DIGITAL INVESTIGATION XXX (2009) 1–9

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 un-encrypted 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

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

packet-length information. Both of these techniques do not disclose the identity of the speakers beyond their language of communication.

3. Main observation

The basic idea behind this work stems from observing the relationship between the speakers' identity and the length of the packet carrying their VoIP speech contents. In order to save bandwidth, especially in case where, on average, 63% of the time one of the two channels in VoIP calls is idle, variable bit rate (VBR) encoding has been introduced (Chu, 2003). VBR techniques allow the codec to change its bit rate dynamically to adapt to the acoustics of the audio being encoded. Sounds like vowels and high-energy transients require a higher bit rate to achieve good quality, while fricatives such as the *s* and *f* sounds can be coded comparatively with fewer bits. For this reason, for a given bandwidth, VBR can achieve lower bit rates for the same quality. As demonstrated in Wright et al. (2007), the bit rate used for encoding speech and the length of the packets carrying it are in perfect synchronization. The reason for this is the fixed frame length (10–30 ms, typically 20 ms) in the case of speech compression and encoding mechanisms used in VoIP communications. For example, a VBR encoder operating at 13.6 kbps will produce a packet length of 34 bytes, excluding the header, for speech sampled at 8000 samples per second.

In order to determine the correlation of the speaker's identity with packet lengths of the encoded speech, we conducted some experiments on the AN4 low vocabulary speech recognition database (Acero, 1993) which consists of identical phrases uttered by different speakers. Fig. 1 shows the histogram of the packet lengths with Gaussian curve fitting for the same phrase uttered by three different speakers randomly selected from the AN4 speech recognition corpus and encoded by the VBR encoder, Speex (Valin and Montgomery, 2006), which encodes this phrase with eight distinct bit rates resulting into eight different packet lengths. It is interesting to note that different speakers produce different distributions for the same eight packet sizes. Various experiments were conducted with similar speech contents but uttered by different speakers. Throughout our experiments, the distribution of frame length sequences produced by different speakers were visually distinguishable. This dependence of the frame length sequencing on the speaker's identity encouraged us to model each speaker with respect to the corresponding frame length sequences.

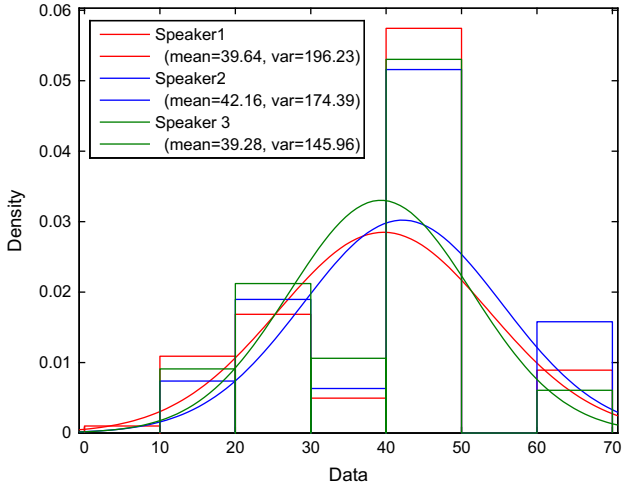

Fig. 1 – Histogram of the packet lengths of letters "P I T T S B U R G H" spoken by different speakers.

In order to understand the effect of encryption on the frame lengths, we studied the practically in-vogue security mechanisms in VoIP communications. One of the techniques proposed to secure real time speech communication over the Internet is to tunnel VoIP over IPSec but this proposal has a serious limitation of inducing unacceptable delays on the real time traffic (Barbeieri et al., 2002). In order to address the issue of confidentiality and privacy in VoIP communications without compromising the quality of service, the real time transport protocol (RTP) for multimedia applications was replaced by the secure real time transport protocol (SRTP) (Baugher and McGrew, 2003). SRTP standardizes only a single encryption algorithm namely the Advanced Encryption Standard (AES) which can be used in two cipher modes: the segmented integer counter mode and the f8 mode. As clearly mentioned in Baugher and McGrew (2003) "none of the pre-defined encryption transforms uses any padding; the RTP and SRTP payload sizes match exactly." Hence the packet-length information remains unchanged after encryption and all exploitation techniques based on this information remain as valid after encryption as they are before encryption.

4. Problem statement

The break up of the speaker recognition as a forensic investigation problem is given by the following two sub-problems.

4.1. *Speaker identification*

Given a set of n suspected speakers $\{S_1, \ldots, S_n\}$ and a disputed anonymous encrypted speech segment O, the investigator is asked to identify the speaker $S \in \{S_1, \ldots, S_n\}$ of the anonymous speech segment. The identified speaker is the one which gives the maximum value for $p(S|O)$. It is assumed that the investigator has access to m speech segments $V = \{v_1, \ldots, v_m\}$ for each suspect and the true speaker is assumed to be one of the suspected speakers.

4.2. *Speaker verification*

Given an observation sequence in the form of a speech segment O, and a hypothesized speaker S, speaker verification is a basic hypothesis test between H_0 (O is from the hypothesized speaker S) and H_1 (O is not from the hypothesized speaker S). If $p(O|H_0)/p(O|H_1) \geq \theta$, we accept H_0, otherwise reject H_0 (accept H_1) where $p(O|H_i)$, $i = 0, 1$ is the probability density function for the hypothesis H_i evaluated for the observed speech segment O and θ is the decision threshold for accepting or rejecting H_0.

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

5. Proposed approach

Fig. 2 shows an overview of our proposed approach. First, the packet-length information from encrypted VoIP conversations is extracted and used to create suitable models for the different speakers. The unknown communication can then be classified using some classification/regression techniques on the basis of the pre-trained models of each speaker. These steps are further explained throughout the rest of this section.

5.1. *Packet length extraction*

The packet-length information can be extracted from the encrypted VoIP conversation using one of the numerous open source packet sniffing tools or any of the techniques

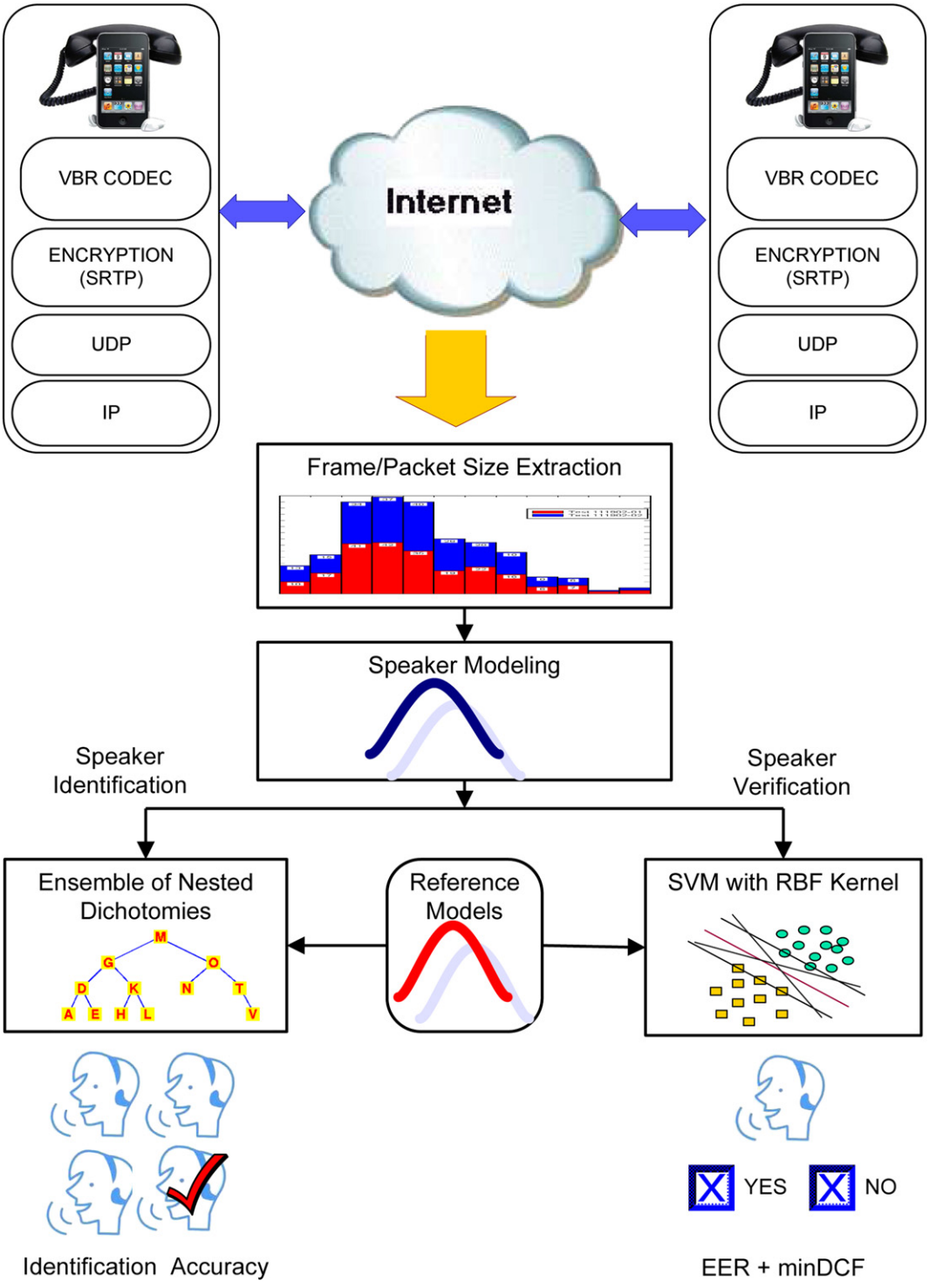

Fig. 2 – Overview of the proposed approach for speaker identification from encrypted VoIP communications.

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

mentioned in Wright et al. (2007, 2008). In addition, multimedia container formats like the Ogg (Pfeiffer, 2003) and MPEG-4 (Koenen, 2002) are also used for carrying multimedia information in packets over the Internet. For the voice transmission over packet switched networks, Ogg uses the Speex encoder which handles voice data at low bit rates (8–32 kbps/channel) and the Vorbis encoder which handles general audio data at mid to high-level variable bit rates (16–500 kbps/channel) (Valin and Montgomery, 2006). To cater for the VBR encoding mechanism, the Ogg page headers reserve a specific field in the page header as a segment table. On one hand, this provides for efficient multiplexing and seamless integration of multimedia data independent of the underlying compression and encoding mechanisms. On the other hand, the segment length information can be retrieved from the page header without even inspecting the internal packet contents.

5.2. *Modelling and classification*

For modelling the speaker's identity with respect to the variable packet-length information, various modelling approaches are explored. In case of conventional speaker recognition approaches, the observations on which the models are based are multidimensional vectors generally in the form of Cepstral features extracted per frame (Bimbot et al., 2004). In the encrypted VoIP scenario, we only know the length of the frame, or for that matter the packet, which is a scalar value per frame. We used discrete hidden Markov models (HMMs) to create a model for each speaker based on the sequence of packet-length information. An HMM is characterized by two model parameters, M and N, and three probability measures A, B and π, where N denotes the number of states $\{S_1, \ldots, S_N\}$ in the model, M denotes the number of distinct observation symbols per state, A is the state transition probability distribution, B is the observation symbol probability distribution and π is the initial distribution. HMMs use the Baum Welch expectation maximization algorithm (Baum et al., 1970) for calculating the model parameters and the Viterbi search (Forney, 1973) for finding the most likely sequences of hidden states.

The other approach which is also used in Wright et al. (2007) for language identification is to create tri-gram models out of the frame lengths for each speaker. Using tri-gram probabilities, speakers are modelled as GMMs with different model orders. Mathematically, a GMM $\lambda(\mathrm{y})$ is given by (Campbell et al., 2006)

$$\lambda(\mathrm{y}) = \sum_{i=1}^{M} \alpha_i G(\mathrm{y}; m_i; \Sigma_i) \quad (1)$$

where $G(\mathrm{y}; m_i; \Sigma_i)$ is a Gaussian model with mean m_i and covariance Σ_i. The weight of each mixture is represented by α_i and M denotes the total number of Gaussians. Different expectation maximization algorithms are used to find the optimum values of mixture weights (Dempster et al., 1977) where each Gaussian in the mixture has its own mean and covariance matrix that has to be estimated separately (McLachlan, 1988).

In addition to the HMMs and GMMs described above, we have also tested several other classifiers. The classifier which obtained the best accuracy for speaker identification was based on the ensemble of nested dichotomies (END) (Frank and Kramer, 2004) which is a recently introduced statistical technique for tackling multi-class problems by decomposing it into multiple two-class problems. Using the C4.5 decision tree (Quinlan, 1992) and logistic regression (Agresti, 2002) as base learners, the ensemble of nested dichotomies shows better classification accuracies compared to the case where we apply these learners directly to multi-class problems. The probability estimates produced by binary classifiers are multiplied together, considering these to be independent, in order to obtain multi-class probability estimates. Nested dichotomies can be represented as binary trees. At each node, we divide the set of classes A associated with the node into two subsets, B and C, that are mutually exclusive such that B and C together contain all the classes in A. The nested dichotomies root node contains all the classes of the corresponding multi-class classification problem. Let L_{i_1} and L_{i_2} be the two subsets of class labels produced by a decomposition of the set of classes L_i at internal node i of the tree and let $p(l \in L_{i_1} | y, l \in L_i)$ and $p(l \in L_{i_2} | y, l \in L_i)$ be the conditional probability distribution estimated by the two class model at node i for a given instance y. Then the estimated class probability distribution for the original multi-class problem is given by

$$p(l = L|y) = \prod_{i=1}^{n-1} \left(I(l \in L_{i_1}) p(l \in L_{i_1} | y, l \in L_i) + I(l \in L_{i_2}) p(l \in L_{i_2} | y, l \in L_i) \right) \quad (2)$$

where $I(\cdot)$ is the indicator function and the product is over all the internal nodes of the tree.

In the speaker verification domain, various classification and modelling approaches can be applied in order to determine whether a particular speech utterance belongs to a particular speaker or not. We used several classification techniques to verify the speaker identity using the tri-grams of the packet lengths of the speech encoded by VBR encoding mechanism. One way to deal with the speaker verification problem is to consider it as a two-class classification problem and assign the probability of the likelihood of the model as the true and imposters scores in order to calculate the equal error rates and minimum detection cost function. Another technique we used is to tackle the speaker verification problem as a regression problem (Smola et al., 2003) by assigning different numerical values to the true speakers and the imposters during the training process. Then we set a threshold for the binary decision. For any unknown utterance, we calculate the values via the same regression modelling technique and attribute the utterance to the target speaker depending on the threshold and the calculated value. The classifier we used for this approach is the SVM with RBF kernel (Burges, 1998; Hsu et al., 2009).

6. Experimental evaluation

Our experiments were conducted on the CSLU speaker recognition database (Cole et al., 1998) which consists of speech files from 91 speakers recorded in twelve sessions over a period of two years. Each session consists of 96 files each of 2–20 s duration. We first encoded the wave files of

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

the database with Speex using variable bit rate encoding in the narrow band mode. This encodes the files with eight different bit rates depending on the speech contents in each 20 ms frame. The VBR encoder encoded the speech files with eight distinct bit rates resulting into different frame sizes each of 6, 10, 15, 20, 28, 38, 46 and 62 bytes. The sequence of these frame sizes depends on the content as well as the acoustics of the underlying speech. A Matlab application was developed for extracting the packet-length information from the encrypted files. Various techniques were then employed to correlate the frames sizes with the speaker identity.

6.1. *Speaker identification*

In case of speaker identification, the job of the forensic investigator is to identify the potential perpetrator in a group of potential suspects. The number of potential suspects may vary from one case to another. We conducted experiments on groups of 5, 10, 15 and 20 suspects. Various modelling and classification approaches were used for the experimental evaluation of identifying the potential perpetrator from the packet-length information extracted from encrypted VoIP conversations. The following are the worth mentioning experiments.

HMM Tests: The HTK toolkit (Young et al., 2003) was used to develop HMM models for different speakers based on their corresponding sequence of frame sizes. 300 speech files per speaker were used for training and 100 files were used for testing. The HMM modelling approach achieves an identification accuracy of 54% for a group of ten speakers.

GMM Tests: For the GMM tests, we calculated the tri-gram probability of the eight symbols (frame lengths) thereby creating a 512 component vector for each speech file. For example, the first component of the 512 component vector corresponds to the probability of the sequence (6,6,6) in the sequence of symbols, the second component corresponds to (6,6,10) and so on. These probability distributions of the tri-grams of each speaker were modelled as a GMM. Again, we used 300 files per speaker for training and 100 files for testing. Various model orders, from 8 to 64, were used but no significant improvement could be observed with the increase in the models orders. The speaker identification accuracies obtained through the GMM approach showed slightly better results as compared to the HMMs. For example, the identification accuracy in case of 10 speakers increased from 54.2% to 58.9% in the case of GMMs with model order 16.

Table 1 – Speaker identification accuracies of various classification methods.

No. of speakers	Identification accuracies (%)			
	HMM	GMM	BayesNet	ENDs
5	60.5	63.7	68.4	74.9
10	54.2	58.9	59.2	72.4
15	43.7	41.7	51.7	59.8
20	38.4	39.6	43.2	51.2

Table 2 – Average Precision, Recall and F-measure values of speaker identification for a group of 10 speakers.

Classification technique	Precision	Recall	F-measure
HMM	0.557	0.542	0.553
GMM	0.599	0.589	0.591
Bayes Net	0.601	0.592	0.598
ENDs	0.734	0.724	0.723

Bayes Net and ENDs Tests: The WEKA toolkit (Witten and Frank, 2005) was used to investigate the performance of different types of classifiers and regression techniques in our speaker identification problem. Among various classification techniques available, the Bayesian network classification technique (Heckerman, 1995) showed results better than both HMMs and GMMs. For the training and test data we used the tri-gram probability as features and the speakers identity as classes. We conducted experiments using 10 fold cross validation evaluation to avoid any biases in the data. Furthermore, the use of meta classification has showed significant improvement over simple classification techniques. In this case, the discriminative power of various base classifiers is combined to achieve better classification accuracies. We conducted some tests using various meta classifiers with different combinations of base classifiers. The ensemble of nested dichotomies (ENDs) showed significant improvement over all other classification methods. For the training and testing data we again used the tri-gram probability distributions of the frame length sequencing to compare the accuracies with the same data set and features.

6.1.1. *Results*

Table 1 shows a summary of the identification accuracies of the above mentioned four significant classification techniques for a group of 5, 10, 15 and 20 speakers.

We also measured the *precision* and *recall* of all the above proposed techniques. The average values of the precision and recall for 10 speakers obtained throughout the above mentioned four classification techniques are presented in

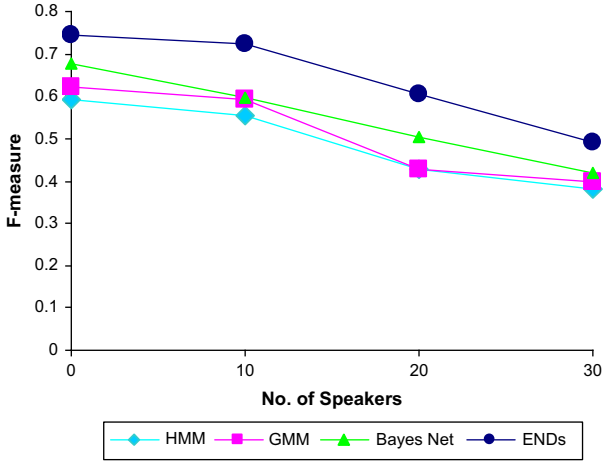

Fig. 3 – F-Measure Variation with Number of Speakers and Classification Techniques.

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

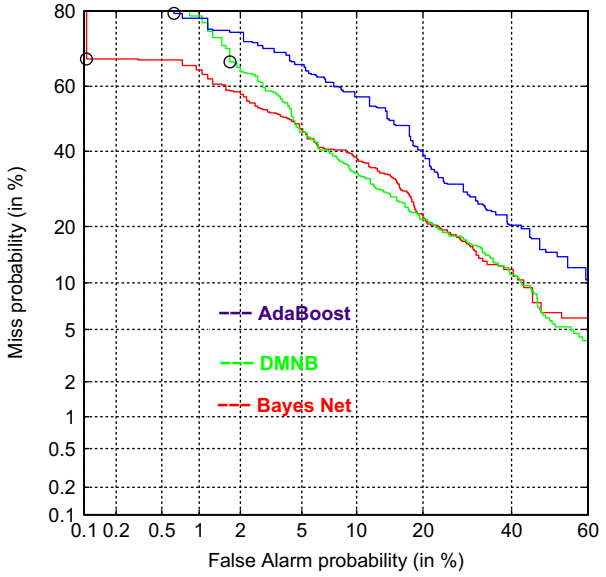

Fig. 4 – A typical DET plot of speaker verification with different classification and regression techniques.

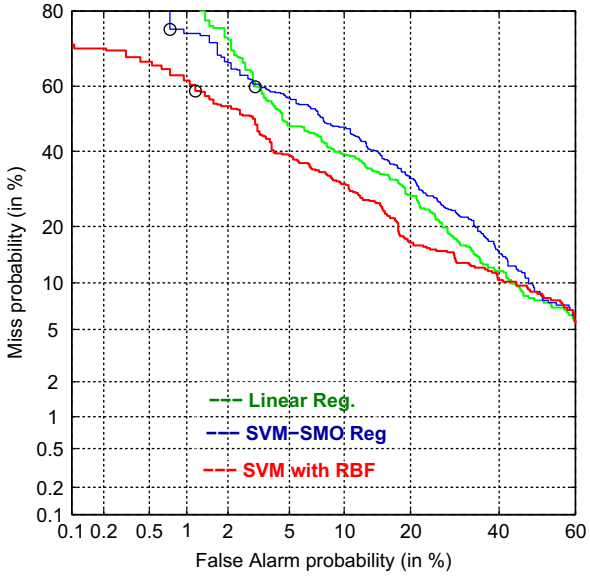

Fig. 5 – A typical DET plot of speaker verification with different classification and regression techniques.

Table 2. The variation of the F-measure, which is defined as the harmonic mean of precision and recall, with the number of speakers for the different classification techniques is depicted in Fig. 3.

6.2. *Speaker verification*

Speaker verification can be thought of as a two-class classification problem. We conducted experiments on the same data set of speakers as the speaker identification but in this case we first used the classification techniques to model two classes, one for the target speaker and one for the cohort or background model. Two types of cohort models were created, one for the male group and another one for the female group. The two cohort models were trained using the complete data sets of male and female speakers respectively. For the target speaker, we used four hundred files per speaker. To evaluate our classification methods, we used the 10 fold cross validation approach to avoid any bias in the evaluation experiments and to judge the classification method over the entire database reserving 90% for training and 10% for testing. The NIST toolkit was used to calculate the equal error rates. For the true speakers and imposters' scores, we used the probability scores of our classifier as input.

6.2.1. *Speaker verification via classification*

For classification, we used three different classification techniques, namely Adaboost.M1 (Yoav and Schapire, 1996), discriminative multinomial naive Bayes (DMNB) (Su et al., 2008) and the Bayesian network (Goldszmidt et al., 1977) classifiers. Fig. 4 shows a typical DET plot of one speaker from our database using the three classification techniques. The point on the DET plot which gives the minimum cost detection function is marked on each curve.

6.2.2. *Speaker verification via regression*

In this case, we again used three different regression techniques. These are linear regression (Witten and Frank, 2005), SVM with sequential minimum optimization (SMO) (Platt, 1999), and SVM with RBF kernel (Burges, 1998). We used the regression scores of the true speakers and the imposters for calculating the equal error rates and minimum detection cost function. The regression approach via SVM with RBF kernel produced lower EER as compared to the Bayesian network classifier. Fig. 5 shows the typical DET plots of one randomly selected speaker from our database when using the three regression techniques described above.

Table 3 shows the mean EER and minimum CDF obtained when using the above discussed classification and regression methods.

Table 3 – Speaker verification accuracies of various classification methods.

Verification	Classification			Regression		
	A. Boost	DMNB	Bayes	SVM-SMO	Lin. Reg	SVM-RBF
EER (%)	23.4	20.1	19.5	22.3	21.3	17.1
minDCF	0.0856	0.0838	0.0690	0.0901	0.0830	0.0681

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

7. Conclusion

With the advancement in VoIP applications, in which the un-encrypted speech communication is diminishing, access to un-encrypted speech can prove to be a very difficult task for investigators. Therefore, future forensic applications need to look into the possibility of identifying perpetrators from encrypted speech segments. This paper is an endeavor in this direction. Several techniques for forensic speaker recognition from encrypted VoIP conversations were presented. It has been shown that the variable packet-length information in case of variable bit speech encoding mechanism can be exploited to extract speaker dependent information from encrypted VoIP conversations. Although the identification and verification accuracies achieved in our experiments are not comparable to the ones achieved in the un-encrypted speech domain, these are by far superior to random guessing as one would expect in case of encrypted communication. It should also be noted that while the current state of the art speaker recognition techniques have not matured enough to be produced in a court as the sole source of evidence against a suspect, these techniques are nonetheless valuable tools that can facilitate forensic investigations. In the same context, the computational data obtained by our experiments, obviously, cannot be used as a complete forensic evidence. However, together with other sources of evidence they can provide some clues for further directions to the forensic investigators.

REFERENCES

Acero A. Acoustical and environmental robustness in automatic speech recognition. Foundations and Trends in Signal Processing March 1993;1(3):195–304.

Advisory Panel on White House Tapes. The Executive Office Building Tape of June 20, 1972: Report on a technical investigation. Technical report, United States District Court for the District of Columbia, May 1974.

Agresti A. Categorical data analysis. John Wiley and Sons; 2002.

Barbeieri R, Bruschi D, Rosti E. Voice over IPSec: analysis and solutions. In: Proceedings of the 18 annual computer security applications conference; December 2002. p. 261–70.

Baugher M, McGrew D. RFC 3711: SRTP: the Secure Real Time Transport Protocol. IETF; July 2003.

Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic factions of Markov chains. Annals of Mathematics and Statistics 1970;4(1).

Bimbot F, Bonastre JF, Fredouille C, Gravier G, Chagnolleau IM, Megnier S, et al. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing 2004;4:430–51.

Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 1998;2:121–67.

Campbell W, Sturim D, Reynolds D, Solomonoff A. SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceedings of the international conference on acoustics, speech and signal processing; 2006. p. 1–97.

Chu WC. Speech coding algorithms. John Wiley and Sons; 2003.

Cole R, Noel M, Noel V. The CSLU speaker recognition corpus. In: Proceedings of the international conference on spoken language processing. Australia; November 1998. p. 3167–70.

Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society 1977;39.

Forney GD. The viterbi algorithm. Proceedings of the IEEE 1973; 61(3).

Frank E, Kramer S. Ensembles of nested dichotomies for multi-class problems. In: ICML '04: Proceedings of the twenty-first international conference on machine learning; 2004.

Goldszmidt M, Friedman N, Geiger D. Bayesian network classifiers. Machine Learning 1977;29.

Heckerman D. A tutorial on learning Bayesian networks. Technical report, Microsoft Research Technical Report; 1995.

Hsu CW, Chang CC,, Lin CJ. A practical guide to support vector classification. Taipei, Taiwan: Technical Report, Department of Computer Science National Taiwan University; 2009.

Kersta LG. Voiceprint identification. Nature 1962;196(4861):1253–7.

Kinnun T, Saastamoinen J, Hautamaki V, Vinni M, Franti P. Comparing maximum a posteriori vector quantization and Gaussian mixture models in speaker verification. In: Proceedings of the IEEE international conference on acoustics speech and signal processing. Taiwan; 2009. p. 145–8.

Koenen R. ISO/IEC JTC1/SC29/WG11: coding of moving pictures and audio; March 2002.

Koolwaaij J, Boves L. On the Use of automatic speaker verification systems in forensic casework. In: Proceedings of audio and video-based biometric person authentication; 1999. p. 224–9.

Martin A, Przybocki M. The NIST speaker recognition evaluation series. National Institute of Standards and Technology Web Site; June 2009.

McLachlan G. Mixture models. New York: Marcel Decker; 1988.

Pfeiffer S. RFC 3533: the Ogg encapsulation format version 0. IETF; May 2003.

Platt JC. Fast training of support vector machines using sequential minimal optimization; 1999. p. 185–208.

Provos N. Voice Over Misconfigured Internet Telephones (VOMIT). http://vomit.xtdnet.nl/.

Przybocki MA, Martin AF, Le AN. NIST speaker recognition evaluation chronicles, part 2. In: IEEE Odysee, ISCA speaker recognition workshop; June 2006.

Przybocki MA, Martin AF, Le AN. NIST speaker recognition evaluations utilizing the mixer corpora-2004, 2005, 2006. IEEE Transactions on Audio Speech and Language Processing September 2007;15(7):1951–9.

Quinlan JR. C4.5: programs for machine learning. Morgan Kaufmann; 1992.

Reynolds DA. An overview of automatic speaker recognition technology. In: Proceedings of the IEEE international conference on acoustics speech and signal processing; 2002.

Reynolds DA, Rose RC. Robust text independent speaker identification using Gaussian mixture speaker models. IEEE Transactions of Speech and Audio Processing 1995;3(1).

Sachs JS. Graphing the voice of terror. Popular Science March 2003:38–43.

Smola AJ, Schlkopf B, Olkopf Bernhard Sch. A tutorial on support vector regression. Technical report, statistics and computing; 2003.

Su J, Zhang H, Ling CX, Matwin S. Discriminative parameter learning for Bayesian networks. In: International conference on machine learning; 2008. p. 1016–23.

Valin JM, Montgomery C. Improved noise weighting in CELP coding of speech-applying the Vorbis Psychoacoustic model to Speex. In: Audio Engineering Society Convention; May 2006.

Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Margan Kaufmann; 2005.

Wright CV, Ballard L, Monrose F, Masson GM. Language identification of encrypted VoIP traffic: Alejandra y Roberto or Alice and Bob? In: SS'07: Proceedings of 16th USENIX security symposium on USENIX security symposium; 2007.

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001

Wright CV, Ballard L, Coull SE, Monrose F, Masson GM. Spot me if you can: uncovering spoken phrases in encrypted VoIP conversations. In: SP '08: Proceedings of the 2008 IEEE symposium on security and privacy; 2008. p. 35–49.

Yoav F, Schapire RE. Experiments with a new boosting algorithm. In: International conference on machine learning; 1996. p. 148–56.

Young SJ, Everman G, Hain T, Kershaw D, Moore GL, Odell JJ, et al. The HTK book. Cambridge: Cambridge University; 2003.

Please cite this article in press as: Khan LA, et al., Speaker recognition from encrypted VoIP communications, Digit. Investig. (2009), doi:10.1016/j.diin.2009.10.001