, and statistical analysis is performed to calculate the false positive rate of tamper detection.

## 1. INTRODUCTION

Tamper detection for multimedia data is increasingly essential to secure communications since modern digital media editing and processing technologies allow high quality forgery to be created at a low cost. While most previous work focused on image/video data, speech integrity is just as valuable for many applications. In fact, it is much easier to tamper with speech without leaving perceptible artifacts. The semantic meaning of speech could be altered by simply reordering several sentences or dropping out a few words. Thus, judging the authenticity of speech data by human perception alone is inadequate.

While traditional cryptographic integrity schemes could be applied to speech data by simply treating it as a binary stream, speech data integrity should be aimed at the protection of the *content* (or called the *semantic meaning*), not the bitstream itself. Integrity verification algorithms that focus on the perceptual content allow channel noise and *content preserving* operations to exist in the transmission channel without triggering the authentication alarm[1]. Such ability makes *content integrity verification* systems compatible with applications such as Internet servers that automatically adjust audio signal volume or perform transcoding.

In order to detect malicious content alterations correctly and efficiently, we have argued [2] that an effective content integrity protection scheme should satisfy all of the following requirements:

1. Compatibility with lossy compression before transmission is necessary for efficient use of network bandwidth.
2. Content integrity verification should be performed without the need of the original data.
3. The introduced noise, if there is any, should be small enough to be imperceptible.

4. Content preserving operations in the transmission channel should be tolerated to facilitate flexible applications.
5. The size of auxiliary authentication data, if there is any, should be significantly smaller than the audiovisual data to be authenticated.
6. Localized malicious modifications should be detected.
7. Low computational cost at both the sender and the receiver side is important to real-time media processing.

## 2. PREVIOUS WORK ON SPEECH INTEGRITY VERIFICATION

### 2.1. Hash Functions

Speech integrity verification could be implemented with traditional cryptographic hash function schemes. As shown in Figure 1(a), the hash function takes the large-size bitstream as the input and produces a small fixed-size *hash code*. Changing any bit in the bitstream would completely alter the hash code so that it is practically impossible to find a way to modify the message yet keeping the hash code remain the same. The hash code is encrypted and sent to the receiver along with the message. The receiver uses the same hash function on the received data and compares the result with decrypted hash code. While hash functions could ensure that every bit in the speech data is unmodified, it cannot tolerate any content preserving operation performed by unknown agents in the transmission channel.

### 2.2. Fragile Watermarking

Fragile watermarking embeds a secret sequence into the host audiovisual data. If the host data are tampered, the secret sequence is also modified. The receiver calculates the correlation between the watermark sequence and the received data. If the correlation is beneath a threshold, it indicates that the data has been modified. While quite a few proposed audio watermarking schemes claim to possess the potential for authentication applications, none of them addresses the problem of speech content integrity directly. We reviewed some of these algorithms in our previous work, e.g. [3, 4].

Currently, we are not aware of any proposed audio/speech fragile watermarking scheme attempting to satisfy all the requirements listed in Section 1. In our evaluation, the most challenging requirement is to detect the fragile watermark in every 0.5 second of speech with a low false positive/negative error probability, especially when modern speech coders compress 0.5 second of speech to less than 400 bytes of data.

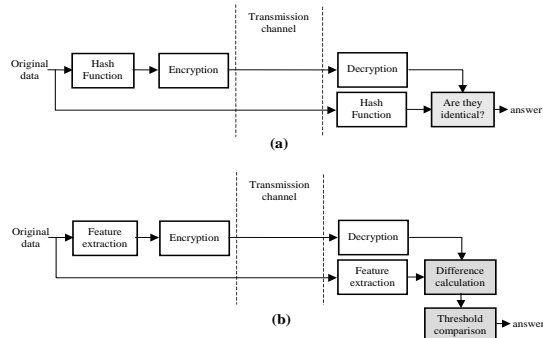


Figure 1: Blockdiagrams of (a) the traditional hash function integrity scheme and (b) the feature extraction method for content integrity verification.

### 2.3. Feature Extraction

As shown in Figure 1(b), the system structure of the feature extraction method is similar to the traditional cryptographic hash function method. It replaces the hash function with feature extraction, and the bit-wise comparison unit with threshold comparison. Depending on the application, the encryption unit in the diagram could be secret-key encryption, public-key encryption or other cryptographic methods. Since the effect of different encryption algorithms has been thoroughly discussed in the cryptographic literature, the focus of this field is on the feature extraction unit. Extracted content features should be insensitive to noise and content preserving operations in the transmission channel so that the result of difference calculation will not exceed the threshold. Furthermore, features should be closely tied with the semantic meaning of speech data so that altering the semantic meaning cannot be done without altering extracted features.

It has been suggested that one possible candidate for the speech content feature is the spoken text, which could be extracted using speech recognition[5]. However, the accuracy of speech recognition is typically only around 90%, which is much lower than the requirement of an authentication system. Moreover, the computation and storage cost of a large-vocabulary speech recognition system is disproportionately larger than any speech coder. In order to avoid high computational costs, Dittmann *et al.* investigated the use of a low-level feature, the signal sign change, as the content feature[5]. The scheme is shown to be tolerant to MP3 compression for some audio pieces, but its reliability varies with different audio categories. Furthermore, it is unlikely to be tolerant to low bit-rate speech coders.

### 3. SELECTION OF SPEECH FEATURES FOR EFFECTIVE TAMPER DETECTION

Since fragile watermarking seems to be incompatible with low bit-rate speech coders, we explore the feature extraction approach for speech content integrity. The selection of speech features should conform to all requirements discussed in Section 1.

In image authentication, popular content features are often low-level features such as the mean intensity of image blocks [6, 7, 8, 9] and the edge information [10, 11]. 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 more focused on the meaning of speech rather than the 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 energy envelope can detect whether any syllable 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.

## 4. PROPOSED ALGORITHM

The proposed algorithm consists of a low cost scheme to re-synchronize the original and the received speech signals, and a speech feature extraction algorithm integrated with CELP speech coders to reduce the computational cost.

### 4.1. 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. If the receiver does not have the knowledge about the de-

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.

tails of all operations that took place in the transmission channel, it cannot anticipate the amount of temporal shifting involved.

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 [12].

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 [13]. The algorithm is basically unchanged from that presented in [13], but the parameters are now tuned for speech signals sampled at the rate of 8kHz. 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 locations 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 the 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.

## 4.2. Integrating Feature Extraction with CELP Speech Coders

Since modern speech coders 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.

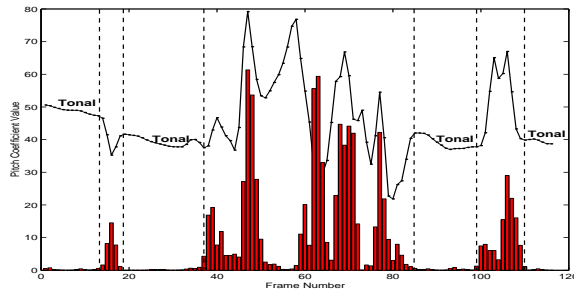


Figure 2: Pitch coefficients (the curve) extracted from a segment of the original speech signal compared with the absolute difference (the bars) between pitch coefficients extracted from the original and the received signals. Pitch information calculation is integrated with the AMR coder at both ends of the transmission channel.

### 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 these two kinds of features 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 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 the sample energy) and one addition operation (to perform summation) for each speech sample.

### 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 ten-second 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 perturb 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 2, the absolute differences between pitch coefficients extracted from the original and the received signals are small in tonal regions and large in non-tonal regions. The tonal regions could be identified as the 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

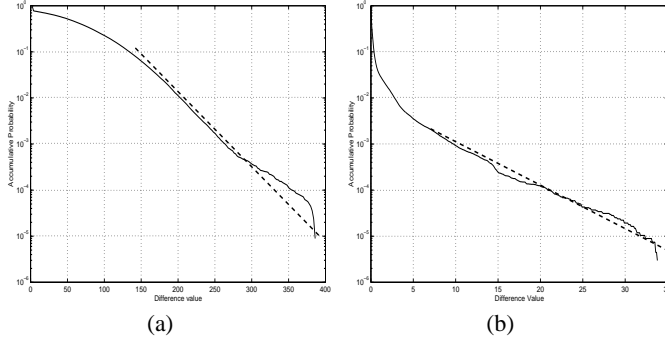


Figure 3: The effect of transcoding on false positive rates: (a) the cumulative probability function of the LSP coefficient difference between the original and the received signals and (b) the cumulative probability function of the pitch coefficient difference between the original and the received signals.

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.

## 5. EXPERIMENTAL RESULTS

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 found [2] 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 caused by 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 3(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 3(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.

## 6. CONCLUSION

The importance of tamper detection for speech data is rapidly increasing since new applications enabled by modern speech communication and processing technologies need integrity protection

to reach their full potential. A speech authentication system focusing on protecting the semantic meaning of speech data was presented in this work. It is tolerant to various kinds of content preserving operations so that the benefit of digital processing technologies is not inhibited. The system is integratable with CELP speech coders to achieve a very low computational cost in feature extraction. Techniques such as re-synchronization with salient points and identification of silent and tonal regions are used to ensure the correctness of authentication without incurring much additional computation. There is still work to be done in the future. For example, we would like to perform more experiments with various CELP speech coders and lower the false detection rate of the system furthermore.

## 7. 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. *IEEE International Conference on Information Technology: Coding and Computing*, pages 680–684, April 2001.
- [3] 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*, 3971:382–392, Jan. 2000.
- [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] J. Dittmann, M. Steinebach, I. Rimac, S. Fischer, and R. Steinmetz. Combined video and audio watermarking: Embedding content information in multimedia data. *SPIE International Conf. on Security and Watermarking of Multimedia Contents II*, 3971(14), January 2000.
- [6] D.-C. Lou and J.-L. Liu. Fault resilient and compression tolerant digital signature for image authentication. *IEEE Transactions on Consumer Electronics*, 46(1):31–39, February 2000.
- [7] 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*, pages 7/1–7/6, April 2000.
- [8] M. Schneider and S. F. Chang. A robust content based digital signature for image authentication. *Proceedings of IEEE International Conference on Image Processing (ICIP'96)*, pages 227–230, 1996.
- [9] M. Wu and B. Liu. Watermarking for image authentication. *Proceedings of International Conference on Image Processing*, 2:437–441, October 1998.
- [10] 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.
- [11] M. Steinder, S. Iren, and P. D. Amer. Progressively authenticated image transmission. *MILCOM 1999*, 1:641–645, November 1999.
- [12] A. Oppenheim and R. Schaffer. *Discrete-Time Signal Processing*. Prentice-Hall, 1989.
- [13] C.-P. Wu, P.-C. Su, and C.-C. J. Kuo. Robust audio watermarking for copyright protection. *SPIE 44th Annual Meeting*, July 1999.