Signal Processing 134 (2017) 9–22

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Steganalysis of adaptive multi-rate speech using statistical characteristics of pulse pairs

Hui Tian[a,*], Yanpeng Wu[a], Chin-Chen Chang[b], Yongfeng Huang[c], Yonghong Chen[a], Tian Wang[a], Yiqiao Cai[a], Jin Liu[a]

[a] College of Computer Science and Technology, National Huaqiao University, Xiamen 361021, China
[b] Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan
[c] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Steganalysis
Steganography
Adaptive multi-rate speech
Pulse-pair statistical characteristics
Adaptive boosting
Support vector machine

ABSTRACT

In this work, we concentrate on the steganalysis of adaptive multi-rate (AMR) speech, which widely exists in mobile Voice-over-IP services. Compared with the state of the arts, the most significant advantage of this work is that more accurate and more complete steganalysis features are presented. To avoid the impact of the possible interchange of pulse positions in one track, we characterize AMR speech exploiting the statistical properties of pulse pairs. Specifically, we employ the probability distributions of pulse pairs as long-term distribution features, extract Markov transition probabilities of pulse pairs as short-term invariant features, and adopt joint probability matrices of pulse pairs as features based on track-to-track correlations. Moreover, to optimize the feature set and reduce its dimension, we introduce a feature selection mechanism using adaptive boosting technique. Exploiting the well-selected features, we further present a steganalysis of AMR speech based on support-vector-machine. The proposed method is evaluated with a good supply of AMR-encoded speech samples, and compared with the state-of-the-art methods. The experimental results demonstrate that the proposed method can effectively detect the state-of-the-art steganography methods for AMR speech, and significantly outperforms the state-of-the-art steganalysis methods on detection accuracy, false-positive rate and false-negative rate for any given embedding rate or speech samples with any given length. In particular, the proposed method can provide accurate detecting results for the existing steganographic methods in a timely manner, and thereby be applied in the steganalysis scenario for real-time speech streams.

1. Introduction

Steganography is an efficient means of hiding secret messages into seemingly innocent carriers (e.g. image [1, 2], video [3, 4], audio [5, 6], text [7, 8]) without perceptible distortion, and hence popularly employed to achieve covert communications [9]. In the latest years, with the rising popularity of VoIP (short for Voice over IP, also called IP telephony) and the fast proliferation of its extensions, such as mVoIP (mobile VoIP) and VoIM (voice over instant message), steganography based on VoIP has become one of hot research topics [10–20]. Compared with traditional steganographic carriers, VoIP possesses many particular advantages, including instantaneity, a large mass of carrier data, high covert bandwidth and flexible conversation length [10, 18–20]. Thus, VoIP-based steganography is believed to be a promising solution for secure communications [10, 12, 14, 16, 18–20]. However, like many other security techniques, VoIP-based steganography might also be employed by lawbreakers, terrorists and hackers for illegitimate purposes, which would facilitate cybercrimes and become a serious threat to cybersecurity [10, 20], because unauthorized secret data with its assistance can be easily smuggled through network firewalls and monitors [11, 14]. Thus, for the cybercrime preventation, it is crucial to develop the countermeasure technique of VoIP-based steganography to detect illegal covert behaviors effectively, i.e., steganalysis of VoIP.

In general, there are two basic types of approaches to achieve VoIP-based steganography [18–20]. One utilizes the specific protocols involved in VoIP as the carrier [11–14], and the other conceals secret messages in speech streams [15–20]. The latter type, by contrast, has attracted most attention in the research community over the past years [10]. In this paper, we mainly concentrate on detecting steganography on adaptive multi-rate (AMR) speech streams, which widely exist in mobile VoIP services.

With the swift progress of mobile communication techniques, wireless communication based on mobile networks has already become

* Corresponding author.
E-mail address: htian@hqu.edu.cn (H. Tian).

http://dx.doi.org/10.1016/j.sigpro.2016.11.013
Received 16 June 2016; Received in revised form 2 October 2016; Accepted 17 November 2016
Available online 21 November 2016

an important communication mode in people's daily lives [21]. In the current wireless world, adaptive multi-rate (AMR) speech codec [22–25], because of its exellent charateristics of adaptive rates and high speech quality, is employed as the standard codec for 3G and 4G services by both 3GPP (3rd Generation Partnership Project) and ITU-T (International Telecommunication Union Telecommunication Standardization Sector) [23–27], and also widely employed in VoIP services by many popular apps for mobile instant messaging (e.g. WeChat, WhatsApp and Snapchat). Moreover, AMR is also a standard file format used for recording spoken audio, extensively supported by almost all smart phones. Due to its widespread usage, AMR speech naturally attracts the attention of the steganographic community, and fruitful studies have been conducted up to now [28–31].

For analysis-by-synthesis (ABS) codecs, fixed codebook indices (FCIs), accounting for a high percentage in each speech frame, are popularly considered to be good candidates for steganographic covers in practice [32–34]. This idea is also extended to ABS-based AMR codec [28, 31]. Geiser and Vary [28] proposed an alternative fixed-codebook-search strategy for joint data hiding and speech coding in the AMR bitstream, in which two bits of secret data can be concealed into a given track pulse by restricting the second FCI searching within two out of eight possible values. For the AMR speech codec at 12.2 kbit/s mode, the experimental results show that this method can achieve a steganographic rate of 2 kbit/s with a negligible effect on the subjective speech quality and reasonable computational complexity. Based on the similar principle, Miao et al. [29] further presented an adaptive suboptimal pulse combination constrained method to achieve information hiding in the FCI searching phase of the AMR codec. Differing from Geiser's steganography, this method introduces an embedding factor to regulate embedding capacities, and by properly setting it, can strike a good trade-off between steanographic transparency and embedding capacity.

On the other hand, to detect potential steganography in AMR speech streams, some steganalysis studies have been carried out in the recent years. Specifically, Miao et al. [30] proposed two methods for steganalysis of AMR speech, Markov-based method and Entropy-based method. The former employs Markov transition probabilities to assess the relationship between pulse positions in each track, while the latter uses the joint entropy and conditional entropy to measure the uncertainty of pulse positions [30]. The experimental results show that both the two methods can detect Geiser's steganography [28] and Miao's steganography [29], while the former can reach higher accuracy than the latter. However, the above two methods ignore the fact that the pulse positions may often be interchanged in the AMR encoding process, because the order of two pulses in the same track depends on their signs and positions. In other words, the pulse positions in one track should be interchanged when their order does not conform to their signs [25]. In this sense, the statistical features for pulse positions presented by Miao et al. [30] are not accurate enough for characterizing AMR speech. In addition, Ren et al. [31] presented a set of features for detecting AMR-based steganography exploiting probabilities of same pulse positions (SPP). Their steganalysis method (called Fast-SPP) based on these features can also detect the two exisiting AMR-based steganography methods [28, 29], and outperforms Miao's steganalysis methods [30] on detection rate in some cases. In the work, SPP features for cover samples and steganographic ones with the embedding rate of 100% are proved to be different. However, these features only describe the distributions of two track-pulses being in the same position, and are thereby not complete enough to characterize AMR speech. As the embedding rates decrease, the distribution difference between the cover SPP features and the steganographic ones would become not apparent any more. Thus, the steganalysis using the SPP features can not achieve high accuracy for the steganography with low embedding rates. Extremely, if one conducts a sophisticated steganography that avoids the change for the distributions of two track-pulses being in the same position (it is easy to realize this steganograhy by abandoning the track-pulses with the same positions and the ones that would become the same after the embedding operation), Ren's steganalysis would not detect any abnormities.

Motivated by the existing problems in the state of the arts [30, 31], we present more accurate and more complete features for steganalysis of AMR speech in this paper. Considering the possible interchange of pulse positions in one track, we characterize AMR speech using the statistical properties of pulse pairs, each of which is two pulses in one track. To be more specific, the probability distributions of pulse pairs are employed as long-term distribution features, Markov transition probabilities of pulse pairs are adopted in the light of the short-term invariant characteristic of speech signals, joint probability matrices of pulse pairs are utilized to characterize the track-to-track correlation. Moreover, to optimize the feature set and reduce its dimension, we conduct feature slection before training using adaptive boosting (AdaBoost) technique [35, 36]. Further, using the optimized feature set, we design a support-vector-machine (SVM) based steganalysis of AMR speech. The proposed method is evaluated with a large quantity of AMR-encoded speech samples, and compared with the state of the arts. The experimental results demonstrate that the proposed method significantly outperforms the previous ones on detection accuracy, false positive rate and false negative rate for any given embedding rate or speech samples with any given length.

The rest of this paper is structured as follows. To make this paper self-contained, Section 2 first introduces the background, including the principle of AMR codec and an overview of the exisiting AMR-based steganography methods, and then reviews the state of the arts for steganalysis of AMR speech. Section 3 presents three types of features based on statistical characteristics of pulse pairs for steganalysis of AMR speech. Section 4 first describes the feature selection mechanism based on AdaBoost and then presents the steganalysis scheme based on SVM, followed by the evaluation and experimental results reported in Section 5. Finally, the concluding remarks are given in Section 6.

2. Background and related work

In this section, we first introduce the principle of AMR codec in brief, then give an overview of the steganographic methods for AMR speech [28, 29], finally review the state of the arts for AMR-based steganalysis [30, 31] and analyze their deficiencies.

2.1. Principle of AMR codec

AMR codec now has two categories, AMR-Narrowband (AMR-NB) and AMR-Wideband (AMR-WB), which are intended for different voice frequency bandwidths [26, 27]. The former uses a speech bandwidth of 300–3400 Hz, whereas the bandwidth of the latter is 50–7000 Hz. However, both of them employ the same coding algorithm, ACELP (short for algebraic code-excited linear prediction). In this algorithm, the encoder obtains the excitation signal by selecting two appropriate code vectors respectively from adaptive and fixed codebooks. The speech is synthesized by feeding the two chosen vectors through a linear prediction synthesis filter [25]. Moreover, an analysis-by-synthesis search procedure for choosing the optimum excitation sequence is involved to minimize the perceptually weighted error between the original and synthesized speeches, and accordingly gain the best speech quality. In the following text, we typically take the AMR-NB at 12.2 kbps mode for example to introduce the principle of the AMR codec and AMR-based steganography.

Table 1 shows the bit allocation of the AMR-NB at 12.2 kbps mode [25]. For each 20 ms speech frame, there are 244 bits produced, of which 140 bits are assigned for the fixed codebook indices (FCIs, also called algebraic codebook indices). Apparently, FCIs occupy a singnificant proportion (57.38%) of all frame bits, which is one of the advantages for the steganographic cover. Table 2 shows the structure of the algebraic codebook for the AMR-NB (12.2 kbps), which adopts an interleaved single-pulse permutation (ISPP) design [25]. That is, 40