RESEARCH ARTICLE OPEN ACCESS

A Privacy Preserving Approach to Analyze Security in VoIPSystem

S.Prashantini[1], T.J.Jeyaprabha[1], Dr.G.Sumathi[2]
[1]Dept of Electronic and Communication, Sri Venkateswara College Of Engineering,Chennai
[1]Assistant Professor, Dept of Electronic and Communication,Sri Venkateswara College Of Engineering, Chennai
[2]Professor &Head-IMS, Department Of Information and Technology,Sri Venkateswara College Of Engineering,Chennai.

ABSTRACT
Pre-processing of Speech Signal serves various purposes in any speech processing application. It includes Noise Removal, Endpoint Detection, Pre-emphasis, Framing, Windowing, Echo Canceling etc. Out of these, silence portion removal along with endpoint detection is the fundamental step for applications like Speech and Speaker Recognition. The proposed method uses multi-layer perceptron along with hierarchal mixture model for classification of voiced part of a speech from silence/unvoiced part. The work shows better end point detection as well as silence removal. The study is based on timing-based traffic analysis attacks that can be used to reconstruct the communication on end-to-end voipsystems by taking advantage of the reduction or suppression of the generation of traffic whenever the sender detects avoice inactivity. Removal of the silence in between the inter packet time in order to obtain a clear communication. The proposed attacks can detect speakers of encrypted speech communications with high probabilities. With the help of different codecs we are going to detect speakers of speech communications In comparison with traditional traffic analysis attacks, the proposed traffic analysis attacks do not require simultaneous accesses to one traffic flow of interest at both sides.

Keywords–Speech recognition, silent suppression, passive analysis.

I. INTRODUCTION

Speech recognition (SR)

Speech Recognition is the translation of spoken words into text. Some SR systems use "speaker independent speech recognition" while others use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "speaker independent" systems. Systems that usetraining are called "speaker dependent" systems.Speech recognition applications include voice user interfaces such asvoice dialing, call routing, appliance control, and search simple data entry.

Speaker recognition is the identification of the person who is speaking by characteristics of their voices (voice biometrics), also called voice recognition. There are two major applications of speaker recognition technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") whereas speaker identification is a 1:N match where the voice is compared against N templates.

To guarantee bandwidth for VoIP packets, a network device must be able to identify VoIP packets in all the IP traffic flowing through it. Network devices use the source and destination IP address This identification and grouping process is called classification and it is the basis for providing any QoS.

Packet classification can be processor-intensive, so it should occur as far out toward the edge of the network as possible. Because every hop still needs to make a determination on the treatment a packet should receive, you need to have a simpler, more efficient classification method in the network core. This simpler classification is achieved through hidden markov model.

HMMs [2] are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of *n*-dimensional real-valued vectors

(with *n* being a small integer, such as 10), outputting one of these every 10 milliseconds. The vectors would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

The term silence suppression is used in telephony to describe the process of not transmitting information over the network when one of the parties involved in atelephone call is not speaking, thereby reducing bandwidth usage.Voice is carried over a digital telephone network by converting the analog signal to a digital signal which is then packetized and sent electronically over the network. The analogue signal is re-created at the receiving end of the network [6]. When one of the parties does not speak, background noise is picked up and sent over the network. This is inefficient as this signal carries no useful information and thus, bandwidth is wasted.Given that typically only one party in a conversation speaks at any one time, silence suppression can achieve overall bandwidth savings in the order of 50% over the duration of a telephone call.

In this paper, we detect silence/unvoiced [14] part from the speech sample using multi-layer perceptron algorithm. The algorithm uses statistical properties of noise based on packets as well as packets based on threshold energy. The experiments are done in which the speech is transferred through internet and categorized as packets. The result shows better classification for the proposed method in both the cases when compared against conventional silence/unvoiced detection methods. We assume that background noise present in the utterances are Gaussian in nature, however a speech signal may also be contaminated with different types of noise. In such cases the corresponding properties of the noise distribution function are to be used for detection purpose.

This paper is organized as follows. In section 2 it describes about the theoretical background. Section 3 presents the method i.e. the algorithm along with description regarding its model and defining the measure of correctness. The results are presented in section 4 and section 5 describes the conclusion.

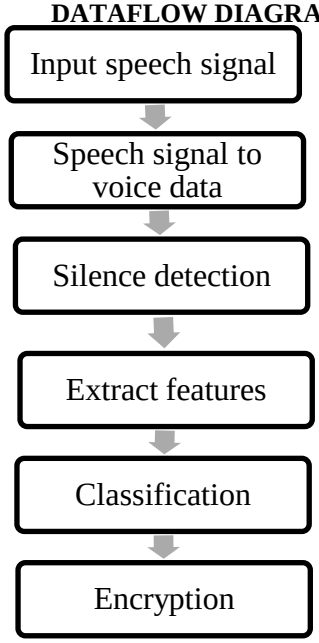
II. THEORITICAL BACKGROUND

2.1 SPEECH SIGNAL AND ITS BASIC PROPERTIES

The proposed system consist of a new class of traffic analysis attacks to encrypted speech communications with the goal of detecting speakers of encrypted speech communications. These attacks are based on packet timing information only and the attacks can detect speakers of speech communications made with different codecs. We evaluate the proposed attacks with extensive experiments over different type of networks including commercial anonymity networks and campus networks. The experiments show that the proposed traffic analysis attacks can detect speakers of encrypted speech communications with high accuracy.

The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time, its characteristics are fairly stationary; however, over long periods of time the signal characteristics change to reflect the different speech sounds being spoken. Usually first 200 msec or more (1600 samples if the sampling rate is 8000 samples/sec) of a speech recording corresponds to silence (or background noise) because the speaker takes some time to read when recording starts.

2.2 DATAFLOW DIAGRAM



2.3 MODULES USED

Speech Coding

In speech communications, an analog voice signal is first converted into a voice data stream by a chosen codec. Typically in this step, compression is used to reduce the data rate. The voice data stream is then packetized in small units of typically tens of

milliseconds of voice, and encapsulated in a packet stream over the Internet. In this paper, we focus on constant bit rate codecs since most codecs used in current speech communications are CBR codecs.

Silence detection

In this, silence period in the voice signal is detected .This was done bychecking threshold range which are matched to silence level.Then this was eliminated and combine the voice into feature values.

Voice Signal Waveform

The packet train generated by feeding the voice signal to X-Lite, a popular speech communication tool. Here it is easily observed that the correspondence between the silence periods in the voice signal and the gaps in the packet train. The length of a silence period is different from the length of the corresponding gap in the packet train.

Speaker Detection

The inputs to this step are the Multi-Layer Perception (MLP) classifier with (FFBP) trained in the previous step and the feature vectors generated from a pool of raw speech communication traces of interest [12]. The output of this step is the intermediate detection result, ie speakers from the candidate pool with talk patterns closest to another talk pattern. The detection step can be divided into two phases: First, the likelihood of each feature vector is calculated with the trained MLP. The trace with the highest likelihood is declared if the intersection step is not used. To improve the detection accuracy, the intermediate detection results can be fed into the optional intersection attack step.

Cross – Codec Detection

In this set of experiments, the training traces and the traces to be detected are generated with different codecs[12]. We believe this set of experiments is important because: Practically training traces and the traces to be detected can be collected from speech communications made with different codecs. Since speech packets are encrypted and possibly padded to a fixed length, adversaries may not be able to differentiate speech communications made with different codecs.

III. METHOD

The algorithm described is divided into two parts. First part classifies the voice packets and identifies the delay by using mlpalgorithm while the second part classifies each packet according to the timing delay by using Hmm model.

Multilayer perceptron (MLP)-

Multi-layer perceptron is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network.MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable. A learning rule is applied in order to improve the value of the MLP weights over a training set T according to a given criterion function.

Back Propagation Algorithm

Back Propagation Algorithm is used to train MLNN. The training proceeds in two phases:

In the **forward phase**, the synaptic weights of the network are fixed and the input signal is propagated through the network layer by layer until it reaches the output.

In the **backward phase**, an error signal is produced by comparing the output of the network with a desired response. The resulting error signal is propagated through the network, layer by layer but the propagation is performed in the backward direction. In this phase successive adjustments are applied to the synaptic weights of the network.The key factor involved in the calculation of the weight adjustment is the error signal at the output neuron j. As we see the credit-assignment problem arises here. In this context we may identify two distinct cases:

Case #1: Neuron *j* is an output node:

The error signal is supplied to the neuron by its own from equation

Case #2: Neuron *j* is a hidden node:

When a neuron j is located in a hidden layer of the network, *there's no* specified desired response for that neuron.

Accordingly, the error signal for a hidden neuron would have to be determined recursively and working backwards in terms of the error signals of all the neurons to which that hidden neuron connected.

The final back-propagation formula for the local gradientWhere*k* represents the number of neurons that are connected to hidden neuron *j*.

As a summary the correction is applied to the synaptic weight connecting neuron I to j

Any *activation function* that is used in multilayer neural networks should be continuous

The most commonly used activation function is sigmoidal nonlinearity**.**

3.1 ACTIVATION FUNCTION

It's preferred to use a sigmoid activation function that's an odd function *in its* arguments.The hyperbolic sigmoid function is the recommended one.Target values (desired) are chosen within the range of the sigmoid activation function.

3.2 HMM TRAINING MODEL

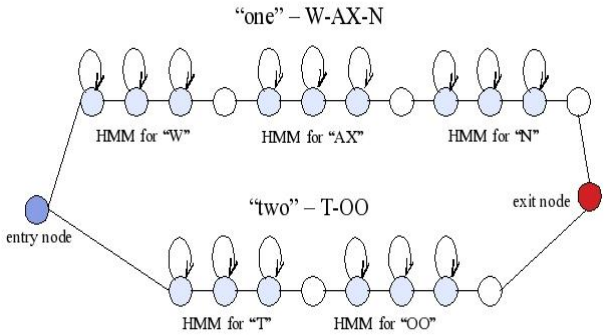


Fig 3.1: HMM model diagram

By combining decisions probabilistically at all lower levels, and making more deterministic decisions only at the highest level [3]. Speech recognition by a machine is a process broken into several phases. Computationally, it is a problem in which a sound pattern has to be recognized or classified into a category that represents a meaning to a human. Every acoustic signal can be broken in smaller more basic sub-signals. As the more complex sound signal is broken into the smaller sub-sounds, different levels are created, where at the top level we have complex sounds, which are made of simpler sounds on lower level, and going to lower levels even more, we create more basic and shorter and simpler sounds. The lowest level, where the sounds are the most fundamental, a machine would check for simple and more probabilistic rules of what sound should represent. Once these sounds are put together into more complex sound on upper level, a new set of more deterministic rules should predict what new complex sound should represent. The most upper level of a deterministic rule should figure out the meaning of complex expressions. In order to expand our knowledge about speech recognition we need to take into a consideration neural networks.

Modern general-purpose speech recognition systems are based on Hidden Markov Models. These are statistical models that output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scales (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes.

HMMs [2] are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of *n*-dimensional real-valued vectors (with *n* being a small integer, such as 10), outputting one of these every 10 milliseconds. The vectors would consist of coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians, which will give a likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

Decoding of the speech (the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence) would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model, which includes both the acoustic and language model information, and combining it statically beforehand (the finite state transducer, or FST, approach).

Filter design is the process of designing a filter which meets a set of requirements. It is essentially an optimization problem. The difference between the desired filter and the designed filter should be minimized. A general goal of a design is that the complexity ("complexity" means "complexity in terms of the number of multipliers" in this paper) to realize the filter is as low as possible.

A codec is a device or computer program capable of encoding or decoding a digitaldata stream or signal.[1][2][3].The word *codec* is a portmanteau of "coder-decoder" or, less commonly, "compressor-decompressor". A codec (the program) should not be confused with a coding or compression format or standard – a format is a document (the standard), a way of storing data, while a codec is a program (an *implementation*) which can read or write such files. In practice, however, "codec" is sometimes used loosely to refer to formats.

A codec encodes a data stream or signal for transmission, storage or encryption, or decodes it for playback or editing. Codecs are used in videoconferencing, streaming media and video editing applications. A video camera's analog-to-

digital converter (ADC) converts its analog signals into digital signals, which are then passed through a video compressor for digital transmission or storage. A receiving device then runs the signal through a video decompressor, then a digital-to-analog converter (DAC) for analog display. The term *codec* is also used as a generic name for a videoconferencing unit.

Speaker recognition [4] is the identification of the person who is speaking by characteristics of their voices (voice biometrics), also called voice recognition. There are two major applications of speaker recognition technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") whereas speaker identification is a 1:N match where the voice is compared against N templates.

IV. RESULTS

The Scenario of training and test of the proposed approach are generated along with few comparison of different voice signals by using MATLAB Simulation software. MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numerical computation. Using MATLAB, you can solve technical computing problems.

In the proposed system the traces of speech is given and being simulated in order to obtain an encrypted speech.Thus the bandwidth of the speech is reduced for which the simulated samples are given.

The various simulation outputs describes about the encrypted speech along with delay and also the correctness data acquired after the reduction of delay. Figure 4.1 describes about the input speech signal along with its encrypted output of value 3 degree. Figure 4.2 consists of a periodogram input which in order to describe the bandwidth reduction.

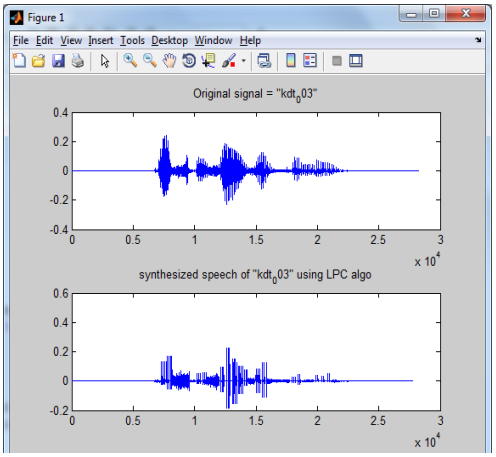


Fig 4.1: Input signal along with encrypted Output

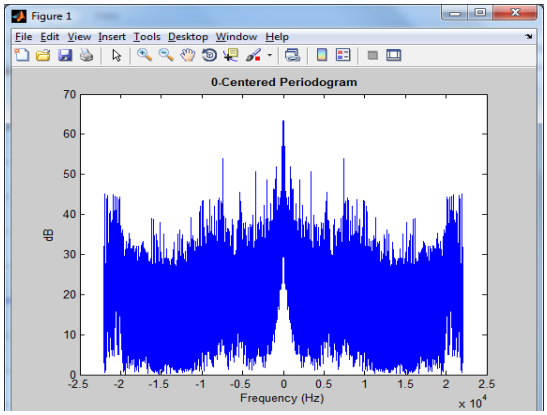


Fig 4.2: Periodogram input of large input

Figure 4.3 represents the input signal along with silence ie delay is present in the transmission. Figure 4.4 represent the signal after the delay is being remover and hence output obtained is free of silence .

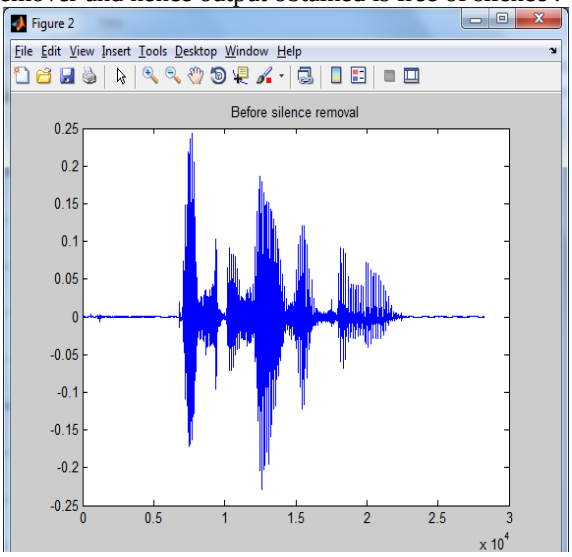


Fig 4.3:before silence removal

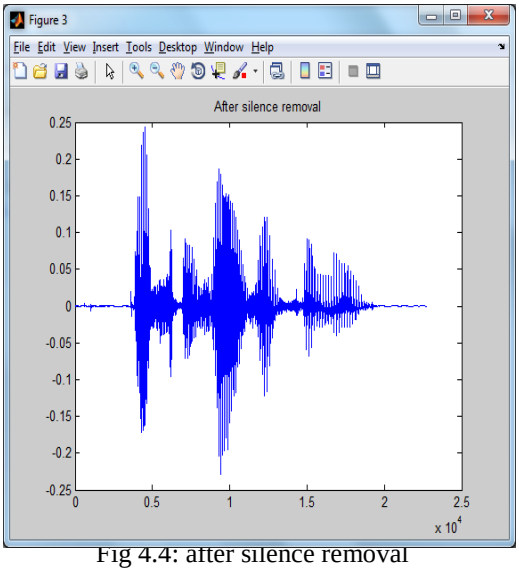


Fig 4.4: after silence removal

The various graphical representations are given which explains about the rate of input signal along with respective output signals. Figure 4.5 describes about the graph of an input signal along with initial conditions of silence threshold value. Figure 4.6 represents the comparison of detected silence and original signal values and 4.7 gives the detected rate of signal transmissions.

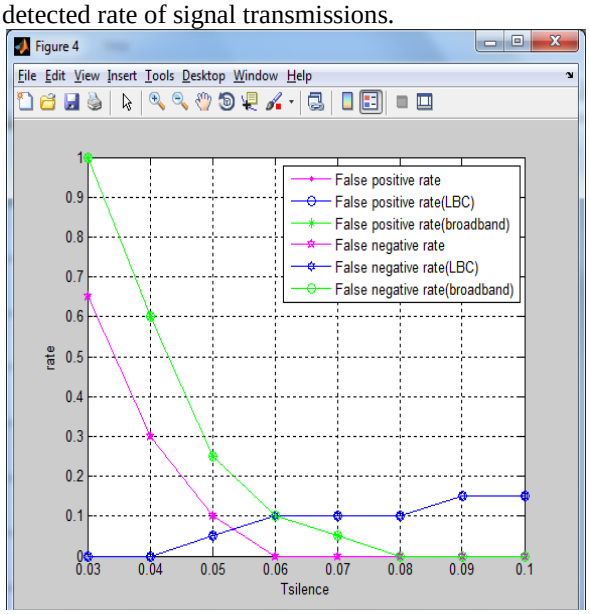


Fig 4.5: graphical representation of silence threshold

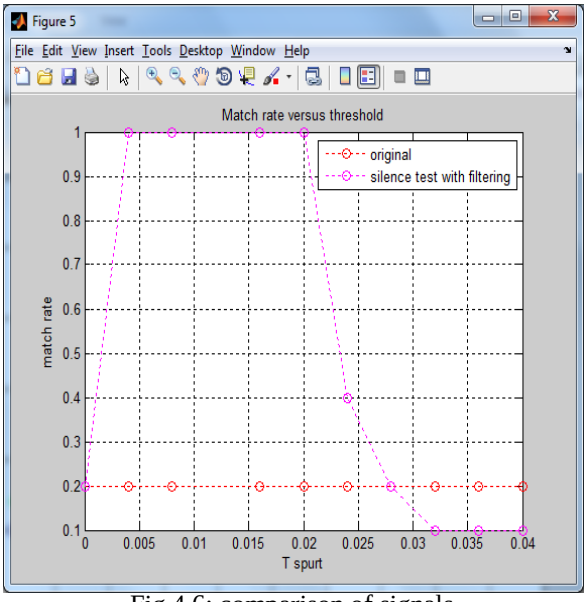


Fig 4.6: comparison of signals

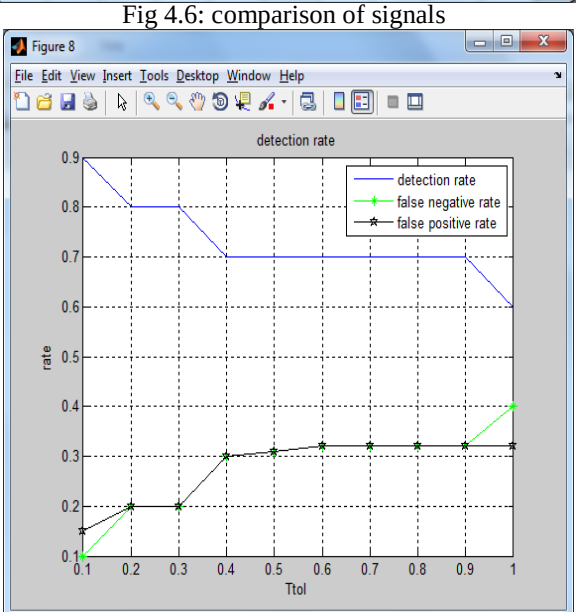


Fig 4.7: Detection rate

V. CONCLUSION

The input consist of speech signal as input and removed the silence present in between in order to give an effective communication by providing reduced bandwidth Hidden markov model has been used effectively in order obtain the voice traces. The proposed attacks is carried out by extensive experiments over different types of networks including periodogram inputs. The experiments show that the proposed traffic analysis attacks can detect speakers of encrypted speech communications with high detection rates based on speech communication traces. The threshold energy used in this method is uniquely specified and depends on the packet transfer. It is shown to be computationally efficient for real time applications and it performs better

thanconventional methods for speech samples collected from noisy as well as noise free environment.

VI. ACKNOWLEDGEMENT

I would like to thank my guide Mrs.T.J.Jeyaprabha, Assistant professor, who guided me to complete this project and also thank everyone those who supported me to achieve my target.

REFERENCES

[1] Y.J. Pyun, Y.H. Park, X. Wang, D.S. Reeves, and P. Ning, "Tracing Traffic through Intermediate Hosts that Repacketize Flows," Proc.IEEE INFOCOM '07, May 2010.

[2] J.W. Deng and H.T. Tsui, "An HMM-Based Approach for Gesture Segmentation and Recognition," Proc. Int'l Conf. Pattern Recognition(ICPR '00), pp. 679-682, 2000.

Bayes Classifier," Proc. ACM Workshop Cloud Computing Security (CCSW '09), pp. 31-42, 2009.

[3] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, vol. 77, no. 2,pp. 267-296..

[4] R. Bakis, "Continuous Speech Recognition via CentisecondAcousticstates," J. the Acoustical Soc. of Am.

[5] G. Danezis and A. Serjantov, "Statistical Disclosure or Intersection Attacks on Anonymity Systems," Proc. Sixth Information Hiding Workshop (IH '04), pp. 293-308, May 2004.

[6] O. Berthold and H. Langos, "Dummy Traffic Against Long Term Intersection Attacks," Proc. Privacy Enhancing Technologies Workshop(PET '02), pp. 110-128, Apr. 2002.

[7] X. Wang, S. Chen, and S. Jajodia, "Tracking Anonymous Peer-to-Peer Voip Calls on the Internet," Proc. 12th ACM Conf. Computer and Comm. Security (CCS '05), pp. 81-91, 2005.

[8] "Audio Signals Used for Experiments, "http:// academic. csuohio. edu/zhu _y/ isc 2010/instruction.txt, 2011.

[9] Q. Sun, D.R. Simon, Y.-M. Wang, W. Russell, V.N. Padmanabhan,and L. Qiu, "Statistical Identification of Encrypted Web Browsing Traffic," Proc. IEEE Symp. Security and Privacy (SP '02), pp. 19-30,2002.

[10] D. Herrmann, R. Wendolsky, and H. Federrath, "Website Fingerprinting: Attacking Popular Privacy Enhancing Technologies with the Multinomial Naı¨ve-

[11] L. Lu, E.-C. Chang, and M. Chan, "Website Fingerprinting and Identification Using Ordered Feature Sequences," Proc.15th European Conf. Research in Computer Security (ESORICS),D. Gritzalis, B. Preneel, and M. Theoharidou, eds. pp. 199-214.2010.

[12] C.V. Wright, L. Ballard, S.E. Coull, F. Monrose, and G.M. Masson,"Spot Me if You Can: Uncovering Spoken Phrases in EncryptedVoip Conversations," Proc. IEEE Symp. Security and Privacy(SP '08), pp. 35-49, 2008.

[13] C.V. Wright, L. Ballard, F. Monrose, and G.M. Masson, "Language Identification of Encrypted Voip Traffic: Alejandra y Roberto orAlice and Bob?," Proc. 16th USENIX Security Symp. USENIX Security Symp., pp. 4:1-4:12, http://portal.acm.org/citationcfm?id=1362903.1362907, 2007.

[14] C. Wright, S. Coull, and F. Monrose, "Traffic Morphing: An Efficient Defense against Statistical Traffic Analysis," Proc. Network and Distributed Security Symp. (NDSS '09), Feb. 2009

[15] Network Security M. Backes, G. Doychev, M. Du¨ rmuth, and B. Ko¨ pf, "Speaker Recognition in Encrypted Voice Streams," Proc. 15th European Symp. Research in Computer Security (ESORICS '10), pp. 508-523, Sept. 2010.