

Effiziente Identifikation von Telefon-Spam

Julian Strobl^{1,2}, Frank Kurth¹, Gary Grutzek², Heiko Knospe²

¹ Fraunhofer FKIE, 53343 Wachtberg, Email: {julian.strobl|frank.kurth}@fkie.fraunhofer.de

² Fachhochschule Köln, Institut für Nachrichtentechnik, 50679 Köln, Email: {gary.grutzek|heiko.knospe}@fh-koeln.de

Einleitung

Telefon-Spam, etwa in Form automatisiert und massenhaft eingespielter Werbeanrufe, kann in Zukunft zu einem relevanten Problem werden. Das Projekt VIAT [2] (Verfahren zur Identifikation und Abwehr von Telefon-SPAM) untersucht die Möglichkeiten zur automatischen Analyse und Erkennung der bei der Telefonie übertragenen Audiodaten. Dafür werden Abschnitte von wenigen Sekunden des Signals mit zuvor eingegangenen Anrufen verglichen. Bei Identifikation von wiederholt eingespielten Sprachdaten kann der Anruf als Telefon-Spam betrachtet und entsprechende Maßnahmen eingeleitet werden. Die technische Herausforderung bei diesem Vorgehen besteht im effizienten und gleichzeitig robusten Abgleich der Audiodaten mit denen vorheriger Anrufe. Die vorliegende Arbeit stellt hierzu ein Verfahren vor, bei dem in einem Vorverarbeitungsschritt robuste Audiomerkmale aus den Anrufen extrahiert werden, die dann als Fingerabdrücke zum Signalvergleich herangezogen werden. Ein indexbasiertes Suchverfahren erlaubt die schnelle Identifikation vorliegender Fingerabdrücke auch bei einer sehr großen Zahl erfasster Anrufe.

Merkmalsextraktion

Zur zuverlässigen Identifikation von Duplikaten müssen charakteristische und robuste Merkmale der aufgezeichneten Audiodaten extrahiert werden, die den Fingerabdruck bilden. Der verwendete Fingerabdruck basiert auf dem Spektrum kurzer, überlappender Ausschnitte der Audiodatei. Die Merkmalsvektoren werden bandbegrenzt von 300 Hz bis 2 kHz und über eine Mel-Filterbank in einzelne Bänder zerlegt.

Die Merkmalsvektoren eignen sich nicht direkt für das hier verwendete indexbasierte Suchverfahren [1]. Für eine Indexierung müssen diese Merkmalsvektoren $V := \{v(1), v(2), \dots, v(N)\}$ jeweils in eine von \tilde{R} Merkmalsklassen quantisiert werden.

Zur Quantisierung werden im Falle der hier vorliegenden \tilde{R} -dimensionalen Merkmalsvektoren Pegelspitzen herangezogen. Die Quantisierung ordnet dann jedem Merkmalsvektor $v(i) = (v_1, v_2, \dots, v_{\tilde{R}})$ die Position \tilde{r} des maximalen Eintrags $v_{\tilde{r}}$ zu. Diese Position der Spitzen ist besonders robust gegenüber Störungen und kann prinzipiell direkt zur Indexierung herangezogen werden. Um jedoch die Auftrittswahrscheinlichkeit gleicher Spitzen in anderen Dateien und somit die Wahrscheinlichkeit falscher Positive zu reduzieren, werden mehrere Pegelspitzen zu Mustern kombiniert und über eine Tabelle einer von R verschiedenen Merkmalsklassen $r \in [1, R]$ zugeordnet.

Implementierung

Die Fenstergröße zur Merkmalsextraktion liegt bei 128 ms mit einer Überlappung von 96 ms. Bei einer Aufzeichnungsdauer von sechs Sekunden werden somit über 180 Merkmalsvektoren extrahiert. Von jedem Merkmalsvektor wird der maximale Eintrag bestimmt. Jeweils zwei Maxima im Abstand von fünf Fenstern werden kombiniert, um ein Muster zu erzeugen. Aus Gründen der Robustheit werden nur solche Maxima verwendet, die aus energiereichen Fenstern extrahiert wurden. Daraus ergeben sich pro Fingerabdruck ungefähr 30 bis 50 Werte aus $R = 841$ Klassen.

Suchverfahren

Das verwendete effiziente Suchverfahren ist listen- bzw. indexbasiert [1]. Aus den Sprachdaten d_1, \dots, d_N werden

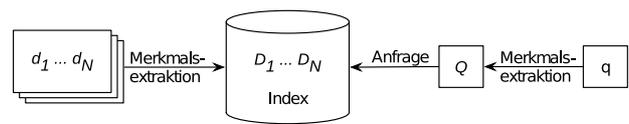


Abbildung 1: Merkmalsextraktion zum Aufbau des Index und zur Anfragebehandlung.

Merkmale extrahiert und als Merkmalsdokumente D_i gespeichert. Aus den Sprachdaten der Anfrage q (Anruf) werden ebenfalls Merkmale extrahiert und als Anfragedokumente Q zum Abgleich mit dem Index benutzt.

Die Merkmalsdokumente sind organisiert in *invertierten Listen* $L(r)$, die zu einer Merkmalsklasse r Tupel bestehend aus (id, t) enthalten. Dabei ist id eine eindeutige Identifikationsnummer und $t \in [0, T]$ die (zeitliche) Position, an der das entsprechende Merkmal r auftritt. Alle Merkmalsdokumente bilden den Index:

$$\text{Index} := (L(r))_{r \in [1, R]} \quad (1)$$

Suchvorgang

Nach der Merkmalsextraktion liegt eine Suchanfrage der Form $Q = \{(q_1, t_1), (q_2, t_2), \dots, (q_n, t_n)\}$ mit $q_i \in [1, R]$ vor. Jedes q_i liefert eine Liste $L(q_i)$ aus dem Index. Diese Listen werden durch Subtraktion der Positionen t_i normiert und dann geschnitten (2). Durch die Normierung wird bei einem Treffer die richtige Konstellation der Merkmale garantiert.

Toleranz gegenüber Fehlstellen

Das beschriebene Vorgehen bei der Suche hat zur Folge, dass nur Anfragen einen Treffer erzielen, bei denen Merkmale und Positionen mit dem Index komplett

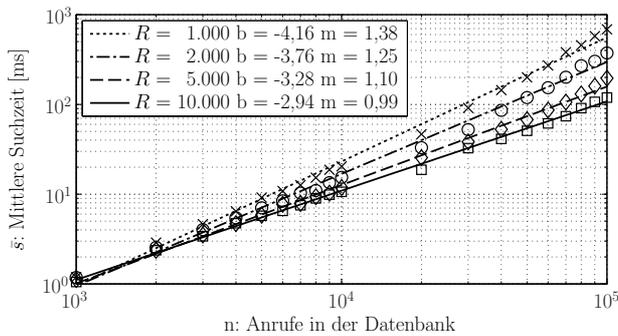


Abbildung 2: Messwerte der mittleren Suchzeit und deren Regressionsgeraden $g_R(n) = 10^b \cdot n^m$ bei unterschiedlicher Anzahl von Merkmalsklassen mit gleichverteilten Zufallsdaten.

übereinstimmen. Um eine festgelegte Anzahl an Fehlstellen k (Kredit) zuzulassen, also Positionen an denen nicht das aus dem Index geforderte Merkmal auftritt, müssen die Listen *fehlertolerant* geschnitten werden (k -Mismatch). D.h. beim Schnitt wird ein Listenelement, das nicht in beiden Listen vorkommt, erst entfernt, wenn es keinen Kredit $k \geq 0$ mehr hat:

$$L_{Ergebnis} = (L(q_1) - t_1) \cap_k \dots \cap_k (L(q_n) - t_n). \quad (2)$$

Der Vorteil dieses Vorgehens ist eine höhere Fehlertoleranz gegenüber verrauschten Audioquellen. Offensichtlich ist die Trefferbestimmung unabhängig gegenüber der Reihenfolge der Schnittmengenbildung. Das kann zur Effizienzsteigerung beitragen, wenn zunächst „kleine“ Listen geschnitten werden.

Fuzzy-Suche

Durch die Fensterung kann es vorkommen, dass ein Merkmal nicht an der gewünschten Position, sondern zeitlich um eine Position verschoben auftritt. Die (zeitliche) Fuzzy-Suche wirkt diesem Effekt entgegen, indem sie das geforderte Merkmal neben der aktuellen Position zulässt. Dazu werden die Listen zunächst vereinigt und danach erst mittels (2) geschnitten:

$$L_i^{fuzzy} = (L(q_i) - t_i) \cup (L(q_i) - (t_i \pm 1)). \quad (3)$$

Der Vorteil ist, dass ähnliche Fingerabdrücke besser getroffen werden. Der Nachteil, dass der Suchvorgang dadurch im Allgemeinen länger dauert.

Evaluierung

Zur Evaluierung des Verfahrens wurde in zwei Testszenarien zum einen die Suchgeschwindigkeit und zum anderen die Erkennungsrate untersucht.

Der Geschwindigkeitstest wurde mit gleichverteilten, zufällig erzeugten Daten durchgeführt. Zur Messung der Erkennungsrate wurde ein Testkorpus aus realen Sprachdaten verwendet. Die Telefon-Spam-Daten wurden durch künstliche Variation und Degradierung (z.B. durch verschiedene Codecs) erzeugt.

Insgesamt wurden Fingerabdrücke von 200 Spam- und 400 Nicht-Spam-Anrufen extrahiert. Tabelle 1 zeigt die

Tabelle 1: k -Mismatch-Suche in 600 Anrufen

Fehlstellen [%]	Suchzeit [ms]	Trefferrate [%]	Fehlerrate [%]
0	1,403	6,35	0,000
25	5,532	41,57	0,001
50	10,203	87,63	0,010
75	15,390	99,77	0,052

Ergebnisse der Suche in allen 600 Anrufen. Wie erwartet, steigt die Suchzeit mit der Anzahl der zugelassenen Fehlstellen, sowie auch die Anzahl falscher Positiver (Fehlerrate). Die Suchzeit liegt bei der Kombination aus k -Mismatch- und Fuzzy-Suche (Tabelle 2) generell höher. Jedoch wird ein vergleichbares Ergebnis schon mit weniger erlaubten Fehlstellen erreicht. Außerdem verbessert sich das Verhältnis von Treffer- zu Fehlerrate. Abbildung 2 zeigt, dass eine Erhöhung der Merkmalsklas-

Tabelle 2: k -Mismatch-Suche kombiniert mit Fuzzy-Suche in 600 Anrufen

Fehlstellen [%]	Suchzeit [ms]	Trefferrate [%]	Fehlerrate [%]
0	2,477	7,43	0,000
25	8,393	67,68	0,004
50	15,753	96,93	0,012
75	22,103	100,00	0,160

sen R prinzipiell die Suchzeit verkürzt. Allerdings führt die Erhöhung der Merkmalsklassen durch Kombination auch zu einer Erhöhung der auftretenden Fehlstellen.

Schlussfolgerungen und Ausblick

Das indexbasierte Verfahren ist zur Identifikation geeignet und vermindert den Suchaufwand gegenüber einer naiven, nicht-indexbasierten Suche erheblich.

Um die Suchzeit weiter zu verkürzen, sollte: i) die Anzahl der Merkmalsklassen erhöht werden, ohne die Robustheit des Fingerabdrucks zu verringern. ii) Der Einfluss der nach der Listenlänge sortierten Suchanfrage soll untersucht werden und iii) soll eine Lastverteilung auf mehrere Prozessoren bzw. Systeme erfolgen.

Um die praktische Umsetzbarkeit zu untersuchen, sind mehr Anrufe in der Datenbank, sowie ein größerer Testkorpus aus realen Sprachdaten notwendig. Auch eine prototypische Integration in eine reale Telefonumgebung ist geplant. Dabei wird zunächst nicht aktiv in den Telefonbetrieb eingegriffen, es werden lediglich Messdaten erhoben.

Literatur

- [1] Kurth, F.; Müller, M.: Efficient Index-Based Audio-Matching. IEEE Transactions on audio, speech, and language processing 16 (2008), 2, 382-395
- [2] Verfahren zur Identifikation und Abwehr von Telefon-SPAM, URL: <http://viat.fh-koeln.de/>