EIN NEUES VERFAHREN ZUR IDENTIFIKATION UND ABWEHR VON TELEFON-SPAM

H. Knospe[1], C. Pörschmann[1]

[1]Gefördert vom Bundesministerium für Bildung und Forschung (BMBF), Förderkennzeichen 1736X09

Institut für Nachrichtentechnik
Fachhochschule Köln
E-mail: heiko.knospe@fh-koeln.de

SUMMARY: The introduction of IP-based networks for telephony (Voice over IP, VoIP) provides many possibilities and cost advantages but also causes new threats. It is for example easily possible to import pre-recorded calls. These calls are generally unrequested. The IP-based telephone SPAM is usually called SPIT and it may evolve into a similar problem as E-Mail SPAM.
There are a number of approaches to identify and to hold off SPIT. The existing mechanisms usually analyze and potentially extend the call signaling exchange. This paper explicates a new method which is based on the analysis of audio signals. The incoming audio data are inspected for identical voice segments. Thereby similar voice data is also detected so that the method is robust with respect to speech coding, noise and other variances. An acoustic fingerprint is computed which has also been used for music identification. The comparison of the audio signals is based on the computation of distances of high-dimensional spectral feature vectors. A Matlab-based implementation of the method showed already promising results.

ZUSAMMENFASSUNG: Der Einsatz von IP-basierten Netzen für Telefonie (Voice over IP, VoIP) bietet viele Möglichkeiten und Kostenvorteile, führt aber auch zu neuen Bedrohungen. So können beispielsweise automatisiert und mit geringem Aufwand aufgezeichnete Anrufe eingespielt werden. Diese Anrufe sind in der Regel unerwünscht. Der IP-basierte Telefon-SPAM wird als SPIT (Spam over IP Telephony) bezeichnet und kann in Zukunft ein ähnliches Problem wie E-Mail SPAM darstellen.

Es gibt zahlreiche Ansätze, SPIT zu identifizieren und abzuwehren. Die existierenden Vorschläge und Verfahren basieren üblicherweise auf einer Auswertung und gegebenenfalls auf einer Erweiterung der Gesprächs-Signalisierung. Dieser Beitrag erläutert ein neues Verfahren, welches auf einer Audiosignalanalyse basiert; die eingehenden Audiodaten werden hierbei auf identische Sprachabschnitte untersucht. Dabei werden auch ähnliche Sprachdaten erkannt, so dass das Verfahren robust gegenüber der Sprachcodierung, Rauschen oder anderen Veränderungen ist. Hierzu wird ein akustischer Fingerabdruck berechnet, wie dies in einer ähnlichen Form bereits bei der Musik-Identifikation eingesetzt wird. Der Vergleich der Audiodaten basiert auf Abstandsbestimmungen von hochdimensionalen spektralen Merkmalsvektoren. Eine Matlab-basierte Implementierung des Verfahrens zeigte bereits Erfolg versprechende Ergebnisse

EINLEITUNG

Im Rahmen der Migration der bisherigen Daten- und Telekommunikationsnetze auf IP-basierte Next Generation Networks (NGNs) wird die klassische leitungs- und vermittlungsorientierte Telefonie durch Voice-over-IP (VoIP) abgelöst. Für die Telekommunikations-Operator bietet VoIP zumindest auf mittlere und lange Sicht Kostenvorteile, da die spezialisierte Vermittlungstechnik durch Standard-Komponenten ersetzt werden kann und Sprache und Daten über ein gemeinsames Kernnetz auf Basis des Internet Protokolls übertragen werden.
Für die Vermittlung der Sprachkommunikation über IP hat sich das Session Initiation Protocol (SIP) [1] gegenüber dem konkurrierenden H.323 weitgehend durchsetzen können. Die Umstellung der Festnetz-Telefonie auf VoIP/SIP hat bereits begonnen und auch im Mobilfunk wird zukünftig VoIP/SIP verwendet werden („Long Term Evolution“). Es bleibt noch abzuwarten, welche weitere Entwicklung proprietäre aber populäre VoIP-Dienste wie Skype nehmen werden.
SIP lehnt sich bezüglich des Nachrichtenaufbaus an das bewährte HTTP Protokoll an, verwendet aber in der Regel UDP als Transportprotokoll. Die Verbindung wird zwischen den Teilnehmern (bzw. zwischen den Teilnehmern und Proxy-Servern) mit SIP-Requests (z.B. INVITE) und Responses (z.B. Trying, Ringing, OK) aufgebaut. Gleichzeitig werden die Medienparameter (z.B. Audio-Codecs) mit Hilfe des Session Description Protocols (SDP) ausgetauscht. Die eigentlichen Mediendaten werden mit dem Real-Time Transport Protocol (RFC 3550) [2] übertragen.
Mit der Umstellung von vergleichsweise abgeschirmten klassischen Telekommunikationsnetzen auf IP-basierte Netze und insbesondere das Internet ergeben sich neue Sicherheitsrisiken. Diese sind seit einiger Zeit Gegenstand umfangreicher Untersuchungen, die auch mögliche Sicherheitsmaßnahmen aufzeigen (z.B. [3]).
Eine spezielle Bedrohung ist die Möglichkeit, automatisiert Gespräche aufzubauen und dann

aufgezeichnete Sprachnachrichten einzuspielen (Robocalls). In IP-basierten Netzen sind Aufwand und Kosten für solche (in der Regel unerwünschten) Anrufe gering. Der IP-basierte Telefon-SPAM, der in der Regel als SPIT bezeichnet wird, kann in Zukunft ein ernstes Problem darstellen, das vergleichbar mit E-Mail SPAM ist. Es gibt derzeit zahlreiche Aktivitäten in den Standardisierungsgremien, SPIT zu identifizieren und den Empfänger vor diesen Anrufen zu schützen (z.B. [4]). Weiterhin nehmen sich auch die politischen Gremien und der Gesetzgeber dem Problem an. So wurde jüngst ein Gesetz in Deutschland verabschiedet, das für unerlaubte Telefonwerbung und Werbung mit unterdrückter Telefonnummer hohe Geldbußen vorsieht [5].
SPIT wird durch den Aufbau einer Sitzung mit einer SIP INVITE Nachricht eingeleitet. Wenn der betroffene Teilnehmer (SIP Response 200 OK) antwortet, so spielt der Angreifer eine Sprachnachricht (z.B. Werbung) über das RTP Protokoll ein. Der Angriff kann automatisiert gegen eine größere Zahl von Teilnehmern ablaufen. Hierfür stehen Tools wie SIPp [6] zur Verfügung. Schon bei einer Netzanbindung mit 500 kBit/s (upstream) sind nach Berechnungen in [4] mindestens 3 erfolgreiche Anrufe pro Sekunde möglich.

VORSCHLÄGE ZUR ABWEHR VON SPIT

Seit einiger Zeit werden Maßnahmen zur Abwehr von SPIT diskutiert, entwickelt und teilweise auch bereits in kommerziell verfügbare Produkte integriert (z.B. [7]). Es gibt insbesondere die folgenden Verfahren:

- Black-Lists: Anrufe von bestimmten Teilnehmern werden abgelehnt.
- White-Lists: Es werden nur Anrufe bestimmter Teilnehmer angenommen.
- Reputation oder Ranking: Nur Anrufer mit guter Bewertung werden akzeptiert.
- Behavioral Detection: Auffällige Anrufer (z.B. viele Anrufe in kurzer Zeit, kurze Gesprächsdauer) werden identifiziert und gesperrt.
- Challenge-Response oder Testverfahren: Der Anrufer muss zunächst Antworten auf Fragen liefern.
- Zahlung: Anrufer zahlen pro Anruf einen Betrag, der zurückerstattet wird, wenn der angerufene Teilnehmer erkennt, dass es sich nicht um SPIT handelt.

Diesen Verfahren bzw. die Vorschläge basieren meistens auf einer Analyse oder Erweiterung der SIP Signalisierung. Eine zentrale Rolle spielt dabei eine zuverlässige, authentifizierte Identität der Teilnehmer, die für eine Freigabe oder wirksame Sperrung von großer Bedeutung ist. Für E-Mail konnte dies bislang nicht erreicht werden, was einen wesentlichen Grund für den massenhaften E-Mail Versand mit gefälschten Absender-Adressen darstellt. Allerdings unterscheidet sich die Situation bei VoIP und E-Mail: obwohl es technisch möglich ist, so erfolgt doch wenig direkte Übernahme von VoIP- Calls von nicht authentifizierten Teilnehmern aus dem Internet. Die Gespräche werden derzeit in der Regel

- innerhalb von lokalen Netzen (z.B. über VoIP Telefonanlagen von Unternehmen), oder
- über einen Proxy (oder Session Border Controller) eines Telekommunikations-Operators

vermittelt. In beiden Fällen wird eine Challenge-Response Authentifikation unter Verwendung eines vorab vereinbarten Passworts durchgeführt.

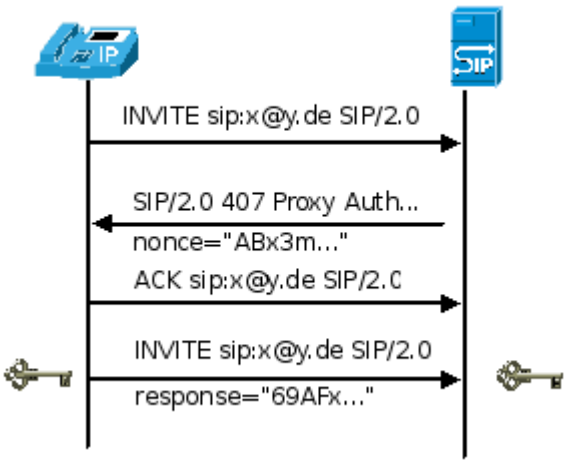

Abbildung 1: Challenge-Response Authentifikation (SIP Digest).

Die Übernahme einer gefälschten Identität ist damit erheblich erschwert und gewährleistet die Identität des Anrufers zumindest gegenüber dem für die Authentifikation zuständigen Proxy- oder Registrar Server, sofern das verwendete User-Passwort geschützt und ausreichend komplex ist. Für Sicherung der Identität bei Übergabe von Gesprächen zwischen verschiedenen Netzen steht auch die Möglichkeit der Signatur von Header-Informationen zur Verfügung (RFC 4474 [8]).

EIN NEUES VERFAHREN ZUR SPIT IDENTIFIKATION

Die bisherigen Vorschläge und Verfahren basieren in der Regel auf der Auswertung von Signalisierungsdaten und beinhalten keine Analyse der eigentlichen Gesprächsdaten. Diese werden erst übertragen, wenn das Gespräch vermittelt und insofern auch ein SPIT-Anruf bereits erfolgreich war. Dennoch ist eine Analyse der Gesprächsdaten a posteriori sinnvoll, um eingespielte Nachrichten zu erkennen und weitere SPIT-Anrufe bereits bei Signalisierung mit Hilfe einer Black-List zu unterbinden.

Der Ansatz des neuen Verfahrens [9] ist daher der Vergleich von Audiodaten mit Hilfe eines akustischen Fingerabdrucks. Ähnliche Verfahren werden bereits zur

Musik-Identifikation eingesetzt [10]. Bereits beim zweiten SPIT-Anruf kann die Wiedereinspielung erkannt und die Anrufer-ID zumindest in eine Beobachtungs-Liste aufgenommen werden. Weitere SPIT-Anrufe erhöhen die Zuverlässigkeit der Erkennung und begründen die Aufnahme des Anrufers in eine Black-List.

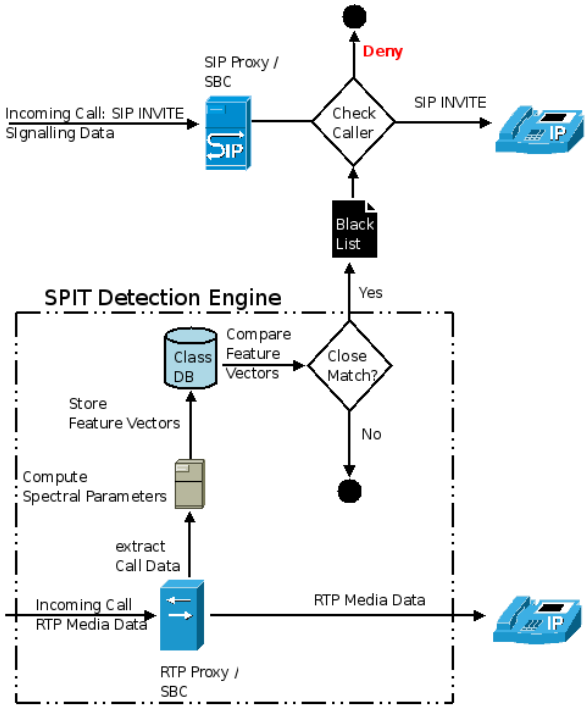

Abbildung 2: Schematische Darstellung des Verfahrens

Die eingehenden Mediendaten (oder zumindest eine Anfangssequenz von z.B. 10 Sekunden) werden beim Diensteanbieter extrahiert (kopiert) und dann spektrale Feature-Vektoren ermittelt, die keine Rekonstruktion der Inhalte zulassen. Die Audiodaten werden anschließend unmittelbar wieder gelöscht (Fernmeldegeheimnis). Die Vermittlung des Gesprächs an den angerufenen Teilnehmer findet wie gewöhnlich statt. Die Feature-Vektoren werden in einer Datenbank gespeichert. Anschließend werden die neuen Feature-Vektoren mit den gespeicherten Daten verglichen und ggf. eine Übereinstimmung mit früheren Anrufen festgestellt. In diesem Fall wird der Anrufer in eine Beobachtungs-Liste bzw. in eine Black-List aufgenommen, um künftige Anrufe zu unterbinden.

BESTIMMUNG UND VERGLEICH VON FEATURE VEKTOREN

Der Kern des Verfahrens ist die Berechnung des akustischen Fingerabdrucks und der anschließende Vergleich von Merkmalsvektoren. Die Auswahl der spektralen Parameter und die Entwicklung eines effektiven Verfahrens für eine robuste Erkennung von identischen oder ähnlichen Audio-Sequenzen ist Gegenstand des vom BMBF (Bundesminsterium für Bildung und Forschung) geförderten Projektes VIAT (Verfahren zur Identifikation und Abwehr von Telefon-SPAM).

Für die Bestimmung der Feature-Vektoren werden die Audiodaten eingelesen, in kurze Zeitabschnitte (30 ms) zerlegt (mit jeweils 10 ms Verschiebung) und eine Fensterfunktion (Hamming) angewendet.

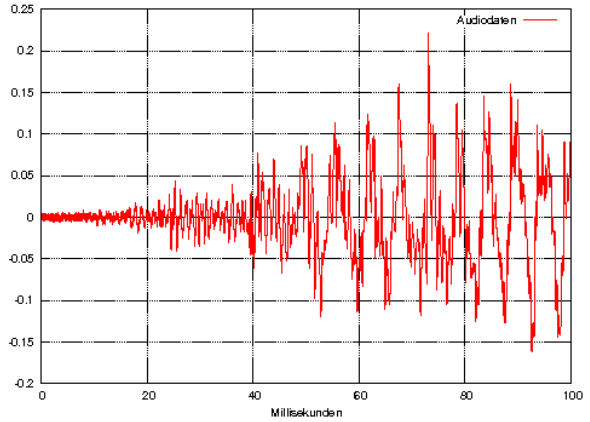

Abbildung 3: Beispiel für Audiodaten im Zeitbereich.

Anschließend wird die diskrete Fourier Transformation (FFT) auf die abgetasteten und gefensterten Daten angewendet und das Spektrum für jedes Fenster bestimmt.

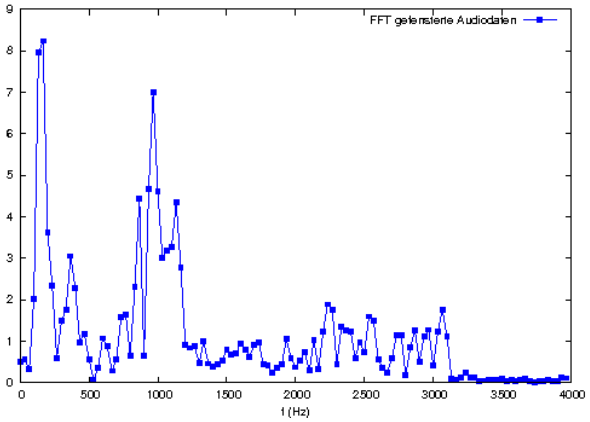

Abbildung 4: Spektrum der Audiodaten aus Abbildung 3 im Zeitbereich 30 – 60 Millisekunden mit Hamming Fensterfunktion.

Das Spektrum zwischen 350 Hz und 4 kHz wird dann logarithmisch in 14 Bänder zerlegt und je Band das Spectral Flatness Measure (SFM) berechnet. Das SFM ist definiert als Quotient von geometrischem und arithmetischem Mittelwert. Die SFM Werte liegen zwischen 0 und 1 und bestimmen die Flachheit des Spektrums; ein Wert nahe 0 deutet auf einen Sinuston, ein Wert nahe 1 auf Rauschen hin. SFM ist einer der Low-Level Audio-Deskriptoren im MPEG-7 Standard [11].

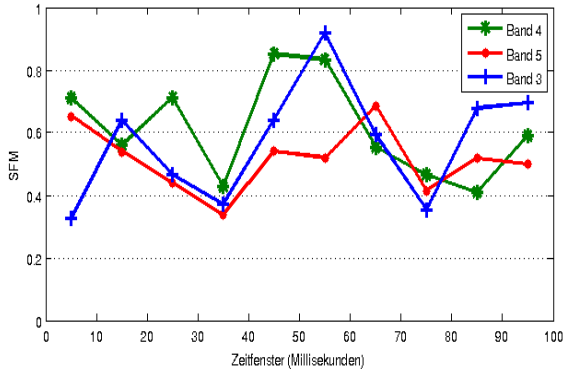

Abbildung 5: SFM-Werte der Audiodaten aus Abbildung 3 für drei verschiedene Frequenzbänder.

Anschließend findet eine gleitende Mittelung von jeweils 50 SFM-Werten statt. Schließlich liegt je 10 Millisekunden ein (14-dimensionaler) SFM Vektor vor. Da es sich ggf. um einige Tausend Vektoren handelt, wird noch ein Vektor-Quantisierer angewendet [12], der die Daten auf 256 Merkmalsvektoren komprimiert. Insgesamt kann ein SFM-basierter -Fingerabdruck eines Anrufs dann bei einer 8-Bit Auflösung der spektralen Merkmalswerte durch 3584 Bytes dargestellt werden. Weitere spektrale Größen wie Spectrum Crest Factor (SCF), der Quotient von maximalem Spektralwert und arithmetischem Mittelwert (je Band), oder Maxima im Frequenzzeitverlauf, können hinzukommen.
Für einen Vergleich der Audiodaten von zwei Anrufen wird nun der Abstand zwischen den Merkmalsvektoren der beiden Sets ermittelt. Ein geringer Abstand deutet auf eine Ähnlichkeit der Audiodaten für eine Zeitdauer von mindestens 500 ms hin. Die genauen Parameter müssen noch bestimmt werden, um eine robuste Erkennung (auch bei verrauschten Audiodaten und anderen Veränderungen) zu gewährleisten, aber andererseits auch eine möglichst geringe Zahl von falschen SPIT-Identifikationen zu erhalten.

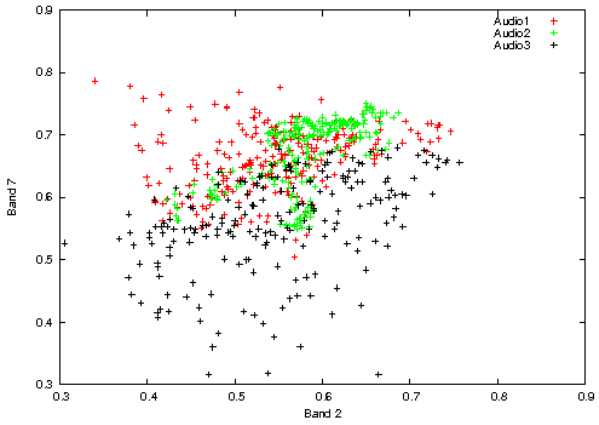

Abbildung 6: Merkmalsvektoren in zwei Dimensionen; Audio 1 und Audio 2 sind ähnlich, Audio 3 unterschiedlich.

Erste Ergebnisse haben ergeben, dass eine Erkennung auch bei den üblichen Modifikationen des Audiosignals (Änderung der Amplitude und der Abspielgeschwindigkeit, Hochpass- / Tiefpass-filterung, Verzerrungen, Ausschneiden von Abschnitten) gegeben ist, das Hinzufügen von stärkerem Rauschen (z.B. SNR=6 dB) aber noch problematisch ist.

SCHLUSSFOLGERUNGEN

Zur Identifikation und Abwehr von eingespielten SPAM-Anrufen existieren verschiedene Vorschläge und Verfahren, die vor allem auf einer Analyse der Signalisierungsinformationen und der Filterung von Anrufern basieren. In einem neuen Verfahren werden die Mediendaten ausgewertet, wobei eingehende Anrufe mit ähnlichen Audiodaten erkannt werden sollen. Hierfür wird ein akustischer Fingerabdruck bestimmt, der die Audiodaten charakterisiert ohne aber eine Rekonstruktion zu erlauben. Hierzu werden spektrale Merkmalsvektoren berechnet (u.a. Spectrum Flatness Measure), in einer Datenbank gespeichert und dann für einen Vergleich der Audiodaten herangezogen.
Eine weitere Parametrisierung und Optimierung sowie die Entwicklung einer SPIT Detection Engine und die anschließende Gesamtintegration des Verfahrens soll im Rahmen des vom BMBF geförderten Forschungsprojekt VIAT durchgeführt werden.

LITERATUR

[1] Rosenberg, A. et al. (2002). SIP: Session Initiation Protocol (RFC 3261).
[2] Schulzrinne, H. et al. (2003). RTP: A Transport Protocol for Real-Time Applications (RFC 3550).
[3] Kuhn, D. et al. (2005). Security Considerations for Voice Over IP Systems. Recommendations of the National Institute of Standards and Technology, NIST Special Publication 800-58.
[4] Rosenberg, J. et al. (2008). The Session Initiation Protocol (SIP) and Spam (RFC 5039).

[5] Gesetzesbeschluss des Deutschen Bundestages (2009). Gesetz zur Bekämpfung unerlaubter Telefonwerbung und zur Verbesserung des Verbraucherschutzes bei besonderen Vertriebsformen, Drucksache 353/09.
[6] SIPp Test Tool. http://sipp.sourceforge.net .
[7] IPTEGO PALLADION. http://www.iptego.com/palladion .
[8] Peterson, J., Jennings, C. (2006). Enhancements for Authenticated Identity Management in the Session Initiation Protocol (SIP) (RFC 4474).
[9] Pörschmann, C., Knospe, H. (2008). „Analysis of Spectral Parameters of Audio Signals for the Identification of Spam Over IP Telephony," in: The Fifth Conference on Email and Anti-Spam – CEAS 2008, Mountain View.
[10] Allamanche, E., Cremer M., Fröba, B., Hellmuth, O., Herre J., Kastner, T. (2001). Content based Identification of Audio Material Using MPEG-7 Low Level Description, 2nd Annual International Symposium on Music Information Retrieval.
[11] MPEG-7 (2002). Information Technology – Multimedia Content Description Interface – part 4, ISO/IEC FDIS 15938-4.
[12] Linde, Y., Buzo, A., Gray R.M. (1980). An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, 702-710.