A Voice Spam Filter to Clean Subscribers' Mailbox

Seyed Amir Iranmanesh[1], Hemant Sengar[2], and Haining Wang[1]

[1] Department of Computer Science, College of William and Mary, Williamsburg, VA 23187, USA
{sairan,hnw}@cs.wm.edu
[2] Technology Development Department, VoDaSec Solutions, Fairfax, VA 22030, USA
hsengar09@gmail.com

Abstract. With the growing popularity of VoIP and its large customer base, the incentives of telemarketers for voice spam has been increasing in the recent years. If the threat of voice spam remains unchecked, it could become a problem as serious as email spam today. Compared to email spam, voice spam will be much more obnoxious and time consuming nuisance for telephone subscribers to filter out. In this paper, we propose a content-based approach to protect telephone subscribers voice mailboxes from voice spam. In particular, based on Dynamic Time Warping (DTW), we develop a speaker independent speech recognition system to make content comparison of speech messages. Using our system, the voice messages left on the media server by callers are matched against a set of spam filtering rules involving the study of *call behavioral* pattern and the analysis of *message content*. The uniqueness of our spam filtering approach lies in its independence on the generation of voice spam, regardless whether spammers play same spam content recorded in many different ways, such as human or machine generated voice, male or female voice, and different accents. We validate the efficacy of the proposed scheme through real experiments, and our experimental results show that it can effectively filter out spam from the subscribers' voice mailbox with 0.67% false positive rate and 8.33% false negative rate.

Key words: VoIP, voice spam, content filtering, Dynamic Time Warping

1 Introduction

IP telephone service providers are moving fast from low-scale toll bypass deployments to large-scale competitive carrier deployments; thus giving an opportunity to enterprise networks for supporting less expensive single network solution rather than multiple separate networks. The broadband-based residential customers also switch to IP telephony due to its convenience and cost effectiveness. On the contrary to traditional telephone system in which the end devices are dumb, the VoIP architecture pushes intelligence towards the end devices like PCs and IP phones, creating many new services. This flexibility coupled with the growing number of subscribers has attracted attackers for malicious resource abuse. As the number of

VoIP subscribers hits a critical mass, it is expected that voice spam will emerge as a serious threat. In fact in Japan where VoIP market is much more mature than USA, has witnessed some recent voice spam attacks. The SoftbankBB, a VoIP service provider with 4.6 million users has reported three incidents of spam attacks within its own network [20]. These incidents include unsolicited messages advertising an adult website, scanning of active VoIP phone numbers and requesting personal information of users. Similarly, Columbia University experienced a voice spam attack, with someone accessing the SIP proxy server and "war dialing" a large number of IP phone extensions [21]. There are many reported incidents of spam messages on Google voice too [7]. Evidently, the effectiveness of telephone calls presents strong incentives for spammers to establish voice channels with many subscribers at the same time. Such machine generated unsolicited bulk calls known as SPIT (Spam over Internet Telephony) may hinder the deployment of IP telephony, and if the problem remains unchecked then it may become as potent as email spam today. In many aspects, the voice spam is similar to an email spam. Moreover, voice spam will be much more obnoxious and harmful than email spam. The ringing of telephone at odd time, answering of spam calls, phishing attacks and inability to filter spam messages from the voicemail box without listening each one are real nuisance and waste of time.

In the past, a number of anti-spam solutions have been proposed. Both academic and industry research groups have made some efforts to address the voice spam problem. Most of the ideas are borrowed from the data security field, using the techniques such as intrusion detection systems, black and white lists, Turing tests, computational puzzles, reputation systems, and rate throttling at the gatekeeper. These solutions generally distinguish a legitimate subscriber from a spammer using only SIP signaling messages. However, in this paper we take a radically different approach. Instead of analyzing the SIP signaling messages and identifying the spam originating source(s) or ascertaining the real identity of spammers, we try to avoid spam message deposition on the subscribers' voice mailboxes. The goal of the proposed approach is two-pronged. First, we allow only legitimate messages to be deposited on the subscribers' mailbox account, unsolicited spam messages are blocked at the media server itself. Secondly, the proposed approach also provides a way to identify spamming sources based on spam messages. To the best of our knowledge, this is a first attempt to clean subscribers' voice mailboxes from voice spam messages.

Beyond the basic observation that SIP signaling messages needs to be analyzed for its source and caller identification, we make three additional observations that are central to our approach. First, the spammers would prefer to see high hit ratio for their spamming attacks. Thus, most of the spamming attacks are expected to occur in bulk (i.e., as much spam as possible within a short duration of time) and most of the spam messages will be delivered to voice mailboxes. Second, during the spam attack instance, a spammer will play pre-recorded messages to many of the spam victims at the same time. Third, the originating spam source is expected to be some sort of interactive voice response (IVR) system, which can interact with the users if the calls are answered and it should also be able to leave a voice mail if the calls are not answered. However, it should be noted that in most of

the spam attacks the voice stream originating from the spam source is machine generated. Based on these observations, we design and develop a voice mailbox filtering approach.

In our approach, we first segment voice messages in their voiced segments using a silence removal technique. Our silence removal technique is based on two audio features; the *signal energy* and the *spectral centroid*. After calculating the partial similarity between each pair of voiced segments coming from two different voice messages, we can determine how similar are the two voice messages content-wise. To measure the similarity between two voiced segments as a metric for content comparison, we use the technique of Dynamic Time Warping (DTW) to compute the cosine similarity between two sequences of speech feature matrices. A popular speech feature representation known as RASTA-PLP (Relative Spectral Transform - Perceptual Linear Prediction) is used to extract speech feature matrices from voice messages. After a message is left on the server by a caller, it is divided into voiced segments using our segmentation method and RASTA-PLP spectra for its voiced segments being calculated. Using our DTW based system, the RASTA-PLP matrix is then matched against a set of spam signatures. If a match is not found, our system is further coupled with Bayesian filtering to reveal the hidden spam words/phrases within a voice message to show how closely (probabilistically) it matches with the known spam messages seen in the past. Normally during a spam attack, many of the deposited voice messages share the same content, we finally use our speaker independent speech recognition technique to find how many similar messages (in content) are deposited within a predefined time interval of ΔT.

We conduct two sets of experiments to evaluate the effectiveness of our proposed solution against realistic spam attack scenarios. In the first experiment, we investigate the most generic spam attack scenario, where a spammer repeatedly sends the same spam message to many of the subscribers at the same time. Three hundred voice messages in various size are deposited from thirty speakers with different accents (such as American, British, or Indian English), different sex and ages to form the scenario. In the second set of experiment, we investigate the power of our method to classify voice messages as spam and non-spam, in which the deposited voice messages include spam words/phrases. Our experimental results show that our approach is computationally efficient, and speaker independent to identify a common segment of voice message out of a database of known spam signatures and classify the voice message correctly.

The remainder of the paper is structured as follows. The basic VoIP architecture, SIP-based IP telephony, voice message deposition process, and a brief overview of the proposed approach are presented in Section 2. In Section 3, we describe the technical details on voice message signature construction. In Section 4, we detail spam detection methodology. Section 5 analyzes the performance of the proposed solution. Section 6 surveys related work. Finally, Section 7 concludes the paper.

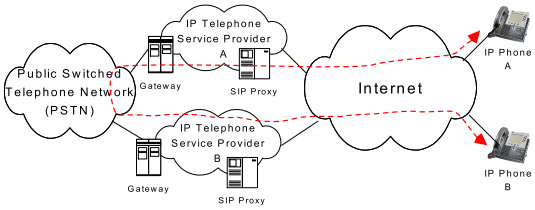


Fig. 1. Island-based SIP VoIP Deployment

2 Background

Voice spam is an extension of email spam in the VoIP domain. The technical know-how and execution style of email spam can easily be adapted to launch voice spam attacks. For example, a voice spammer first harvests user's SIP URIs or telephone numbers from the telephone directories or by using spam bots crawling over the Internet. Then, a compromised host is used as a SIP user agent (UA) that sends out call setup request messages. Finally, the established sessions are played with a pre-recorded .wav file. However, voice spam is much more obnoxious and harmful than email spam. The ringing of telephone at odd time, answering of spam calls, phishing attacks and inability to filter spam messages from the voicemail box without listening each one are real nuisance and waste of time.

Before we delve into voice spam problem, we briefly describe the basic VoIP architecture as it serves two purposes: first, it explains as why we do not hear much of voice spam attacks today as compared to email spam; second, it also describes as why it could be a serious problem for VoIP subscribers in the near future.

2.1 VoIP Architecture

As shown in Fig. 1, in today's IP telephony world most of the VoIP service providers (such as Vonage, AT&T Callvantage, and ViaTalk) operate in partially closed environments and are connected to each other through the public telephone network. VoIP service providers allow only their own authenticated subscribers to access SIP proxy server resources. The authentication of call requests is feasible because user accounts are stored locally on the VoIP service provider's SIP servers. However, in general the threat of spam calls is associated with the open architecture of VoIP service, where VoIP service providers interact with each other through the IP-based peering points. It provides an ability for individual subscribers to connect with each other without traversing the PSTN cloud. Therefore, it is quite possible that an `INVITE` message received by a VoIP service provider from another service provider (through IP network) for one of its subscriber may not have any type of authentication credentials for the calling party.

Recently, we are witnessing a large demand for SIP trunks. A SIP trunk is a service offered by a VoIP service provider permitting business subscribers to reach beyond the enterprise network and connect to the PSTN through IP-based connections. Generally most of the SIP trunks are set up without authentication. Only few of the service providers use TLS or IPSec to secure SIP signaling. In

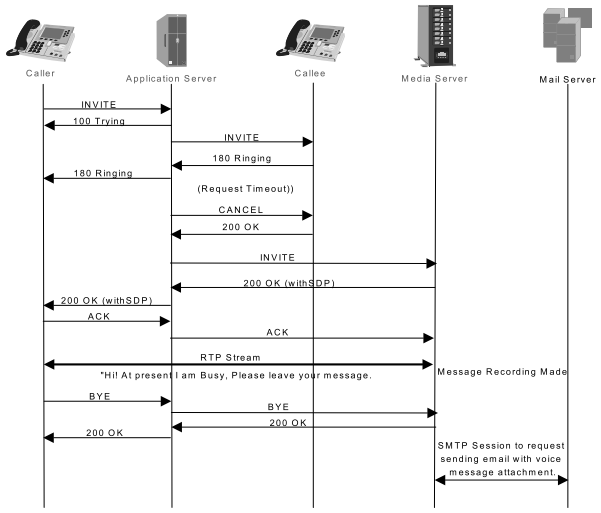


Fig. 2. Voice Message Deposition

this scenario, a spam attack can be launched from within the enterprise network (e.g., a corporate network is infected with malicious worm) or by a man-in-the-middle where SIP signaling is transported over the Internet in plaintext without any encryption.

2.2 SIP-based IP Telephony

The Session Initiation Protocol (SIP) [15], belonging to the application layer of the TCP/IP protocol stack, is used to set up, modify, and tear down multimedia sessions including telephone calls between two or more participants.

SIP-based telecommunication architectures have two types of elements: end devices referred to as user agents (UAs) and SIP servers. Irrespective of being a software or hardware phone, UAs combine two sub-entities: the connection requester referred as the user agent client (UAC) and the connection request receiver referred to as the user agent server (UAS). Consequently, during a SIP session, both UAs switches back and forth between UAC and UAS functionalities.

SIP messages consisting of request-response pairs are exchanged for call set up, from six kinds including `INVITE, ACK, BYE, CANCEL, REGISTER`, and `OPTIONS` - each identified by a numeric code according to RFC 3261 [15].

2.3 Voice Mail Deposition

A simple voice message deposition scenario is shown in Fig. 2. A caller calls a callee who is busy and unable to take phone call, in this particular case, the call is answered by a voice messaging system. The call is set up between caller and callee's voice messaging system that plays a "busy" greeting message and asks the caller to leave a voice message. The caller records the voice message and then hangs up.

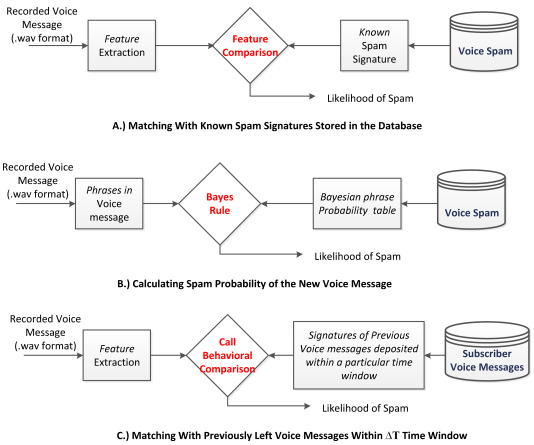


Fig. 3. Overview of Spam Filtering Approach

With the *SendMail* command, the application (i.e., call control) server requests the media server to deliver the recorded voice message to the callee's inbox. The media server sends email with the recorded message as an attachment (in .wav file format) to the user account on SMTP mail server.

2.4 Overview of Spam Filtering Approach

As shown in Fig. 3, our spam filtering approach can be briefly described as a three-step process. Given a recorded voice message, we first verify if it matches with any of the known spam signatures stored in the database. For example, when a caller leaves a voice message for a callee, media server records the RTP stream and converts it into a .wav file. The *feature extraction* process takes this .wav file as an input and extracts few features from the corresponding spectrogram. This set of features is searched in the database to find a match with known spam signatures. In the second step, even if a match is not found with known spam signatures, we observe the words and phrases and their spamicity. The overall spam score of the message determines its likelihood of being a spam message. In the third step, we observe how many similar messages (in content) are deposited within a predefined time interval of ΔT.

3 Voice Message Signature Construction

This section provides technical details as how we can extract some specific features from a recorded message on the media server, which later on can be used to construct a signature of the deposited message.

3.1 Visual Representation of a Voice Message

Now assume that a telemarketer has left a voice message in one of the callees voice mailbox saying:

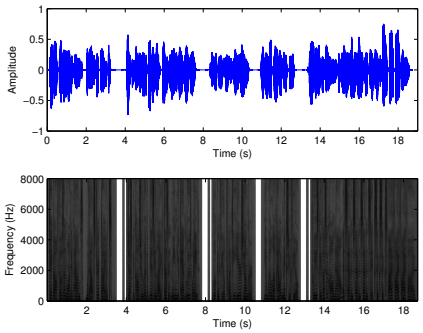


Fig. 4. Speech Waveform and Spectrogram (US Female Speaker)

Take off those unwanted pounds - without strict diets. Just because you live a busy life doesn't mean you can't lose weight. Look and feel 20 years younger. You will Love how it makes you feel. Please give us a call now at 777 666 5555

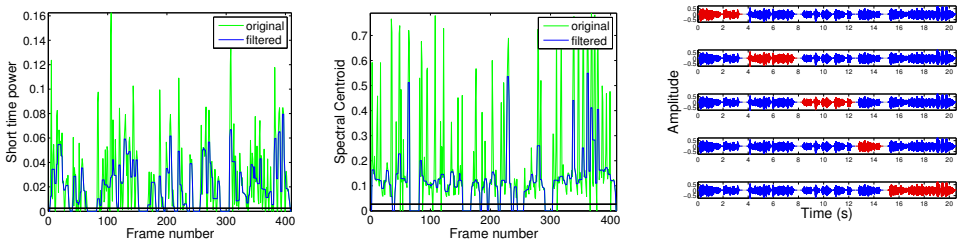
When we analyze the recorded .wav file, Fig. 4 shows the visual representation of human speech vibrations in the form of *waveform* and *spectrogram*. At the top, the waveform tracks variation in pressure as a function of time for a given point in space. Although we can learn quite a lot by a visual inspection of a speech waveform, it is impossible to detect individual speech sounds from waveforms because a speech consists of vibrations produced in the vocal tract. The vibrations themselves can be represented by speech waveforms. To read the *phonemes* in a waveform, we need to analyze the waveform into its frequency components, i.e., a spectrogram which can be deciphered (the bottom of Fig. 4). In the spectrogram, the darkness or lightness of a band indicates the relative amplitude or energy present at a given frequency.

3.2 Silence Removal From Deposited Voice Message

In our spam content analysis, we are interested in only voiced portions of the deposited message. Therefore, we need a method to remove all silence periods and segment the deposited message in voided segments. We use a method based on two simple audio features, namely the *signal energy* and the *spectral centroid*. In order to extract the feature sequences, the signal is first broken into non-overlapping short-term-windows (frames) of 50 msec. length. For each frame, the two features, described below, are calculated, leading to two feature sequences for the whole deposited voice message.

Signal Energy: Let us assume that the deposited voice message's i^{th} frame has N audio samples $x_i(n), n = 1, 2, ..., N$. The i^{th} frame energy is calculated as:

$$E(i) = \frac{1}{N} \sum_{n=1}^{N} |x_i(n)|^2 \tag{1}$$



(a.) Signal Energy (b.) Spectral Centroid (c.) Voiced Segments

Fig. 5. Detected voiced segments from a deposited voice message

Spectral centroid: The spectral centroid, C_i , of the i^{th} frame is defined as the center of gravity of its spectrum

$$C_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N}|X_i(k)|^2} \tag{2}$$

where $X_i(k)$, k=1,2,...,N, is the Discrete Fourier Transform (DFT) coefficients of the i^{th} short-term frame, where N is the frame length.

Estimating two thresholds – $T1$ and $T2$, the two feature sequences are compared with their respective thresholds. The voiced segments are formed by successive frames for which respective feature values are larger than their thresholds. The detailed description of the method can be found in [6]. We use the same example spam message recorded by Crystal, a US native English speaker and apply silence removal method. Fig. 5 (a) and (b) show energy and spectral centroid sequences and their threshold values, respectively. The detected voice segments are shown in Fig. 5 (c). These individual voiced segments serve as fundamental units to build our spam detection methodology.

3.3 RASTA-PLP Spectrogram Characterization

As the first step towards comparing two voiced segments, Short-time Fourier transform (STFT) can be adopted. Using STFT features, the sinusoidal frequency and phase content of local sections of a signal as it changes over time, can be determined. Since STFT, similar to most of speech parameter estimation techniques, is easily influenced by the frequency response of the speech channel, e.g. from a telephone line, we use another popular speech feature representation known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. PLP is a speech analysis technique for warping spectra to minimize the differences between speakers while preserving the important speech information [9]. RASTA was proposed to make PLP more robust to linear spectral distortions. RASTA applies a band-pass filter to the energy in each frequency subband to remove any constant offset resulting from steady-state spectral factors of the speech channel and to tolerate short-term noise variations [10]. After a deposited message is segmented to voiced segments, RASTA-PLP spectra for all

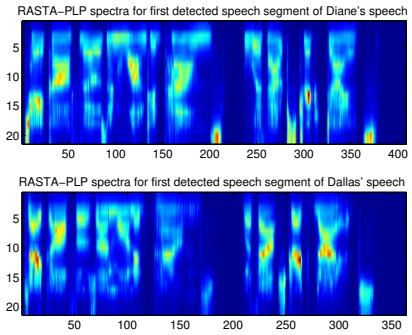


Fig. 6. RASTA-PLP spectral features for the first voiced segment of Diane (Female) and Dallas (Male) Native Speakers

voiced segments of the voice message is calculated. For each spam voice message, its RASTA-PLP spectral matrices, corresponding to its voiced segments, are stored in the spam signature database. Fig. 6 shows the RASTA-PLP spectrograms for the first voiced segment ("Take off those unwanted pounds without strict diets.") of two deposited messages from different speakers, Diane (Female English speaker) and Dallas (Male English speaker), with the same content.

3.4 Matching Process

The spam filtering architecture can work in a standalone or distributed collaborative manner. In the standalone mode, the voice messages left by the callers are undergoing through the behavioral analysis and signature matching based on the locally stored signatures. However, in the collaborative distributed mode, a group of disparate VoIP service providers work together. A centralized spam database can be queried as per need basis by individual service providers for signature matching, and at the same time newly found spam message is made available to the database so that it can be signaturised and used by the other service providers.

For signature matching and call behavior analysis, the newly arrived voice message is divided to voiced segments and corresponding RASTA-PLP matrices are calculated. The database of known spam signatures is queried to find the voice spam message that has similar content to the newly arrived voice message. If the computed cosine distance between the newly arrived and an already known spam message is less than a threshold, we confidently declare that a match has been found. However, in case there is no match found, then we perform call behavior analysis. Within a predefined time interval of ΔT (say 5 minutes), we segment all of the voice messages left on the media server to their voiced segments and calculate their corresponding RASTA-PLP matrices to observe how many messages are of similar content. Beyond a threshold value (say 3 messages per 5 minutes), the matched messages are considered to be a part of an impending spam attack and demand further analysis. The unmatched messages are deposited to their respective user accounts (i.e., mailboxes).

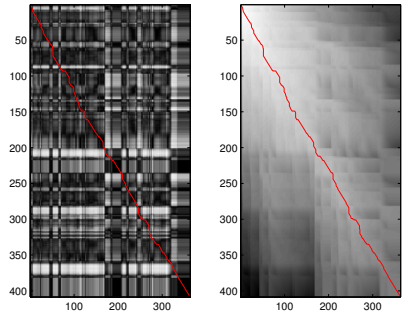


Fig. 7. Using DTW to find similarity between constructed scores matrices for the first voiced segment of Diane's speech and Dallas's speech

4 Detection Methodology

To either find if the newly arrived voice message has similar content to a spam signature or observe as how many similar messages (content-wise) are recorded on the media server within a predefined time interval, we propose a speaker independent speech recognition method. The newly deposited message is first divided into small voiced segments using the silence removal technique described in Section 3.2. For each of the voiced segments, we create RASTA-PLP matrices. As a similarity measure, we use Dynamic Time Warping (DTW) method and calculate the cosine distance for each pair of voiced segments coming from two different voice messages. Based on these partial scores for the corresponding speech segments, we finally determine if the two voice messages are similar enough and a match is found. The details of these phases are presented as follows.

4.1 Scoring similarity between two speech segments

Constructing scores matrix Cosine Similarity is considered here as the similarity measure between two speech segments. We calculate the cosine distance between every pair of frames from RASTA-PLP spectral matrices for two segments, and then we construct the *local match* scores matrix. The left side of Fig. 7 shows spectrogram-like scores matrix for the first voiced segment ("Take off those unwanted pounds without strict diets.") of two speech snippets of Diane (female) and Dallas (male) native speakers. High similarity values can be seen as a dark stripe approximately down the leading diagonal in the figure.

Dynamic Time Warping (DTW) Although two different voice segments (speaker's utterances) with same content have more or less the same sounds in the same order, the durations of each sub-segment (words and letters) may not match. As a consequence, matching between two voice segments without temporal alignment may fail. To cope with different speaking speeds and differences in timing between two segments, we use a dynamic programming method named Dynamic Time Warping (DTW) [5]. Considering a 2D space with the X-axis of time

frames from one segment and the Y-axis of time frames from another segment, DTW tries to find the path through this 2D space that maximizes the local match between the aligned time frames. The total *similarity cost* found by DTW can be considered as a proper indication of how well these two segments match. The right side of Fig. 7 illustrates how DTW finds the lowest-cost path between the opposite corners of the scores matrix. As we can see in the right side of Fig. 7, the path on the scores matrix follows the dark stripe depicted in the left side of Fig. 7.

Similar to other dynamic programing, the bottom right corner of the *minimum-cost-to-this point* matrix returns the cost of minimum-cost alignment of the two speech segments. This value as a partial score, can be used as our similarity measure. The smaller is a partial score, the closer are the two corresponding segments of different voice messages. Since the value of the partial score has a relationship with the size of spectral matrices (duration of voiced segments), we divided the partial score by the minimum duration of two segments to define a more comparable weighted partial score. To specify a threshold to find if two segments are similar enough, the method against many different voice messages is tested. Hence, we empirically found 10 as the proper threshold for acceptance or rejection of similarity between two segments.

4.2 Voice message content matching

To find if two speech messages are similar enough, weighted partial scores for all pair of corresponding segments of both messages are calculated. After comparing the weighted partial scores to the threshold value of 10 for each pair of corresponding segments, we can determine if the two segments have same content. If a certain number of corresponding segments for both messages have same content, the two whole speech messages are also similar enough and a match has been found.

4.3 Bayesian content filter

Based on the idea of Bayesian filtering for email spam, we propose a similar method for voice spam filtering. In this method, we have a database of known spam words named spam speech database. In the training phase, the spam words are converted to speech using text-to-speech (TTS) system and stored in the spam speech database. Speech words here can be a single word, a combination of words (i.e., phrase), phone number or URL address with high spamicity. In other words, we transform the known email spam database and its probabilities to voice spam world. Since there is no speaker independent speech segmentation method (without language-specific knowledge) to perfectly segment speech messages at the word level, we take an alternative approach. In our approach, entries of the spam speech database are tested against the voice message to find if the voice message includes an entry of the database. As an example, suppose Mike left a voice message, “Free mortgage consultations available now”, for his friend. To check if the deposited message is spam, entries of the database are tested against this voice message. Assuming that “mortgage” is an entry in the spam speech database, that was previously detected from another speaker (Crystal) and stored in the database,

we try to find if the voice message includes this speech word "mortgage". Starting from the beginning of the voice message, a frame in size of the entry of the database (speech word "mortgage") traverses the waveform of the speech message. While the frame traversing the message, the dissimilarity of the current frame of the speech message and the speech word from database ("mortgage") is calculated using DTW. Reaching the end of the speech message, the frame of speech message with maximum similarity is the determiner if the message includes the spam word ("mortgage"). This similarity score is compared to a threshold to find if the speech message includes that spam word. Using Bayes' Formula and based on the number and spamicity of spam words from the database that the spam message contains, we can decide if the the speech message is spam or not. To justify the threshold, as the most important part of this method, we have tested the method for different words and phrases in different sizes. Hence, it is empirically found that the similarity score using DTW is tightly related to the size of speech words. For example, DTW similarity score for word in size of "mortgage" is about 4.5-5, and for word in size of "777 5555 666" (as a phone number) is about 50. Therefore, the threshold is set in a dynamic way based on the size of the speech word to be tested.

4.4 Searching

As explained in Section 3, we construct two separate databases to store RASTA-PLP matrices; Spam Signature Database for spam signatures, and Spam Speech Database for spam words and phrases with high spamicity. After voice messages are left on the media server by callers, the Spam Signature Database is first queried to find a match. Entries in the Spam Signature Database can be organized in categories based on VoIP service providers where they have been locally stored from to speed up the search process. In case a match is not found (i.e., signature does not exist in the Spam Signature Database), entries of the Spam Speech Database are searched against the voice message to find if the voice message includes that entry of the database. After performing this search, Bayesian spam filtering is used to determine the final probability of the voice message being spam. To reduce the search time, we propose a cluster-like structure for the Spam Speech Database, where cluster heads are speech words with the highest probabilities in each cluster. For example, two clusters of the database are described here:

- Cluster 1:
  - cluster head: Viagra
  - cluster members: sex, cheap, night, www.buyviagraonline.com
- Cluster 2:
  - cluster head: Mortgage
  - cluster members: 100% free, lower interest, "555 666 7777" (phone number)

To perform a search, we start with cluster heads. If none of the cluster heads matches, the voice message is classified as non-spam. If one of the cluster heads matches, we narrow our search to the corresponding cluster to consider all other

Table 1. False Positive and False Negative rates of Voice Message Content Filtering

Case	Correct	False Positive	False Negative
#1	91	0	9
#2	87	1	12
#3	95	1	4

relevant words in relevance order. The Baye's Formula will take care of calculating the probability of being spam based on the number of spam segments it contains and their spamicities.

5 Performance Evaluation

We conduct a series of experiments to evaluate the performance of our solution. In our experiments, we left voice messages on Google voice [7] and then later on analyzed for their legitimacy and spam detection rate. In addition to these manually deposited voice messages, three popular TTS systems are used to generate various voice messages with different speakers in different sizes. Eight speakers were selected from AT&T Natural Voices® TTS system [1]. Twelve speakers were selected from Cepstral engine [3], as a TTS system that makes realistic synthetic voices. Moreover, ten speakers were selected from PlainTalk [22], the advanced built-in TTS technology of Mac OS. These thirty selected speakers have different accents (such as American, British or Indian English), different sex (male and female) with ages ranging from 10 to 60 years old.

5.1 Arrival of Same Content Voice Messages

This is a most generic spam attack scenario where a spammer repeatedly sends the same spam message to many of the subscribers at the same time. If a newly arrived voice message matches with any of the signatures stored in the database, the message is categorized as a spam message.

Ten totally different text messages with different size and content were converted to voice messages spoken by the thirty above mentioned different speakers to form 10 different sets of 30 voice spam messages with same content. All of these 300 different voice messages were first segmented into small voiced speech segments. Then the RASTA-PLP spectral matrices for all segments were calculated as well. After randomly selecting 3 voice messages of different speakers out of total 30 messages from each set of speech messages (with same content), a database with 30 entries were generated. For each sub-experiment, this process was repeated 10 times and each time one voice message from one of 10 sets is selected to check if it is spam. Iterating the sub-experiment 10 times forms a complete experiment. To take average, the complete experiment was conducted three times and the results are summarized in Table 1:

In our experiments, we found that our speaker independent spam detection algorithm can detect similar content message with 91% accuracy while generating

Table 2. Cluster details of the Spam Speech Database

Cluster	# of cluster members	# of special elements
Employment	24	4
Financial (Business)	15	2
Financial (Personal)	18	2
Marketing	35	5
Medical	18	3
Calls-to-Action	9	2

0.67% false positive rate and 8.33% false negative rate. However, if the newly arrived message does not match with any of the spam signatures stored in the database, we recorded its signature and observed if this signature matches with any of the future deposited messages within a predefined time interval of ΔT ($\simeq 5$ minutes). The similar message count beyond a threshold value within a time period can be categorized as an impending spam attack and needs further analysis.

We are aware that there are some legitimate applications that can generate calls in bulk. For example, it is possible that an emergency response system within a company, city or college may call many of the telephone numbers at the same time alarming about some untoward incidents. It is also possible for a credit card company to send a prerecorded generic message at a particular time to many of its customers regarding fraudulent activity in their accounts. In all such cases, there will be a number of matches (beyond a defined threshold value) within a predefined time interval ΔT and therefore possibly be labeled as spam messages without delivering to their respective mailboxes.

These legitimate call scenarios may cause false positives. To avoid such false positives, before labeling these legitimate voice messages as spam, our Bayesian content filtering method is used to calculate the probability of being spam for one of the newly deposited voice messages. Moreover, if we are provided with the calling numbers and the originating source IP addresses used by these bulk call applications in advance, then combining the SIP signaling information and content filtering approach can also avoid such false positives.

5.2 Hiding Spam Words/Phrases Within A Voice Message

In this set of experiments, the Spam Speech Database was built with 137 entries in five clusters: Employment, Financial (Business and Personal), Marketing, Medical, and Calls-to-Action. In addition to having one or more cluster heads, each cluster has several cluster members converted from email spam trigger words/phrases, and some special elements, such as URL address, email address and phone number, which have been extracted from our Spam Signature Database. Table 2 summarizes the details of the clusters in our Spam Speech Database.

To evaluate the efficiency of the proposed Bayesian based content filtering method, we recorded 30 various voice messages in different size from mentioned speakers with different accents, genders, and ages. This set of voice messages includes three types of voice messages as follows:

(1) *Spam voice message*: a voice message that includes at least one cluster head and either at least one special element or significant number of relevant cluster members. This type of voice messages should be classified as spam.

(2) *Doubtful voice message*: a voice message that includes at least one cluster head but neither special element nor significant number of relevant cluster members. Although this type of voice messages could be classified as either spam or non-spam, our system classifies it as non-spam to reduce the false positives. In other words, a few relevant words/phrases from a cluster of the Spam Speech Database do not classify a deposited message as spam. There have to be enough words/phrases with a high spamicity to outweigh the rest of the voice message that includes words/phrases with a low spamicity. For example, a voice message from your spouse taking out a second mortgage on the house should not be misclassified as spam.

(3) *Non-spam voice message*: a voice message that does not include even one cluster head. This type of voice messages should be classified as non-spam.

Our Bayesian based spam detection method is used to classify the test set of voice messages. The results show that the method can correctly classify 83.33% of voice messages while 13.33% of either non-spam or doubtful voice messages are misclassified as spam and 20% of spam voice messages are not detected. We further looked into the results and details of the method to find the causes of these false positives and false negatives. It is discovered that the problem arises when voice messages are deposited by speakers with accents rather than US English, such as British or Indian English. Since the entries of our Spam Speech Database are converted by Crystal, a US native English speaker from spam email world, the dissimilarity score computed by our DTW based algorithm is not dependable enough to compare the small-size speech words of those speakers with different accents.

6 Related Work

The SIP IETF working group has published a couple of informational drafts proposing (1) *computational puzzles* to reduce spam in SIP environments and (2) an extension of SIP protocol to send user's feedback information to the SPIT identification system [12, 14]. To some extent, the combination of user's whitelist with the Turing tests or computational puzzles can prevent spam calls. However, the capability of a SIP UA to solve the computational puzzle relies on its computing resources. Therefore, it cannot be ignored that a spammer can potentially have significantly more resources than a normal user. The solving of audio Turing tests requires caller's time and manual intervention. Still, the Turing tests cannot be a solution for deaf (or blind) users and can be thwarted by employing cheap labor. Recently, a number of products such as Sipera's IPCS [19] and NEC's VoIP SEAL [11] incorporate audio Turing test to solve the voice spam problem. However, an attacker may abuse these security devices as reflectors and amplifiers to launch a stealthy DDoS attack [16]. Now we review some other related work on SPIT prevention.

Inferring Spoken Words. The closest work to our approach is a method in which the spam detection module detects spoken words within an established voice stream. The most intuitive way to detect a spam message is to use *"speech-to-text"* engine, where deposited voice messages can be converted to text format and then the well-known email filtering approaches can be used for detection. However, the performance of speech-to-text engine is largely depends on speaker, speaking style, ambient environment, and language. Because of the high error rate, this approach is still far away to become a commercially viable solution to filter voice spam messages.

Collaborative Approach. Google Voice [7] has a feature to report calls as spam and block future calls from that number. This is a reactive approach requiring spam call to be received by a user and then block that number. It has a few drawbacks to be applicable in telecommunication networks: (1) what will happen if the spam message is generated from a spoofed number, e.g., every time a new telephone number is used to send a spam message; (2) the current generation of hardphones do not provide any button to send feedback about received spam calls; (3) it is based on inferring spoken words and thus suffers from the same drawback as discussed above; and (4) there is no previous study on what will happen if the message content itself mutates (i.e., spam messages use different accents or male/female speakers), making it difficult to infer spoken words.

Content analysis. The *V-Priorities* [8] system developed by Microsoft is explored to filter spam calls. V-Priorities works on three levels: first, analysis examines the prosody – rhythm, syllabic rate, pitch, and length of pauses – of a caller's voice; secondly, rudimentary word and phrase recognition is done to spot target words that could indicate the nature of a call; and finally, at the third level analysis involves metadata, such as the time and length of a message. The voice content analysis does not require maintenance of caller's call history and remains independent of signaling. However, this approach suffers from scalability issue since it is difficult to monitor hundreds of voice streams simultaneously. The real-time content analysis is an exceedingly difficult task. By the time, calls are analyzed to be spam calls, it has already affected the receiver (human recipient or voice mailbox). The prosody analysis of machine generated voice may give different results compared to human generated voice. As mentioned earlier, inferring spoken words makes it error-prone and its success largely depends upon users, ambient environment, and language.

Black/Whitelists, trust and reputation system. The unwanted callers and domains are blacklisted so that their future calls can be filtered as spam calls. By contrast, the known callers are put in a whitelist and the calls from such callers are given preference by allowing them to go through. The trust and reputation system is used in conjunction with black/whitelists. The *social network* mechanism is used to derive a reputation value for a caller. Dantu et al. [4] used the Bayesian algorithm to compute the reputation value of a caller based on its past behavior and callee's feedback. Rebahi et al. [13] derived caller's reputation value by consulting SIP repositories along the call path from call's source to its destination. As an anti-spam solution, Sipera's IPCS [19] also relies on caller's reputation value. These solutions can block the spam call during the call setup

phase. However, the derivation of caller's reputation value requires building a social network. The notion of user's feedback requires the modification of SIP clients and an extension of SIP protocol [12]. The construction of a whitelist suffers from the *introduction problem* and the calculation of a reputation value is vulnerable to "bad-mouthing attacks", where malicious users may collude and provide unfair ratings for a particular caller. Furthermore, these schemes rely on caller's identity which can be spoofed.

Call duration-based Approach. Sengar et al. [17] observed the significance of call duration in spam detection and raised a fundamental question about how small it could be for normal conversations. Their proposed statistical approach lacks the consideration of those calls that are hidden behind a firewall, SBC or B2BUA agents. Balasubramaniyan et al. [2] used the call duration to develop call credentials. A caller provides a call credential to the callee when he makes a call. However, a spammer could set up at least two accounts to build call credentials by calling each other and then later on use these trusted accounts to launch spam attacks.

Recently, Wu et al. [23] proposed a spam detection approach involving user-feedback and semi-supervised clustering technique to differentiate between spam and legitimate calls. However, the current generation of telephone sets do not provide an option to give feedback of a call to service provider's system. Sengar et al. [18] used callers calling behavior (day and time of calling, call duration etc.) to detect an onslaught of spam attack. However, it is difficult to capture calling pattern for each of the subscribers and, being an after-the-fact method, by the time we detect a spam attack many of the subscribers must have already been affected by the spam.

7 Conclusion

Although there are very few reported incidents of voice spam today, with the growth of VoIP and its openness, the voice spam could become a serious threat in the near future. The heart of the problem lies in the fact that a spammer can send unsolicited advertisements and messages with low or no cost while being anonymous. Unfortunately, many of the mechanisms which work for email spam fail completely in the context of VoIP. Most of the previous solutions against voice spam are proposed to distinguish a legitimate subscriber from a spammer using SIP signaling messages. Instead of analyzing the SIP signaling messages to identify the spammer, this paper proposes a speaker independent speech recognition scheme for content filtering to avoid spam message deposition on the subscribers' voice mailboxes. Being a speaker independent, computationally efficient, and scalable solution, our proposed approach can effectively protect subscribers' voice mailboxes from spam messages. Our work is evaluated in real-world experiments. The experimental results show that our spam filtering approach can successfully classify a voice message into spam with 91% accuracy, while having 0.67% false positive rate and 8.33% false negative rate.

References

1. AT&T Labs Research. At&t natural voices® text-to-speech system. Website, http://www2.research.att.com/ ttsweb/tts/.
2. V. Balasubramaniyan, M. Ahamad, and H. Park. CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation. In *The Fourth Conference on Email and Anti-Spam*, 2007.
3. Cepstral®. Cepstral text-to-speech engine. Website, http://www.cepstral.com/.
4. R. Dantu and P. Kolan. Detecting spam in voip networks. In *Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop*, 2005.
5. D. Ellis. Dynamic time warp (dtw) in matlab. Web resource, http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/, 2003.
6. T. Giannakopoulos. A method for silence removal and segmentation of speech signals, implemented in matlab. Web resource, http://www.mathworks.com/matlabcentral/fileexchange/authors/30223, 2010.
7. Google. Google Voice. Website, www.google.com/voice, 2011.
8. D. Graham-Rowe. A Sentinel to Screen Phone Calls. Website, http://www.technologyreview.com/communications/17300/?a=f, 2006.
9. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.
10. H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.
11. NEC Corporation. NEC Develops World-Leading Technology to Prevent IP Phone SPAM. Product News, http://www.nec.co.jp/press/en/0701/2602.html, 2007.
12. S. Niccolini, S. Tartarelli, M. Stiemerling, and S. Srivastava. SIP Extensions for SPIT identification. draft-niccolini-sipping-feedback-spit-03, IETF Network Working Group, Work in Progress, 2007.
13. Y. Rebahi and A. Al-Hezmi. Spam Prevention for Voice over IP. Technical report, http://colleges.ksu.edu.sa/ComputerSciences/Documents/NITS/ID143.pdf, 2007.
14. J. Rosenberg and C. Jennings. The Session Initiation Protocol (SIP) and Spam. RFC 5039, IETF Network Working Group, 2008.
15. J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, IETF Network Working Group, 2002.
16. H. Sengar. *Beware of New and Readymade Army of Legal Bots.* USENIX ;login:, October 2007.
17. H. Sengar. Voice Spam (SPIT) Problem. Website, http://www.vodasec.com/, March 2007.
18. H. Sengar, X. Wang, and A. Nichols. Call Behavioral Analysis to Thwart SPIT Attacks on VoIP Networks. In *SecureComm*, 2011.
19. SIPERA. *Sipera IPCS: Products to Address VoIP Vulnerabilities.* http://www.sipera.com/index.php?action=products,default, April 2007.
20. VOIPSA. Confirmed cases of SPIT. Mailing list, http://www.voipsa.org/ pipermail/voipsec_voipsa.org/2006-March/001326.html, 2006.
21. VOIPSA. VoIP Attacks in the News. Website, http://voipsa.org/blog/ category/voip-attacks-in-the-news/, 2007.
22. Wikipedia. Plaintalk. Website, http://en.wikipedia.org/wiki/PlainTalk.
23. Y.-S. Wu, S. Bagchi, N. Singh, and R. Wita. Spam Detection in Voice-Over-IP Calls through Semi-Supervised Clustering. In *IEEE Dependable Systems and Networks Conference (DSN 2009)*, June-July 2009.