Eng. & Tech., Vol.25, No.10, 2007

A Content-Based Authentication Using Digital Speech Data

Dr. Hana'a M. Salman*
Received on: 3/7/2007
Accepted on: 4/11/2007

Abstract

A watermarking technique for speech content and speaker authentication scheme, which is based on using abstracts of speech features relevant to semantic meaning and combined with an ID for the speaker is proposed in this paper. The ID which represents the watermark for the speaker, is embedded using spread spectrum technique while the extracted abstracts of speech features are used to represent the watermark for the speech, embedded in the original speech file using secret key. The abstracts speech feature is implemented using B-spline curve interpolation. The paper provides a background knowledge for the concept of speaker watermarking and content-fragile watermarking based on digital speech data. Then, the suggested feature based authentication scheme is developed and the results from the evaluation are presented. Show that the suggested scheme is successful in combining speech and speaker watermark authentication.

Keywords: speech content authentication, digital watermarking for speaker authentication, content-fragile watermarking authentication, spline interpolation, wavelet transform.

الخلاصة

في هذا البحث تم اقتراح مخطط لتقنية العلامة المائية لمحتوى الكلام وتخويل المستخدم، والذي مبني على استخدام ملخصات عن مبرزات خصائص الكلام ذات العلاقة بالمعنى مع معرف المستخدم. إن معرف المستخدم يمثل العلامة المائية بالنسبة إلى المستخدم، والتي تغمر بواسطة استخدام تقنية الانتشار الطيفي، بينما ملخصات عن مبرزات خصائص الكلام ذات العلاقة بالمعنى تعتبر العلامة المائية بالنسبة إلى الكلام، والتي تغمر بصورة مباشرة في ملف الكلام بواسطة مفتاح سري.تم استخدام تقنية توليد المنحنيات الأساسية لتوليد ملخصات عن مبرزات خصائص الكلام. البحث يوفر قاعدة معرفة لمفهوم العلامة المائية للمستخدم والعلامة المائية ذات المحتوى الرقيق المبني على أساس الكلام. تم بناء مخطط التخويل باستخدام المبرزات المقترح ومن ثم تمت عملية تقيم للنتائج وتثبيتها. أن النتائج تشير إلى فعالية المخطط المقترح لدمج تخويل العلامة المائية لكلا من الكلام مع المستخدم.

1. Introduction

Speech watermarking is a technique of embedding a digital signal into speech signal using techniques that render the signal imperceptible [1]. Digital watermarking technique has an edge over speech processing for speaker authentication since a fraudster can mimic the voice of the user. A digital watermark can be created from user or transaction specific information, which can be embedded in the speech. The embedded information can then be

*Department of Computer Science & Information System
University of Technology-Baghdad, Iraq
salmanhana2007@yahoo.com

detected and verified at the receiver side to authenticate the speaker [1]. Most multimedia signals today are in digital formats which are easy to reproduce and modify without any trace of manipulations [2]. On other hand a certain loss of trust in media data can be observed. As a result the need for security increases particularly in the field of multimedia. Acording to the structure and complexity of multimedia, the purpose of applications, possibility and the way of applying security mechanisms to multimedia data is determined.

The security requirements such as integrity (unauthorized modification of data) or data authentication (detection of origin and data alterations) can be met by the succeeding security measures using cryptographic mechanisms and digital watermarking techniques [3]. Digital watermarking techniques based on steganoraphic systems embed information directly into the media data. Besides cryptographic mechanisms, watermarking represents an efficient technology to ensure both data integrity and data origin are authentic. All manipulations on multimedia can be classified into two categories, incidental and malicious manipulations [2]:

(1) ـ Incidental manipulations: Incidental manipulations do not change the authenticity of the perceptual content of multimedia, and should be accepted by an authentication system. Common ones include format conversions, lossless and high-quality lossy compression, A/D and D/A conversions, re-sampling, etc.

(2) ـ Malicious manipulations: Manipulations in this category change the perceptual quality or semantic meaning to a user, and thus should be rejected. They include cropping, dropping, inserting, replacing, reordering perceptual objects or video frames, etc.

Different applications may have different criteria to classify manipulations, hence different types of authentication are used. Multimedia authentication can be classified according to integrity criteria into three types: hard, soft, and content-based authentications. Hard authentication detects any modification to the content representation of digital multimedia. The only incidental manipulation accepted by the hard authentication is lossless compression or format conversions in which the visual pixel values or audio samples do not change. Soft authentication detects any manipulations that lower the perceptual quality below an acceptable level. Content-based authentication detects any manipulations that change the semantic meaning of the content to a user. This is normally achieved by authenticating perceptual features extracted from the media [2].

All the proposed authentication methods can be classified into two types of approaches: external signatures and watermarking. External signature approaches attach

the authentication data generated from a digital multimedia signal to the media by concatenation or in the format's header field. Most approaches of this type do not change the media they authenticate, but some do, especially for soft authentication. Watermarking, on the other hand, embeds authentication data into the media to be authenticated, due to the redundancy and irrelevancy contained in multimedia signals. Three types of watermarking have been developed: fragile watermarking which is designed to be fragile to any modification to the content, semi-fragile watermarking that is robust to some perceptual quality preserving manipulations but fragile to others, and robust watermarking which is robust to both signal processing and intentional attacks. Most proposed algorithms for hard or soft authentication are based on fragile or semi-fragile watermarking, and those for content-based authentication are based on either external signatures or robust/semirobust watermarking. The watermarking used for content-based authentication is called content-fragile watermarking. We note here that hard or content-based authentication can use either external signatures or watermarking approaches, but it is more convenient for soft authentication to use watermarking approaches [2].

The most important properties of digital watermarking techniques are: robustness, security, imperceptibility/transparency, complexity, speed, capacity, and possibility of verification and invertibility [2].

Our contribution focuses mainly on the design of a content-fragile speech and speaker watermarking scheme, by combining fragile feature extraction and robust speech watermarking.

The following subsections, give a review of the art of basic background knowledge for the concept of content-fragile watermarking based digital speech data. Secondly, a description of the general suggested scheme for is given in Sections 3. In Section 4, a test for the proposed method,is followed by a conclusions in Section 5.

2. The Concept of Content-Fragile Speech Watermarking

Fragile watermarking embeds a secret sequence into the host audiovisual data. If the host is tampered the secret sequence is also modified. The receiver calculates the correlation between the watermark sequence and the received data. If the correlation is below a threshold, it indicates that the data has been modified [4].

In the following sub sections: content-based authentication, and content-fragile watermarking concept are introduced.

2.1 Content-Based Authentication

A set of perceptual characteristics or features of a multimedia signal is called content. Content determines how human beings interpret a multimedia signal (The semantic meaning of media). Content-based

authentication is used to authenticate the content extracted from a multimedia signal instead of the signal itself. Content is represented by a vector called the feature vector. Content authentication is gauged by a distance of the feature vector of the original signal S0 from the feature vector of the test signal St whose authenticity is to be tested [2]:

$$d = \| feature\ (S_0) - feature\ (S_t) \| \quad (2.1)$$

If the distance d is larger than a preset, application-dependent threshold T, then the content is modified, otherwise the content is authentic. By measuring localized feature distances, a content-based authentication scheme may be able to give local tamper measurement [2].

All content-based authentication approaches need to first extract content from a multimedia signal. This is called feature extraction. These features are then used as the hash value in the classical signature-based authentication approach or embedded directly into the signal by applying some robust or semi-robust watermarking schemes, or both. The goal of content-based authentication schemes is to accept all manipulations that preserve the content of a multimedia signal while rejecting all other manipulations that modify the content [2]. The fulfilling of this goal depends mainly on choosing a set of features which adequately describes the content of a multimedia signal, while meeting the desired property that the features are fragile to the set of content modifying manipulations but robust to content preserving manipulations for a certain application.

2.2 Content-Fragile Watermarking Concept

Content-fragile watermarking can also be used for content-based multimedia authentication. Features extracted from digital media can be embedded into the media itself with a robust or semi-robust watermarking scheme instead of generating an external digital signature. Although it uses different ways to place the feature data, content-fragile watermarking shares other steps with its external signature counterpart. In content-fragile watermarking, the whole space of signal coefficients is divided into the three subspaces: signature generating, watermarking, and ignorable subspaces. Signal coefficients in the signal generating subspace are used to generate content-features which will be converted to authentication bits to be embedded into the signal itself. The coefficients in the watermarking subspace will be used to embed the authentication bits into the signal. Those coefficients in the ignorable subspace will not be used in either way. The division of subspaces should be kept secret or determined pseudo-randomly based on some secret key. The signature generating subspace can be overlapped with the watermarking subspace. If they are not overlapped, then there is no interference between the two procedures. This is much simpler than the overlapping case. The drawback is that it may be less

accurate in locating tamper regions. The overlapping approach is much more complex since the watermarking procedure modifies the signal slightly which may affect the accuracy of feature extraction from the watermarked signal. Care has to be taken to make sure that the watermarking procedure does not distort the feature extraction significantly enough to have a false alarm, especially. when the signal is under acceptable manipulations. A general approach to addressing this problem for the overlapping case is to carefully design the feature extraction algorithm so it is robust to the watermarking procedure, and/or to design the watermarking procedure which does not distort the feature extraction much. Another common approach is to use some iterative procedures to guarantee the extracted signatures, before and after watermarking, match each other [1].

3. The Proposed Content-Fragile Speech Watermark

In this section, the proposed content-fragile speech watermarking based on the concepts introduced in Section2 is presented, and Synchronization is between sent and received speech signals.

3.1 Content-Fragile Authentication

With an idea to use content feature vector as an indicator for manipulations of speech and also to embed a secure ID in the feature vector which, represent a speaker. The ID is a combination of the binary number representation for the user code number, and additional bit called the parity bit. The speech file represents the cover carrier, while the summarized extracted feature vector represents the watermark for the speech file. The general proposed content-fragile watermarking schematic is implemented. The embedding and extraction algorithm is presented bellow.

Proposed Embedding Algorithm

Input: Speech file, secret key , and ID

Output: Watermarked speech file.

Process:

Step1: A feature vector extraction for the input speech file, using the bellow sub steps:

Step 1.1: Multiply the absolute value of the input speech file by a hamming window with the same size of the speech file.

Step 1.2: Apply a daubechies four wavelet (db4), level 7, see [5] for more information

Step 1.3 Extract the approximation, which represents the feature vector, see Fig. (1).

Step2: Summarize the result feature vector by using B-spline interpolation, with a secret key "the feature vector is divided into frames of size 256 sample, and from each frame one value is taken. The positions pointing to the selected feature vector values, represent the secret key" the result is the summarized feature vector "checksum", which represents the watermark for the speech file.

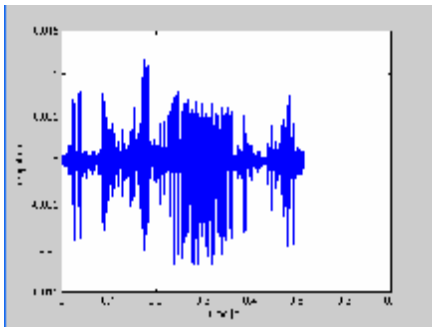
Step3: Use the result checksum to embed the ID using the below steps"

Step 3.1: Apply the Fast Fourier Transform to the result checksum

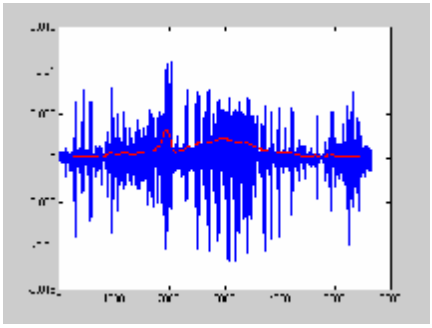
Step3.2: Extracted the real part
Step 3.3: Embed the ID bits into the LSBs
Step 3.4: Apply the Inverse Fast Fourier Transform
Step4: Embedded the result of step 3 into the original speech file using the secret key.
Step5: End

Proposed Extracting Algorithm

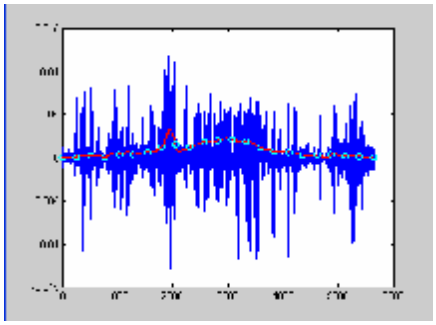
Input: Watermarked speech file, secret key , and ID
Output: Alarm for correct/modified, and fraud speech and speaker..
Process:
Step1: By using the secret key extract the visual speech data that correspond to it. The result is the summarized feature vector with the embedded ID in the LSBs.
Step2: For the result of step one extracted the LSBs
Step3: Compare the extracted ID, with original one. if they are identical the user is identified correctly, else observe the parity bit if it is a fraud one, else perform a synchronization operation, else an alarm to notify of a fraud user.
Step4: If the result of the step3 is correct, compare the extracted feature vector with original one after embedding the ID in it. If the result of comparison is the same then the speech is confirmed, else perform a synchronization operation, else an alarm to notify of a modified speech file.
Step5: End



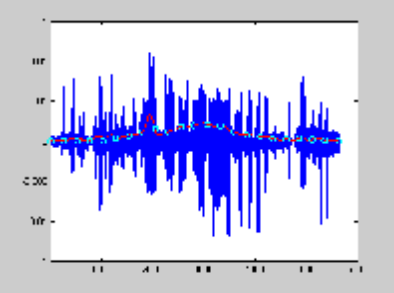
(a) The speech data



(b) The feature vector presented in red color



(c) The summarized feature vector



The embedding algorithm main stages

Figure(1)(a-d): The embedding algorithm main stages

3.2 Synchronization Between Sent and Received Speech Signals

Miss-synchronization could be caused by content preserving operations such transcending or D/A-A/D conversion. If the receiver does not have the knowledge about the details 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 "by using FFT to calculate the cross-correlation function" between this encrypted segment and the received speech to find out the amount of mis-synchronization and realignment received signals [1].

4. Test Results

A prototype implementation based on the proposed watermarking algorithm is implemented. The test results are concluded using the basic idea for tests, which can be described in the following steps:

1. Select a speech file as a cover to be secured according to specific user.
2. Select a feature that describes speech file and ID that represents a specific person.
3. Retrieve the features for a given amount of time.
4. Create a summarized feature vector as watermark for the speech file to be embedded in.
5. Embed the ID in the summarized feature as a watermark for speaker identification.
6. Embed the result feature checksum as a watermark for speech content.
7. Attack the cover.
8. Retrieve the watermark from the attacked cover.
9. Retrieve the features from the attacked cover and generate the checksums.
10. Retrieve the ID from the generated checksums.
11. Compare both to decide if a content-change has occurred or ID is fabricated.

For the same ID, and 10 Speech files the following results are attained, which represent the main requirements :

Audibility of the speech file:

The audibility of the watermark speech files is tested and found that, the embeded ID does not affect the

watermark speech file, the speech file is still audible effectively.

Removal of the ID:

The use of the secret key, spread spectrum for embedding, and the party bit gives the technique its robust agents removal.

Noise attack:

When a white Gaussian noise is added to the watermarked speech file the ID is extracted successfully even if the signal-to-noise ratio per sample is 100 dB. The test is implement with increase in 5dB, each time. Also the speech feature is extracted correctly.

Temporal attack:

The way in which the feature is extracted " by using the spectral envelop of the speech file" controls the temporal location of each syllable and the number of syllables in each time frame. Controlling the spectral envelop of the speech file can detect whether any syllable which has been added /deleted.

5. Summary and Conclusions

In general the fraud star can mimic the voice of the user so, the watermarking technique has an edge over speech processing for speaker authentication. on the other hand , the semantic meaning of speech could be altered by simply recording several sentences or by dropping out a few words. Content based watermarking technique for authentication preserve the semantic meaning of speech data. Finding such a scheme that combins speech and the speaker watermarking authentication technique is demanding. The proposed watermarking technique is suitable for various applications such as phone banking, flight voice communication and other similar applications.

The concept for digital speech content authentication and speaker identification is introduced. Content-fragile watermarking is based on combining robust watermarking and fragile content features. The robust nature of the watermark, due to the use of spread spectrum technique, and the right choice of content features and their compressions provide tolerance to such operations while they still enable us to identify content changes and the speaker.

The way in which, B-spline interpolation is implemented in the proposed method as abstraction to the feature vector since it performs lossless compression on the feature vector, also control points position as a secret key for the facture vector value selection and the spread spectrum for the ID over the abstracted feature vector.

The number of bits needed for the speaker ID depends mainly on the used number of control points, so the payload for the proposed method is good.

If a modification to ID holds, then the parity check bit could identify it, also if a modification to the speech file, "addition or/and cutting" the feature identify it, also the embedding ID could identify the modification.

The computation 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 used method for checksum gave the proposed method its cheapest cost, and its suitability for an efficient use of network bandwidth

The ID is inaudible and the used secret key spreads it in the frequency domain. ensures that only authorized users can detect the hidden information.

The extracted content features is insensitive to noise even if the signal to noise ratio is 100dB.

Content integrity verification for the Speaker is performed without the need of the original data. The size of authentication data (ID), is significantly smaller than the audiovisual data to be authenticated.

The used spectrum technique makes the watermarking technique more robust against attacks such as watermark removal and impairment attacks.

6. References

[1]: Shilpa A. etal. "Adaptive Spread Spectrum Based Watermarking of Speech". School of engineering Nanyang Technological University, Singapore, 2002.

[2]: Bin B. and etal. ,"Multimedia Authentication and Watermarking". Chapter 7, page no. 149-178, 2002.

[3]: Chung P. W. etal, "Speech Content Authentication Integrated with CELP

Speech Coders," IEEE Int. Conf. on Multimedia and Expo (ICME2001), pp. 1009-1012, Aug, 2001.

[4]: Martin S., etal. "Watermarking–Based Digital Audio Data Authentication",

Eurasip Journal on Applied Signal Processing, 2003

[5]: Burrus C. S. Gopenath, R. A., etal. "Introduction to Wavelet and Wavelet Transform",. Practice Hall, Inc., 1998.