Vergleich spektraler Merkmale zur Identifikation von Telefon–SPAM

Gary Grutzek[1], Christoph Pörschmann[2], Heiko Knospe[3]

[1] *Fachhochschule Köln, Institut für Nachrichtentechnik, 50679 Köln, Deutschland, Email: gary.grutzek@fh-koeln.de*

[2] *Fachhochschule Köln, Institut für Nachrichtentechnik, 50679 Köln, Deutschland, Email: christoph.poerschmann@fh-koeln.de*

[3] *Fachhochschule Köln, Institut für Nachrichtentechnik, 50679 Köln, Deutschland, Email: heiko.knospe@fh-koeln.de*

Einleitung

Da die Kosten für Telefongespräche, besonders in IP-basierten Netzen stetig sinken, wird SPAM over IP-Telephony (SPIT) zunehmend interessanter. SPIT-Calls sind Anrufe, die automatisiert, mehrfach eingespielt werden. Das Projekt VIAT (Verfahren zur Identifikation und Abwehr von Telefon-SPAM, gefördert vom Bundesministerium für Bildung und Forschung (BMBF), Förderkennzeichen 1736X09) beschäftigt sich mit der Erkennung und Bekämpfung von Telefon-SPAM [1]. Ziel des Projektes VIAT ist es, diese Robocalls von realen Telefongesprächen zu unterscheiden, um SPIT frühzeitig und automatisiert zu blockieren. Zur SPIT-Erkennung wird aus den Geprächsdaten eines jeden eingehenden Anrufs ein so genannter Fingerabdruck erzeugt und in einer Datenbank des Telefon-Providers abgelegt. Sind zwei Fingerabdrücke so ähnlich, dass man davon ausgehen kann, dass es sich um dieselbe automatisierte Ansage handelt, wird die Caller-ID des Anrufers auf eine Blacklist gesetzt. Die im Folgenden erläuterten Untersuchungen dienen der Bestimmung eines für diese Anwendung geeigneten Fingerabdrucks.

Merkmalsextraktion

Spektrale Merkmale sind extrahierte Eigenschaften des Spektrums einer Audiodatei, welche über die Fourier-Transformation kurzer, überlappender und gefensterter Auschnitte des Audiosignals s berechnet werden (1). Die Fensterlänge N und die Schrittweite M müssen zur Anpassung an die jeweilige Anwendung empirisch ermittelt werden.

$$S(k,l) = \sum_{n=0}^{N-1} w(n)s(n+lM)e^{-j(2\pi/N)kn} \qquad (1)$$

Das Kurzzeitspektrums S mit dem Frequenzindex k wird mit Hilfe einer Filterbank in mehrere Subbänder geteilt. Für jedes Subband und zu jedem Zeitpunkt l wird ein Merkmalsvektor extrahiert. Die Folge aller Merkmalsvektoren einer Audiodatei bildet einen einfachen Fingerabdruck.
Im Zuge des Projekts VIAT wurden folgende Merkmale untersucht:

- Spectral Flatness Measure (SFM) [2]
- Spectral Crest Factor (SCF) [2]
- Mel frequency Cepstrum Coefficients (MFCC) [3]
- Spektraler Schwerpunkt, Schwankung, Wölbung und Schiefe

SFM und SCF wurden sowohl mit einer linearen Filterbank als auch mit einer Mel-Filterbank berechnet. Die Mel-Filterbank ist der nicht-linearen Frequenzwahrnehmung des Menschen nachempfunden [3]. MFCC werden durch Kosinus-Transformation des logarithmierten, melgefilterten Spektrums gewonnen. Durch Subtraktion des cepstralen Mittelwerts kann der Einfluss des Kanals minimiert werden [3]. Daher wurden auch Koeffizienten aus dem mittelwertfreien Cepstrum gebildet und verglichen.

Vergleich der Merkmale

Die Testumgebung

Der Testkorpus besteht aus 20 verschiedenen Audiosamples, davon 5 desselben Sprechers. Die Samples entstammen kurzen Ausschnitten verschiedener Podcasts, freien Hörbüchern und aufgezeichneten Teleshopping-Sendungen. Jedes Sample wurde auf vier verschiedene Arten manipuliert um eine mehr oder weniger realistische Degradierung durch Telefon-Übertragung zu simulieren. Die Manipulationen im Einzelnen sind:

- weißes Rauschen mit einem SNR von $20dB$,
- $150ms$ Echos mit $-10dB$,
- fullrate GSM-Codec,
- 10% Paketverlust.

Die Audiosamples wurden in 16 Bit, 8 kHz PCM Wav-Dateien konvertiert, mit einem 300 Hz bis 3.4 kHz Bandpass gefiltert, normalisiert und von Sprachpausen befreit. Aus den ersten sechs Sekunden wurden anschließend die Merkmale extrahiert. Dies entspricht einer realistischen Länge, da SPIT-Anrufe häufig frühzeitig vom Angerufenen abgebrochen werden.
Der durchgeführte Test dient dem qualitativen Vergleich einiger spektraler Merkmale und deren Parametrierung in Bezug auf die Unempfindlichkeit gegenüber Manipulationen der Audiosamples. Die Fenstergröße wurde von 256 Samples bis 32768 Samples, also ungefähr $30ms$ bis $4s$ bei einer Abtastrate von $8kHz$, variiert. Die Überlappung der Fenster betrug 25%, 50% oder 75%. Die Merkmale wurden aus 16 und 32 Subbändern extrahiert. Als Maß für die Ähnlichkeit wurde die *sum of minimum distances* [4] gewählt.

$$d_{md}(S_1,S_2) = \frac{1}{2}\left(\sum_{e\in S_1} d_{min}(e,S_2) + \sum_{e\in S_2} d_{min}(e,S_1)\right) \qquad (2)$$

d_{min} bezeichnet die minimale euklidische Distanz eines Merkmalsvektors e aus einem Merkmalsraums S zu einem Vektor des zu vergleichenden Merkmalsraums. Zur

Berechnung der Erkennungsrate, welche als Grundlage des Merkmalsvergleichs dient, wird für jedes Merkmal die geringste Distanz d_{md} zwischen allen Audiodateien als Schwellwert gesetzt, so dass die Falschakzeptanzrate FAR = 0 ist. Von der Falschrückweisungsrate (FRR), also den Dateien gleichen Ursprungs mit einer Distanz über dem Schwellwert, kann man die Erkennungsrate ableiten.

Testergebnisse

Die folgenden Diagramme zeigen die Erkennungsraten der Merkmale in Abhängigkeit von Fenstergröße und Anzahl der Subbänder. Eine Überlappung von 75% führte durchgehend zu den besten Ergebnissen. Aus Gründen der Übersichtlichkeit werden daher nur jene illustriert. Abb.1 zeigt die Erkennungsrate von SFM und SCF mit variierter Fenstergröße und 75% Überlappung, jeweils mit 16 und 32 Subbändern. Wird zur Merkmalsextraktion eine Mel-Filterbank anstatt einer linearen genutzt, ergibt sich für SFM eine bessere Erkennungsrate, für SCF verschlechtert sich allerdings das Ergebnis.

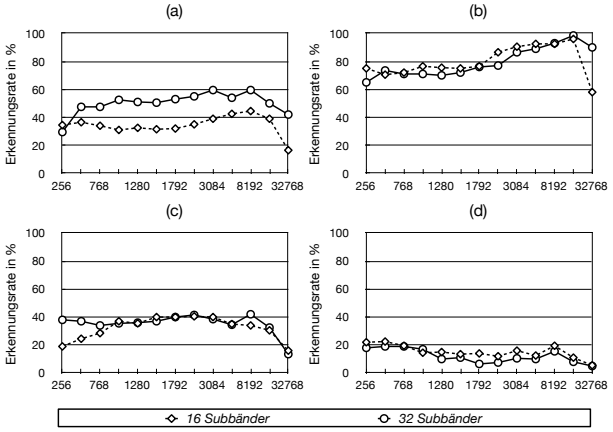

Abbildung 1: a) SFM mit linearer Filterbank, b) SFM mit Mel-Filterbank, c) SCF mit linearer Filterbank, d) SCF mit Mel-Filterbank

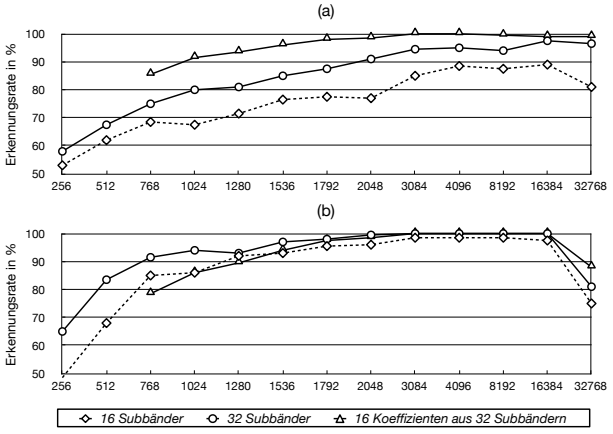

Abbildung 2: MFCC (a) und mittelwertfreie MFCC (b)

Bei MFCC führt die Subtraktion des cepstralen Mittelwerts zu einer wesentlich besseren Erkennung ähnlicher Samples. Zwar bewirken 32 Subbänder eine Verbesserung gegenüber 16 Subbändern, Abb.2 zeigt allerdings auch, dass die Erkennungsrate sogar besser ist, wenn zwar eine Mel-Filterbank mit 32 Subbändern genutzt wird, aber nur die ersten 16 Koeffizienten des Cepstrums für den Fingerabdruck verwendet werden.

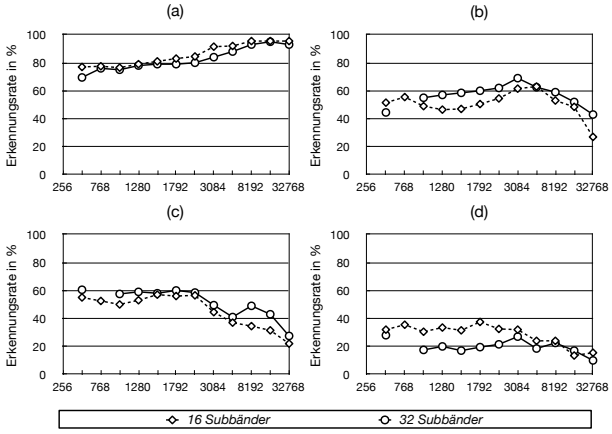

Abbildung 3: a) Schwerpunkt b) Schwankung c) Schiefe d) Wölbung

Von den statistischen Momenten hat sich der Schwerpunkt des Spektrums als robustes Merkmal erwiesen (Abb.3, fehlende Punkte sind auf ungültige Ergebnisse bei der Merkmalsextraktion zurückzuführen).

Schlussfolgerungen

In diesem Artikel wurde gezeigt, dass die optimalen Parameter zur Merkmalsextraktion hochgradig von der jeweiligen Anwendung abhängen. So unterscheiden sich die als optimal erachteten Parameter stark von denen der Spracherkennung. Die meisten Merkmalsvektoren zeigen die beste Erkennungsrate bei Fenstergrößen von 1-2 Sekunden. Ein gut geeigneter Fingerabdruck ließe sich aus mittelwertfreien MFCC mit einer Fenstergröße von 16384 Samples und einer Überlappung von 75% bilden. Setzt man eine Auflösung von 8 Bit und 16-dimensionale Vektoren an, ergibt sich damit eine Größe des Fingerabdrucks von nur $9 \cdot 16 \cdot 8\ Bit = 1152\ Bit = 144\ Byte$.

Literatur

[1] Pörschmann, C. & Knospe, H.: Spectral Analysis of Audio Signals for the Identification of Spam over IP Telephony. NAG/DAGA 2009, 1027-1029

[2] Herre, J. & Allamanche, E. & Hellmuth, O.: Robust matching of audio signals using spectral flatness features, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001, 127-130

[3] Pfister, B. & Kaufmann, T.: Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung, Springer, 2008

[4] Eiter, T. Mannila, H.: Distance measures for point sets and their computation, Acta Informatica Vol. 34, 1997, 109-133