available at www.sciencedirect.com

ScienceDirect

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

Digital Investigation

Speaker recognition from encrypted VoIP communications

L.A. Khan [a], M.S. Baig [b], Amr M. Youssef [a,*]

[a] Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada H3G1M8
[b] Centre for Cyber Technology and Spectrum Management, NUST, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 10 June 2009
Received in revised form
30 August 2009
Accepted 15 October 2009

Keywords:
Forensic investigation
Speaker identification
Speaker verification
VoIP
Encryption
Classification

ABSTRACT

Most of the voice over IP (VoIP) traffic is encrypted prior to its transmission over the Internet. This makes the identity tracing of perpetrators during forensic investigations a challenging task since conventional speaker recognition techniques are limited to unencrypted speech communications. In this paper, we propose techniques for speaker identification and verification from encrypted VoIP conversations.

Our experimental results show that the proposed techniques can correctly identify the actual speaker for 70–75% of the time among a group of 10 potential suspects. We also achieve more than 10 fold improvement over random guessing in identifying a perpetrator in a group of 20 potential suspects. An equal error rate of 17% in case of speaker verification on the CSLU speaker recognition corpus is achieved.

1. Introduction

Recent statistics show shrinking market share for traditional public switched telephone networks (PSTNs). This decline of the PSTN market share is a direct result of the substitution from voice platforms as fixed wire-line operators migrate customers to all-IP voice platforms and as consumers opt for mobile voice platforms, which will also become all-IP. Unlike traditional telephony systems where calls are transmitted through dedicated networks, voice over IP (VoIP) calls are transmitted through the Internet, a mix of public and private networks, which presents a threat to the privacy and confidentiality of VoIP communications. In order to overcome this problem, VoIP traffic is usually encrypted prior to its dispatch over the Internet (Provos). Encrypting VoIP traffic, on one hand, helps to preserve the privacy and confidentiality of legitimate users, but on the other hand might be exploited in criminal activities. Scammers, terrorists and blackmailers may abuse the end-to-end encryption facility to conceal their true identity. To address the problem of anonymity and to identify or confirm the true speaker of a disputed anonymous speech, the area of speaker recognition has long been the focus of forensic investigations. Research in speaker recognition has a relatively long history starting from introducing the term *voiceprint* identification (Kersta, 1962) in 1962 to the tremendous development in the field of automatic speaker recognition during the last decade (Reynolds, 2002) which is marked by the National Institute of Standards and Technology (NIST) evaluation campaigns (Martin and Przybocki, 2009; Przybocki et al., 2006, 2007). Famous cases include the 1974 investigation of a conversation of the former US president, Richard Nixon, and the former chief of staff, Harry Haldeman, which was recorded in the executive office building in 1972 (Advisory Panel on White House Tapes, 1972). The authentication of the speech recordings of Osama Bin Laden and other terrorists (Sachs, 2003) using modern automatic speaker recognition techniques has also been used as the last resort to provide some forensic evidence in these recent cases.

* *Corresponding author*. Tel.: +1 514 848 2424/5441; fax: +1 514 848 3171.
E-mail address: youssef@ciise.concordia.ca (A.M. Youssef).
1742-2876/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2009.10.001

Automatic speaker recognition can be divided into two categories: *identification* and *verification*. In the former scenario, given a set of suspected speakers together with their recorded speech segments, the problem is to determine the likelihood that a disputed encoded speech segment belongs to one of these suspects. In the latter scenario, a forensic investigator is given a disputed speech segment along with a set of recordings of a potential perpetrator and is asked to check if both sets of the speech segments originate from the same individual (Koolwaaij and Boves, 1999). Both scenarios are addressed in this paper but from the perspective of encrypted VoIP communications. Existing speaker identification and verification techniques are employed for analyzing un-encrypted speech only. To the best of our knowledge, there is no such study available for *encrypted* speech.

Variable bit rate (VBR) encoding techniques, which result in variable length VoIP packets, have been introduced to preserve the network bandwidth. The encryption techniques currently in use in order to preserve privacy of the calling and called parties do not change the packet length (Baugher and McGrew, 2003). Hence any exploitation mechanism based on the packet-length information remains valid for the encrypted communication. In this paper, we propose speaker identification and verification techniques based on using the packet-length information without even knowing the contents of the encrypted VoIP conversations. We demonstrate that the packet-length information, being extracted from either the file headers (in case of multimedia container formats) or being physically monitored during a VoIP conversation, can be used to identify or verify the speaker. In particular, we use discrete hidden Markov models to model each speaker by the sequence of packet lengths produced from their conversation in a VoIP call. Tri-gram probabilities of the packet length sequences were also used to create Gaussian mixture models and decision trees, based on these probability distributions, for each speaker. Various statistical modelling and classification/regression techniques were also applied, out of which the ensemble of nested dichotomies (ENDs) achieved more than 10 fold improvement over random guessing in identifying a speaker from a group of 20 suspects. In case of speaker verification, an equal error rate of 17% was obtained using support vector machine (SVM) based regression techniques.

The significant contributions of our approach are:

(1) We are the first, to the best of our knowledge, to apply speaker identification and verification to *encrypted* VoIP conversations.
(2) The recently developed container formats which are used to store and carry multimedia information over the Internet are explored from the perspective of speaker recognition in case of encrypted communications.
(3) Our experimental results indicate that different types of classification and regression techniques, that are usually used in data mining and machine learning applications, outperform both the Gaussian mixture models and the hidden Markov models, the classifiers which perform very accurately in the conventional speaker recognition studies.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in the area of speaker recognition as well as the packet-length information exploitation in encrypted VoIP conversations. The basic idea behind our work is discussed in Section 3. The problem statement of our work is presented in Section 4 and the proposed approach is explained in Section 5. Section 6 presents the experimental evaluation and the paper is concluded in Section 7.

2. Related work

Although significant work has been done in the area of speaker recognition, throughout this section, we only focus on two pertinent approaches: the Gaussian mixture model universal background model (GMM-UBM) (Reynolds and Rose, 1995), and the mixed GMM-UBM and SVM technique (Campbell et al., 2006). These models are commonly used in text-independent speaker recognition problems especially in speaker verification or source confirmation disputes. The mixed GMM-UBM and SVM approach combines the modelling efficacy of Gaussian mixtures and the discriminative power of SVMs and has shown significant improvement in terms of identification accuracies. In the case of speaker identification, the accuracy measurement is simple and can be termed as the ratio of the correctly identified speech segments to the total number of segments in a group of speakers. This accuracy measure is greatly dependent on the potential number of suspects; increasing the population size reduces the accuracy. Speaker verification, being a two-class classification problem, can generate two types of errors, namely false rejection (rejecting a valid speaker) and false acceptance (accepting an invalid speaker). The probabilities of these two events are denoted as P_{fr} and P_{fa}, respectively. Both errors depend on the value of the threshold set for classification. It is, therefore, possible to represent the performance of the system by plotting P_{fa} versus P_{fr}, a curve that is generally known as the detection error trade-off (*DET*) curve. In order to judge the performance of speaker verification systems, different performance measures are in place, among which the equal error rate (*EER*) and minimum detection cost function (*minDCF*) are the most popular ones. The *EER* corresponds to the point where $P_{fa} = P_{fr}$. The *minDCF* punishes strictly the false acceptance rate and is defined as the minimum value of $0.1 \times \textit{false rejection rate} + 0.99 \times \textit{false acceptance rate}$ (Kinnun et al., 2009). Another noticeable work in the field of speaker recognition is the national institute of standards and technology (NIST) speaker recognition evaluation (SRE) framework used to evaluate different text-independent speaker recognition techniques and models (Przybocki et al., 2007). It started in 1996 and continues untill this paper was published.

As of now, there is no study available as far as speaker recognition from encrypted speech is concerned. However, Wright et al. (2007, 2008) have studied the utilization of the packet-length information in extracting some crucial information about encrypted VoIP traffic. In particular, the authors were able to identify the spoken language of the transmitted encrypted media with an average accuracy of 66%. In the second case, partial speech contents were extracted using the