TEXTABHÄNGIGE SPRECHERERKENNUNG

Peter Geßler
Matthias Wolff

BTU Cottbus-Senftenberg
peter.gessler@tu-cottbus.de

Kurzfassung: An der Professur für Kommunikationstechnik der BTU Cottbus-Senftenberg findet zur Zeit die Entwicklung eines automatischen Stimmenauthentifizierungssystems statt. Das System basiert auf einer textabhängigen Sprechererkennung, die zur Modellierung von berechtigten Personen *Gaussian Mixture Models* (GMM) verwendet. Im vorliegenden Beitrag werden neben dem Systemaufbau eines Prototypen zur Berechnung von Erkennungsquoten Untersuchungen vorgestellt, inwieweit sich für die Spracherkennung optimierte Merkmale ebenfalls für die textabhängige Sprechererkennung eignen. Zum Vergleich dienen weitere in der Sprechererkennung bekannte Merkmaltypen.

1 Einleitung

Das Sprachlabor der BTU Cottbus-Senftenberg dient zur Erforschung von Interaktionen des Menschen mit medialen Systemen sowie deren Entwicklung. Zur Zeit entwickeln wir für den Zugang zum Hauptprogramm ein Stimmenauthentifizierungssystem, welches eine Person mittels textabhängiger Sprechererkennung autorisiert oder abweist. Der vorhandene Spracherkenner verwendet eine auf einer Mel Filterbank, dynamischen Merkmalen und PCA basierenden Merkmalanalyse, die für Spracherkennung optimiert ist. In diesem Beitrag wird primär untersucht, ob dieser Merkmaltyp zur textabhängigen Sprechererkennung geeignet ist. Zum Vergleich dienten die primären UASR Merkmale *pfv_30* sowie *mfcc_30*, *lfcc_30*, *mfcc_30_ms* und *lfcc_30_ms*. Die Implementierung eines Prototypen zur Berechnung der Erkennungsquoten erfolgte in der Entwicklungsumgebung *MATLAB* und dem Experimentiersystem *Unified Approach to Speech Synthesis and Recognition* (UASR). Des Weiteren fand in einem zweiten Test die Untersuchung der Erkennungsquote auf ihre Leistungsfähigkeit bei einem vorhandenen zeitlichen Unterschied von sechs Monaten zwischen Trainings- und Arbeitsdaten statt. Parallel ist in beiden Untersuchungen die Anzahl M der Normalverteilungen innerhalb eines Sprechermodells (GMM) variiert worden.

2 Systemablauf

Die Echtzeitausführung des Systems war bei den vorgenommenen Untersuchungen nicht relevant. Dies ermöglichte eine unabhängige Durchführung von Sprachaufnahmen und Berechnung der Erkennungsquoten. Der sequenzielle Ablauf des implementierten Prototypen ist aus Abbildung 1 zu entnehmen. Während der Trainingsphase des Systems werden von einer berechtigten Person Sprachsignale mit der sprecherspezifischen Passphrase durch ein externes Aufnahmeprogramm aufgezeichnet und als Datei vom Typ *.wav* hinterlegt. In der Komponente *Merkmalanalyse* transformiert das System die aufgenommenen Sprachsignale zunächst in einzelne Merkmalvektorfolgen. Der Aufbau sowie die Anzahl der Merkmalvektoren in einer Merkmalvektorfolge sind vom verwendeten Merkmaltyp abhängig. Für die Abbildung einer berechtigten

Person verwendet das System eine multidimensionale Gaußsche Mischverteilung. Nach der Initialisierung der Gaußschen Mischverteilung erfolgt eine Anpassung der Parameter durch den *Expectation-Maximization* Algorithmus. Damit eine Klassifikation in der Arbeitsphase stattfinden kann, kennzeichnet das System die Gaußsche Mischverteilung mit dem Namen der Person. Anschließend wird das GMM in der Systemdatenbank hinterlegt. Neben den sprecherspezifischen GMM's existiert ein *Backoff-Modell* für die Ablehnung von nicht berechtigten Personen und ungültigen Passphrasen. Dieses ist ebenfalls eine Gaußsche Mischverteilung und berechnet sich aus den Merkmalvektorfolgen der Trainingsdaten aller berechtigten Personen.

In der Arbeitsphase sieht das System den Sprecher als unbekannt an. Das Sprachsignal der unbekannten Person, die berechtigt oder nicht berechtigt ist, liegt für die Untersuchungen als klassifizierte *.wav* Datei vor. Nach der Merkmalanalyse berechnet das System im *Klassifikationsschritt* den negativ logarithmischen Likelihood (NLL)-Wert der generierten Merkmalvektorfolge zu jeder hinterlegten Gaußschen Mischverteilung. Der Klassifikator entscheidet sich im Folgenden für die Klasse mit dem kleinsten NLL-Wert. Eine Aussage zur Korrektheit der Entscheidung erfolgt durch den Vergleich der gewählten Klasse und der Annotation der Merkmalvektorfolge. Dieser Ablauf wiederholt sich für alle zu überprüfenden Sprachsignale. Im letzten Schritt findet eine Auswertung zu den Aussagen des Klassifikators statt.

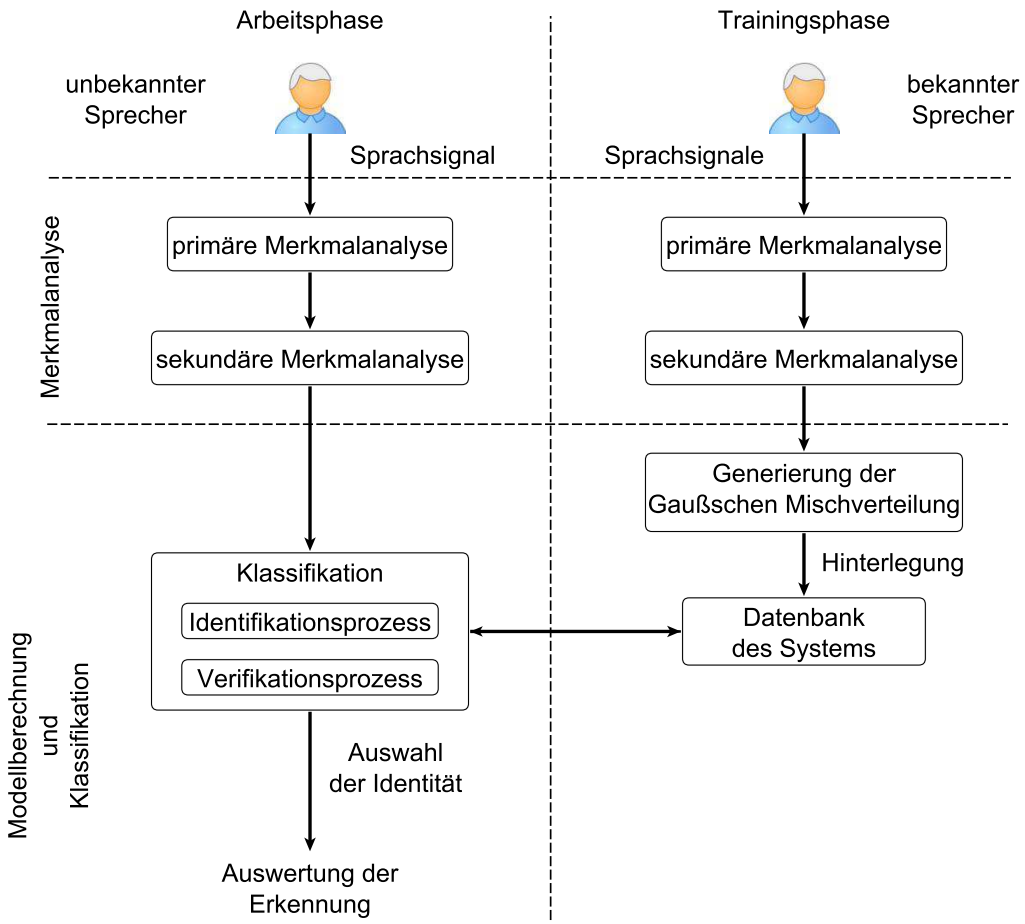

Abb. 1 Ablaufdiagramm des implementierten Prototypen

3 Merkmaltypen

Bevor eine Merkmalextraktion vom System vorgenommen wird, erfolgt eine Segmentierung, Fensterung und Höhenanhebung (Preemphasis) des Sprachsignals. In den Untersuchungen wählten wir bei einer Abtastfrequenz von 16 kHz ein Analyseintervall $x_\tau(k)$ von $K = 160$ Abtastwerten. Die Überlappung der einzelnen Segmente beträgt dabei 78,125%. Zur Minimierung des Leck-Effektes verwendet der Prototyp das Blackman-Fenster. Durch die Anwendung der Diskreten Fourier Transformation sowie der Bildung des Betragsspektrums ist der Ausgangspunkt aller zu generierenden Merkmaltypen vorhanden.

3.1 pfv_30 & mfcc_30

Primary Feature Vectors (pfv) und *Mel Frequency Cepstral Coefficients* (mfcc) werden durch die Bewertung des Betragsspektrums mit einer Mel-Filterbank gebildet. Die Filterbank besitzt in unserem Fall 30 überlappende Dreiecksbandpässe. Während bei pfv_30 vor der Bewertung mit den Dreiecksbandpässen eine Logarithmierung des Betragsspektrums und danach eine cepstrale Glättung erfolgt, führt das System die Logarithmierung bei mfcc_30 nach der Bewertung durch die Dreiecksbandpässe aus. Des Weiteren findet keine cepstrale Glättung statt, sondern eine Überführung in das Cepstrum mittels Diskreter Cosinus Transformation (DCT). Die geltenden mathematischen Beziehungen bezüglich Mel-Filterbank, cepstrale Glättung, Logarithmierung und DCT sind in [5], [2] und [1] nachzulesen.

3.2 sfv_24

Der Merkmaltyp *Secondary Feature Vectors* (sfv) leitet sich aus dem Merkmaltyp pfv_30 ab. Er wurde explizit für die Spracherkennung entwickelt und besitzt $D = 24$ Dimensionen. Die nachfolgenden Erläuterungen finden exemplarisch an einem Merkmalvektor $\vec{x}(k)$ einer Merkmalvektorfolge $\vec{\mathbf{x}}$ mit Folgenindex $k = 1,\ldots,K$ vom Typ pfv_30 statt. Zunächst führt das System aus numerischen Gründen eine Vektorstandardisierung aus. Anschließend berechnet es von den ungeraden Dimensionen des Merkmalvektors $\vec{x}(k)$ Delta- und Delta-Delta-Merkmale. Die Beschränkung auf ungerade Dimensionen hat ebenfalls numerische Gründe bezüglich der anschließenden *Hauptkomponentenanalyse*.

$$\vec{x}'(k) = \vec{x}(k+3) - \vec{x}(k-3) \tag{1}$$

$$\vec{x}''(k) = \vec{x}'(k+3) - \vec{x}'(k-3) = \vec{x}(k+6) + \vec{x}(k-6) - 2\vec{x}(k) \tag{2}$$

Die bei der Berechnung der dynamischen Merkmale auftretenden Zeitindizes $k \leq 0$ beziehungsweise $k > K$ werden durch zyklische Rotation der Merkmalvektorfolge ergänzt. Somit ergibt sich zunächst der folgende Supervektor

$$\vec{y}(k) = \begin{pmatrix} \vec{x}(k) \\ \vec{x}'(k) \\ \vec{x}''(k) \end{pmatrix} \qquad \text{mit} \qquad k: \text{ Folgenindex} \tag{3}$$

mit einer Anzahl von $D = 60$ Dimensionen.
Zur Reduzierung der Dimensionen verwendet das System eine Hauptkomponentenanalyse (engl. principal component analysis, PCA). Das UASR-System wählt dabei die 24 Dimensionen des Vektors $\vec{y}(k)$ mit der höchsten Streuungserklärung und generiert den neuen Merkmalvektor $\vec{o}(k)$.

3.3 lfcc_30

Die bisher vorgestellten Merkmaltypen dienten der Darstellung von Informationen, die sprecherunabhängig sind. In der Sprechererkennung ist es jedoch entscheidend, dass der Merkmaltyp die sprecherspezifischen Eigenschaften innerhalb der Stimme repräsentiert. Nach [4] reflektieren die höheren Formanten die sprecherspezifischen Teile eines Vokaltraktes. Zur Realisierung schlägt *Hertlein* die Verwendung einer linearen Filterbank anstatt einer Mel-Filterbank [1] vor. Die von uns verwendete lineare Filterbank der *HTK-Toolbox* besitzt ebenfalls wie die verwendete Mel-Filterbank 30 Dreiecksbandpässe. Nach der Bewertung erfolgt äquivalent zu den mfcc_30 eine Logarithmierung und Fourier-Rücktransformation mittels DCT. In der Literatur ist dieser Merkmaltyp als *Linear Frequency Cepstral Coefficients* bekannt.

3.4 mfcc_30_ms & lfcc_30_ms

Die Kennzeichnung *ms* steht in diesem Beitrag für *Mittelwertsubtraktion*. Diese ermöglicht nach [1] eine Kompensation der eventuell vorhandenen Kanalstörungen unter der Annahme, dass sich die Eigenschaften des Übertragungskanals im zeitlichen Verlauf nicht ändern. Mit der Anwendung des Merkmalextraktionsverfahrens mfcc oder lfcc erhalten wir Merkmalvektoren $\hat{\vec{o}}(k)$, die sich aus dem Quellsignal $\vec{c}(k)$ und dem Einfluss des Übertragungskanals $\tilde{G}$ additiv zusammensetzen.

$$\hat{\vec{o}}(k) = DCT\left\{10\log\left(\tilde{G}\right)\right\} + \vec{c}(k) \qquad \text{wobei } \tilde{G} = G(n) \approx const. \text{ gilt} \tag{4}$$

Die Eliminierung der hinzugekommenen Eigenschaften des Übertragungskanals erfolgt mit

$$\vec{o}_{Ms}(k) = \hat{\vec{o}}(k) - \vec{\mathbf{c}}_{Mittelwert} \qquad \text{mit } \vec{\mathbf{c}}_{Mittelwert} = \frac{1}{K}\sum_{k=1}^{K}\hat{\vec{o}}(k). \tag{5}$$

Mildner verweist jedoch darauf, dass durch die Verwendung der cepstralen Mittelwertsubtraktion neben dem Kanal ebenfalls Anteile des Quellsignals kompensiert werden könnten.

4 Modellbildung & Klassifikation

4.1 Gaussian Mixture Models & EM-Algorithmus

Unter der Annahme, dass die zeitliche Struktur der Merkmalvektorfolge (abgesehen von den dynamischen Merkmalen) nicht entscheidend zur Klassifikation beiträgt, reicht die Betrachtung der Gesamtverteilungsdichte aller Merkmalvektoren einer Person aus [3]. Die entsprechende Verteilungsdichtefunktion ergibt sich durch die additive Zusammensetzung einzelner Gaußverteilungen. Eine Gaußsche Mischverteilung q, welche die einzelnen Gaußverteilungen m repräsentiert, setzt sich aus den Parametern λ_m Mischungsgewichten, $\vec{\mu}_m$ Mittelwerten und $\mathbf{C}_{\vec{o}\vec{o},m}$ Kovarianzmatrizen zusammen, $q = \{\lambda_m, \vec{\mu}_m, \mathbf{C}_{\vec{o}\vec{o},m}\}$. Es gilt zu beachten, dass es sich um eine D-dimensionale Gaußsche Mischverteilung handelt und das System volle Kovarianzmatrizen für die Berechnungen verwendet. Zur Abbildung eines Sprechers c mittels Gaußsche Mischverteilung q_c fordern wir im Folgenden eine optimale Anpassung der Modellparameter q an die gegebenen Trainingsdaten der Person. Wegen der nicht geschlossenen Lösbarkeit dieses Optimierungsproblems verwendet das System zur Anpassung der Parameter den Expectation-Maximization Algorithmus nach Dempster. Innerhalb des implementierten Prototypen ist die Ausführung vom Algorithmus auf $N = 10$ Iterationsschritte eingeschränkt.

4.2 Klassifikation & Bewertung

In dem verwendeten Klassifikator kommen Wahrscheinlichkeitsdichtefunktionen der Form $p(\vec{\mathbf{o}}|q_c)$ für die Klassifizierung zum Einsatz. Diese gibt an, wie hoch die Wahrscheinlichkeit ist, dass eine Merkmalvektorfolge $\vec{\mathbf{o}}$ einem gegebenen Sprechermodell q_c zugeordnet wird [5]. Der Klassifikator entscheidet sich für die Person c, deren Sprechermodell q_c die höchste Wahrscheinlichkeitsdichte im Vergleich zu den anderen Sprechermodellen aufweist.

$$s = \underset{q_c=0\leq c\leq C}{\arg\max}\ p(\vec{\mathbf{o}}|q_c) \qquad \text{mit} \quad \begin{array}{l} q_0 := \text{Backoff-Modell} \\ q_c := \text{spezifisches Sprechermodell für } c \geq 1,\ c \in \mathbb{N} \end{array} \tag{6}$$

Entscheidet sich der Klassifikator in der Arbeitsphase für $s = 0$ (Wahl des Backoff-Modells), erfolgt eine Ablehnung der Person. Wählt der Klassifikator $s \neq 0$, folgt eine Autorisierung der Person. Nach jeder vorgenommenen Erkennung überprüft der implementierte Prototyp, ob die Kennzeichnung der Merkmalvektorfolge und der ausgewählten Klasse übereinstimmt. Ist dies der Fall, wertet das System die Erkennung als korrekt $v = 1$. Im Gegensatz dazu wählt das System $v = 0$ bei nicht korrekter Übereinstimmung. Mit der Verwendung von negativ logarithmischen Likelihoods sowie der Annahme, dass die Merkmalvektoren der Merkmalvektorfolge statistisch unabhängig voneinander sind, folgt für die Auswahl einer Person

$$s = \underset{q_c=0\leq c\leq C}{\arg\min}\ -\log p(\vec{\mathbf{o}}|q_c) = \underset{q_c=0\leq c\leq C}{\arg\min}\ -\sum_{n=1}^{N} \log p(\vec{o}_n|q_c). \tag{7}$$

5 Datenbasis

5.1 Korpus & Struktur

Im Folgenden wird ausschließlich der Korpus für die durchgeführten Untersuchungen vorgestellt, an denen 29 Personen teilgenommen haben. Die Einteilung der Probanden erfolgte zufällig in die Gruppen *berechtigte Personen* und *nicht berechtigte Personen*.

→ Geschlecht ↓ Gruppe	männlich	weiblich
berechtigte Personen	7	4
nicht berechtigte Personen	11	7

Tabelle 1: Verteilung der Probanden auf die Gruppen

Die Sprachaufnahmen mit den Probanden fanden an drei unterschiedlichen Sitzungsterminen statt. Der Abstand von Sitzung 1 zu 2 beträgt einen Monat und von Sitzung 2 zu 3 sieben Monate. Mitglieder der Gruppe *nicht berechtigte Personen* nahmen jedoch nur an Sitzung 1 und 3 teil. Jede berechtigte Person erhielt vor der ersten Sitzung eine Passphrase mit der Struktur: *Autorisierung - Nachname* der Person - *erstes Wort - zweites Wort*. Das erste und zweite Wort am Ende der Passphrase sind dabei aus dem deutschen Funkalphabet *DIN5009* entnommen. Ein Proband der Gruppe *nicht berechtigte Person* sprach für die Untersuchungen jeweils fünfmal die Passphrase jeder berechtigten Person. Von einem Probanden der Gruppe *berechtigte Personen* nahmen wir in jedem Sitzungstermin 30 mal seine eigene Passphrase und 20 Sprachäußerungen ohne gültige Passphrase auf. Der verwendete Korpus besitzt somit eine Gesamtgröße von 3630 .wav-Dateien.

→ Passphrasen ↓ Person	korrekte	ungültige	Gesamtanzahl pro Gruppe
berechtigte	90	60	1650
nicht berechtigte	110	-	1980

Tabelle 2: Verteilung der gesprochenen Sprachproben aus Sitzung 1, 2 und 3

5.2 Sprachaufnahmen

Alle Aufnahmen erfolgten an einem Arbeitsplatz im Sprachlabor der BTU-Cottbus-Senftenberg. Zur Aufnahme verwendeten wir ein *Rode NT-1 A* Mikrofon mit vorangestelltem Popup-Filter. Die Digitalisierung der Aufnahmen erfolgte durch eine externe Soundkarte (*Focusrite Scarlett 8i6*). Mit Hilfe der Software *Scarlett Mixcontrol* ist der Eingangspegel auf einen Headroom von maximal -3 dbFs festgelegt worden. Damit die Probanden den minimalen Aussteuerungspegel von -9 dbFs Headroom erreichen, legten wir ein Abstand von 5 bis 15 cm zwischen Kopf und Mikrofon fest. Zur Umsetzung eines annähernd realen Szenarios baten wir die Probanden, alle getätigten Sprachäußerungen mit einer Variation in der Artikulation zu versehen.

6 Erkennungsergebnisse & Auswertung

Eine Aussage zur Leistungsfähigkeit des Systems, unter Verwendung der Erkennungsquote, ist vom gewünschten Sicherheitsgrad abhängig. In unserem Fall gilt das System als leistungsfähig bei einer Erkennungsquote von über 90,0% in Untersuchung 1. Eine Überprüfung der Erkennungsquote des Systems, bei einem zeitlichen Abstand von über sechs Monaten zwischen den Aufnahmen der Trainings- und Testdaten erfolgt in Untersuchung 2. Die Abweichung der Erkennungsquoten darf dort nicht über 10,0% liegen. Die Anforderungen an Tests und Berechnungen zu statistischen Auswertungen von Erkennungsquoten sind [5] und [6] zu entnehmen. Zum Training eines sprecherspezifischen Modells q_c verwendeten wir 10 generierte Merkmalvektorfolgen von einer berechtigten Person aus Sitzung 1.

6.1 Erkennungsergebnisse

Untersuchung 1

→ M ↓	1		2		4		8	
pfv_30	$75{,}9^{+2,6}_{-2,7}$	$21{,}3^{+2,6}_{-2,5}$	$91{,}1^{+1,7}_{-1,9}$	$21{,}3^{+2,6}_{-2,5}$	$92{,}0^{+1,6}_{-1,8}$	$21{,}3^{+2,6}_{-2,5}$	$94{,}1^{+1,4}_{-1,6}$	$21{,}3^{+2,6}_{-2,5}$
sfv_24	$85{,}1^{+2,1}_{-2,3}$	$21{,}3^{+2,6}_{-2,5}$	$91{,}8^{+1,6}_{-1,8}$	$21{,}3^{+2,6}_{-2,5}$	$93{,}5^{+1,4}_{-1,7}$	$21{,}3^{+2,6}_{-2,5}$	$\mathbf{95{,}0^{+1,3}_{-1,5}}$	$21{,}3^{+2,6}_{-2,5}$
mfcc_30	$80{,}9^{+2,4}_{-2,5}$	$21{,}3^{+2,6}_{-2,5}$	$90{,}4^{+1,7}_{-2,0}$	$21{,}3^{+2,6}_{-2,5}$	$92{,}4^{+1,5}_{-1,8}$	$21{,}3^{+2,6}_{-2,5}$	$94{,}0^{+1,4}_{-1,6}$	$21{,}3^{+2,6}_{-2,5}$
lfcc_30	$85{,}4^{+2,1}_{-2,3}$	$21{,}3^{+2,6}_{-2,5}$	$90{,}9^{+1,7}_{-1,9}$	$21{,}3^{+2,6}_{-2,5}$	$93{,}2^{+1,5}_{-1,7}$	$21{,}3^{+2,6}_{-2,5}$	$94{,}7^{+1,3}_{-1,5}$	$21{,}3^{+2,6}_{-2,5}$
mfcc_30_ms	$85{,}0^{+2,1}_{-2,3}$	$21{,}3^{+2,6}_{-2,5}$	$91{,}2^{+1,7}_{-1,9}$	$21{,}3^{+2,6}_{-2,5}$	$92{,}8^{+1,5}_{-1,7}$	$21{,}3^{+2,6}_{-2,5}$	$94{,}7^{+1,3}_{-1,5}$	$21{,}3^{+2,6}_{-2,5}$
lfcc_30_ms	$87{,}8^{+1,9}_{-2,2}$	$21{,}3^{+2,6}_{-2,5}$	$92{,}1^{+1,6}_{-1,8}$	$21{,}3^{+2,6}_{-2,5}$	$94{,}3^{+1,3}_{-1,6}$	$21{,}3^{+2,6}_{-2,5}$	$\mathbf{95{,}0^{+1,3}_{-1,5}}$	$21{,}3^{+2,6}_{-2,5}$

Tabelle 3: Erkennungsquoten in Prozent der untersuchten Merkmaltypen.
Linke Spalte: mit Rückweisung - rechte Spalte: ohne Rückweisung.

Die Testdaten von Untersuchung 1 setzen sich aus 220 Sprachaufnahmen von berechtigten Personen mit korrekter Passphrase, 594 mit korrekter Passphrase von nicht berechtigten Personen und 220 von berechtigten Personen mit ungültiger Passphrase aus Sitzung 1 zusammen.

Untersuchung 2

Die in Tabelle 4 dargestellten Erkennungsquoten basieren auf getesteten Sprachäußerungen (20 pro Sprecher) mit korrekter Passphrase von berechtigten Personen, die von Sitzung 1 stammen und dienen als Referenz. Tabelle 5 zeigt dagegen die Erkennungsquote bei der Verwendung von Testdaten aus Sitzung 3. Die Auswahlstruktur ist dabei identisch. Der signifikante Abstand aller Erkennungsquoten beim Vergleich von Tabelle 4 und 5 veranlasste uns zu einer weiteren Berechnung von Erkennungsquoten (Tabelle 6), in der die Trainingsdaten aus Sitzung 2 stammen und die Testdaten aus Sitzung 3.

$\rightarrow M$ $\downarrow$	1		2		4		8	
pfv_30	$95{,}5^{+2{,}3}_{-3{,}7}$	$100^{+0{,}0}_{-1{,}7}$	$98{,}2^{+1{,}3}_{-2{,}8}$	$100^{+0{,}0}_{-1{,}7}$	$98{,}2^{+1{,}3}_{-2{,}8}$	$100^{+0{,}0}_{-1{,}7}$	$93{,}2^{+3{,}0}_{-4{,}2}$	$100^{+0{,}0}_{-1{,}7}$
sfv_24	$99{,}5^{+0{,}4}_{-2{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$99{,}5^{+0{,}4}_{-2{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$98{,}6^{+1{,}1}_{-2{,}6}$	$100^{+0{,}0}_{-1{,}7}$	$97{,}3^{+1{,}7}_{-3{,}1}$	$100^{+0{,}0}_{-1{,}7}$
mfcc_30	$95{,}5^{+2{,}3}_{-3{,}7}$	$100^{+0{,}0}_{-1{,}7}$	$97{,}3^{+1{,}7}_{-3{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$94{,}5^{+2{,}6}_{-3{,}9}$	$100^{+0{,}0}_{-1{,}7}$	$90{,}0^{+3{,}6}_{-4{,}7}$	$100^{+0{,}0}_{-1{,}7}$
lfcc_30	$93{,}2^{+3{,}0}_{-4{,}2}$	$100^{+0{,}0}_{-1{,}7}$	$93{,}6^{+2{,}8}_{-4{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$93{,}6^{+2{,}8}_{-4{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$89{,}5^{+3{,}7}_{-4{,}8}$	$100^{+0{,}0}_{-1{,}7}$
mfcc_30_ms	$99{,}5^{+0{,}4}_{-2{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$\mathbf{100}^{\mathbf{+0{,}0}}_{\mathbf{-1{,}7}}$	$100^{+0{,}0}_{-1{,}7}$	$98{,}2^{+1{,}3}_{-2{,}8}$	$100^{+0{,}0}_{-1{,}7}$	$94{,}5^{+2{,}6}_{-3{,}9}$	$100^{+0{,}0}_{-1{,}7}$
lfcc_30_ms	$99{,}5^{+0{,}4}_{-2{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$99{,}5^{+0{,}4}_{-2{,}1}$	$100^{+0{,}0}_{-1{,}7}$	$99{,}1^{+0{,}8}_{-2{,}3}$	$100^{+0{,}0}_{-1{,}7}$	$96{,}4^{+2{,}1}_{-3{,}4}$	$100^{+0{,}0}_{-1{,}7}$

Tabelle 4: Erkennungsquoten in Prozent mit Trainings- und Testdaten aus Sitzung 1. Linke Spalte: mit Rückweisung - rechte Spalte: ohne Rückweisung.

$\rightarrow M$ $\downarrow$	1		2		4		8	
pfv_30	$28{,}2^{+6{,}4}_{-5{,}8}$	$72{,}3^{+5{,}8}_{-6{,}4}$	$28{,}2^{+6{,}4}_{-5{,}8}$	$76{,}8^{+5{,}4}_{-6{,}1}$	$4{,}1^{+3{,}5}_{-2{,}2}$	$75{,}0^{+5{,}6}_{-6{,}3}$	$0{,}5^{+2{,}1}_{-0{,}4}$	$70{,}9^{+5{,}9}_{-6{,}5}$
sfv_24	$26{,}4^{+6{,}3}_{-5{,}7}$	$75{,}0^{+5{,}6}_{-6{,}3}$	$27{,}7^{+6{,}4}_{-5{,}8}$	$78{,}2^{+5{,}3}_{-6{,}0}$	$8{,}6^{+4{,}5}_{-3{,}4}$	$74{,}1^{+5{,}7}_{-6{,}3}$	$0{,}5^{+2{,}1}_{-0{,}4}$	$71{,}8^{+5{,}8}_{-6{,}4}$
mfcc_30	$25{,}9^{+6{,}3}_{-5{,}7}$	$70{,}9^{+5{,}9}_{-6{,}5}$	$25{,}9^{+6{,}3}_{-5{,}7}$	$75{,}9^{+5{,}5}_{-6{,}2}$	$5{,}5^{+3{,}9}_{-2{,}6}$	$71{,}8^{+5{,}8}_{-6{,}4}$	$0{,}0^{+1{,}7}_{-0{,}0}$	$56{,}8^{+6{,}6}_{-6{,}8}$
lfcc_30	$25{,}5^{+6{,}3}_{-5{,}6}$	$63{,}2^{+6{,}4}_{-6{,}7}$	$25{,}5^{+6{,}3}_{-5{,}6}$	$65{,}5^{+6{,}3}_{-6{,}7}$	$15{,}0^{+5{,}4}_{-4{,}4}$	$65{,}0^{+6{,}3}_{-6{,}7}$	$0{,}0^{+1{,}7}_{-0{,}0}$	$53{,}6^{+6{,}7}_{-6{,}8}$
mfcc_30_ms	$\mathbf{45{,}0}^{\mathbf{+6{,}8}}_{\mathbf{-6{,}7}}$	$89{,}5^{+3{,}7}_{-4{,}8}$	$44{,}5^{+6{,}8}_{-6{,}7}$	$94{,}1^{+2{,}7}_{-4{,}0}$	$26{,}8^{+6{,}4}_{-5{,}7}$	$\mathbf{94{,}5}^{\mathbf{+2{,}6}}_{\mathbf{-3{,}9}}$	$7{,}7^{+4{,}4}_{-3{,}2}$	$85{,}5^{+4{,}4}_{-5{,}4}$
lfcc_30_ms	$33{,}6^{+6{,}7}_{-6{,}2}$	$86{,}8^{+4{,}2}_{-5{,}2}$	$34{,}1^{+6{,}7}_{-6{,}2}$	$86{,}8^{+4{,}2}_{-5{,}2}$	$29{,}1^{+6{,}5}_{-5{,}9}$	$90{,}5^{+3{,}5}_{-4{,}7}$	$11{,}8^{+5{,}0}_{-4{,}0}$	$90{,}0^{+3{,}6}_{-4{,}7}$

Tabelle 5: Erkennungsquoten in Prozent mit Trainingsdaten aus Sitzung 1 und Testdaten aus Sitzung 3. Linke Spalte: mit Rückweisung - rechte Spalte: ohne Rückweisung.

$\rightarrow M$ $\downarrow$	1		2		4		8	
pfv_30	$50{,}9^{+6{,}8}_{-6{,}8}$	$83{,}6^{+4{,}6}_{-5{,}6}$	$55{,}5^{+6{,}7}_{-6{,}8}$	$88{,}2^{+4{,}0}_{-5{,}0}$	$26{,}4^{+6{,}3}_{-5{,}7}$	$86{,}8^{+4{,}2}_{-5{,}2}$	$15{,}5^{+5{,}5}_{-4{,}5}$	$86{,}8^{+4{,}2}_{-5{,}2}$
sfv_24	$80{,}5^{+5{,}0}_{-5{,}9}$	$96{,}4^{+2{,}1}_{-3{,}4}$	$\mathbf{81{,}4}^{\mathbf{+4{,}9}}_{\mathbf{-5{,}8}}$	$97{,}7^{+1{,}5}_{-3{,}0}$	$50{,}5^{+6{,}8}_{-6{,}8}$	$97{,}3^{+1{,}7}_{-3{,}1}$	$39{,}1^{+6{,}8}_{-6{,}5}$	$91{,}8^{+3{,}3}_{-4{,}4}$
mfcc_30	$54{,}1^{+6{,}7}_{-6{,}8}$	$96{,}8^{+1{,}9}_{-3{,}3}$	$54{,}5^{+6{,}7}_{-6{,}8}$	$97{,}3^{+1{,}7}_{-3{,}1}$	$23{,}2^{+6{,}1}_{-5{,}4}$	$92{,}7^{+3{,}1}_{-4{,}3}$	$4{,}5^{+3{,}7}_{-2{,}3}$	$92{,}7^{+3{,}1}_{-4{,}3}$
lfcc_30	$27{,}3^{+6{,}4}_{-5{,}8}$	$89{,}1^{+3{,}8}_{-4{,}9}$	$29{,}5^{+6{,}5}_{-5{,}9}$	$90{,}9^{+3{,}4}_{-4{,}6}$	$10{,}5^{+4{,}8}_{-3{,}7}$	$86{,}4^{+4{,}2}_{-5{,}3}$	$3{,}6^{+3{,}4}_{-2{,}1}$	$80{,}0^{+5{,}1}_{-5{,}9}$
mfcc_30_ms	$76{,}4^{+5{,}5}_{-6{,}2}$	$99{,}1^{+0{,}8}_{-2{,}3}$	$77{,}7^{+5{,}3}_{-6{,}1}$	$99{,}1^{+0{,}8}_{-2{,}3}$	$62{,}7^{+6{,}4}_{-6{,}8}$	$99{,}5^{+0{,}4}_{-2{,}1}$	$37{,}7^{+6{,}8}_{-6{,}4}$	$99{,}5^{+0{,}4}_{-2{,}1}$
lfcc_30_ms	$57{,}3^{+6{,}6}_{-6{,}8}$	$\mathbf{99{,}5}^{\mathbf{+0{,}4}}_{\mathbf{-2{,}1}}$	$59{,}5^{+6{,}5}_{-6{,}8}$	$98{,}6^{+1{,}1}_{-2{,}6}$	$46{,}8^{+6{,}8}_{-6{,}7}$	$99{,}5^{+0{,}4}_{-2{,}1}$	$23{,}2^{+6{,}1}_{-5{,}4}$	$99{,}1^{+0{,}8}_{-2{,}3}$

Tabelle 6: Erkennungsquoten in Prozent mit Trainingsdaten aus Sitzung 2 und Testdaten aus Sitzung 3. Linke Spalte: mit Rückweisung - rechte Spalte: ohne Rückweisung.

6.2 Auswertung & Ausblick

Unter der Bedingung, dass Trainings- und Testdaten einer berechtigten Person aus einer Aufnahmesitzung stammen, erfüllt der Merkmaltyp sfv_24 ab $M = 2$ die geforderte Bedingung einer Erkennungsquote von 90%, wenn ein Backoff-Modell vorhanden ist. Die Verbesserung der Erkennungsquoten durch die Erhöhung der Anzahl von Normalverteilungen ist in Tabelle 3 ebenfalls ersichtlich. Eine generelle Aussage über die Leistung eines bestimmten Merkmaltyps ist allerdings wegen Überlappung der Konfidenzintervalle ab $M = 2$ nicht mehr möglich. Mit einem tabellarischen Vergleich der Erkennungsquoten in Untersuchung 2 zeigt sich jedoch, dass das bisher implementierte System mit Backoff-Modell nicht leistungsfähig ist. Ein Vergleich mit den Erkennungsquoten ohne Backoff-Modell sowie der bekannten Literatur führte zur Erkennung von drei prägnanten Schwachstellen im System.

Das Backoff-Modell ist im Gegensatz zu einem Sprechermodell unspezifisch und bildet mit einer höheren Wahrscheinlichkeit Beobachtungen, die in den Testdaten aus Sitzung 3 vorkommen, ab. Dieses Problem lässt sich mit der Überprüfung eines Sperrbereiches bei der Wahl des Backoff-Modells beheben. Der Sperrbereich ist eine vorgegebene Wahrscheinlichkeitsdichte, die zwischen dem Backoff-Modell und dem „erstbesten“ Sprechermodell existieren muss. Wenn die Differenz zwischen der Wahrscheinlichkeitsdichte des spezifischen Sprechermodells und dem Backoff-Modell kleiner ist, erfolgt eine Entscheidung zu Gunsten des Sprechers.

Wegen der Vernachlässigung einer zeitlichen Struktur innerhalb der Merkmalvektorfolge bei GMM's nimmt der Klassifikator ebenfalls Passphrasen mit veränderter Struktur an. Die zeitliche Struktur kann jedoch mit dem Hinzufügen von Delta- und Delta-Delta-Merkmalen teilweise zurückgewonnen werden. Als letzte Schwachstelle ist die nicht vollständige Trennung von Sprache und Sprecher bei allen verwendeten Merkmaltypen zu erwähnen. Ein Verfahren zur vollständigen Trennung ist zur Zeit jedoch noch nicht bekannt. Neben den geschilderten Schwachstellen sind weitere Fragen in Bezug zur erstellten Datenbasis offen. In der uns bekannten Literatur zur Sprechererkennung fehlen genauere Beschreibungen zum Aufbau der vorgenommenen Untersuchungen. Es ist dementsprechend von Interesse, einen Vergleich der Erkennungsquoten existierender textabhängiger Sprechererkennungssysteme mit der erstellten Datenbasis durchzuführen. Mit den vorgenommenen Untersuchungen kann jedoch die Aussage getroffen werden, dass sich der Merkmaltyp sfv_24 neben der Spracherkennung ebenfalls für die textabhängige Sprechererkennung eignet.

Literatur

[1] HERTLEIN, H. R.: *Fusion von Klassifikationssystemen für die automatische Sprechererkennung*. Dissertation. Universität Nürnberg-Erlangen, 2010.

[2] HOFFMANN, R. und M. WOLFF: *Intelligente Signalverarbeitung 1: Signalanalyse*. 2.Auflage, Springer Vieweg Verlag, 2015.

[3] RHODENBURG, T.: *Klassifikation von Audio-Signalen*. Diplomarbeit. Universität Bremen, Arbeitsbereich Nachrichtentechnik, 2003.

[4] ROSE, P.: *Forensic speaker identification*. CRC Press, 2002.

[5] WOLFF, M.: *Akustische Mustererkennung*. TUDpress Verlag, 2011.

[6] WOLFF, M. Technical report. Brandenburgische Technische Universität, Professur für Kommunikationstechnik, 2014.