

Hauptseminar – Biometrische Passwortsysteme

Sprechererkennung

24.01.2001 Nikolaus Kimmel

1. Einleitung

Zunehmend werden immer mehr biometrische Systeme zur Identifikation bzw. Verifikation eingesetzt. Eines davon ist die sogenannte Sprechererkennung. Dieses Merkmal eines Menschen ist dazu gut geeignet, da die Sprache die natürlichste Form der Kommunikation ist und bei den Anwendern auf eine hohe Akzeptanz stößt. Außerdem ist es durch Fortschritte in der digitalen Signalprozessortechnologie und Sprachforschung nun möglich ein schnelles, kostengünstiges Sprachverifikationssystem zu erstellen. Die Identifikation, bei der ein Sprecher einem Sprechermodell aus einer bekannten Menge zugeordnet wird, wird als „closed set“ Situation bezeichnet. Die Verifikation, auf die wir uns im folgenden beschränken, bei der entschieden werden muss, ob der Sprecher der ist, der er zu sein vorgibt, wird als „open set“ Situation bezeichnet. Bei der Sprechererkennung gibt es noch mal eine Entscheidung, die zu treffen ist, nämlich ob man ein textabhängiges oder –unabhängiges System wünscht. Textabhängige Systeme haben den Vorteil, dass zum trainieren und erkennen nur etwa 2-3 Sekunden Sprache nötig sind, während bei den unabhängigen ca. 20-30 Sekunden zum Training und ca. 10 Sekunden zum erkennen nötig sind.

2. Überblick über die Verfahren

Die Verfahren können grob in drei Gruppen unterteilt werden, Template Matching Verfahren, generative Modelle und neuronale Netze. Des weiteren gibt es auch Kombinationen von verschiedenen Ansätzen. Bei den Templat Matching Verfahren gibt es die Vektorquantisierung (VQ), die lernende Vektorquantisierung (LVQ) und das Dynamic Time Warping Verfahren (DTW). Beim DTW findet ein dynamischer Vergleich zwischen einem Test- und einem Referenzmodell statt, bei dem der Abstand zwischen den Modellen berechnet wird. Der Vergleich findet aber nicht beschränkt von einem Frame des Testmodells auf einen des Referenzmodells statt, sondern kann mehrere Frames umfassen, da gesprochene Worte in der Dauer variieren können. Zu den generativen Modellen zählen die Gaussian Mixture Models (GMM) und die Hidden Markov Models (HMM). Die HMM's repräsentieren das Spracherzeugungsmodell als einen Zustandsautomaten, der sich in endlich vielen Zuständen befinden kann. Die Zustandsübergänge sind mit Wahrscheinlichkeitsverteilungen belegt. Das Modell versucht nun die zeitliche Struktur von Sprache zu mehren Kurzzeitbeobachtungen zu verknüpfen. Die Zustände des Automaten können z.B. aus Worten oder einzelnen Phonemen bestehen. Weit verbreitet ist z.B. eine links-rechts Topologie der HMM's, die sich wie folgt darstellt:

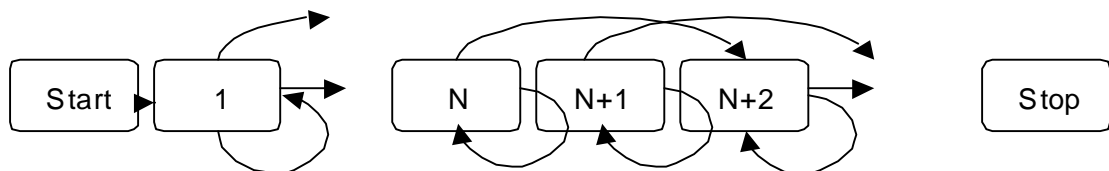


Abbildung 1: Left to right HMM structure

Der nächste Ansatz ist der über neuronale Netze, den sogenannten Self organizing maps, kurz SOM. SOM's betrachten als Sprechermodell ein Kodebuch und bauen beim unüberwachten lernen ein topologisch geordnetes Kodebuch auf. Die Dichte der Kodevektoren nähert sich der Wahrscheinlichkeitsdichte der Eingabevektoren während des Trainings an. Es wird also eine nichtlineare Projektion des Feature-raums auf das neuronale Netz vorgenommen.

3. Sprachdatenbanken

Um die Sprechererkennungsverfahren zu testen, sind derzeit vier Datenbanken öffentlich verfügbar. Als erstes gibt es die TIMIT Datenbank. Diese entstand als eine Zusammenarbeit von Texas Instruments und dem MIT. Sie bietet Sprachsamples in idealer Qualität an. Von insgesamt 630 Sprechern wurden jeweils 10 phonetisch reiche Sätze gesprochen und aufgezeichnet. Dieselbe Datenbank ist auch unter dem Namen NTIMIT verfügbar, wobei der Unterschied hier ist, dass alle Sprachdaten durch ein Telefonnetzwerk gesendet wurden und dann nochmals aufgezeichnet wurden. Durch das bei den Aufnahmen hinzugekommene Rauschen ist die Erkennung hier schon schwieriger, aber es liegen immer noch recht gute Bedingungen vor, da immer dieselbe Telefonanlage und Leitungslänge verwendet wurde. Die näher an der Realität liegende Variante ist die Switchboard Datenbank. Sie besteht aus 129 männlichen und 125 weiblichen Stimmen, von denen 1309 Telefongespräche mit einer Dauer von ungefähr 5-6 Minuten aufgezeichnet wurden. Die vierte oft verwendete Datenbank ist die YOHO Datenbank. Sie beinhaltet Codewörter der Art 24 72 54 von denen 156 Männern und 30 Frauen jeweils 96 Codewörter zum trainieren und 40 zum testen bereit gestellt haben. Die Aufzeichnungen entstanden in einem Zeitraum von 3 Monate unter realen Bürobedingungen:

4. Vorverarbeitung

Da alle Verfahren die Sprache prinzipiell gleich vorverarbeiten, hier nun die Vorverarbeitung allgemein. Zuerst wird die Sprache segmentiert in Frames, die um die 10 ms lang sind. Die gewonnenen Frames werden dann zu einem Speech Activity Detector (SAD) weitergegeben. Dieser dient dazu, Ruhe- und Lärmframes zu entfernen. Dies ist wichtig, damit der Sprecher und nicht die Umgebung detektiert wird. Nun werden Cepstralfeaturevektoren erzeugt, dies kann man mit Jialong He's spkrtool machen, das man kostenlos herunterladen kann (<http://www.speech.cs.cmu.edu/comp.speech/Section6/Verification/jialong.html>). Anschließend folgt eine Normalisierung, bei der der durchschnittliche Cepstralvektor von dem Featurevektor abgezogen wird. Nötig ist die Normalisierung vor allem, um gute Ergebnisse zu erzielen, wenn die Aufnahmen der Sprache von verschiedenen Mikrofonen stammen.

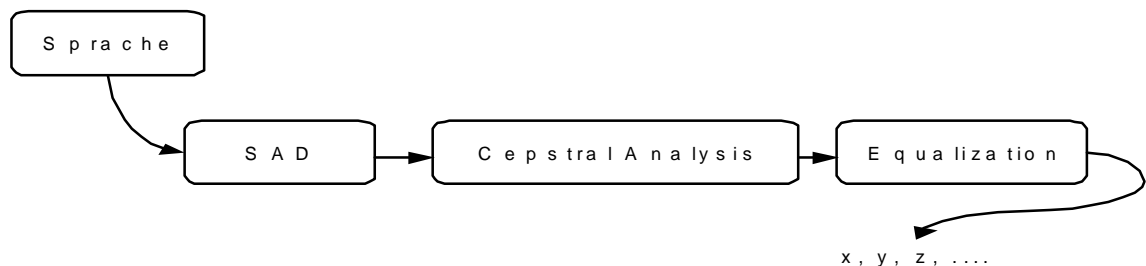


Abbildung 2: Vorverarbeitungskette

5. VQ und GMM im Detail

5.1 VQ

Vektorquantisierung ist eine schnelle Methode, die zur Sprecherverifikation verwendet wird. Die Methode benützt Abhängigkeiten der Featurevektoren, um den Merkmalsraum in Zellen einzuteilen. Dies geschieht unüberwacht und folgt dem Linde-Buzo-Gray Algorithmus (LBG). Das hier zu Grunde liegende Sprachmodell ist ein Kodebuch, das aus Kodevektoren besteht, von denen jeder jeweils aus mehreren ähnlichen Einzelvektoren (z.B. Silben) berechnet wird. Ziel ist es nun, den Sprecher dadurch zu erkennen, dass man beim Einsatz herausfindet, wie die gegebenen Sprache geclustert ist. Bei der erweiterten Form, dem sogenannte LVQ hat man nun vordefinierte Klassen und gelabelte Testdaten vorhanden. Die Klassengrenzen werden nach der Nearest Neighbour Regel definiert. Es handelt sich hier um eine überwachte Version des VQ, bei der versucht wird das Set von Prototypen zu finden, die eine Klasse am besten repräsentieren.

5.2 GMM

Hier wird eine Mischung von Gaussdichten verwendet um die Verteilung der Featurevektoren jedes Sprechers zu modellieren. Das Problem ist, dass das Training recht aufwendig ist. Außerdem benötigt man viele Testdaten, da man sonst Gefahr läuft, dass das Verfahren numerisch instabil wird, da die beim Verfahren benutzten Kovarianzmatrizen dann nicht immer invertierbar sein müssen. Für einen D-Dimensionalen Featurevektor x ist nun die Mixturdichte für den Sprecher S :

$$p(x | \lambda_s) = \sum_{i=1}^M p_i^s b_i^s(x)$$

Die Dichte ist gerichtete Linearkombination von M Komponenten uni-modaler Gaussdichten, $b_i^s(x)$, jeder parametrisiert von einem Mittelvektor, M_i^s , und der Kovarianzmatrix Σ_i^s . Hierbei ist

$$b_i^s(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i^s|^{1/2}} \times \exp\left\{-\frac{1}{2}(x - M_i^s)' (\Sigma_i^s)^{-1} (x - M_i^s)\right\}$$

Die Mixturgewichte p_i^s erfüllen die Bedingung: $\sum_{i=1}^m p_i^s = 1$

Die größten Wahrscheinlichkeiten der Sprachmodellparameter werden mit dem iterativen Expectation-Maximization (EM) Algorithmus geschätzt.

Bei der Verifikation nun die Wahrscheinlichkeit eines Eingabeworts berechnet um zu entscheiden, ob der Sprecher akzeptiert oder abgelehnt wird. Nach anwenden der Regel von Bayes und dem Weglassen der konstanten priori Wahrscheinlichkeiten erhält man folgende Formel:

$$\Lambda(X) = \log p(X | \lambda_c) - \log p(X | \lambda_{\bar{c}})$$

Wobei X das Wort ist, von dem nun entschieden wird, ob es zum Sprachmodell λ_c gehört.

5.3 EVQ

Bei der Vektorquantisierung wird zum Trainieren nun eine deterministische Version des LBG Algorithmus mit euklid'scher Abstandmessung eingesetzt. Das Splitten der Cluster wird nun nicht durch zufällige Variation der Mittelvektoren gemacht, sondern die Variationen sind +/- 20% der Standardabweichung der betrachteten Komponente. Der sogenannte „utterance score“, anhand dem entschieden wird zu welchem Cluster der Featurevektor gehört, berechnet sich wie folgt:

$$D_{utt}(\vec{X} | \lambda) = \frac{1}{N} \sum_{i=1}^N d_{frame}(\vec{x}_i | \lambda)$$

Er ist das arithmetische Mittel der N euklidischen Abstände $d_{frame}(\vec{x}_i | \lambda)$ zwischen den

Featurevektoren x_i und dem nächsten Zentrum des bestimmten Sprachmodells λ bei Frame i .

Beim GMM System wird beim EM Algorithmus eine bessere Stabilität durch ein niedriges Limit für das Absolute jeder Komponente der Kovarianzmatrizen erreicht. Die Initialisierung der Mittelvektoren werden durch den VQ Algorithmus vorgenommen. Die Wahrscheinlichkeit für ein GMM M 'ter Ordnung für einen Frame mit einem D -dimensionalen Featurevektor x_i beträgt dann:

$$p_{frame}(x_i | \lambda) = \sum_{j=1}^M w_j b_j$$

mit $b_j = \frac{1}{(2\pi)^{D/2} |\sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}_i - \vec{\mu}_j)^T \sigma_j^{-1} (\vec{x}_i - \vec{\mu}_j)\right\}$

den Gewichten w_j , den Mittelvektoren μ_j und der Kovarianzmatrix σ_j . Der „utterance score“ im GMM Fall für Modell λ ist:

$$P_{utt}(\vec{X} | \lambda) = \frac{1}{N} \sum_{i=1}^N p_{frame}(\vec{x}_i | \lambda)$$

Das VQ verfahren wurde zum Training verwendet und das GMM dann zur Auswertung. Dies ist möglich, da beide Ansätze ähnlich sind, sie stellen beide eine Verteilung der Datenvektoren im Merkmalsraum dar, nur auf unterschiedliche Weise. Das Problem auf das man trifft ist, dass ein GMM Modell mit der gleichen Anzahl von Zentren wie ein VQ Modell mehr Parameter benötigt. Zusätzlich zu den Mittelvektoren brauchen GMM's noch die zum Zentrum korrespondierende Kovarianzmatrix σ_j und Gewichte w_j . Daher ist ein zusätzlicher Schritt beim VQ Verfahren nötig. Nämlich wird für jeden Featurevektor im Trainingsset nach dem Zentrum mit der geringsten Distanz gesucht und der Vektor mit dem Zentrum assoziiert. Mit diesen Vektoren wird dann für jedes Zentrum die Kovarianzmatrix ausgerechnet und die relative Frequenz der assoziierten Vektoren, welche dann als Schätzung des Gewichtes eines Zentrums verwendet wird. Mit den so gewonnenen fehlenden Parametern, wird dann das GMM benützt. Das so trainierte System wurde nun mit der YOHO Datenbank getestet. Hierzu wurden 73 Sprecher für das Training verwendet und 33 um das Hintergrundmodell zu trainieren. Die Tests wurden mit 2 verschiedenen Thresholds durchgeführt. Zum einen mit einem lokalen Threshold, d.h. jeder Sprecher hat seinen eigenen Threshold mit $FR = FA$. Die andere Variante war ein globaler Threshold, was bedeutet, dass alle Sprecher den gleichen Threshold haben, aber $FR > FA$, wobei die Summe aller FA's und FR's gleich ist.

Tabelle 1: Equal error rates (EER) in percent for speaker verification (local thresholding)

method	Model order (No of centers)				
	4	8	16	32	64
VQ	11,4	9,6	9,0	6,8	5,4
GMM	4,3	3,1	2,3	-	-
EVQ	5,4	4,2	3,3	2,8	2,8

Tabelle 2: Equal error rates (EER) in percent for speaker verification (global thresholding)

method	Model order (No of centers)				
	4	8	16	32	64
VQ	12,9	11,1	9,6	8,1	6,9
GMM	5,3	4,1	3,3	-	-
EVQ	6,3	5,2	4,5	5,9	4,2

6. Schluss

Die Sprechererkennung funktioniert, je nach Anwendungsfall recht annehmbar, hat aber mit einigen Problemen zu kämpfen, die zum einen durch die Benutzer auftreten, wie z.B. schlechte Aussprache, Krankheit, Durst, dem emotionalen Befinden des Sprechers und der Alterung, da dies alles Faktoren sind, die die Stimme beeinflussen. Das Problem, das bei anderen biometrischen Systemen nicht in der Art gegeben sind, ist, dass sich Sprache z.B. durch eine Erkältung oder einen Streit sehr kurzfristig verändern kann und eine Erkennung dadurch beeinträchtigt werden kann. Im kommerziellen Bereich gibt es einige Firmen, die Sprachverifikationssysteme anbieten, ich will hier nur Voice Security Systems nennen, da hier auf der Firmenwebseite ein Applet zum testen zur Verfügung steht (<http://www.voice-security.com>). Auch sind derzeit zahlreiche Systeme für den Hausgebrauch preiswert erhältlich, um seinen eigenen PC oder auch Firmenrechner, durch eine Sprechererkennung zu schützen. Als einziges, von mir gefundenes, System das in einem größeren Rahmen eingesetzt werden soll ist das Projekt der Manhattan Chase Bank. Hier soll eine Verifikation der Kunden durch Sprache erfolgen, da in Umfragen herausgefunden wurde, dass die Kunden mit 95% für eine Verifikation durch die Stimme sind.

7. Literaturverzeichnis

- Combination of vector quantization and Gaussian mixture models for speaker verification with sparse training data; Guido Kolano, Dr. Peter Regel-Brietzmann, Proc. EuroSpeech 1999, pp. 1203-1206
- Text-Dependent Speaker Identification Based on Input/output HMM's; Ke Chen, Dahong Xie, Huisheng Chi; Neural Processing Letters, Vol. 3; No 2, 1996, pp.81-89
- Speaker identification and verification using Gaussian mixture speaker models; Daouglas A. Reynolds; Speech Communication 17 (1995) 91-108
- Speaker Verification: A Tutorial; Jayant M. Naik; IEEE Communicatinos Magazine, January 1990, 42-48