Brandenburgische Technische Universität
Cottbus - Senftenberg

Lehrstuhl für Kommunikationstechnik
Fakultät 3

Bachelorarbeit

Stimmenauthentifizierung

Voice Authentication

Autor: Peter Geßler
Matrikel-Nr.: 3002706
E-Mail: gessler.peter@outlook.de

Betreuer: Prof. Dr.-Ing. habil. Matthias Wolff

Brandenburgische
Technische Universität
Cottbus - Senftenberg

Lehrstuhl
Kommunikationstechnik

Eidesstattliche Erklärung

Der Verfasser erklärt, dass er die vorliegende Arbeit selbständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt hat. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Ort, Datum: Cottbus, den 30. Juli 2014 .....

(Unterschrift)

Einführung

Der Zugang zu hochsensiblen Einrichtungen als auch technisch relevanten Systemen erfordert einen geeigneten Schutz gegenüber unbefugten Benutzern. Heutzutage wird dieser durch wissensbasierte elektronische Kontrollsysteme gewährleistet. Der Benutzer benötigt einen - abhängig vom System - generierten Code, um sich authentifizieren zu können. Als Beispiel sei hier der standardisierte Bankautomat genannt. Die vorhandene Bankkarte dient zur Identifikation der Bank sowie des Kontos. Die Eingabe der Personal Identification Number (PIN) dient der Verifikation und soll sicherstellen, dass kein unbefugter Benutzer einen Kontozugriff erhält.
Beide Komponenten sind jedoch nur indirekt vom rechtmäßigen Benutzer abhängig. Eine unautorisierte Person, welche im Besitz der Bankkarte und der PIN ist, kann ebenso wie ein rechtmäßiger (autorisierter) Benutzer auf das Konto zugreifen.
Die Trennung zwischen Identifikation und Verifikation zur Authentifizierung an der Mensch-Maschine-Schnittstelle stellt somit eine relevante Sicherheitslücke dar. Eine direkte Abhängigkeit zwischen autorisiertem Benutzer und generiertem Code kann nur durch die Überprüfung von personengebundenen Merkmalen erfolgen.
Die Wissenschaft beschäftigt sich seit Anfang des 20. Jahrhunderts damit, eine mittels Biometrik (automatisiertes Messen eines oder mehrerer spezifischer Merkmale eines Lebewesens) spezifizierte Person von anderen unterscheidbar zu machen. Zu den bekanntesten Authentifizierungsmethoden gehört, neben Fingerbild- oder Iriserkennung, die in dieser Arbeit thematisierte Sprechererkennung. Dieses Verfahren wird inzwischen nicht mehr nur für die Zugangskontrolle geschützter Bereiche eingesetzt, sondern auch in der Forensik zur Identifikation einer Person.
Mit dieser Arbeit wird ein Verfahren zur textabhängigen Sprechererkennung entworfen, implementiert und getestet, welches für die Nutzerauthentifizierung im Sprachlabor des Lehrstuhls Kommunikationstechnik zuständig ist.
Der Schwerpunkt liegt dabei in der Untersuchung von Verfahren der Merkmalextraktion und Nachbearbeitungsschritten, die eine Unterscheidungsfähigkeit zwischen den Sprachsignalen der Sprecher ermöglichen. Eine Klassifikation der Sprachsignale erfolgt mittels *Gaussian Mixture Models*. Die Vorarbeiten und Berechnungen der Testergebnisse wurden in dieser Arbeit durch die Software Matlab durchgeführt. Eine endgültige Implementierung und Verifikation erfolgt im Experimentiersystem *Unified Approach to speech Synthesis and Recognition (UASR)*.

Inhaltsverzeichnis

1 Grundlagen

1.1 Stimmenauthentifizierungssystem - Prozesse

Das Ziel der automatischen Stimmenauthentifizierung besteht in der korrekten Entscheidungsfindung über Annahme oder Zurückweisung einer Person, welche sich am System authentifizieren möchte. Die Entscheidung wird dabei durch ein elektronisches System zur Sprechererkennung getroffen, weshalb der Begriff *automatisch* verwendet wird.
Die Stimme eignet sich dabei besonders durch ihre große Vielfalt von statistischen Eigenschaften sowie sprecherspezifischen Merkmalen. Neben den physiologischen Attributen wie Lage und Länge der Stimmbänder, Ausprägung des Artikulationstraktes oder Größe der Nasenhöhlen spielen weitere Komponenten wie Lautstärke oder Sprechergeschwindigkeit eine entscheidende Rolle für die spezifische Artikulation von Sprache durch eine Person.
Der wesentliche Unterschied zur oft im gleichen Kontext genannten Spracherkennung besteht in der Auswertung von Eigenschaften des Sprachsignals. Während in der Spracherkennung eine inhaltliche Überprüfung des Sprachsignals stattfindet, werden bei der Stimmenauthentifizierung die Ermittlung sowie der Vergleich von sprecherspezifischen Eigenschaften des Sprachsignals vorgenommen. Im Allgemeinen spricht man von einem Mustervergleich.
Ein System zur automatischen Stimmenauthentifizierung lässt sich im konkreten Anwendungsaufbau in die Prozesse *Sprecheridentifikation* und *Sprecherverifikation* einteilen.

Sprecheridentifikationsprozess

Bei einer Sprecheridentifikation gilt die Annahme, dass die Identität des Sprechers dem System nicht bekannt ist. Das Ziel der Sprecheridentifikation (Abb. 1.1) liegt damit in der Auswahl einer hinterlegten Identität s_i aus der Klassenmenge $S = \{s_1, \ldots, s_I\}$ mit $I \in \mathbb{N}$ bekannten Identitäten, welche eine maximale Ähnlichkeit zur unbekannten Identität aufweist.

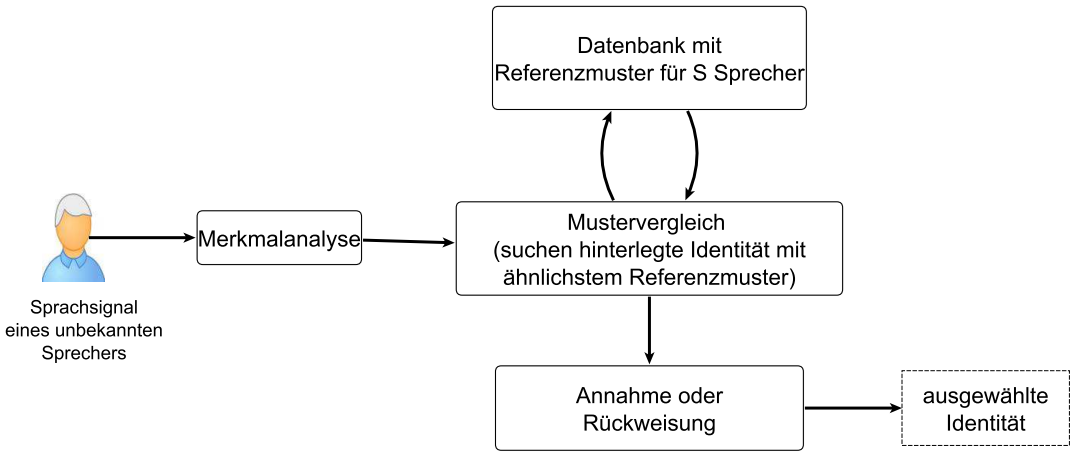

Abb. 1.1. Allgemeines Blockschaltbild zum Sprecheridentifikationsprozes in der Arbeitsphase [Fli95].

Sprecherverifikationsprozess

Der Sprecherverifikationsprozess setzt hingegen eine zuvor bekanntgegebene Identität voraus[1]. Es erfolgt ausschließlich eine Überprüfung der Identität mit dem dazugehörigen hinterlegten Referenzmuster der Sprecherklasse s_i. Die Entscheidung über Annahme oder Zurückweisung einer Person kann daher vereinfacht als *Zweiklassenproblem* angesehen werden [Fli95].

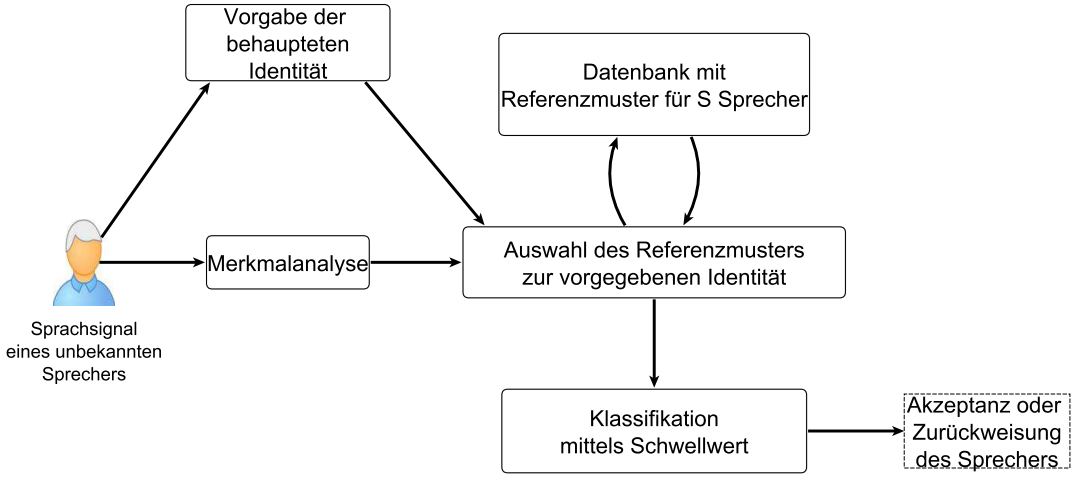

Abb. 1.2. Allgemeines Blockschaltbild zum Sprecherverifikationsprozess in der Arbeitsphase [Fli95].

Dieser Sachverhalt stellt ein entscheidendes Kriterium für die Entwicklung eines Stimmenauthentifizierungssystems dar und soll deshalb an einem einfachen Szenario zur Nutzerauthentifizierung noch einmal erläutert werden. Dabei ist zu beachten, dass das dargestellte Authentifizierungssystem unabhängig von dem zu entwickelnden System ist.

Beispiel zum Sprecheridentifikations- und verifikationsprozess:

Es existiert ein Kontrollzugang (Stimmauthentifizierungssystem) zu einer gesicherten Einrichtung. I berechtigte Personen hinterlegen in einer *Trainingsphase* jeweils ein akustisches Referenzmuster im System. Eine unbekannte Person versucht, sich in der *Arbeitsphase* mit einer akustischen Passphrase am Authentifizierungssystem anzumelden. Das System wählt im ersten Schritt die hinterlegte Identität s_i aus, bei der eine maximale Ähnlichkeit zwischen der gesprochenen Passphrase und einem der hinterlegten Referenzmuster existiert (Identifikationsprozess). Im zweiten Schritt wird überprüft, ob die Passphrase eine genügend große Ähnlichkeit zum Referenzmuster der ausgewählten Identität aufweist (Verifikationsprozess).
Sollte dies der Fall sein, würde ein Zugang zur gesicherten Einrichtung erfolgen. Wenn die Ähnlichkeit jedoch nur gering ist, erfolgt eine Abweisung vom Authentifizierungssystem und der Zugang bleibt verwehrt.

[1] Dies erfordert einen entsprechenden Systemaufbau, in dem die Identität vorher bekanntgegeben wird.

1.2 Stimmenauthentifizierungssystem - Komponenten und Formen

Ein System zur Stimmenauthentifizierung lässt sich des Weiteren, wie schon im vorangegangenen Beispiel erwähnt, in die getrennten Phasen *Trainingsphase* und *Arbeitsphase* unterteilen. Die primäre Aufgabe der Trainingsphase besteht dabei in der Berechnung eines Sprechermodells oder eines Referenzmusters des zu hinterlegenden Sprechers i. In der Arbeitsphase wird hingegen die Entscheidungsfindung ν zur Auswahl einer hinterlegten Identität und - Annahme oder Zurückweisung (Verifikation) der unbekannten Identität durchgeführt.
Beide Phasen enthalten jedoch eine identische Merkmalanalyse.

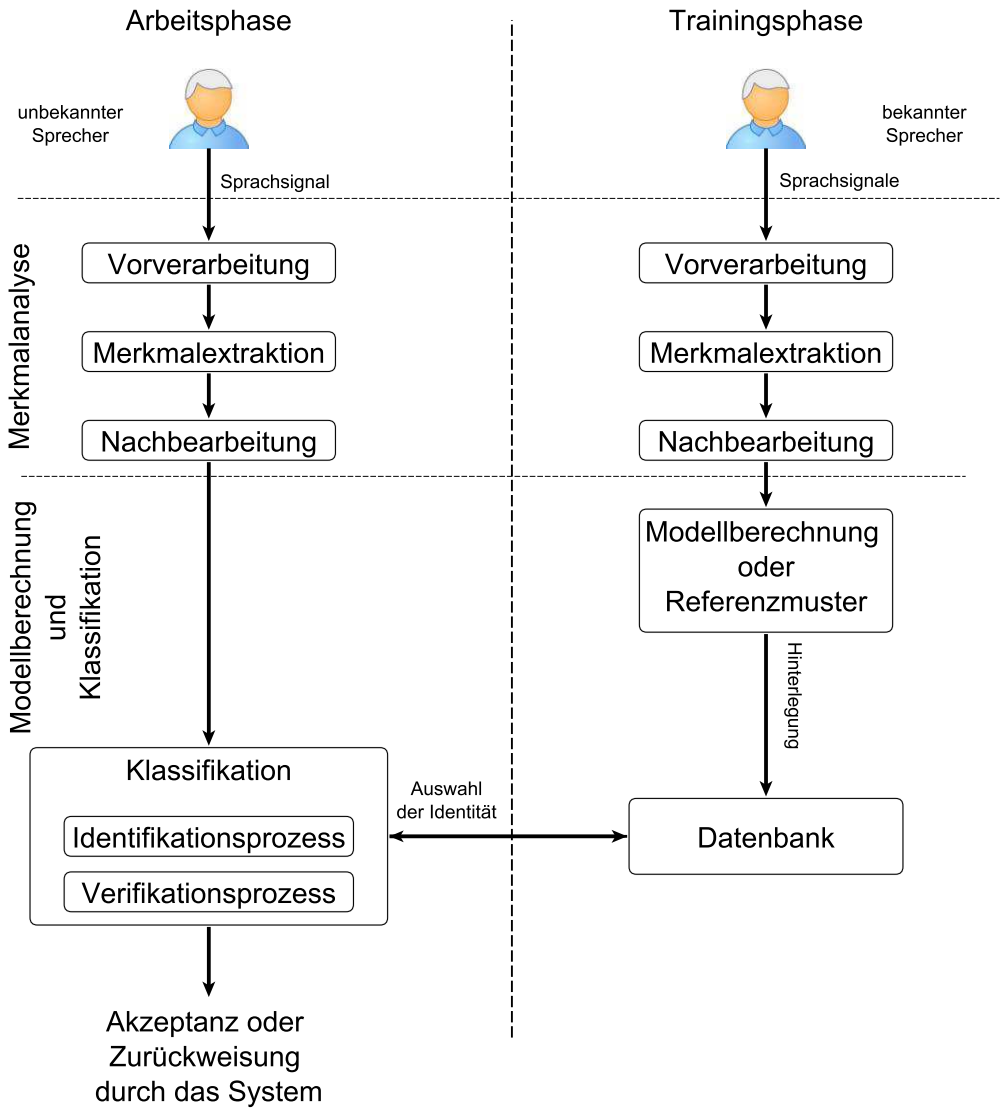

Abb. 1.3. Allgemeines Blockschaltbild Stimmenauthentifizierungssystem getrennt in Arbeits- und Trainingsphase

Ein weiterer Aspekt, der beachtet werden muss, ist das Verfahren zur Verarbeitung des Sprachsignals. In dem zu entwickelnden System besitzt jeder Sprecher eine eindeutig zugeordnete Passphrase. Wir sprechen in diesem Zusammenhang von einem textabhängigen Verfahren. Trotzdem sollen nachfolgend nochmals die Unterschiede zwischen den Varianten textabhängiger, textunabhängiger und text-promted Stimmenauthentifizierung erläutert werden.

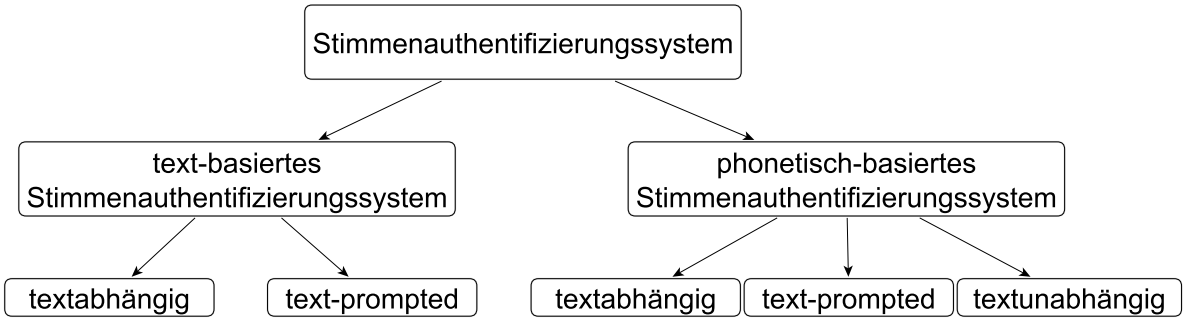

Abb. 1.4. Formen zur Realisierung von Stimmenauthentifizierungssystemen ([Har01], S. 9).

Bei textabhängigen Stimmenauthentifizierungssystemen wird eine Passphrase oder ein Code vorgegeben, welcher zur Authentifizierung notwendig ist. Während einer Trainingsphase wiederholt die autorisierte Person den Code 1 bis N mal. Die einzelnen Observationen werden für die Modellbildung oder Bildung des Referenzmusters der Sprecherklasse s_i verwendet[2].
Das Verfahren ist jedoch mit der Offenlegung des Codes bei der Authentifizierung verbunden. Ein gezielter Täuschungsversuch durch Imitation von sprecherspezifischen Eigenschaften ist möglich. Untersuchungen über erfolgreiche Täuschungsversuche textabhängiger Sprechererkennungssysteme existieren bisher jedoch nicht ([Har01], S. 10).
Textunabhängige Stimmenauthentifizierung, phonetisch basiert, bietet hingegen die Möglichkeit einer kurzen Lernphase. Der Benutzer kann sich unabhängig von der Passphrase am System anmelden und benötigt somit kein zusätzliches Wissen mehr. Des Weiteren kann bei Verbindung mit einem Spracherkennungssystem die Stimmenauthentifizierung direkt in den Sprachdialog integriert werden. Eine Täuschung des Stimmenauthentifizierungssystems ist hierbei, mit einer beliebigen Tonbandaufnahme eines autorisierten Benutzers, wiederum leicht zu realisieren.
Eine der sichersten Methoden gegen Täuschungsversuche ist das text-prompted Verfahren, welches das Sprechen einer vom System generierten Passphrase erfordert. Der Imitator hat in diesem Fall keine Zeit für ein Training der neuen Passphrase mit den sprecherspezifischen Eigenschaften eines autorisierten Benutzers. Die Imitation wird dementsprechend erschwert, weil ein Fehler des Imitators während des Täuschungsversuches wahrscheinlicher ist.

[2] Abhängig vom gewählten Klassifikationsverfahren

1.3 Mustererkennung

Dieser Abschnitt erläutert kurz die für diese Arbeit relevanten Grundbegriffe der akustischen Mustererkennung. Die aufgeführten Definitionen und Begrifflichkeiten sind dabei in gekürzter Form aus [Wol11] und [Hof98] entnommen worden.

Klasseneinteilung von Signalen

Die Zuordnung des Signals x_1 zu einem Sprecher s_i der Klassenmenge S erfolgt durch einen Mustervergleich. Wenn dabei nur „geringe“ Unterschiede zwischen x_1 und einem Referenzsignal x_2 des Sprechers s_i vorhanden sind, spricht man von einer Äquivalenzrelation $x_1 \sim x_2$.
Durch die Einführung von Äquivalenzrelationen lässt sich folglich jedes Signal einer Signalmenge $\mathcal{X}$ einem Sprecher s_i zuordnen. Die Klassenmenge S kann unter diesen Bedingungen als Menge von Äquivalenzklassen angesehen werden.

Analysetransformation & Merkmalvektorfolge

Ein direkter Vergleich von Sprachsignalen und die damit verbundene Zuordnung ist in einem Stimmenauthentifizierungssystem nicht ausreichend. Die Filterung von relevanten sprecherspezifischen Eigenschaften wird deswegen mit einer Analysetransformation f durchgeführt. Zunächst erfolgt die Überführung des vorhandenen Sprachsignals vom Zeitbereich in den Bildbereich. Der zweite Analyseabschnitt ist für die Filterung und Aufbereitung der für uns relevanten Informationen (Merkmale) des Sprachsignals verantwortlich.
Durch die Analysetransformation, welche in der Literatur auch als *Merkmalanalyse* bekannt ist, wird in unserem Fall eine Merkmalvektorfolge $\vec{\mathbf{o}}$ generiert. Diese besteht aus $N \in \mathbb{N}$ Merkmalvektoren $\vec{o}_n$, welche einen bestimmten Zeitbereich des Signals repräsentieren. In der Sprachverarbeitung ist es üblich, dass die Anzahl der Dimensionen innerhalb eines Merkmalvektors zwischen 10 und 30 beträgt.

$$\vec{\mathbf{o}} = (\vec{o}_1, \vec{o}_2, \ldots, \vec{o}_n, \ldots, \vec{o}_N) = (f(x(t,\tau_1)), f(x(t,\tau_2)), \ldots, f(x(t,\tau_n)), \ldots, f(x(t,\tau_N)))$$

$$\begin{aligned} f &: \text{Funktion zur Analysetransformation} \\ N &: \text{Segmentanzahl} \\ h(t) &: \text{Fensterfunktion} \\ x(t,\tau_n) &= x(t) \cdot h(t-\tau_n) \\ \tau_n &= n \cdot \Delta\nu \\ \Delta\nu &\approx 10 \text{ ms} \end{aligned} \tag{1.3.1}$$

Eine detaillierte Darstellung der Merkmalanalyse wird in Kapitel 2 vorgenommen.

Merkmalraum & Abstandsberechnung

Unter der Voraussetzung, dass ein D-dimensionaler Merkmalvektor $\vec{o}_n$ die extrahierten Merkmale eines Zeitabschnitts $x(t, \tau_n)$ repräsentiert, ist dieser genau als ein Punkt in einem D-dimensionalen Vektorraum dargestellt. In der Mustererkennung wird üblicherweise davon ausgegangen, dass die Merkmalvektoren Elemente eines *Merkmalraums* $\mathcal{O}$ sind.
Eine Aussage über die Ähnlichkeit zweier Merkmalvektoren $\vec{o}$ und $\vec{r}$ [3] ist somit über eine skalare Abstandsfunktion $d(\vec{o}, \vec{r})$ möglich. Zur Veranschaulichung der Abstandsberechnung wird in diesem Teilabschnitt davon ausgegangen, dass ein Sprachsignal nur durch einen Merkmalvektor dargestellt wird. Die Grundlage der Abstandsberechnung stellt dabei die MINKOWSKI-Norm in einem D-dimensionalen Merkmalraum dar.

$$d(\vec{o}, \vec{r}) = \sqrt[g]{\sum_{d=1}^{D} (o_d - r_d)^g} \qquad \text{mit } g \in \mathbb{N} \text{ und } g \geq 1 \tag{1.3.2}$$

$\vec{o}$: erster Merkmalvektor

$\vec{r}$: zweiter Merkmalvektor

D : Dimension der Merkmalvektoren

Eine Voraussetzung der Abstandsberechnung nach [Wol11] in einem Merkmalraum ist die Homogenität der physikalischen Einheiten einer Merkmalvektorkomponente. Diese Annahme kann in der Sprechererkennung jedoch nicht prinzipiell vorausgesetzt werden. Ein einfaches Beispiel dieser Problematik ergibt sich schon, wenn beispielsweise die Artikulationsgeschwindigkeit und der Lautstärkepegel als Merkmalparameter angenommen werden.
Ein weiterer Nachteil besteht in der nicht skalierten Gewichtung, mit der die Komponenten in die Abstandsberechnung eingehen. Spezifische Merkmale mit kleineren Werten werden in der anschließenden Klassifikation nur geringfügig bis gar nicht beachtet. Es wird vorgeschlagen, einen geeigneten Gewichtungsfaktor w_D zur prioritäts-basierten Gewichtung der jeweiligen Merkmalkomponente hinzuzufügen.
Für die Bildung des gewichteten euklidischen Abstands mit $g = 2$ ergibt sich beispielsweise folgende Gleichung:

$$d(\vec{o}, \vec{r}) = \sqrt{\sum_{d=1}^{D} \frac{(o_d - r_d)^2}{w_d^2}} \tag{1.3.3}$$

Ein weiterer Vorteil liegt in der Hinzufügung einer physikalischen Einheit, wodurch die Kürzung der „Dimensionseinheit“ erfolgen kann.

[3] Die beiden Merkmalvektoren repräsentieren unterschiedliche Sprachsignale.

1.4 Klassifikation & Leistungsbewertung

Klassifikation bedeutet in der vorliegenden Arbeit, eine eindeutige Auswahl ν aus einer definierten Anzahl von Entscheidungsmöglichkeiten vorzunehmen.
Im Identifikationsprozess wird die Entscheidung zu einer Sprecherklasse s_i aus der Klassenmenge S getroffen $\nu = s_i$. Der Verifikationsprozess enthält die Auswahlmöglichkeiten $\nu = 0$ - Zurückweisung des Sprechers oder $\nu = 1$ - Annahme des Sprechers (Detektion).

Modifikation der Klassenmenge S

Unter der Voraussetzung, dass der unbekannte Sprecher eine korrekte Passphrase wiedergibt, ist es für den Sprecheridentifikationsprozess in einem ersten Ansatz sinnvoll, dass genau ein Sprecher s_i aus der Klassenmenge S ausgewählt wird.
Wird hingegen eine ungültige Passphrase wiedergegeben, hat der Klassifikator im Identifikationsprozess den Sprecher zurückzuweisen. Dies gilt ebenso bei einer „unsicheren“ Entscheidung des Klassifikators. Dieser Vorgang wird durch das Einfügen einer zusätzlichen Sprecherklasse s_0 realisiert. Eine Entscheidung $\nu = s_0$ ist vom Klassifikator zu wählen, wenn keine oder nur eine geringe Ähnlichkeit zwischen $\vec{\mathbf{o}}$ und den hinterlegten Referenzmustern oder Sprechermodellen $\vec{r}_i$ besteht.

Es existiert somit eine weitere Entscheidungsmöglichkeit - die der Zurückweisung $\nu = s_0$

$$\rightarrow \nu = \underset{i=0,\ldots,I}{\arg \operatorname{ext}} \; d(\vec{\mathbf{o}}, \vec{\mathbf{r}}_i) \qquad \text{mit } \vec{\mathbf{r}}_i : \text{Referenzmuster des } i\text{-ten Sprechers} \tag{1.4.1}$$

Gleichung 1.4.1 stellt die allgemeine Entscheidungsregel nach der Modifikation dar. Eine Ersetzung des Argumentes durch die Operatoren min oder max ist von der Arbeitsweise des Klassifikators abhängig.

Detektion

Die Detektion ist ein spezielles Klassifikationsverfahren [Wol11]. Als Einsatzgebiet sei hier beispielsweise der Prozess Sprecherverifikation genannt. Das Verfahren berücksichtigt genau eine Sprecherklasse s_i und überprüft ausschließlich, ob die generierte Merkmalvektorfolge vom Sprachsignal des unbekannten Sprechers mit der des ausgewählten Referenzmusters ähnlich ist. Wenn dies der Fall ist, wird der Sprecher vom System angenommen $\nu = 1$. Ähnelt die Merkmalvektorfolge nicht der des Referenzmusters, wird der Sprecher zurückgewiesen $\nu = 0$.

Statistische Klassifikation

Eine weitere Möglichkeit neben der Berechnung des Abstands nach Abschnitt 1.3 besteht in der statistischen Betrachtung der Zugehörigkeit einer Merkmalvektorfolge $\vec{\mathbf{o}}$ zu einen Sprecher i. Die Darstellung dieses Zusammenhangs ist mittels kontinuierlicher Dichtefunktionen realisierbar und wird nachfolgend anhand des Bayes-Klassifikators erläutert.

Bayes-Klassifikator. Nach [Wol11] ist die größte a-posteriori-Wahrscheinlichkeit $\max P(s_i|\vec{\mathbf{o}})$ für die Zuordnung zu einer Sprecherklasse maßgebend. Die a-posteriori-Wahrscheinlichkeit steht dabei für die Auswahl des Sprechers i unter der Bedingung, dass $\vec{\mathbf{o}}$ vorliegt.

$$\nu = \arg\max_i \; P(s_i|\vec{\mathbf{o}}) \tag{1.4.2}$$

$P(s_i|\vec{\mathbf{o}})$ ist jedoch nicht bekannt, weshalb eine Umformung mit dem Satz von Bayes $p(s_i, \vec{\mathbf{o}}) = p(s_i|\vec{\mathbf{o}}) \cdot p(\vec{\mathbf{o}}) = p(\vec{\mathbf{o}}|s_i) \cdot P(s_i)$ vorgenommen wird. Mit der Umformung erhalten wir die Gleichung

$$\nu = \arg\max_i \; \frac{p(\vec{\mathbf{o}}|s_i) \cdot P(s_i)}{p(\vec{\mathbf{o}})}.$$

$$\begin{aligned}
\vec{\mathbf{o}} &: \text{ Merkmalvektorfolge der gesprochenen Phrase} \\
s_i &: \text{ Klasse des Sprechers } i \\
p(\vec{\mathbf{o}}) &: \text{ Wahrscheinlichkeit von } \vec{\mathbf{o}} \\
P(s_i) &: \text{ a-priori Wahrscheinlichkeit} \\
p(\vec{\mathbf{o}}|s_i) &: \text{ Auftretenswahrscheinlichkeit von} \\
&\quad \vec{\mathbf{o}} \text{ unter } s_i \text{ (Likelihood-Funktion)}
\end{aligned} \tag{1.4.3}$$

Die Wahrscheinlichkeit der Merkmalvektorfolge $p(\vec{\mathbf{o}})$ ist dabei unabhängig vom Sprecher und für die weitere Entscheidung irrelevant. Der Bayes-Klassifikator wird somit durch die Gleichung

$$\nu = \arg\max_i \; p(\vec{\mathbf{o}}|s_i) \cdot P(s_i) \tag{1.4.4}$$

repräsentiert.

Maximum-Likelihood-Klassifikation bedeutet hingegen, dass die Entscheidungsfindung ausschließlich von der maximalen Likelihood Funktion abhängt.

$$\nu = \arg\max_i \; p(\vec{\mathbf{o}}|s_i) \tag{1.4.5}$$

Es wird bei der Maximum-Likelihood-Klassifikation von einer Gleichverteilung der a-priori-Wahrscheinlichkeiten ausgegangen $P(s_i) = 1/I$.

1.5 Bewertung von Systemen zur Stimmenauthentifizierung

Für eine Bewertung des Stimmenauthentifizierungssystems werden vorerst eine Trainingsstichprobe zum Trainieren der Sprechermodelle (siehe Kapitel 3) und eine Teststichprobe für den Klassifikationsvorgang benötigt. Diese sind zueinander disjunkt und in unserem Fall klassifiziert. Es kann somit bei jedem Sprachsignal der Teststichprobe eine Aussage getroffen werden, ob eine korrekte Entscheidung des Klassifikators vorliegt. Wenn eine Aussage zur Leistungsfähigkeit des Stimmenauthentifizierungssystems getroffen wird, erfolgt implizit eine Bewertung der Klassifikatoren des Identifikations- und Verifikationsprozesses.

Verwechslungsmatrix

Durch das Wissen über die objektive Klassenzugehörigkeit der Merkmalvektorfolge $\vec{\mathbf{o}}$ zur Klasse s_i ist eine einfache Bewertung über das Abzählen der korrekten Entscheidung gegenüber den fehlerhaften Entscheidungen möglich.
Eine konkrete Darstellung des Ergebnisses erhalten wir durch eine *Verwechslungsmatrix*.

$$\mathbf{H} = (\mathrm{H}\{(\nu, s_i)\}) = \begin{array}{c} \\ \downarrow \nu \end{array} \begin{array}{c} \rightarrow s_i \\ \begin{pmatrix} H(s_0,s_0) & H(s_0,s_1) & \cdots & H(s_0,S_I) \\ H(s_1,s_0) & H(s_1,s_1) & \cdots & H(s_1,S_I) \\ \vdots & \vdots & \ddots & \vdots \\ H(S_I,s_0) & H(S_I,s_1) & \cdots & H(S_I,S_I) \end{pmatrix} \end{array} \tag{1.5.1}$$

$H(\nu, s_i)$ steht in der Matrix für die absolute Häufigkeit, dass eine Entscheidung ν getroffen wurde, wenn die Klasse s_i vorliegt. Die korrekten Entscheidungen können entlang der Hauptdiagonalen abgelesen werden. Des Weiteren werden Zurückweisungen in der Matrix durch die erste Zeile und falsche Zuordnungen durch $H(\nu, s_i) \neq Hauptdiagonale$ repräsentiert.

Zusätzlich können wir durch eine Normierung die relativen Häufigkeiten $h(\nu, s)$ bestimmen.

$$h(\nu, s_i) = \frac{H(\nu, s_i)}{N} \qquad \text{mit } N : \text{Anzahl der Testelemente} \tag{1.5.2}$$

Die relativen Häufigkeiten ermöglichen eine Aussage zur Erkennungsrate des Klassifikators:

$$RR = \sum_{s_i=0}^{S} h(s_i, s_i). \tag{1.5.3}$$

Es ist ersichtlich, dass ein idealer Klassifikator hier den Wert $RR = 1$ annehmen würde.

Bewertungsmaße und Verwechslungsmatrix bei Detektion

Eine Verwechslungsmatrix gibt, wie im vorherigen Teilabschnitt beschrieben, Auskunft über die Häufigkeit der korrekten oder irrtümlichen Entscheidungen zu einer vorliegenden Klasse s_i. Für die Bewertung einer Sprecherverifikation in einem Stimmenauthentifizierungssystem, welche als Detektion anzusehen ist, sind jedoch nur folgende Entscheidungen relevant:

TP	true positive	korrekte Akzeptanz
TN	true negative	korrekte Zurückweisung
FP	false positive	falsche Akzeptanz
FN	false negative	falsche Zurückweisung

Tabelle 1.1 Entscheidungszenarien für Detektion

Es ergibt sich für die Detektion eine spezielle Verwechslungsmatrix:

$$\mathbf{H} = (\mathrm{H}\{(\nu, s)\}) = \begin{pmatrix} H(1,1) & H(1,0) \\ H(0,1) & H(0,0) \end{pmatrix} = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \tag{1.5.4}$$

Der Wert $\nu = 1$ gibt hier die Akzeptanz einer Klasse s an. Bei der Detektion kann es sich dabei um die korrekte Klasse $s = 1$ oder eine falsche Klasse $s = 0$ handeln. Falls der Detektor keine Ähnlichkeit zwischen der Merkmalvektorfolge $\vec{\mathbf{o}}$ und der hinterlegten Sprecherklasse s feststellt, wird eine Zurückweisung $\nu = 0$ getroffen. Die dazugehörigen Bewertungsmaße wurden in einer Arbeit von [DG64] zur Signaldeckungstheorie erstellt und lauten wie folgt:

Größe der Teststichprobe	$N = TP + FP + FN + TN$
Erkennungsrate Klasse richtig erkannt	$ACC = \frac{TP+TN}{N} = 1 - ERR$
Fehlerrate Klasse falsch erkannt	$ERR = \frac{FP+FN}{N} = 1 - ACC$
Sensitivität $s = 1$ und richtig erkannt	$SEN = \frac{TP}{TP+FN} = 1 - FRR (= TPR)$
Spezifität $s = 0$ und richtig erkannt	$SPC = \frac{TN}{FP+TN} = 1 - FAR (= TNR)$
Relevanz $\nu = 1$ und richtig erkannt	$REL = \frac{TP}{TP+FP} (= PPV)$
Fehlalarmrate $s = 0$ und falsch erkannt	$FAR = \frac{FP}{FP+TN} = 1 - SPC (= FPR)$
Fehldetektionsrate $\nu = 1$ und falsch erkannt	$FDR = \frac{FP}{FP+TP}$
Segreganz $\nu = 0$ und richtig erkannt	$SEG = \frac{TN}{TN+FN} (= NPV)$
Fehlrückweisungsrate $s = 1$ und falsch erkannt	$FRR = \frac{FN}{TP+FN} = 1 - SEN (= FNR)$

Tabelle 1.2 Bewertungsmaße für Detektoren

Leistungsfähigkeit eines Detektors (Equal Error Rate)

Eine Bewertung durch die in Tabelle 1.2 genannten Bewertungsmaße ist zwar möglich, jedoch wegen der Abhängigkeit vom eingestellten Schwellwert d_{rej} nicht zielführend.

Entscheidend für die Bewertung eines Stimmenauthentifizierungssystems, bei dem der Verifikationsprozess durch Detektion realisiert wird, ist daher die Frage des letztendlich zugelassenen Risikos einer fehlerhaften Klassifikation. Durch entsprechende Einstellung des Schwellwertes d_{rej} wird erreicht, dass entweder jeder Sprecher angenommen ($\Rightarrow FAR = 1$) oder abgewiesen ($\Rightarrow FRR = 1$) wird.

Für das zu entwickelnde Stimmauthentifizierungssystem ist der Rückweisungsschwellwert auf den Wert einzustellen, bei dem eine Gleichfehlerrate (engl. *equal error rate*, EER) zwischen Falschakzeptanz- und Falschrückweisungsrate auf einer Entwicklungsstichprobe vorliegt. Der EER-Wert ist im optimalen Fall minimal.

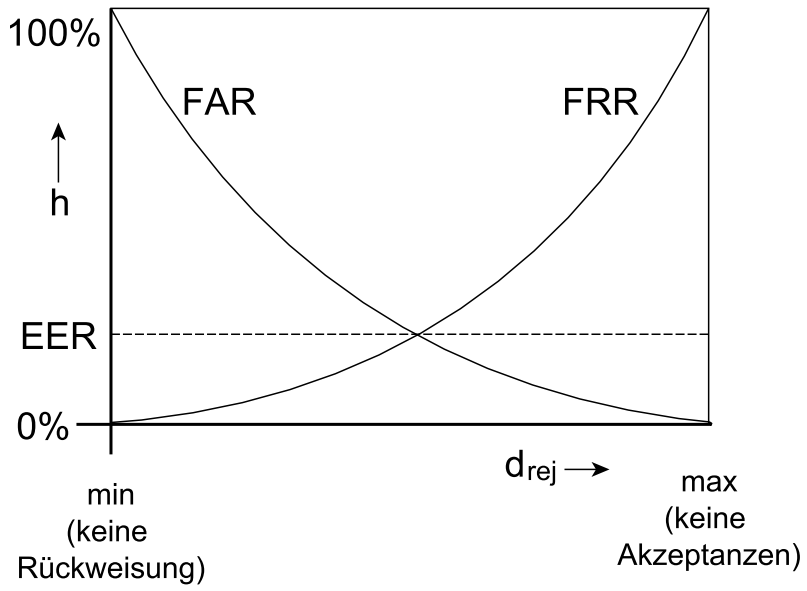

Abb. 1.5. Equal Error Rate - Verlauf von
Falsch-Akzeptanz-Rate (FAR) gegenüber Falsch-Rückweisungs-Rate (FRR).

2 Merkmalanalyse

Die Transformation eines akustischen Signals in eine Merkmalvektorfolge ist die zentrale Aufgabe der Merkmalanalyse. Nachfolgend werden die drei grundlegenden Verarbeitungsschritte eines Analysators erläutert.

2.1 Vorverarbeitung von Sprachsignalen

2.1.1 Quellsignal-Filterung

Zur Vereinfachung ist davon auszugehen, dass die Sprachaufnahme durch einen einzigen Sensor realisiert wird und eine Betrachtung im Zeitbereich stattfindet. Durch Störgeräusche $n(t)$ sowie der Übertragungsfunktion des Kanals $g(t)$ und des Aufnahmesensors wird das Quellsignal $x(t)$ in ein Signal $y(t)$ überführt. Verfahren der Störgeräuschreduktion und/oder Kanalkompensation versuchen in einer ersten Stufe, eine Rekonstruktion des Quellsignals vorzunehmen $y(t) \rightarrow \hat{x}(t) = x(t)$.
Eine exakte Wiederherstellung ist, durch Veränderung von Störgeräuschen und Kanaleigenschaften über die Zeit, nur bedingt realisierbar. Das gefilterte Signal $\hat{x}(t)$ ist somit nur eine optimierte Annäherung an das Quellsignal $x(t)$.

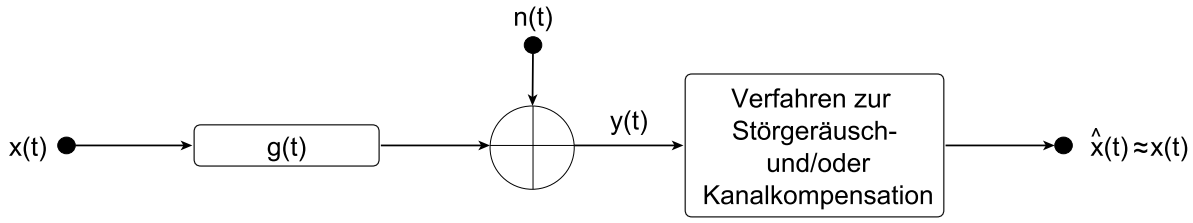

Abb. 2.1. Modell eines linearen zeitinvarianten Systems mit anschließendem Filterverfahren. $x(t)$ ist Nutzsignal, $n(t)$ Störsignal und $g(t)$ Impulsantwort des Kanals. $y(t)$ beschreibt das zu verarbeitende Signal und $\hat{x}(t)$ das angenäherte Quellsignal.

Die Rekonstruktion des Quellsignals $x(t)$ ist Aufgabe der Audio- und Signalverarbeitung und wird im weiteren Verlauf dieser Arbeit nicht näher erläutert werden.
Für eine genauere Betrachtung dieser Problematik sowie mögliche Verfahren zur Störgeräuschreduktion oder Kanalkompensation wird auf das Kapitel 2.1.1 in [Wol11] verwiesen.
In der weiteren Betrachtung liegt das Audio-Signal in einer digital abgetasteten Form $x(k)$ vor. Weiterhin wird die Annahme getroffen, dass Verfahren zur Störgeräuschreduktion
sowie Kanalkompensation bereits durchgeführt wurden. Es gilt lediglich zu beachten, dass dieses zeit- und wertdiskrete Signal auch nur eine Näherung an das Quellsignal $x(t)$ darstellt.

2.1.2 Segmentierung & Fensterfunktionen

Segmentierung beschreibt allgemein die Aufteilung eines Signals in eine endliche Anzahl von Zeitabschnitten τ. Jeder generierte Merkmalvektor stellt somit die Informationen eines bestimmten Zeitabschnitts dar. In der vorliegenden Arbeit wurde bei einer Abtastfrequenz von 16 kHz ein Analyseintervall $x_\tau(k)$ von $K = 160$ Abtastwerten gewählt. Es ergibt sich hierbei eine Analysezeit von 10 ms. Die Anzahl der Spektrallinien ist auf $N = 512$ festgelegt.
Eine Überlappung der Segmente erfolgte ebenfalls und beträgt im implementierten System 78,125%. Die anschließende Multiplikation des zeitdiskreten Signals eines Segments mit einer Fensterfunktion $h(k)$ minimiert in der anschließenden Spektralanalyse die Folgen des sogenannten *Leck-Effekt*[4].

$$\widehat{x}_\tau(k) = x_\tau(k) \cdot h(k) \tag{2.1.1}$$

Zur Minimierung der spektralen Fehlstellen im Frequenzbereich wird in der Literatur ([Har01], [Fli95]) vorwiegend das *Hamming-Fenster* als Fensterfunktion genutzt. Damit eine annähernde Äquivalenz bei der Vorverarbeitungsstufe zum UASR-System gegeben ist, wird für die Matlab-Implementierung das *Blackman-Fenster* verwendet.

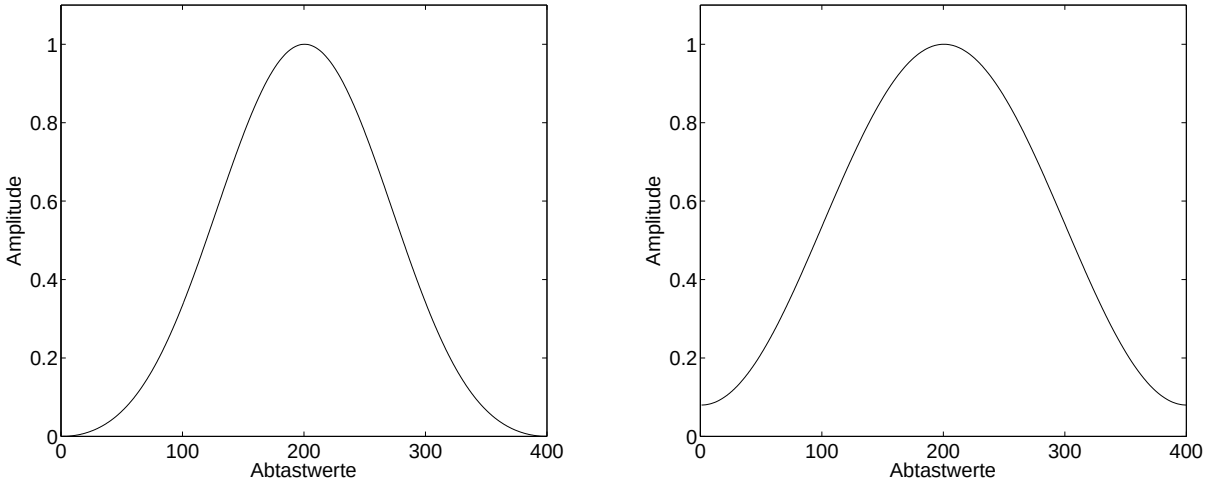

Abb. 2.2. Blackman- (links) und Hamming-Fenster (rechts) im diskreten Zeitbereich.

[4] spektrale Fehlstellen im Spektrum

2.2 Merkmalextraktion

Eine korrekte Klassifizierung des Signals ist neben dem verwendeten Klassifikationsverfahren primär von den Eigenschaften der generierten Merkmalvektoren und deren Parametern abhängig. An die Parameterauswahl werden nach [Wol72] deshalb folgende Bedingungen gestellt:

- **Darstellung des Signals als Merkmalvektorfolge**
  Durch die Beibehaltung der zeitlichen Struktur ergibt sich eine weitere Möglichkeit, die Klassifikation zu optimieren.
- **Robustheit der Parameter**
  Die extrahierten Parameter sollten im Idealfall nicht vom physiologischen und psychologischen Zustand des Sprechers abhängig sein. Störgeräusche oder Kanaleigenschaften dürfen zu keiner signifikanten Veränderung der Parameter führen.
- **Maximierung der *F-Ratio***
  F-Ratio (auch Varianzanalyse genannt) beschreibt das Verhältnis der Inter-Sprecher-Varianz zur Intra-Sprecher-Varianz. Es sind Parameter zu wählen, die einen großen F-Ratio Wert aufweisen.

$$F - Ratio = \frac{Inter - Sprecher - Varianz}{Intra - Sprecher - Varianz} \stackrel{!}{=} Max. \tag{2.2.1}$$

2.2.1 Mel Frequency Cepstral Coefficients (MFCC-Analyse)

In der Literatur gelten die Mel-Frequency-Cepstral Coefficients (MFCC), welche für die Spracherkennung entwickelt wurden, ebenfalls für die Sprechererkennung als *Baseline*-Merkmale [Fel12]. Die Berechnung der Merkmalkoeffizienten setzt sich dabei aus fünf Teilschritten zusammen, die der Abbildung 2.3 entnommen werden können.

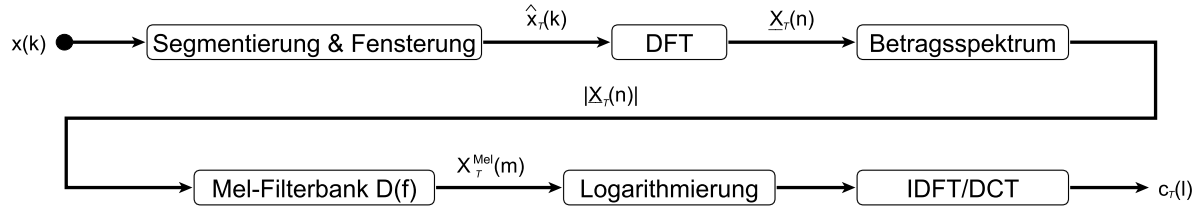

Abb. 2.3. Schritte der MFCC-Analyse inklusive Segmentierung und Fensterung.

Nachfolgend wird die Berechnung eines Merkmalvektors $\vec{o}_n$ aus einem gefensterten Zeitabschnitt $\widehat{x}_\tau(k)$ erläutert.

Diskrete Fourier-Transformation & Betragsspektrum

Das gefensterte zeitdiskrete Signal $\widehat{x}_\tau(k)$ wird zunächst mittels *Diskreter Fourier Transformation* (DFT) oder *Fast Fourier Transformation*[5] in ein frequenzdiskretes Spektrum transformiert.

$$\underline{X}_\tau(n) = DFT\{\widehat{x}_\tau(k)\} = \sum_{k=0}^{K-1} \widehat{x}_\tau(k) \cdot e^{\frac{-j2\pi kn}{N}} \tag{2.2.2}$$

Die Verwendung der DFT basiert auf der Annahme, dass ein Sprachsignal mit einer Länge von bis zu 30 ms als *quasi-stationär*[6] angesehen werden darf. Anschließend werden die reellwertigen Beträge der Koeffizienten gebildet.

$$|\underline{X}_\tau(n)| \tag{2.2.3}$$

Mel-Skala, Mel-Filterbank & Mel-Spektrogram

Für die Berechnung der Mel-Cepstral Koeffizienten wird eine Mel-Filterbank verwendet. Diese besteht aus vordefinierten Dreiecksbandpässen, die sich überlappen und annähernd logarithmisch über das Frequenzband verteilen.
Eine genaue Anzahl der zu definierenden Dreiecksbandpässe hängt dabei von der gewünschten Dimensionsanzahl innerhalb eines Merkmalvektors ab [Her10]. Die Berechnung der Bandpässe geht auf die *Mel-Skala* (Abbildung 2.4) von Stevens und Volkman sowie der dazugehörigen *Mel-Funktion* zurück. Diese rechnet die Frequenz f in die Mel-Skala um [Mil07].

$$f_{mel}(f) = 2595\ log\left(1 + \frac{f}{700\ Hz}\right) \qquad \text{in } mel \tag{2.2.4}$$

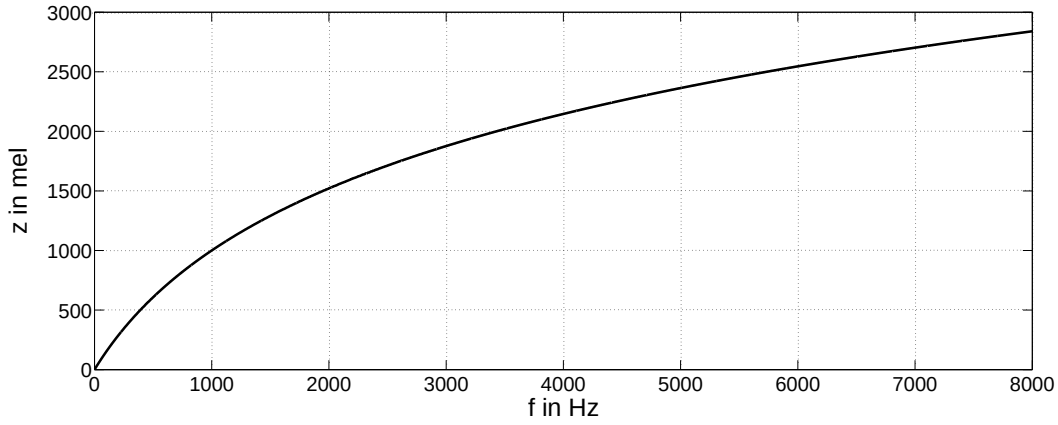

Abb. 2.4. Mel Skala nach (2.2.4)

[5]Die Berechnung der DFT oder FFT wird in [Wen04] erläutert.
[6]Autokorrelationsfunktion ändert sich nicht wesentlich.

Durch die in dieser Arbeit verwendete Abtastfrequenz von $f_A = 16$ kHz befinden sich die Mittenfrequenzen $B_{MF,m}$ und kritischen Bandbreiten b_m eines Bandpasses innerhalb von 0 bis 8 kHz (siehe Abb. 2.5) und können wie folgt berechnet werden:

$$B_{MF,m}^{Mel} = \frac{m}{M+1} \cdot f_{mel}(8\ kHz) \qquad \text{mit } m = 1,2,...,M \text{ Bandpässen.} \tag{2.2.5}$$

Es erfolgt die Rücktransformation in die Einheit Hertz:

$$B_{MF,m} = 700 \cdot \left(10^{\frac{B_{MF,m}^{Mel}}{2595}} - 1\right) \qquad \text{in Hz.} \tag{2.2.6}$$

Für die Berechnung der kritischen Bandbreite b_m eines Dreiecksbandpasses rechnen wir zunächst die Bandmittenfrequenz $B_{MF,m}$ in die Bark-Skala um:

$$B_{MF,m}^{Bark} = 13 \cdot \arctan\ (0{,}00076 \cdot B_{MF,m}) + 3{,}5 \cdot \arctan\left[\left(\frac{B_{MF,m}}{7500}\right)^2\right] \qquad \text{in Bark.} \tag{2.2.7}$$

Die kritische Bandbreite ist durch die Formel

$$b_m = \frac{52548}{{B_{MF,m}^{Bark}}^2 - 52{,}56 \cdot B_{MF,m}^{Bark} + 690{,}39} \qquad \text{in Hz.} \tag{2.2.8}$$

gegeben.

Die Frequenzverläufe der M Dreiecksbandpässen ergeben sich nach [Mil07] wie folgt:

$$D_m^{Mel}(f) = \begin{cases} 1 - \frac{|f - B_{MF,m}|}{b_m} & B_{MF,m} - b_m < f < B_{MF,m} + b_m \\ 0 & \text{sonst} \end{cases}. \tag{2.2.9}$$

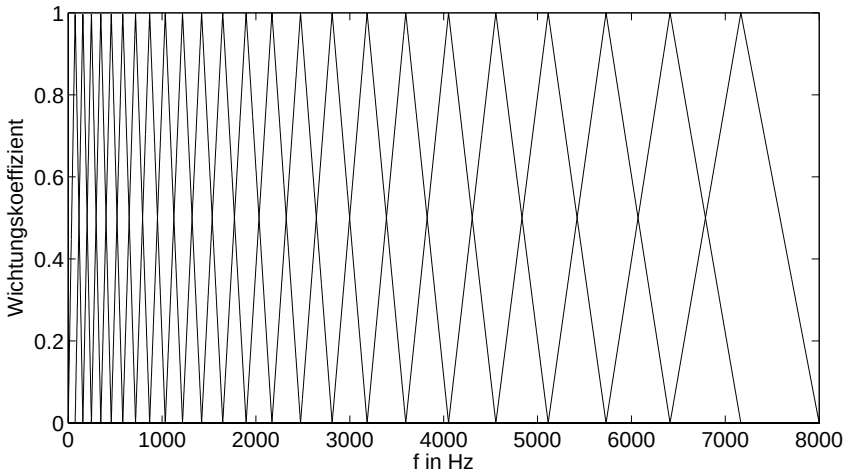

Abb. 2.5. Mel-Filterbank mit 24 Bandpässen nach matlab Toolbox *voicebox* [Bro].

In Abbildung 2.5 ist erkennbar, dass die manuelle Berechnung der Werte nicht mit denen der Abbildung übereinstimmt. Dies liegt daran, dass in den meisten Implementierungen keine eigenständige Berechnung der Bandmittenfrequenzen durchgeführt wird.
Eine Verwendung von statischen Werten, wie in Tabelle 2.1 dargestellt, ist üblich.

Index m	1	2 ...	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$B_{MF,m}$	100	200 ...	900	1000	1149	1320	1516	1741	2000	2297	2633	3031	3482	4000	4595	5278	6063	6964
b_m	100	100 ...	100	124	160	184	211	242	278	320	367	422	484	556	636	734	843	969

Tabelle 2.1: Verteilung der Bandmittenfrequenzen und Bandbreiten für 24 Dreiecksbandpässe bei einer Abtastastfrequenz $f_A = 16$ kHz. Die Angabe der Werte von Bandmittenfrequenzen und kritischen Bandbreiten erfolgt in Hz.

Die Koeffizienten des Betragsspektrums $|\underline{X}_\tau(n)|$, welche sich innerhalb des Frequenzbereiches eines Dreiecksbandpasses befinden, werden anschließend mit dem dazugehörigen Wert des Bandpasses $D_m^{Mel}(n)$ gewichtet, aufsummiert und logarithmiert.
Wir erhalten somit das Mel-Spektrogramm:

$$X_\tau^{Mel}(m) = 10\ log\left(\sum_{n=B_{MF}(m)-b_m(m)}^{B_{MF}(m)+b_m(m)} |\underline{X}_\tau(n)| \cdot D_m^{Mel}(n)\right). \tag{2.2.10}$$

Diskrete Cosinus Transformation (DCT)

Durch die Anwendung einer *Inversen Diskreten Fourier-Transformation* (IDFT) auf das logarithmierte Mel-Spektrum[7] werden die Mel-Frequency-Cepstral Koeffizienten generiert. Anstatt der IDFT wird normalerweise jedoch die DCT verwendet, welche in diesem Zusammenhang das gleiche Ergebnis liefert[8].

Der i-te Cepstral Koeffizient lässt sich dabei wie folgt berechnen:

$$c_\tau(i) = \sum_{m=1}^{M} X_\tau^{Mel}(m) \cdot \cos\left[\frac{\pi}{M} i\left(m - \frac{1}{2}\right)\right] \qquad \text{mit} \quad \begin{matrix} i = 1, ..., I \text{ Cepstralindex} \\ m = 1, ..., M \text{ Bandpässe} \end{matrix} . \tag{2.2.11}$$

Ein Merkmalvektor $\vec{o}_n$ wird anschließend aus den einzelnen Cepstral- Koeffizienten des Zeitabschnitts τ gebildet:

$$\vec{o}_n = \begin{pmatrix} c_\tau(1) \\ \vdots \\ c_\tau(I) \end{pmatrix} \qquad \text{mit} \qquad i = 1, \ldots, I \text{ Koeffizientenindex.} \tag{2.2.12}$$

2.2.2 Linear Frequency Cepstral Coefficients (LFCC-Analyse)

Die im letzten Abschnitt beschriebenen Mel-Frequency-Cepstral Koeffizienten sind für die Spracherkennung entwickelt worden. Es ist davon auszugehen, dass sie Informationen bereitstellen, die im Idealfall unabhängig vom Sprecher sind. Eine Wiederverwendung dieses Merkmalextraktionsverfahrens zur Sprechererkennung ist damit aus theoretischer Sicht unlogisch. [Her10] erläuterte in seiner Veröffentlichung, weshalb eine genauere Abbildung der höheren Formanten im Merkmalvektor durch die Verwendung einer linearen Filterbank ermöglicht wird. Diese reflektieren nach Untersuchungen von [Ros02] die sprecherspezifischen Teile des Vokaltraktes. Die Berechnung der sogenannten LFCC-Merkmale erfolgt annähernd analog zu den MFCC-Merkmalen. Der Unterschied besteht lediglich in der Berechnung der Bandmittenfrequenzen:

$$B_{MF,m} = \frac{m}{M+1} \cdot 8 \text{ kHz} \qquad \text{mit } m = 1,2, ..., M \text{ Bandpässen.} \tag{2.2.13}$$

Die Berechnung der kritischen Bandbreiten b_m verringert sich zu:

$$b_m = 2 \cdot B_{MF,1} \qquad \forall\ m \in \mathbb{N}. \tag{2.2.14}$$

[7] Das Ergebnis dieser Transformation ist in der Literatur als *Cepstrum* bekannt [RW99].
[8] Die Äquivalenz zwischen IDFT und DCT wird in [Mil07] - Abschnitt 2.2.1 näher erläutert.

2.3 Nachbearbeitung

2.3.1 Gewichtetes Liftering

Mittels gewichtetem *Liftering* wird die „Anhebung“ der Cepstral-Koeffizienten im oberen Cepstralbereich erreicht. Die Koeffizienten im unteren Cepstralbereich werden dabei abgeschwächt, wodurch in der Gesamtbetrachtung die Koeffizienten im oberen Bereich einen größeren Einfluss bei der Klassifikation nehmen. In der Spracherkennung wurden von [Tal95] Untersuchungen durchgeführt, wonach eine Auswertung dieser Koeffizienten zu keiner signifikanten Verbesserung der Ergebnisse geführt hat ([Har01], S. 43).
Es erfolgte dementsprechend eine Untersuchung, inwieweit eine Schwächung der unteren Cepstral-Koeffizienten das Erkennungsergebnis in der Sprechererkennung beeinflusst. Nach [Her10] können wir durch (2.3.1) eine Verstärkung der oberen Koeffizienten gegenüber den unteren Koeffizienten vornehmen:

$$c_{\tau,lift}(i) = c_\tau(i) \cdot l(i) \qquad \text{mit} \quad \begin{array}{l} \forall\ i = 1,2,...,I \text{ Cepstralindex} \\ l(i) : \text{Lifterkoeffizient} \end{array} . \tag{2.3.1}$$

Der i-te Lifter-Koeffizient ergibt sich aus:

$$l(i) = 1 + \frac{L}{2} \cdot \sin\left(\frac{\pi \cdot i}{L}\right) \qquad \text{mit} \quad \begin{array}{l} \forall\ i = 1,2,...,I \\ L = 2 \cdot I \end{array} . \tag{2.3.2}$$

Die Einführung eines Lifter-Gewichtes $w \in \mathbb{R}$, das zwischen 0 und 1 liegt, ermöglicht mittels folgender Formel eine zusätzliche Variation der „Anhebung“ und „Schwächung“ beim Liftering:

$$c_{\tau,lift}(i) = w \cdot \frac{2}{2+L} \cdot l(i) \cdot c_\tau(i) + (1-w) \cdot c_\tau(i) \qquad \text{mit} \quad \begin{array}{l} \forall\ i = 1,2,...,I \\ w \leq 1 \\ L = 2 \cdot I \end{array} . \tag{2.3.3}$$

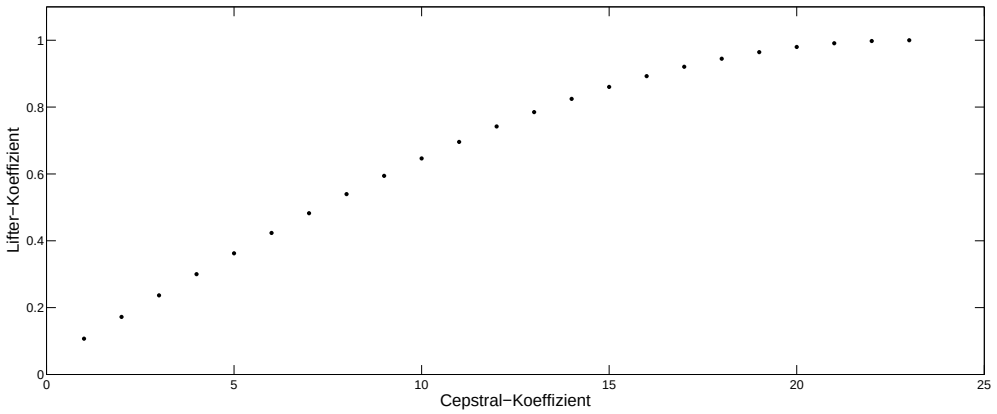

Abb. 2.6. Gewichtetes Liftering mit $w = 1$ nach [Her10]

2.3.2 Cepstrale Mittelwertsubtraktion

In Abschnitt 2.1.1 wurde für die weitere Merkmalanalyse eine bereits durchgeführte Kanalkompensation vorausgesetzt. Diese Annahme wird jetzt zurückgestellt. Es wird hingegen angenommen, dass sich die Eigenschaften des Übertragungskanals über die Zeit nicht ändern.
Es existiert im diskreten Zeitbereich somit folgender Zusammenhang:

$$x_\tau(k) = x_{Q,\tau}(k) * g(k) \tag{2.3.4}$$

Im diskreten Frequenzbereich gilt dementsprechend:

$$\underline{X}_\tau(n) = \underline{X}_{Q,\tau}(n) \cdot \underline{G}(n). \tag{2.3.5}$$

Eine Weiterführung der Merkmalanalyse, wie in Abschnitt 2.2.1 und 2.2.2 beschrieben, führt nach [Mil07] zu folgendem Resultat:

$$\hat{X}_\tau(m) = 10\ log\left(\tilde{G}\right) + X_\tau(m) \qquad \text{mit} \quad \tilde{G} = G(n) \approx const. \tag{2.3.6}$$

Für die Cepstral-Koeffizienten gilt folglich:

$$\hat{c}_\tau(i) = DCT\left\{\hat{X}_\tau(m)\right\} = DCT\left\{10\ log\left(\tilde{G}\right) + X_\tau(m)\right\}. \tag{2.3.7}$$

Durch die Linearität der DCT ist folgende Zerlegung möglich:

$$\hat{c}_\tau(i) = DCT\left\{10\ log\left(\tilde{G}\right)\right\} + c_\tau(i). \tag{2.3.8}$$

Die additive Störung im Merkmalbereich $DCT\left\{10\ log\left(\tilde{G}\right)\right\}$ kann durch eine Mittelwertbefreiung (engl. *Mean Subtraction*) kompensiert werden [Her10]:

$$c_{Ms,\tau} = \hat{c}_\tau - \mathbf{c}_{Mittelwert} \qquad \text{mit} \qquad \mathbf{c}_{Mittelwert} = \frac{1}{T}\sum_{\tau=1}^{T}\hat{c}_\tau. \tag{2.3.9}$$

Nach [Mil07] gilt es jedoch zu beachten, dass eventuell auch Anteile der ungestörten Merkmalvektoren $\vec{o}_n$ gefiltert werden.

2.3.3 Delta-Merkmale

Delta-Merkmale (engl. delta features) enthalten Informationen über die zeitliche Änderung eines Merkmalvektors von einem Analyse-Zeitpunkt zum nächsten. Die Berechnung von Delta-Merkmalen ist durch mehrere sogenannte Regressionsformeln realisierbar.
Die für uns relevante Berechnung ist durch die einfache Differenz gegeben:

$$\vec{o}_n{}' = \frac{\vec{o}_{n+L} - \vec{o}_{n-L}}{2 \cdot L}. \tag{2.3.10}$$

Wir beschränken uns bei der Generierung von MFCC-Delta- und LFCC-Delta-Merkmalvektoren zunächst auf ein Analysefenster mit der Breite $L = 1$.
Die Berechnung der Delta-Koeffizienten innerhalb des UASR-Systems basiert hingegen auf folgender Formel[9]:

$$\vec{o}_n{}' = \sum_{k=-L}^{L} \lambda_k \cdot \vec{o}_{n+k} \qquad \text{mit Koeffizienten } \lambda_k. \tag{2.3.11}$$

Delta-Delta-Merkmale

Die vom UASR-System verwendeten sfv-Merkmale besitzen Koeffizienten, welche in der Literatur auch als Beschleunigungsmerkmale oder Delta-Delta-Merkmale bekannt sind.
Eine Berechnung dieser erfolgt mittels der zweiten zeitlichen Ableitung:

$$\vec{o}_n{}'' = \sum_{k=-L}^{L} \sum_{j=-L}^{L} \lambda_k \cdot \lambda_j \cdot \vec{o}_{n+k+j} \qquad \text{mit Koeffizienten } \lambda_k, \lambda_j. \tag{2.3.12}$$

Supervektoren

Die Bildung von sogenannten „Supervektoren“ $\vec{y}_n$ nach [Wol11] beschreibt die Zusammenfügung einzelner spezifischer Merkmale zu einem Merkmalvektor. In der vorliegenden Arbeit wurden neben Merkmalvektoren mit Original-Koeffizienten ebenfalls Merkmalvektoren, welche mit Delta-Koeffizienten (2.3.10) angereichert sind, untersucht. Der Aufbau der Supervektoren sieht dabei wie folgt aus:

$$\vec{y}_n = \begin{pmatrix} \vec{o}_n(1) \\ \vdots \\ \vec{o}_n(I) \\ \vec{o}_n{}'(1) \\ \vdots \\ \vec{o}_n{}'(I) \end{pmatrix} \qquad \text{mit} \qquad i = 1, \ldots, I \text{ Koeffizientenindex} \tag{2.3.13}$$

[9] Eine Auswahl des korrekten Parameters λ_k in Abhängigkeit der Breite L des Analysefensters kann aus ([Wol11], S. 19) entnommen werden.

3 Modellbildung & Klassifikationverfahren

Im Folgenden wird das GMM-Verfahren *Gaussian Mixture Models* betrachtet, welches in der Arbeit zur Modellbildung und als Klassifikationsverfahren verwendet wurde. Anstatt eines Referenzmusters wird bei diesem Verfahren ein Sprechermodell des zu hinterlegenden Sprechers i gebildet, dass die sprecherspezifischen Eigenschaften darstellt.

3.1 Gaussian Mixture Models (GMM)

3.1.1 Allgemein

Gaussian Mixture Models werden in der Literatur ([Mil07], [RR75], [Roh03]) häufig zur textunabhängigen Sprecherverifikation eingesetzt[10]. Bei der Verwendung von GMM's wird allgemein davon ausgegangen, dass die zeitliche Struktur einer Merkmalvektorfolge nicht entscheidend für die Klassifikation ist. Nach [Roh03] muss deswegen nur die Gesamtverteilungsdichte aller Merkmalvektoren betrachtet werden. Die Zugehörigkeit der Merkmalvektoren zu einer Sprecherklasse wird mittels einer beliebigen Anzahl I von Gaußschen Normalverteilungen beschrieben.

Für eine genauere Betrachtung der einzelnen Arbeitsschritte innerhalb des Verfahrens gehen wir zunächst von einer Merkmalvektorfolge mit eindimensionalen Merkmalvektoren aus:

$$\vec{\mathbf{o}} = (\vec{o}_{1,1}, \vec{o}_{2,1}, \ldots, \vec{o}_{n,1}, \ldots, \vec{o}_{N,1})^\top \qquad \text{mit} \qquad n = 1, \ldots, N \text{ Analyseblock.} \tag{3.1.1}$$

Die Verteilungsdichtefunktion von $\vec{\mathbf{o}}$ wird dabei, wie beispielhaft in der Abbildung 3.1 ersichtlich, durch eine additive Zusammensetzung der einzelnen Gaußverteilungen approximiert.

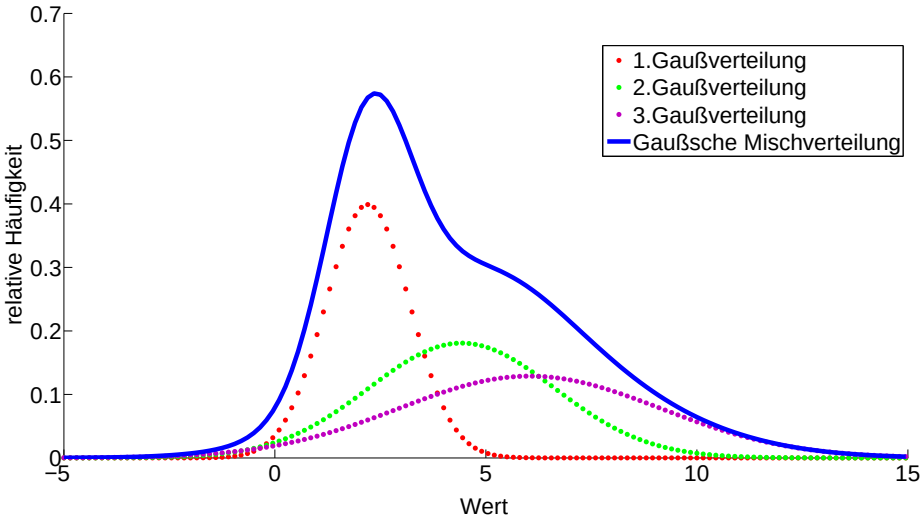

Abb. 3.1. Beispielhafte Verteilungsdichtefunktion einer Merkmalvektorfolge

[10] Eine Verwendung des Verfahrens zur textabhängigen Sprecherverifikation ist jedoch legitim.

Eine eindimensionale Gaußverteilung kann unter den gegebenen Bedingungen durch die folgenden Parameter beschrieben werden:

$$\begin{aligned} p &: Mischungsgewicht, \\ \mu &: Mittelwert, \\ \sigma^2 &: Varianz. \end{aligned} \tag{3.1.2}$$

Liegt eine I-fache Gaußsche Mischverteilung vor, erhalten die drei Parameter jeweils einen Index i, welcher die i-te einzelne Gaußverteilung beschreibt. Diese werden üblicherweise zusammengefasst zu einem Sprecher-/Klassenmodell λ mit:

$$\lambda = \left\{ p_i, \mu_i . {\sigma_i}^2 \right\} \qquad \text{mit} \qquad i = 1, \ldots, I. \tag{3.1.3}$$

Liegen für eine spezifische Sprecherklasse genug statistische Daten vor, kann die Wahrscheinlichkeitsdichte einer Gaußschen Mischverteilung für eine Merkmalvektorfolge, unter Annahme von Ergodizität[11], wie folgt beschrieben werden:

$$p(\vec{\mathbf{o}}|\lambda) = \sum_{i=1}^{I} p_i \cdot b_i(\vec{\mathbf{o}}), \tag{3.1.4}$$

wobei für die Mischungsgewichte

$$\sum_{i=1}^{I} p_i = 1 \tag{3.1.5}$$

gilt.

Eine Einzelverteilung ergibt sich durch folgende Gleichung:

$$b_i(\vec{\mathbf{o}}) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\vec{\mathbf{o}}-\mu_i)^2}{2{\sigma_i}^2}}. \tag{3.1.6}$$

Für jeden autorisierten Sprecher wird in der Trainingsphase mittels Expectation-Maximization (EM)-Algorithmus ein sprecherspezifisches GMM λ_s bestimmt.

[11] Beschreibt in diesem Zusammenhang, dass eine Merkmalvektorfolge $\vec{\mathbf{o}}$ hinreichend genau einen stochastischen Prozess X darstellt.

3.1.2 Expectation-Maximization Algorithmus

Ziel des Lernalgorithmus ist es, durch eine iterative Maximum-Likelihood-Schätzung die Modellparamter $\lambda = \{p_i, \mu_i, {\sigma_i}^2\}$ solange an die Merkmalvektorfolge $\vec{\mathbf{o}}$ der Trainingsdaten anzupassen, bis die Wahrscheinlichkeitsdichte $p(\vec{\mathbf{o}}|\lambda)$ mit dem vorhandenen Gaußschen Mischverteilungs Modell maximal wird.

$$p(\vec{\mathbf{o}}|\lambda) = \prod_{n=1}^{N} p(\vec{o}_n|\lambda) \stackrel{!}{=} Max \tag{3.1.7}$$

Wegen der geringen numerischen Werte bei großer Modellordnung I ist es von Vorteil, die logarithmische Wahrscheinlichkeitsdichte zu berechnen:

$$\log p(\vec{\mathbf{o}}|\lambda) = \sum_{n=1}^{N} \log p(\vec{o}_n|\lambda) \stackrel{!}{=} Max. \tag{3.1.8}$$

In einem ersten Schritt werden für die Initialisierung des GMM's λ beliebige Parameterwerte für p_i, μ_i und ${\sigma_i}^2$ sowie die Anzahl der Verteilungen I unter den geltenden Bedingungen festgelegt [Roh03].

Mit dem Expectation(E)-Schritt berechnen wir zunächst für alle einzelnen Verteilungen $i = 1, \ldots, I$ eine a-posteriori-Wahrscheinlichkeit, die aussagt, wie gut diese zum gesamten Sprechermodell λ passt:

$$p(i|\vec{o}_n, \lambda) = \frac{p_i b_i(\vec{o}_n)}{p(\vec{o}_n|\lambda)} = \frac{p_i b_i(\vec{o}_n)}{\sum_{i=1}^{I} p_i b_i(\vec{o}_n)}. \tag{3.1.9}$$

Die Maximierung der Parameter (M-Schritt) erfolgt über folgende Gleichungen:

Neuberechnung des Mischungsgewichts:

$$p_i = \frac{1}{N} \sum_{n=1}^{N} p(i|\vec{o}_n, \lambda), \tag{3.1.10}$$

Neuberechnung des Mittelwerts:

$$\mu_i = \frac{\sum_{n=1}^{N} p(i|\vec{o}_n, \lambda) \cdot \vec{o}_n}{\sum_{n=1}^{N} p(i|\vec{o}_n, \lambda)}, \tag{3.1.11}$$

Neuberechnung der Varianz:

$${\sigma_i}^2 = \frac{\sum_{n=1}^{N} p(i|\vec{o}_n, \lambda) \cdot \vec{o}_n^{\,2}}{\sum_{n=1}^{N} p(i|\vec{o}_n, \lambda)} - {\mu_i}^2. \tag{3.1.12}$$

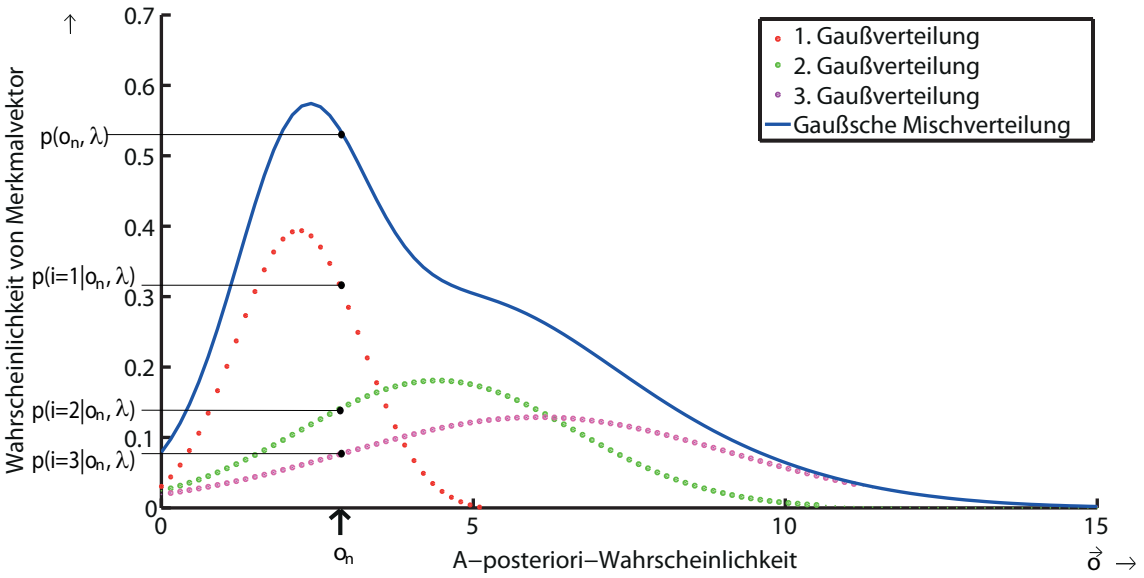

Abb. 3.2. A-posteriori-Wahrscheinlichkeit

Das neu berechnete Sprechermodell wird anschließend in einer Klassifikation dem alten Modell gegenübergestellt. Es gilt nach [Mil07] folgender Zusammenhang zwischen dem alten Modell λ und dem neuen Modell $\overline{\lambda}$:

$$p(\vec{\mathbf{o}}|\overline{\lambda}) \geq p(\vec{\mathbf{o}}|\lambda). \tag{3.1.13}$$

Die Neuberechnung der Parameter wird solange iterativ fortgesetzt, bis ein vordefiniertes Abbruchkriterium erfüllt ist. Dieses könnte nach [Mil07] wie folgt aussehen:

$$\Theta < 10^{-6} \Rightarrow \ \overline{\lambda} \text{ ist optimales Modell} \qquad \text{mit } \Theta = \left| \frac{\log p(\vec{\mathbf{o}}|\overline{\lambda})}{\log p(\vec{\mathbf{o}}|\lambda)} - 1 \right|. \tag{3.1.14}$$

Die Vorgabe einer Anzahl auszuführender Iterationsschritte ist in vielen Implementierungen ebenfalls als Abbruchkriterium realisiert.

GMM Berechnung für N-dimensionale Merkmalvektoren

Die bisherigen Berechnungen beruhten auf der Annahme einer eindimensionalen Merkmalvektorfolge $\vec{\mathbf{o}}$. Wie bereits in den vorigen Abschnitten erläutert, besteht ein Merkmalvektor $\vec{o}_n$ jedoch normalerweise aus 10 bis 30 Koeffizienten oder mehr[12]. Dementsprechend werden die Formeln zur Berechnung der Parameter sowie der Wahrscheinlichkeit einer Einzelverteilung nach [Roh03] modifiziert.

Einzelverteilung:

$$b_i(\vec{o}_n) = \frac{1}{2\pi^{Dim/2} \cdot |\mathcal{C}_{\vec{\mathbf{o}}\vec{\mathbf{o}},i}|^{1/2}} \cdot e^{-\frac{1}{2}\cdot(\vec{o}_n-\mu_i)\cdot\mathcal{C}^{-1}_{\vec{\mathbf{o}}\vec{\mathbf{o}},i}\cdot(\vec{o}_n-\mu_i)^\top} \tag{3.1.15}$$

A-posteriori-Wahrscheinlichkeit:

$$p(i|\vec{o}_n,\lambda) = \frac{p_i b_i(\vec{o}_n)}{\sum_{i=1}^{I} p_i b_i(\vec{o}_n)} \tag{3.1.16}$$

Mischungsgewichte:

$$p_i = \frac{1}{N}\sum_{n=1}^{N} p(i|\vec{o}_n,\lambda) \tag{3.1.17}$$

Mittelwerte:

$$\mu_i = \frac{\sum_{n=1}^{N} p(i|\vec{o}_n,\lambda)\cdot\vec{o}_n}{\sum_{n=1}^{N} p(i|\vec{o}_n,\lambda)} \tag{3.1.18}$$

Kovarianzmatrizen:

$$\mathcal{C}_{\vec{\mathbf{o}}\vec{\mathbf{o}},i} = \frac{\sum_{n=1}^{N} p(i|\vec{o}_n,\lambda)\cdot\vec{o}_n^{\,\top}\cdot\vec{o}_n}{\sum_{n=1}^{N} p(i|\vec{o}_n,\lambda)} - \mu_i^\top\cdot\mu_i \tag{3.1.19}$$

Für (3.2.15) bis (3.2.19) gilt:

$$\begin{aligned} \vec{o}_n &\in \mathbb{R}^{1\times D} \\ \mu_i &\in \mathbb{R}^{1\times D} \\ \mathcal{C}_{\vec{\mathbf{o}}\vec{\mathbf{o}},i} &\in \mathbb{R}^{1\times D\times D} \\ b_i(\vec{o}_n) &\in \mathbb{R}^{1\times 1} \\ p_i &\in \mathbb{R}^{I\times 1} \end{aligned} \tag{3.1.20}$$

.

D beschreibt in den genannten Gleichungen die Anzahl der Koeffizienten/Dimensionen eines Merkmalvektors.

[12] Anreicherung durch Delta-Koeffizienten.

3.1.3 Klassifikation

Die in den vorliegenden Literaturstellen am häufigsten verwendete Klassifikationsvariante in Verbindung mit Gaussian Mixture Models ist in (3.1.21) dargestellt. Sie basiert auf der Voraussetzung, dass eine Merkmalvektorfolge $\vec{\mathbf{o}}$ einer Sprecherklasse s_i mit $i = 1, \ldots, I$ zugeordnet wird. Die Möglichkeit der Zurückweisung $\nu = 0$ ist damit ausgeschlossen. Des Weiteren wird in den Literaturstellen üblicherweise die Sprecherklasse mit der höchsten Wahrscheinlichkeitsdichte ausgewählt (siehe Abschnitt 1.4):

$$\nu = \underset{s_i=1\leq i\leq I}{\arg\max}\ p(\vec{\mathbf{o}}|\lambda_{s_i}). \tag{3.1.21}$$

Durch die Verwendung der negativen logarithmischen Likelihoods ergibt sich

$$\nu = \underset{s_i=1\leq i\leq I}{\arg\min}\ -\log p(\vec{\mathbf{o}}|\lambda_{s_i}). \tag{3.1.22}$$

In der Annahme, dass die Merkmalvektoren der Merkmalvektorfolge unabhängig voneinander sind folgt:

$$\nu = \underset{s_i=1\leq i\leq I}{\arg\min}\ -\log p(\vec{\mathbf{o}}|\lambda_{s_i}) = \underset{s_i=1\leq i\leq I}{\arg\min}\ -\sum_{n=1}^{N}\log p(\vec{o}_n|\lambda_{s_i}). \tag{3.1.23}$$

Diese Variante der Klassifikation führt die besprochenen Prozesse, Identifikation und Verifikation des Sprechers, zusammen aus. Der ausgewählte Sprecher i wird immer angenommen $v = 1$. Das zu entwickelnde System hat jedoch einen Sprecher, der eine ungültige Passphrase spricht oder nicht im System hinterlegt ist, abzuweisen.
Eine Trennung des Identifikations- und Verifikationsprozesses, wie sie in dieser Arbeit erfolgte, wird in Abschnitt 4.2.1 erläutert. Außerdem werden in Abschnitt 4.2.3 zwei Klassifikationsvarianten vorgestellt, die eine Zurückweisung $\nu = 0$ ermöglichen.

4 Datenbasis & Prototyp

Dieses Kapitel stellt eine Übersicht zur Erstellung und Struktur des Sprachkorpus dar. Anschließend folgt eine Beschreibung der einzelnen Verarbeitungsschritte des in Matlab implementierten Prototyps.

4.1 Datenbasis

4.1.1 Sprachaufnahmen

Die Aufnahme der Testdaten erfolgte im Sprachlabor des Lehrstuhls Kommunikationstechnik. Als Aufnahmemikrofon wurde ein *Rode NT-1 A*, mit vorangestelltem Windschutz, in Kombination mit einer externen Soundkarte (*Focusrite Scarlett 8i6*) verwendet. Die mitgelieferte Software *Scarlett Mixcontrol* diente dabei der Kontrolle des Aussteuerungspegels. Eine Übersteuerung der Audio-Signale wurde durch die Wahl des *Headrooms* auf -3 db bis -6 db verhindert.

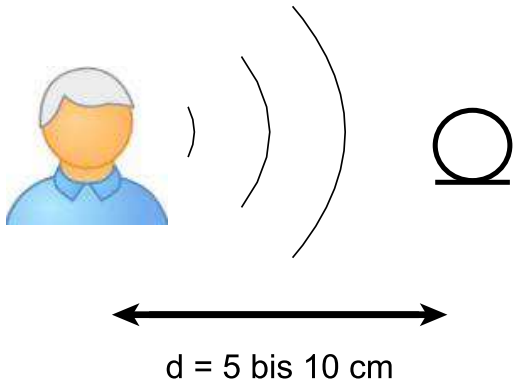

Abb. 4.1. Aufnahme des Probanden mit Abstandsangabe d zwischen Mikrofon und Sprecher.

Im Vorfeld der Sprachaufnahmen wurden die Probanden gebeten, eine Einverständniserklärung zu unterschreiben. Diese beinhaltet die Zusage des jeweiligen Probanden, dass seine Sprachaufzeichnungen für weitere wissenschaftliche Experimente am Lehrstuhl Kommunikationstechnik verwendet werden dürfen[13].

[13] Ein Musterexemplar der Einverständniserklärung befindet sich in Anhang A.1

4.1.2 Korpus

Insgesamt sind für die vorliegende Arbeit vier Sets mit einer Anzahl von 3719 Sprachturns aufgenommen worden. Das Set v001 diente für den ersten Probelauf eines rudimentären DTW-Erkenners[14], welcher zur Einarbeitung in das Thema Sprechererkennung implementiert wurde. Set v002 beinhaltet Sprachaufnahmen, die durch ein Headset der Firma Sennheiser realisiert wurden. Wegen eines zu geringen Pegels konnten sie jedoch nicht weiter verwendet werden.Die Sets v003 und v004 beinhalten 2340 Sprachturns von 31 Probanden und wurden in den Tests (siehe Kapitel 5) verwendet.
Die Einteilung der Probanden erfolgte zunächst in die Gruppen *-autorisierte Person-* oder *-nicht autorisierte Person-*.

Gruppe	männlich	weiblich
autorisiere Person	8	4
nicht autorisierte Person	11	8

Tabelle 4.1: Verteilung der Probanden auf die Gruppen *-autorisierte Person-* und *-nicht autorisierte Person-*

Ein Proband der Gruppe -nicht autorisierte Person- wurde gebeten, jeweils fünfmal die korrekte Passphrase einer autorisierten Person zu wiederholen. Dies ergab bei 12 autorisierten Personen insgesamt 60 Sprachturns pro Proband der Gruppe -nicht autorisierte Person- im Set v003.
Die Sprachturns der Probanden mit der Gruppenzugehörigkeit -autorisierte Person- sind an zwei unterschiedlichen Terminen aufgenommen worden[15]. An beiden Terminen wurde der Proband gebeten, jeweils 30 mal die eigene Passphrase zu wiederholen. Außerdem sind zu jedem Termin 20 weitere Phrasen aufgenommen worden, die keine korrekte Passphrase enthalten. Jedem Probanden aus der Gruppe -autorisierte Person- sind dadurch jeweils 50 Sprachturns im Set v003 und 50 weitere im Set v004 zuzuordnen.
Die korrekten Passphrasen der textabhängigen Autorisierung besitzen folgende Struktur:

Autorisierung - Nachname der Person - erstes Wort - zweites Wort[16]

Eine Aufteilung in Trainings-, Entwicklungs- und Teststichprobe wird erst in einem definierten Testfall (siehe Kapitel 5) vorgenommen.

[14] Die Klassifikation und Arbeitsweise eines DTW-Erkenners wird in [Fli95] erläutert.
[15] Der zeitliche Abstand zwischen den Aufnahmeterminen für Set v003 und Set v004 betrug rund einen Monat.
[16] Das erste und zweite Wort stammen aus dem deutschen Funkalphabet $DIN5009$, welches im Anhang A.2 aufgeführt ist.

4.2 Aufbau

Verzeichnisstruktur

Als Verzeichnisstruktur der zu hinterlegenden Daten wurde eine Äquivalenz zum UASR-System geschaffen; eine allgemeine Übersicht ist dem Anhang B zu entnehmen. Signale und generierte Daten werden dabei in Unterordnern des Verzeichnisses ...*vau**common* abgelegt. Zwischen den Signalen und den daraus generierten Daten ist die Bezeichnung identisch[17].

Die Dateibezeichnung ist wie folgt aufgebaut: *SSS_AUT_PP_NNN.format*
Sprachturns mit unkorrekten Passphrasen: *SSS_____NNN.format*

SSS	Sprecher-ID	drei Großbuchstaben, z.B. *PGE* für „Peter Geßler“
AUT	Auth-ID	drei Großbuchstaben der Authentifizierung, z.B. *PGE* für „Peter Geßler“
PP	Phrasen-ID	Großbuchstaben der Passphrase, z.B. *AT* für „Anton Theodor“
NNN	Wiederholung	Nummer der Wiederholung, gegebenenfalls links mit Nullen auffüllen
	Sonderzeichen	Ä → 1 Ö → 2 Ü → 3 ß → 4 Ch → 5 Sch → 6

Tabelle 4.1: Namenskonvention von Sprachsignalen und generierten Dateien.

Realisierung von Identifikationsprozess und Verifikationsprozess

Die Entwicklung des Stimmauthentifizierungssystems für das UASR-System wird insofern erleichtert, als dass jeder Sprecher eine eigene geordnete Passphrase besitzt.
Eine Überprüfung auf Korrektheit der gesprochenen Passphrase und damit die Zuordnung zu einem hinterlegten Sprechermodell λ_{s_i} kann durch diese Bedingung über den bereits implementierten Spracherkenner des UASR-Systems erfolgen. Der Identifikationsprozess wird somit vom UASR-System übernommen. Eine Verifikation des Sprechers kann anschließend beispielsweise über die Klassifikation mittels Rückweisungsschwellwert oder Backoff-Modell (siehe Abschnitt 4.2.3) vorgenommen werden. Für den implementierten Prototyp wurde in den ersten Tests die Voraussetzung gestellt, dass eine optimale Sprecheridentifikation stattgefunden hat und somit nur korrekte Passphrasen geprüft wurden. Diese Voraussetzung wurde, um die Leistungsfähigkeit des Prototyps zu bewerten, in einem abschließenden Testszenario zurückgestellt.

[17] Eine Relation zwischen den Daten ist somit ersichtlich.

4.2.1 Generierung von Merkmalen

Eine Generierung der „Baseline“ Merkmale MFCC und LFCC erfolgte mittels der Matlab-Toolboxen *Signal Processing* [Dev], *netlab* [Nab] und *voicebox* [Bro]. Die Segmentierung eines Sprachsignals wurde bereits ausführlich in Abschnitt 2.1.2 erläutert. Zu erwähnen gilt, dass beiden Verfahren eine Präemphase (Höhenanhebung) mit dem Koeffizienten $\alpha = 0{,}97$ vorausging. Die Möglichkeit der Nachbearbeitung wurde durch die Implementierung der einzelnen Algorithmen ebenfalls realisiert.

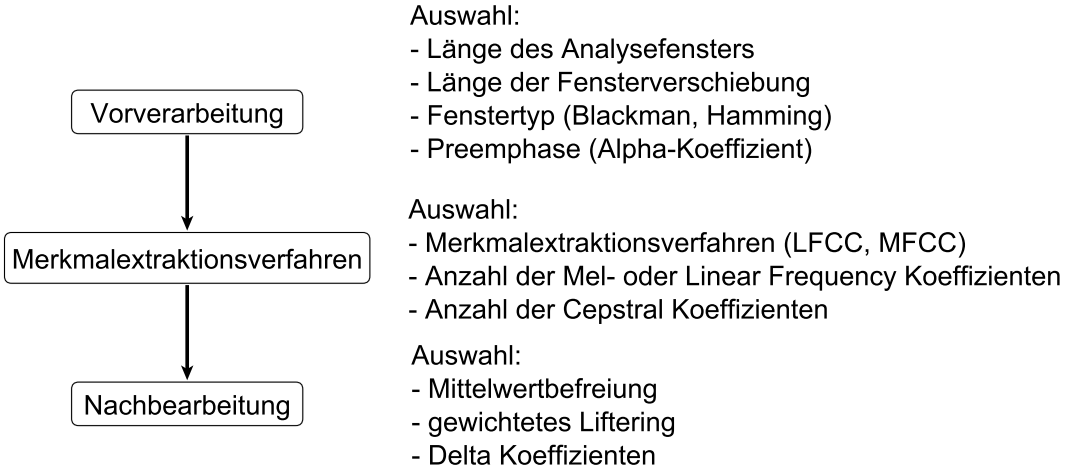

Abb. 4.4. Auswahl der Einstellungen für die Merkmalanalyse des Prototypen

Eine Kombination der optionalen Verfahren zur Nachbearbeitung ist ebenfalls möglich.

Die generierte Merkmalvektorfolge $\vec{\mathbf{o}}$ eines Sprachsignals $x(t)$ wird anschließend im Text-Format hinterlegt. Eine getroffene Auswahl der im Bezug auf ihre sprecherspezifischen Eigenschaften getesteten Merkmale befindet sich in Abschnitt 4.3.

4.2.2 Berechnung eines Sprechermodells

Ein spezifisches Sprechermodell λ_{s_i} wird mittels Gaussian Mixture Modells dargestellt. Als Trainingsdaten dienten 10 Sprachsignale des autorisierten Sprechers mit korrekter Passphrase, die nach der Merkmalanalyse zu einer Merkmalvektorfolge zusammengefasst wurden. Die Modellinitialisierung sowie das Modelltraining (EM-Algorithmus mit 10 Iterationsschritten) wurden mit der netlab-Toolbox [Nab] durchgeführt.
Neben den verschiedenen Merkmaltypen sollte ebenfalls die Auswirkung einer Variation der Gaußschen Mischverteilungsanzahl I auf das Erkennungsergebnis betrachtet werden. Es wurden dementsprechend die Modellordnungen $I = 1/2/3/4$ ausgewählt.

Untersuchungen von [RR75] zeigen in anderen Literaturstellen ([Mil07],[Roh03]) indes, dass erst ab einer Ordnung von $I \geq 16$ eine Verschlechterung des Erkennungsergebnisses[18] eintritt.
Zum einen liegt dies an einem Übertraining des Sprechermodells, andererseits an den zu geringen Trainingsdaten. Die zeitliche Länge der verwendeten Daten zur Berechnung des Sprechermodells, im implementierten System, sind jedoch teilweise um 75% geringer gegenüber denen der Literaturstellen. Ein Versuch der Bildung eines Sprechermodells mit der Modellordnung $I = 8$ scheiterte an zu kleinen Determinanten der Kovarianzmatrizen in der Matlab- Implementierung.

4.2.3 Klassifikation

Klassifikation mit Rückweisungsschwellwert

Die Bestimmung des globalen Rückweisungsschwellwerts d_{rej} erfolgt durch eine Klassifikation auf einer Entwicklungsstichprobe. Diese ist zur Trainings- und Teststichprobe disjunkt. Der Schwellwert wurde angepasst, bis eine Gleichfehlerrate (EER) auf der Entwicklungsstichprobe vorlag. Zu beachten gilt, dass der Rückweisungsschwellwert separat für jeden Merkmaltyp und der dazugehörigen Modellordnung bestimmt werden muss.
Des Weiteren wird diese Variante nur für die Verifikation eines Sprechers eingesetzt. Die Auswahl des korrekten Modells würde über den Spracherkenner des UASR-Systems erfolgen.

Durch die Verwendung von negativen logarithmischen Likelihoods sowie einer Normierung auf die Anzahl N der Merkmalvektoren innerhalb einer Merkmalvektorfolge $\vec{\mathbf{o}}$ erhalten wir folgende Gleichung für die Klassifikation des Verifikationsprozesses:

$$\nu = \begin{cases} 1 \ , \ \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_i})}{N} < d_{rej} \\ \\ 0 \ , \ \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_i})}{N} > d_{rej} \end{cases} . \tag{4.2.1}$$

Klassifikation mit Backoff-Modell

Neben der Klassifikation mittels globalem Rückweisungsschwellwert d_{rej} wurde eine weitere Klassifikationsvariante, die des Backoff-Modells, implementiert. Das Backoff-Modell ist ebenfalls ein GMM und wird mit allen Trainingsdaten der hinterlegten Sprechermodelle λ_{s_i} trainiert. Diese Klassifikationsvariante ermöglicht es dem implementierten Prototypen, die Identifikation sowie Verifikation des Sprechers vorzunehmen.
Für die Identifikation eines Sprechers wird zunächst das hinterlegte Sprechermodell λ_{s_i} ausgewählt, welches die geringste Wahrscheinlichkeitsdichte gegenüber der Merkmalvektorfolge aufweist.

[18] Unter Beibehaltung der restlichen Modellparameter.

$$\nu = \underset{s_i=1 \leq i \leq I}{\arg\min} \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_i})}{N} \tag{4.2.2}$$

Bei der Verifikation wird die Wahrscheinlichkeitsdichte der ausgewählten Identität mit der Wahrscheinlichkeitsdichte des Backoff-Modells, welches im Folgenden als λ_{s_0} gekennzeichnet ist, verglichen. Gilt dabei:

$$\frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_0})}{N} < \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_i})}{N} \tag{4.2.3}$$

wird das Backoff-Modell als Identität ausgewählt. Sollte dies der Fall sein, findet beim Verifikationsprozess eine Rückweisung statt.

$$\nu = \begin{cases} 1\ , \ \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_0})}{N} > \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_i})}{N} \\ \\ 0\ , \ \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_0})}{N} < \frac{-\sum_{n=1}^{N} \log p(\vec{o}_n|\lambda_{s_i})}{N} \end{cases} \tag{4.2.4}$$

Neben der Ausführung beider Prozesse ist auch eine getrennte Betrachtung möglich. Wird der Identifikationsprozess, wie bei der Klassifikation mittels Rückweisungsschwellwert, durch das UASR-System ausgeführt, kann ebenfalls in einer zweiten Stufe die Überprüfung zu einem Backoff-Modell λ_{s_0} stattfinden.

4.3 Merkmale

4.3.1 Generierte Merkmale

Die Auswahl von Merkmalen, welche sich eventuell für die Sprechererkennung eignen, gestaltete sich durch die Anzahl an Kombinationsmöglichkeiten von Schritten zur Nachbearbeitung (siehe Abschnitt 2.3) als schwierig. Dementsprechend werden neben den Merkmalen aus dem UASR-System ausschließlich die unten aufgelisteten Merkmalkombinationen untersucht:

- **(1) mfcc_30 und lfcc_30 Merkmale**
  Zunächst werden die generierten MFCC- und LFCC-Merkmale ohne Modifikation getestet. Die Anzahl der Koeffizienten eines Merkmalvektors wurde an die der pfv-Merkmale des UASR-Systems angepasst.

- **(2) mfcc_ms_30 und lfcc_ms_30 Merkmale**
  Da in der Vorverarbeitung keine Kanalkompensation erfolgte, wird eine Mittelwertbefreiung nach Abschnitt 2.3.2 durchgeführt.

- **(3) mfcc_ms_delta_60 und lfcc_ms_delta_60 Merkmale**
  Das verwendete Verfahren zur Modellberechnung *GMM* beachtet grundsätzlich nicht die zeitliche Struktur der Merkmalvektorfolgen. Eine Anreicherung mit Delta-Koeffizienten (siehe Abschnitt 2.3.3) ist somit als sinnvoll anzunehmen.
  Die Anzahl der Dimensionen eines Merkmalvektors beträgt hier $D = 60$.

- **(4) mfcc_ms_lift_30 und lfcc_ms_lift_30 Merkmale**
  Durch ein gewichtetes Liftering der in (2) generierten Merkmale erreichen wir ein Abschwächen der unteren Cepstral-Koeffizienten. Nach den Ausführungen von [Tal95] und [Ros02] ist folglich die Annahme zulässig, dass durch die „Verstärkung“ der höheren Cepstral-Koeffizienten die sprecherspezifische Teile des Vokaltraktes besser reflektiert werden. In Abschnitt 2.2.1 bis 2.3.2 sind die entsprechenden Gleichungen notiert.

4.3.2 Merkmale vom UASR-System

Die durch das UASR-System generierten Merkmale lassen sich in die drei Gruppen *Primary Feature Vectors* (pfv), *Secondary Feature Vectors* (sfv) und *Least Significant Feature Vectors* (lfv) einteilen. Die sfv-Merkmale wurden dabei explizit für die Spracherkennung entwickelt.
Dieser Abschnitt soll nur eine kurze Übersicht zu den Merkmalen des UASR-Systems geben. Eine ausführliche Dokumentation über die geltenden mathematischen Algorithmen zur Generierung dieser Merkmale befindet sich in [Wol11].

- **(1) pfv_30**
  Die pfv-Merkmale basieren im Allgemeinen auf den vorgestellten MFCC-Merkmalen. Zunächst erfolgt eine Präemphase des abgetasteten Signals $x(k)$. Anschließend wird das diskrete Signal - analog zu der Beschreibung in Abschnitt 2.1.2 - segmentiert und gefenstert. Mittels Fast Fourier Transformation (FFT) erhalten wir das frequenzdiskrete Spektrum. Durch die Bildung des Betragsspektrums wird ein ausschließlich reelles Spektrum erhalten. Im Gegensatz zu den vorgestellten MFCC-Merkmalen wird vor der Bewertung mit den Dreiecksbandpässen eine Logarithmierung des Betragsspektrums zur Basis 10 durchgeführt. Das logarithmierte Spektrum wird anschließend mit 30 Dreiecksbandpässen bewertet. Als letzter Analyseschritt wird eine cepstrale Glättung vorgenommen. Diese eliminiert die sprecherspezifischen Frequenzen innerhalb des Signals.
  Eine mögliche Realisierung des Vorgehens stellt die Tiefpassfilterung des Spektrums dar.

- **(2) sfv_24**
  Für die Spracherkennung werden innerhalb des UASR-Systems die sfv-Merkmalvektoren benutzt. Die pfv-Merkmalvektoren werden dabei mit dynamischen Merkmalen angereichert. Dynamische Merkmale sind Delta- und Delta-Delta und Kontext-Merkmale. Ein Merkmalvektor besitzt somit temporär $D = 60$ Koeffizienten. Nach einer *Hauptkomponentenanalyse* (engl. principal component analysis, PCA) werden die Koeffizienten mit zu geringer Streuungserklärung gelöscht. Ein Merkmalvektor enthält nach der Generierung noch 24 Koeffizienten.

- **(3) lfv_36**
  Least Significant Feature Vectors stellen die Merkmale dar, welche wegen der zu kleinen Streuungserklärung bei den sfv-Merkmalen gelöscht wurden. Diese könnten sprecherspezifische Eigenschaften enthalten. Die Anzahl der Koeffizienten beträgt dabei $D = 36$.

5 Tests und Auswertung

Das entwickelte Simulationssystem sollte unter Bedingungen getestet werden, welche möglichst realistisch gegenüber der späteren Arbeitsumgebung sind. In der Praxis ist davon auszugehen, dass die 10 Sprachturns zum Training des Sprechermodells innerhalb eines kurzen Zeitabschnitts aufgenommen werden.
Auch wenn die Probanden der Gruppe -autorisierte Person- gebeten wurden, eine Variation ihrer Stimmenlage innerhalb der einzelnen Sprachturns vorzunehmen, kann dies in der Test- sowie Entwicklungsstichprobe nicht als Ersatz für Sprachturns, welche zu unterschiedlichen Terminen aufgenommen worden, gelten.
Die Testergebnisse sowie deren Bewertung sind somit nur in dem vorgegeben Rahmen haltbar.

Vertrauensbereich

Die Angabe eines Vertrauensbereichs von Detektionsergebnissen, der in der Literatur auch als Konfidenzintervall bekannt ist, wurde ausführlich in der Arbeit von [Foe14] diskutiert. Es genügt uns an dieser Stelle, die Berechnung des 95% Konfidenzintervall c_{95} zu übernehmen, welches für die Evaluation verwendet wurde.

$$c_{95} \approx \pm 2\sqrt{\frac{m - m^2}{N - 1}} \qquad \text{mit } N = \text{Stichprobengröße} \tag{5.0.1}$$

m stellt das Verhältnis der korrekt klassifizierten Werte $0 \leq C \leq N$ gegenüber der Stichprobengröße N dar:

$$m = \frac{C}{N} \tag{5.0.2}$$

Für eine gute Schätzung sollten die Bedingungen $C < 50 < N - C$ eingehalten werden.

5.1 Testszenario 1 (rudimentäre Identifikation und Verifikation mit Prototyp)

In einem ersten Schritt wurde der Identifikations- und Verifikationsprozess mit der in Literaturstellen vorgegebenen Klassifikationsvariante nach (3.1.23) getestet.

Test 1.1

Als Testdaten wurden zunächst ausschließlich Sprachturns der autorisierten Personen verwendet, die eine korrekte Passphrase enthalten. In der Literatur ist diese Variante als „closed-Set Klassifikation“ bekannt. Der Test wurde unter den folgenden Bedingungen durchgeführt:

- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Sprachset v003, die eine korrekte Passphrase des autorisierten Sprechers enthalten, trainiert.
- **Teststichprobe**
  20 Sprachturns pro autorisierter Person mit korrekter Passphrase vom Sprachset v003.
- **Klassifikation**
  Die Klassifikation des Identifikationsprozesses erfolgt - wie in (3.1.23) beschrieben - durch die Auswahl der Sprecherklasse s_i, dessen Sprechermodell λ_i eine minimale Wahrscheinlichkeitsdichte gegenüber der vorliegenden Merkmalvektorfolge des unbekannten Sprechers aufweist.

Ergebnis der Klassifikation

Modelordnung → Merkmaltyp ↓	1	2	3	4
pfv_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
sfv_24	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
lfv_36	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
mfcc_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
lfcc_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
mfcc_ms_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
lfcc_ms_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
mfcc_ms_delta_60	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
lfcc_ms_delta_60	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
mfcc_ms_lift_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$
lfcc_ms_lift_30	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$	$(100 \pm 0)\%$

Tabelle 5.1 : Ergebnis von Test 1.1 / Angabe der Erkennungsrate

Test 1.2

Die Trainings- und Testdaten in Test 1.1 stammen aus dem Sprachset v003. Es folgte eine Untersuchung, inwieweit sich die Erkennungsrate beim Austausch der Teststrichprobe (Ersetzung von v003 durch v004) gegenüber Test 1.1 unterscheidet. Hier gilt es zu erwähnen, dass diverse Probanden keinen optimalen Gesundheitszustand vorwiesen. Der Test wurde unter den folgenden Bedingungen durchgeführt:

- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Sprachset v003, die eine korrekte Passphrase des autorisierten Sprechers enthalten, trainiert.
- **Teststichprobe**
  20 Testdaten pro autorisierter Person mit korrekter Passphrase vom Sprachset v004.
- **Klassifikation**
  Die Klassifikation erfolgt äquivalent zum Test 1.1.

Ergebnis der Klassifikation

Modelordnung → Merkmaltyp ↓	1	2	3	**4**
pfv_30	(77,92 ± 5,37)%	(82,50 ± 4,92)%	(81,67 ± 5,00)%	(82,92 ± 4,87)%
sfv_24	(84,58 ± 4,67)%	(85,83 ± 4,51)%	(82,92 ± 4,87)%	(81,25 ± 5,05)%
lfv_36	(83,33 ± 4,82)%	(89,17 ± 4,02)%	(91,67 ± 3,58)%	(90,00 ± 3,88)%
mfcc_30	(79,17 ± 5,25)%	(79,17 ± 5,25)%	(75,83 ± 5,54)%	(78,33 ± 5,33)%
lfcc_30	(70,00 ± 5,93)%	(71,67 ± 5,83)%	(83,33 ± 4,82)%	(83,33 ± 4,82)%
mfcc_ms_30	(95,00 ± 2,82)%	(94,17 ± 3,03)%	(86,67 ± 4,40)%	(91,25 ± 3,66)%
lfcc_ms_30	(94,58 ± 2,93)%	(84,58 ± 4,67)%	(92,50 ± 3,41)%	(95,42 ± 2,71)%
mfcc_ms_delta_60	(94,58 ± 2,93)%	(95,00 ± 2,82)%	(93,33 ± 3,23)%	(92,92 ± 3,32)%
lfcc_ms_delta_60	(97,50 ± 2,02)%	(91,67 ± 3,58)%	(96,67 ± 2,32)%	(**97,92 ± 1,85**)%
mfcc_ms_lift_30	(95,00 ± 2,82)%	(95,00 ± 2,82)%	(91,67 ± 3,58)%	(92,50 ± 3,41)%
lfcc_ms_lift_30	(94,58 ± 2,93)%	(85,42 ± 4,57)%	(91,67 ± 3,58)%	(95,00 ± 2,82)%

Tabelle 5.2 : Ergebnis von Test 1.2 / Angabe der Erkennungsrate

5.2 Testszenario 2 (Verifikationsprozess mit Rückweisungsschwellwert)

Dieses Testszenario stellt den grundlegenden Verifikationsprozess des Stimmenauthentifizierungssystems mit globalen Rückweisungsschwellwert d_{rej} dar. Die Bestimmung des Rückweisungsschwellwertes wurde nach Abschnitt 4.2 durchgeführt[19]. Eine Entwicklungsstichprobe setzt sich dabei aus 10 Sprachturns mit korrekter Passphrase pro autorisierten Sprecher und zwei Sprachturns pro nicht autorisierten Sprecher mit korrekter Passphrase jedes autorisierten Sprechers zusammen. Die Gesamtgröße beträgt somit 576 Sprachturns von Set v003.

Test 2.1

Wir gehen von einer optimalen Sprecheridentifikation durch den Spracherkenner des UASR-Systems aus. Das vorliegende Ergebnis stellt die Erkennungsrate unter Verwendung eines globalen Rückweisungsschwellwertes d_{rej} dar. Dieser wurde anhand der Equal Error Rate auf der Entwicklungsstichprobe für jede Modellordnung und jeden Merkmaltyp ermittelt. Der Test wurde unter den folgenden Bedingungen durchgeführt:

- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Set v003, die eine korrekte Passphrase eines autorisierten Sprechers enthalten, trainiert.
- **Entwicklungsstichprobe**
  Die Zusammensetzung der Entwicklungsstichprobe wurde am Anfang des Testszenarios erläutert.
- **Teststichprobe**
  10 Sprachturns pro autorisierter Person, mit korrekter Passphrase vom Sprachset v003.
  Des Weiteren wurden zwei Sprachturns, mit korrekter Passphrase jeder autorisierten Person pro nicht autorisierter Person von v003 hinzugefügt. Die Teststichprobengröße beträgt somit 576 Sprachturns.
- **Klassifikation**
  Klassifikation des Sprecherverifikationsprozesses mit globalen Rückweisungsschwellwert d_{rej} nach (4.2.1).

[19] Die spezifischen Rückweisungsschwellwerte und Gleichfehlerraten auf der Entwicklungsstichprobe können in Anhang C eingesehen werden

Ergebnis der Klassifikation mit eingestellten Rückweisungsschwellwert

Modelordnung → Merkmaltyp ↓	1	2	**3**	**4**
pfv_30	(95,49 ± 1,73)%	(95,83 ± 1,67)%	(96,53 ± 1,53)%	(96,70 ± 1,49)%
sfv_24	(92,88 ± 2,14)%	(96,18 ± 1,60)%	(97,22 ± 1,37)%	(97,92 ± 1,20)%
lfv_36	(93,40 ± 2,07)%	(95,14 ± 1,79)%	(95,49 ± 1,73)%	(95,31 ± 1,76)%
mfcc_30	(95,49 ± 1,73)%	(95,83 ± 1,67)%	(96,35 ± 1,56)%	(97,05 ± 1,41)%
lfcc_30	(96,18 ± 1,60)%	(96,53 ± 1,53)%	(96,35 ± 1,56)%	(96,53 ± 1,53)%
mfcc_ms_30	(93,23 ± 2,10)%	(94,97 ± 1,82)%	(96,01 ± 1,63)%	(96,88 ± 1,45)%
lfcc_ms_30	(95,14 ± 1,79)%	(96,70 ± 1,49)%	(**98,26 ± 1,09**)%	(**98,26 ± 1,09**)%
mfcc_ms_delta_60	(91,32 ± 2,35)%	(94,44 ± 1,91)%	(95,49 ± 1,73)%	(97,05 ± 1,41)%
lfcc_ms_delta_60	(92,88 ± 2,14)%	(95,49 ± 1,73)%	(96,01 ± 1,63)%	(**98,26 ± 1,09**)%
mfcc_ms_lift_30	(93,23 ± 2,10)%	(95,31 ± 1,76)%	(96,35 ± 1,56)%	(97,74 ± 1,24)%
lfcc_ms_lift_30	(95 14 ± 1,79)%	(96,53 ± 1,53)%	(97,17 ± 1,19)%	(**98,26 ± 1,09**)%

Tabelle 5.3 : Ergebnis von Tests 2.1 / Angabe der Erkennungsrate

Test 2.2

Im folgenden Test wurden die 10 korrekten Sprachturns pro autorisierten Sprecher vom Set v003 mit 10 korrekten Sprachturns pro autorisierten Sprecher vom Set v004 getauscht. Die globalen Rückweisungsschwellwerte d_{rej} wurden von Test 2.1 übernommen. Der Test wurde unter den folgenden Bedingungen durchgeführt:

- **optimale Sprecheridentifikation**
  Bestehende Bedingung wird aus Test 2.1 übernommen.
- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Set v003, die eine korrekte Passphrase eines autorisierten Sprechers enthalten, trainiert.
- **Entwicklungsstichprobe**
  Die Zusammensetzung der Entwicklungsstichprobe ist äquivalent zu Test 2.1.
- **Teststichprobe**
  10 Sprachturns pro autorisierter Person, mit korrekter Passphrase vom Set v004.
  Außerdem wurden zwei Sprachturns pro nicht autorisierter Person mit korrekter Passphrase jeder autorisierten Person vom Set v003 hinzugefügt.
- **Klassifikation**
  Klassifikation des Sprecherverifikationsprozesses mit globalen Rückweisungsschwellwert d_{rej} nach (4.2.1).

Ergebnis der Klassifikation mit Rückweisungsschwellwert

Modelordnung → Merkmaltyp ↓	1	2	3	**4**
pfv_30	(80,73 ± 3,29)%	(82,64 ± 3,16)%	(82,47 ± 3,17)%	(82,12 ± 3,20)%
sfv_24	(79,69 ± 3,36)%	(80,03 ± 3,33)%	(79,69 ± 3,36)%	(80,03 ± 3,33)%
lfv_36	(82,19 ± 3,20)%	(82,81 ± 3,15)%	(82,81 ± 3,15)%	(82,19 ± 3,20)%
mfcc_30	(80,21 ± 3,32)%	(80,38 ± 3,31)%	(80,9 ± 3,28)%	(80,56 ± 3,30)%
lfcc_30	(83,33 ± 3,11)%	(81,42 ± 3,24)%	(80,56 ± 3,30)%	(80,56 ± 3,30)%
mfcc_ms_30	(79,51 ± 3,37)%	(78,99 ± 3,40)%	(79,69 ± 3,36)%	(80,38 ± 3,31)%
lfcc_ms_30	(83,85 ± 3,07)%	(80,38 ± 3,31)%	(81,94 ± 3,21)%	(81,60 ± 3,23)%
mfcc_ms_delta_60	(80,73 ± 3,29)%	(81,42 ± 3,24)%	(81,42 ± 3,24)%	(81,60 ± 3,23)%
lfcc_ms_delta_60	(82,99 ± 3,13)%	(82,47 ± 3,17)%	(82,47 ± 3,17)%	(**84,9 ± 2,99**)%
mfcc_ms_lift_30	(79,51 ± 3,37)%	(79,17 ± 3,39)%	(80,21 ± 3,32)%	(80,90 ± 3,28)%
lfcc_ms_lift_30	(83,51 ± 3,10)%	(80,56 ± 3,30)%	(81,25 ± 3,26)%	(81,25 ± 3,26)%

Tabelle 5.4 : Ergebnis von Test 2.2 / Angabe der Erkennungsrate

5.3 Testszenario 3 (Verifikationsprozess mit Backoff-Modell)

Testszenario 3 beinhaltet die Untersuchung, inwieweit sich die Erkennungsraten bei einer reinen Verifikation mittels Backoff-Modell nach (4.2.3) gegenüber den Erkennungsraten in Testszenario 2 verbessern oder verschlechtern. Im vorliegenden Testszenario wird ebenfalls davon ausgegangen, dass eine optimale Sprecheridentifikation durch das UASR-System stattfindet.

Test 3.1

- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Set v003, die eine korrekte Passphrase eines autorisierten Sprechers enthalten, trainiert.
- **Teststichprobe**
  10 Sprachturns pro autorisierter Person, mit korrekter Passphrase vom Sprachset v003. Des Weiteren wurden zwei Sprachturns pro nicht autorisierter Person mit korrekter Passphrase jeder autorisierten Person vom Set v003 hinzugefügt.
- **Klassifikation**
  Sprecherverifikationsprozess durch Klassifikation mit Backoff-Modell nach (4.2.3).

Ergebnis der Klassifikation mit Backoff-Modell

Modelordnung → / Merkmaltyp ↓	**1**	2	3	4
pfv_30	(97,57 ± 1,28)%	(98,96 ± 0,85)%	(98,61 ± 0,98)%	(98,26 ± 1,09)%
sfv_24	(97,92 ± 1,19)%	(98,78 ± 0,91)%	(99,13 ± 0,77)%	(98,44 ± 1,03)%
lfv_36	(98,09 ± 1,14)%	(97,92 ± 1,19)%	(97,57 ± 1,28)%	(96,35 ± 1,56)%
mfcc_30	(98,61 ± 0,98)%	(98,61 ± 0,98)%	(97,74 ± 1,24)%	(97,22 ± 1,37)%
lfcc_30	(98,61 ± 0,98)%	(98,44 ± 1,03)%	(98,44 ± 1,03)%	(97,74 ± 1,24)%
mfcc_ms_30	(98,44 ± 1,03)%	(99,13 ± 0,77)%	(99,31 ± 0,69)%	(98,26 ± 1,09)%
lfcc_ms_30	**(99,65 ± 0,49)%**	(99,48 ± 0,60)%	(99,65 ± 0,49)%	(99,13 ± 0,77)%
mfcc_ms_delta_60	(99,31 ± 0,69)%	(98,96 ± 0,85)%	(97,05 ± 1,41)%	(96,53 ± 1,53)%
lfcc_ms_delta_60	**(99,65 ± 0,49)%**	(99,31 ± 0,69)%	(96,35 ± 1,56)%	(95,66 ± 1,70)%
mfcc_ms_lift_30	(98,44 ± 1,03)%	(99,13 ± 0,77)%	(98,78 ± 0,91)%	(98,09 ± 1,14)%
lfcc_ms_lift_30	**(99,65 ± 0,49)%**	(99,48 ± 0,60)%	(99,48 ± 0,60)%	(98,78 ± 0,91)%

Tabelle 5.5 : Ergebnis von Test 3.1 / Angabe der Erkennungsrate

Test 3.2

Es erfolgte ebenfalls eine Untersuchung der Erkennungsrate, bei der die Sprachturns mit korrekter Passphrase eines autorisierten Sprechers vom Sprachset v003 durch Sprachset v004 ausgetauscht wurden.

- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Sprachset v003, die eine korrekte Passphrase eines autorisierten Sprechers enthalten, trainiert.
- **Teststichprobe**
  10 Sprachturns pro autorisierter Person, mit korrekter Passphrase vom Sprachset v004. Des Weiteren wurden zwei Sprachturns pro nicht autorisierter Person mit korrekter Passphrase jeder autorisierten Person vom Set v003 hinzugefügt.
- **Klassifikation**
  Sprecherverifikationsprozess durch Klassifikation mit Backoff-Modell nach (4.2.3).

Ergebnis der Klassifikation mit Backoff-Modell

Modelordnung → Merkmaltyp ↓	**1**	2	3	4
pfv_30	(86,11 ± 2,88)%	(82,47 ± 3,17)%	(81,60 ± 3,23)%	(80,73 ± 3,29)%
sfv_24	**(88,02 ± 2,71)**%	(84,20 ± 3,04)%	(82,47 ± 3,17)%	(81,25 ± 3,26)%
lfv_36	(86,46 ± 2,85)%	(84,20 ± 3,04)%	(82,64 ± 3,16)%	(80,73 ± 3,29)%
mfcc_30	(84,90 ± 2,99)%	(83,16 ± 3,12)%	(81,42 ± 3,24)%	(80,73 ± 3,29)%
lfcc_30	(85,24 ± 2,96)%	(82,64 ± 3,16)%	(81,42 ± 3,24)%	(80,38 ± 3,31)%
mfcc_ms_30	(87,85 ± 2,73)%	(84,72 ± 3,00)%	(83,68 ± 3,08)%	(82,81 ± 3,15)%
lfcc_ms_30	(87,85 ± 2,73)%	(84,90 ± 2,99)%	(83,68 ± 3,08)%	(82,81 ± 3,15)%
mfcc_ms_delta_60	(86,81 ± 2,82)%	(83,51 ± 3,10)%	(82,29 ± 3,18)%	(80,38 ± 3,31)%
lfcc_ms_delta_60	(86,63 ± 2,84)%	(83,85 ± 3,07)%	(80,56 ± 3,30)%	(79,34 ± 3,38)%
mfcc_ms_lift_30	(87,85 ± 2,73)%	(85,42 ± 2,94)%	(83,85 ± 3,07)%	(82,29 ± 3,18)%
lfcc_ms_lift_30	(87,85 ± 2,73)%	(85,42 ± 2,94)%	(83,51 ± 3,10)%	(82,47 ± 3,17)%

Tabelle 5.6 : Ergebnis von Test 3.2 / Angabe der Erkennungsrate

5.4 Testszenario 4 (Identifikation und Verifikation mit Backoff-Modell)

Abschließend wurde in Testszenario 4 die Ausführung des Identifikations- und Verifikationsprozesses mittels Klassifikationsvariante - Backoff-Modell untersucht.

Test 4.1

- **Trainingsstichprobe**
  Ein Sprechermodell wird mit 10 Sprachturns vom Set v003, die eine korrekte Passphrase eines autorisierten Sprechers enthalten, trainiert.
- **Teststichprobe**
  Alle Sprachturns vom Sprachset v003 mit korrekten und falschen Passphrasen von autorisierten und nicht autorisierten Personen, welche nicht zum Training der hinterlegten Sprechermodelle verwendet wurden. Die Gesamtgröße beträgt 1620 Sprachturns.
- **Klassifikation**
  Sprecheridentifikationsprozess durch Klassifikation nach (4.2.2).
  Sprecherverifikationsprozess durch Klassifikation nach (4.2.3).

Ergebnis der Klassifikation mit Backoff-Modell

Modelordnung → Merkmaltyp ↓	1	2	3	**4**
pfv_30	(78,40 ± 2,05)%	(92,41 ± 1,32)%	(92,96 ± 1,27)%	(94,44 ± 1,14)%
sfv_24	(86,98 ± 1,67)%	(93,70 ± 1,21)%	(94,57 ± 1,13)%	(95,43 ± 1,04)%
lfv_36	(83,21 ± 1,86)%	(93,70 ± 1,21)%	(94,51 ± 1,13)%	(95,12 ± 1,07)%
mfcc_30	(85,31 ± 1,76)%	(91,98 ± 1,35)%	(93,27 ± 1,25)%	(94,44 ± 1,14)%
lfcc_30	(89,94 ± 1,50)%	(92,35 ± 1,32)%	(93,09 ± 1,26)%	(94,51 ± 1,13)%
mfcc_ms_30	(86,79 ± 1,68)%	(92,53 ± 1,31)%	(94,14 ± 1,17)%	(94,88 ± 1,10)%
lfcc_ms_30	(91,48 ± 1,39)%	(93,52 ± 1,22)%	(95,37 ± 1,04)%	(95,93 ± 0,98)%
mfcc_ms_delta_60	(91,73 ± 1,37)%	(94,44 ± 1,14)%	(95,25 ± 1,06)%	(95,62 ± 1,02)%
lfcc_ms_delta_60	(93,02 ± 1,27)%	(95,00 ± 1,08)%	(96,42 ± 0,92)%	**(96,79 ± 0,88)**%
mfcc_ms_lift_30	(86,79 ± 1,68)%	(93,46 ± 1,23)%	(94,07 ± 1,17)%	(94,94 ± 1,09)%
lfcc_ms_lift_30	(91,48 ± 1,39)%	(93,77 ± 1,20)%	(95,37 ± 1,04)%	(95,93 ± 0,98)%

Tabelle 5.7 : Ergebnis von Test 4.1 / Angabe der minimalen Erkennungsrate

Die Teststichprobe enthielt Sprachturns mit falschen Passphrasen. Innerhalb dieser Passphrasen wurden teilweise jedoch nur die Wörter der korrekten Passphrasen ausgetauscht. Wie in Abschnitt 3.1.1 erläutert, ist die Reihenfolge der Merkmalvektoren in einer Merkmalvektorfolge, bei einer Klassifikation mittels GMM, jedoch nur von geringer Bedeutung.
Es ist deshalb davon auszugehen, dass die Erkennungsrate über den berechneten Werten liegt.

5.5 Auswertung der Testszenarien

Im Folgenden wird eine Auswertung der vorgenommenen Tests durchgeführt. Die getroffenen Aussagen und Bewertungen sind ausschließlich auf die Tests bezogen und können nicht als allgemeine Aussage gewertet werden.

Testszenario 1

Die Durchführung des Identifikations- und Verifikationsprozesses durch die Klassifikationsvariante nach (3.1.23) konnte in Test 1.1 zunächst keine Aussage über die Robustheit der getesteten Merkmaltypen und einer Variation der Modellordnung liefern. Die berechneten Werte lassen jedoch die Aussage zu, dass eine Identifikation sowie die implizite Verifikation auf einer Teststichprobe, welche nur autorisierte Personen mit korrekter Passphrase enthält, eine Erkennungsrate von 100% unabhängig der Modellordnung und des Merkmaltyps ermöglicht.

Im Vergleich mit Test 1.1 verschlechtert sich die Erkennungsrate überwiegend signifikant beim Austausch der Sprachturns[20], wie in Test 1.2 vorgenommen. Neben dem schlechten Gesundheitszustand der Probanden im Set v004 könnte ebenfalls die unterschiedliche Körperhaltung der Probanden während der Aufnahmetermine für eine Änderung der Stimmenlage sprechen. Ein Aufnahmefehler kann zu diesem Zeitpunkt ebenfalls nicht ausgeschlossen werden. Als robust gegenüber den genannten Fehlerquellen sind lediglich die *lfcc_ms_delta_60* Merkmale, unabhängig von der Modellordnung, anzusehen.

Testszenario 2

Eine Verifikation mittels Rückweisungsschwellwert d_{rej} auf einer Teststichprobe, die autorisierte und nicht autorisierte Personen jedoch mit korrekter Passphrase aus demselben Sprachset enthält (Test 2.1), ergab ebenfalls nur geringe Unterschiede in den Erkennungsraten der unterschiedlichen Merkmaltypen, unabhängig von der Modellordnung.

Die in Test 2.2 vorgenommene Ersetzung der Sprachturns (siehe Testbeschreibung) ergab ebenfalls eine Verschlechterung der Erkennungsrate, unabhängig vom Merkmaltyp und der Modellordnung. In der Auswertung von Testszenario 1 wurde bereits über mögliche Fehlerquellen gesprochen. Weiterhin ist während der Bestimmung der jeweiligen Rückweisungsschwellwerte durch Auffindung der Gleichfehlerrate auf der Entwicklungsstichprobe aufgefallen, dass eine Probandin überwiegend oberhalb des Rückweisungsschwellwertes lag. Die Probandin hatte gegenüber den anderen Probanden von der subjektiven Wahrnehmung her eine höhere Stimme.

[20] Ersetzung der Sprachturns mit korrekter Passphrase vom autorisierten Sprecher vom Set v003 durch Set v004.

Testszenario 3

Die höchsten Erkennungsraten wurden für die Sprecherverifikation durch eine Klassifikation mittels Backoff-Modell erreicht (siehe Test 3.1). Im vorliegenden Test erwiesen sich dabei die nachbearbeiteten LFCC-Merkmale *lfcc_ms_30*, *lfcc_ms_delta_60* und *lfcc_ms_lift_30* mit $(99{,}65 \pm 0{,}49)\%$ als robust. Eine konkrete Aussage, welcher Merkmaltyp sich besser für die Stimmenauthentifizierung eignet, ist wegen der fehlenden signifikanten Änderungen zwischen den Merkmaltypen -unabhängig der Modellordnung- jedoch ebenfalls nicht möglich.

Test 3.2 zeigte, dass eine Klassifikation mittels Backoff-Modell keine wesentliche Verbesserung der Erkennungsraten beim Austausch der Sprachturns, wie in Test 1.2 und 2.2 vorgenommen, bringt. Ein Verdacht von begangenen Fehlern während der Aufnahmen erhärtet sich damit. Im Gegensatz zu den bisherigen Tests erweist sich der Merkmaltyp *sfv_24* bei Modellordnung $I = 1$ mit einer Erkennungsrate von $(88{,}02 \pm 2{,}71)\%$ als optimal. Dies ist als ungewöhnlich anzusehen, da dieser Merkmaltyp für die Spracherkennung entwickelt wurde.

Testszenario 4

Eine Identifikation sowie Verifikation des Sprechers mittels der Klassifikationsvariante Backoff-Modell ergab unter Verwendung des Merkmaltyps *lfcc_ms_delta_60* mit Modellordnung $I = 4$ eine Erkennungsrate von $(96{,}79 \pm 0{,}88)\%$.
In Test 4.1 wurde bereits angesprochen, dass die Erkennungsrate fehlerhaft ist. Durch die Verwendung von Gaussian Mixture Models wurden ebenfalls Sprachturns von autorisierten Personen, welche nicht korrekte Passphrasen enthalten (Wörter der korrekten Passphrase wurden getauscht), vom System angenommen. Es ist deshalb davon auszugehen, dass die korrekten Erkennungsraten über den ermittelten liegen.

Vergleich - Verifikation mittels Rückweisungsschwellwert/Backoff-Modell

In Abhängigkeit des Merkmaltyps sowie der Modellordnung kann bei einem Vergleich von Test 2.1 mit Test 3.1 kein wesentlich signifikanter Unterschied in den Erkennungsraten festgestellt werden. Eine Aussage über die erhöhte Leistungsfähigkeit einer Klassifikationsvariante gegenüber der anderen ist anhand des Merkmaltyps und der Modellordnung somit nicht möglich. Bei einer genaueren Betrachtung bietet sich jedoch die Klassifikationsvariante mit Backoff-Modell als vorteilhafter an. Die Gründe dafür liegen in der einfacheren Implementierung und einer nicht benötigten Entwicklungsstichprobe zur Ermittlung des Rückweisungsschwellwertes.

Vergleich - Merkmaltypen sfv_24 und lfv_36

In Test 1.2 wird eine signifikante Differenz der Erkennungsraten, zwischen den Merkmaltypen, erst ab einer Modellordnung von $I = 3$ ersichtlich. Die Testszenarien 2 und 3 lassen hingegen keine Aussage über einen signifikanten Unterschied der Erkennungsraten zu.
Unter der Voraussetzung, dass der Merkmaltyp sfv_24 nur die sprachspezifischen Eigenschaften eines Sprachsignals repräsentiert, würde die Erkennungsrate in Testszenario 2.1 und 3.1 maximal 20,83% betragen. Dies wäre der Fall, wenn alle Sprachturns der nicht autorisierten Personen mit korrekter Passphrase als *False Positive* klassifiziert werden würden.
Die generierten lfv_36 Merkmale, welche nach Abschnitt 4.3.2 die sprecherspezifischen Eigenschaften repräsentieren könnten, erreichen in den Tests 2.1 und 3.1 hingegen plausible Erkennungsraten. Sie zeigen jedoch, wie alle anderen getesteten Merkmaltypen, eine Anfälligkeit gegenüber möglichen Fehlerquellen (Vergleich von Test 2.1 mit 2.2 und Test 3.1 mit 3.2).

Punktueller Vergleich von Sprachsignalen der Sprachsets v003 und v004

Wegen der signifikanten Verschlechterung der Erkennungsraten beim Austausch der Sprachturns, wie in Test 1.2, 2.2 und Test 3.2 vorgenommen, wurde ein punktueller subjektiver Vergleich der Sprachsignale einzelner autorisierter Personen vorgenommen.
Es stellte sich dabei heraus, dass in den getesteten Sprachsignalen vom Set v003 die Höhen schlechter zur Geltung kommen als in den Sprachsignalen vom Set v004. Eine Erklärung dafür ist durch die Veränderung des Abstandes vom Windschutz zum Mikrofon während der Aufnahmetermine zu erklären. Ein Vergleich der Tests innerhalb der einzelnen Testszenarien ist somit nur bedingt zulässig. Generell ist davon auszugehen, dass die Erkennungsraten in Test 1.2, 2.2 und 3.2 über den ermittelten liegen.

5.6 Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurden insgesamt 11 verschiedene Merkmaltypen auf ihre Fähigkeit, sprecherspezifische Eigenschaften zu repräsentieren, getestet. Weiterhin wurden zwei Klassifikationsvarianten für den Verifikationsprozess des Systems entwickelt und getestet.
Es stellte sich dabei in Test 2.1 und 3.1 heraus, dass der Merkmaltyp *lfcc_30* in einer modifizierten Form die höchsten Erkennungsraten für den Sprecherverifikationsprozess erreicht. Dies gilt ebenfalls für die in Test 4.1 vorgenommene Kombination von Identifikation und Verifikation durch den Prototypen. Die Frage der Robustheit eines Merkmaltyps bleibt jedoch wegen eventuell begangener Fehler zwischen den Aufnahmeterminen offen. Unter den gegebenen Voraussetzungen lässt sich jedoch festhalten, dass die *lfcc_ms_delta_60* mit Modellordnung $I = 4$ in Test 1.2 und 2.2 am geringsten von den vorhandenen Störungen beeinflusst werden.
Bei einem Verifikationsprozess mittels Backoff-Modell setzte sich hingegen der Merkmaltyp *sfv_24* mit Modellordnung $I = 1$ durch (siehe Test 3.2). Dies ist, wie schon des Öfteren angesprochen, wegen der primären Aufgabe des Merkmaltyps *sfv_24*, die sprachspezifischen Eigenschaften eines Sprachsignals zu repräsentieren, als ungewöhnlich anzusehen.

Die generierten Merkmaltypen sowie entwickelten Klassifikationsvarianten für das Stimmenauthentifizierungssystem weisen durchaus Schwachstellen in ihrer Robustheit beziehungsweise in der Implementierung und Handhabung auf.
Der besprochene Faktor, welcher zum Unterschied zwischen Sprachset v003 und v004 geführt hat, kann einerseits als Fehler gelten und andererseits als stark veränderte Aufnahmebedingung gewertet werden. Unter diesem Aspekt kann keiner der generierten Merkmaltypen als robust angesehen werden. Des Weiteren sind die globalen Rückweisungsschwellwerte in Testszenario 2 auf die vorhandenen hinterlegten autorisierten Personen optimiert. Dementsprechend ist es mit den durchgeführten Tests nicht möglich, eine Aussage zur Veränderung der Erkennungsraten beim Hinzufügen einer weiteren Person zu treffen. Weiterhin bleibt die Frage offen, inwieweit eine Reduktion oder Erweiterung der Koeffizientenanzahl innerhalb eines Merkmalvektors der untersuchten Merkmalvektortypen die Erkennungsrate beeinflussen würde.

Abschließend kann jedoch unter den gegebenen Bedingungen folgende Aussage getroffen werden: Unter der Voraussetzung, dass sich lediglich die Dynamik und Stimmenlage einer Person ändert[21], erhalten wir bei der Ausführung des Identifikations- und Verifikationsprozesses mit Modellordnung $I = 4$ und unabhängig vom Merkmaltyp eine Erkennungsrate von mindestens 93,3%. In der Weiterentwicklung des Systems muss neben den genannten Faktoren ebenfalls geklärt werden, wie eine Bewertung von stark veränderten Aufnahmebedingungen aussieht.
Zusätzlich wurde in Testszenario 4 deutlich, dass eine Erhöhung der Modellordnung zu höheren Erkennungsraten führte. Es stellt sich die Frage, ob eine weitere Erhöhung $I > 4$ zu höheren Erkennungsraten führt.

[21] Trainings-,Entwicklungs- und Teststichprobe stammen aus demselben Sprachset

A Einverständniserklärung & Funkalphabet

A.1 Musterexemplar Einverständniserklärung

Informationen zur Teilnahme an Testaufnahmen
zur Bachelorarbeit „Stimmauthentifizierung“

Bitte lesen sie sich die unten aufgeführten Informationen sorgfältig durch. Eine Kopie des Teilnahmeformulars kann bei Bedarf erstellt werden.

Studientitel:	Stimmauthentifizierung (Voice Authentication)
Forschungsleiter:	Peter Geßler
Institut:	Lehrstuhl Kommunikationstechnik BTU
Zeit:	30 – 40 min
Ort:	Raum 321 / LG3A / BTU Cottbus-Senftenberg

Beschreibung der Forschung und ihrer Teilnahme:
Im Rahmen der Bachelorarbeit Stimmauthentifizierung wird ein Klassifikationssystem entwickelt, welches anhand des biometrischen Merkmals „Stimme“ ein Sprecher akzeptiert oder abweist. Im Folgenden werden Audioaufnahmen von ihnen aufgezeichnet, welche zur Ermittlung des Schwellwertes und der Leistungsfähigkeit vom System dienen. Die Testpersonen werden für die Untersuchung in die beiden Gruppen „autorisierte Person“ und „nicht-autorisierte Person“ aufgeteilt. Autorisierte Personen sprechen 30-mal die ihnen zugehörige Passphrase sowie 20 Turns ohne gültige Passphrase. Nicht-autorisierte Personen sprechen je fünf Mal die Passphrase der autorisierten Personen. Die benötigten Autorisierungsphrasen werden ihnen in einem weiteren Dokument vorgelegt.

Gefahren des vorliegenden Experimentes:
Es gibt zurzeit keine bekannten Gefahren, für das vorliegende Forschungsexperiment.

Schutz der Vertraulichkeit:
Die Aufnahmen beinhalten die Nachnamen der autorisierten
Personen und können daher nicht als anonym gewertet werden. Eine Nutzung sowie Analyse der Audiodaten ist ausschließlich dem Forschungsleiter und dem Lehrstuhl Kommunikationstechnik / Brandenburgischen Technischen Universität Cottbus-Senftenberg vorbehalten.

Erweiterte Nutzung der Aufnahmen:
Die Aufnahmen werden im weiteren Verlauf der Arbeit in eine Sprachdatenbank integriert. Für die Reproduzierbarkeit der Ergebnisse, wird die Datenbank öffentlich zugänglich sein.Des Weiteren werden die Sprachaufnahmen (v003) vom _____ im Nachhinein ebenfalls der Datenbank hinzugefügt.

Kontaktinformationen:
Wenn sie Fragen bezüglich der Audioaufnahmen haben, setzen sie sich mit dem Forschungsleiter unter peter.gessler@tu-cottbus.de in Kontakt.

Zustimmung:
Ich bin mindestens 18 Jahre alt. Ich habe die Informationen zur Teilnahme an diesem Experiment gelesen und bin über die Risiken aufgeklärt worden.

Name des Teilnehmers	Unterschrift des Teilnehmers	Datum

A.2 Funkalphabet DIN5009

A	Anton	**O**	Otto
Ä	Ärger	**Ö**	Ökonom
B	Berta	**P**	Paula
C	Cäsar	**Q**	Quelle
Ch	Charlotte	**R**	Richard
D	Dora	**S**	Samuel
E	Emil	**Sch**	Schule
F	Friedrich	**ß**	Eszett
G	Gustav	**T**	Theodor
H	Heinrich	**U**	Ulrich
I	Ida	**Ü**	Übermut
J	Julius	**V**	Viktor
K	Kaufmann	**W**	Wilhelm
L	Ludwig	**X**	Xanthippe
M	Martha	**Y**	Ypsilon
N	Nordpol	**Z**	Zacharias

Tabelle A.2 : Funkalphabet nach DIN-Norm 5009.

B Verzeichnisstruktur - Testkorpus

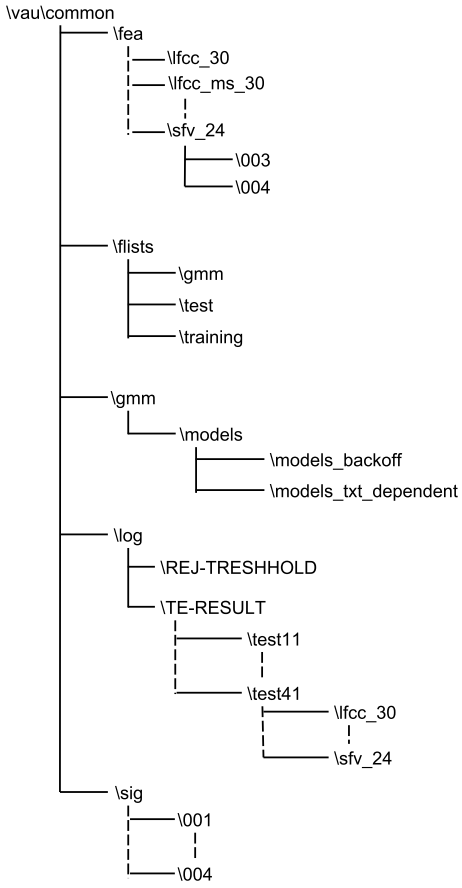

Abb. B: Verzeichnisstruktur des Testkorpus.

\fea	enthält die generierten Daten der Merkmaltypen aller Sprachturns
\flists	enthält File Listen, von den Trainings-, Entwicklungs- und Teststichproben auf die innerhalb des Systems zugegriffen wird
\gmm	beinhaltet die berechneten Gaussian Mixture Models für die einzelnen Sprecher sowie die generierten Backoff-Modelle
\log	in \log werden alle Test Resultate sowie die berechneten Rückweisungs-schwellwerte in die jeweiligen Unterordner hinterlegt
\sig	enthält die aufgenommenen Sprachturns, welche in den Tests verwendet wurden, als .wav Datei in den Unterordnern 003 und 004

C Rückweisungsschwellwerte & Erkennungslisten

Eingestellte Rückweisungsschwellwerte auf Entwicklungsstichprobe

Modelordnung → / Merkmaltyp ↓	1	2	3	4
pfv_30	2,4309	1,7551	1,4417	1,5647
sfv_24	26,9243	25,7246	25,3712	25,1146
lfv_36	−30,82	−31,5	−31,2497	−31,1729
mfcc_30	27,1649	26,5963	26,729	26,598
lfcc_30	26,1864	25,0374	25,543	26,0007
mfcc_ms_30	24,7396	24,4007	24,1583	24,4585
lfcc_ms_30	23,7462	22,8747	23,0079	22,834
mfcc_ms_delta_60	26,7624	25,2045	25,0278	25,2782
lfcc_ms_delta_60	24,5342	23,9823	24,0596	24,7422
mfcc_ms_lift_30	7,8069	7,329	7,172	7,5394
lfcc_ms_lift_30	6,7523	5,96	5,8408	6,0692

Tabelle C.1 : Globale Rückweisungsschwellwerte d_{rej} für Test 2.1 und Test 2.2

Equal Error Rate auf Entwicklungsstichprobe

Modelordnung → / Merkmaltyp ↓	1	2	3	**4**
pfv_30	3,82%	3,82%	3,13%	2,78%
sfv_24	7,29%	3,82%	1,39%	0,69%
lfv_36	4,86%	3,82%	3,82%	3,82%
mfcc_30	3,82%	3,47%	2,78%	2,43%
lfcc_30	3,47%	3,13%	2,78%	2,43%
mfcc_ms_30	4,86%	2,78%	1,04%	0,69%
lfcc_ms_30	3,13%	<u>1,74%</u>	0,35%	**0**%
mfcc_ms_delta_60	6,25%	3,82%	2,78%	1,04%
lfcc_ms_delta_60	5,21%	2,78%	1,04%	0,69%
mfcc_ms_lift_30	4,86%	2,78%	0,69%	0,69%
lfcc_ms_lift_30	3,13%	2,08%	0,35%	0,35%

Tabelle C.2 : Ermittelte Gleichfehlerrate (EER) auf der Entwicklungsstichprobe. Unterstrichener Wert, weicht von EER um eine Fehlklassifikation ab.

Erkennungslisten von Entwicklungsstichprobe

pfv_30

Modelordnung → / Erkennung ↓	1	2	3	4
TP	109	109	111	112
TN	445	445	447	448
FP	11	11	9	8
FN	11	11	9	8

sfv_24

Modelordnung → / Erkennung ↓	1	2	3	4
TP	99	109	116	118
TN	435	445	452	454
FP	21	11	4	2
FN	21	11	4	2

lfv_36

Modelordnung → / Erkennung ↓	1	2	3	4
TP	106	109	109	109
TN	442	445	445	445
FP	14	11	11	11
FN	14	11	11	11

mfcc_30

Modelordnung → / Erkennung ↓	1	2	3	4
TP	109	110	112	113
TN	455	446	448	449
FP	11	10	8	7
FN	11	10	8	7

lfcc_30

Modelordnung → / Erkennung ↓	1	2	3	4
TP	110	111	112	113
TN	446	447	448	449
FP	10	9	8	7
FN	10	9	8	7

mfcc_ms_30

Modelordnung → / Erkennung ↓	1	2	3	4
TP	106	112	117	118
TN	452	448	453	454
FP	14	8	3	2
FN	14	8	3	2

lfcc_ms_30

Modelordnung → / Erkennung ↓	1	2	3	4
TP	111	116	119	120
TN	447	451	455	456
FP	9	5	1	0
FN	9	4	1	0

mfcc_ms__delta_60

Modelordnung → / Erkennung ↓	1	2	3	4
TP	102	109	112	117
TN	438	445	448	453
FP	18	11	8	3
FN	18	11	8	3

lfcc_ms_delta_60

Modelordnung → Erkennung ↓	1	2	3	4
TP	105	112	117	118
TN	441	448	453	454
FP	15	8	3	2
FN	15	8	3	2

mfcc_ms_lift_30

Modelordnung → Erkennung ↓	1	2	3	4
TP	106	112	118	118
TN	442	448	454	454
FP	14	8	2	2
FN	14	8	2	2

lfcc_ms_lift_30

Modelordnung → Erkennung ↓	1	2	3	4
TP	111	114	119	119
TN	447	450	455	455
FP	9	6	1	1
FN	9	6	1	1

D DVD-Inhalt & Bedienungsanleitung

Die zur vorliegenden Bachelorarbeit erstellte DVD enthält alle[22] benötigten Daten für die Reproduktion der Testergebnisse. Auf der DVD befinden sich die folgenden Ordner:

matlab	enthält alle programmierten Skripte sowie die Toolboxen *netlab* und *voicebox*
vau	enthält die Verzeichnisstruktur des Testkorpus, wie in Anhang B beschrieben
backup	enthält die in dieser Arbeit vorgestellten Testergebnisse sowie die Ergebnisse vom Vergleich mit dem UASR System (siehe Anhang E)
audiofiles	enthält alle aufgenommen Sprachturns der Sprachsets v001, v002, v003, v004
trl	enthält alle Transliterationen der aufgenommen Sprachturns von den Sprachsets v002, v003, v004

D.1 Integration des Prototyps

Der auf dieser DVD vorhandene Prototyp des Stimmenauthentifizierungssystems wurde auf einer Windows 7 Professional 64-bit Plattform unter der Version Matlab R2013b entwickelt, implementiert und getestet.

Für die Inbetriebnahme des Prototyps muss zunächst der Ordner *vau* in ein beliebiges Verzeichnis des Rechners kopiert werden. Anschließend ist der Ordner *matlab* in das Unterverzeichnis des Arbeitsverzeichnisses von Matlab zu kopieren.
Damit die entsprechenden Unterordner des Prototypen geladen werden können, öffnet man die in *...\matlab\config* hinterlegte Konfigurationsdatei *cfg.m*. Im nächsten Stritt erfolgt die Anpassung der *homedir* Variable. Durch das Ersetzen der drei Punkte mit dem Verzeichnispfad, in dem der Ordner *vau* hinterlegt wurde, werden die generierten Daten in den entsprechenden Unterordnern abgelegt.

[22] Ausgeschlossen der *Signal Processing Toolbox* und der Software *Matlab* selbst.

D.2 Bedienungsanleitung

Nachdem der Prototyp in das vorhandene System integriert wurde, können die geschriebenen Skripte im Verbund zur Reproduktion der Testergebnisse genutzt werden. Zunächst wird das Programm Matlab gestartet und in das Unterverzeichnis *...\matlab\Tools* des Prototyps gewechselt. Dort befinden sich die Ordner *Classification*, *FeatureGenerator*, *GMM* sowie die Datei *SimulateAll.m*.

SimulateAll.m

Nach Ausführung des Befehls *SimulateAll* in der Matlab Console werden zunächst die „Baseline“ Merkmale mfcc_30 und lfcc_30 aus den vorhanden Signalen im Verzeichnis *...\vau\common\sig\003* und *\004* generiert. Im Anschluss werden die generierten Daten in den jeweiligen Unterordner des Verzeichnisses *...\vau\common\fea\feature_name\003* und *\004* abgelegt. Die generierten Daten werden im weiteren Verlauf des Skripts genutzt, um die modifizierten Merkmaltypen zu erzeugen[23]. In einer zweiten Stufe werden die Gaussian Mixture Models der autorisierten Sprecher sowie die Backoff-Modelle für jeden Merkmaltyp und der Modellordnungen $I = 1\backslash 2\backslash 3\backslash 4$ generiert und hinterlegt. Abschließend findet die Berechnung der Erkennungsraten durch die vorgestellten Klassifikationsvarianten statt. Die Resultate befinden sich nach Abschluss der Berechnungen in dem Verzeichnis *...\vau\common\log\TE-Results\testname*. Der Ordner des jeweiligen Tests enthält zur besseren Übersicht ebenfalls die Struktur *testname \ feature_name*. Innerhalb des Ordners befinden sich die einzelnen Testergebnisse des Merkmaltyps zu den getesteten Modellordnungen.

[23] Ausgenommen sind hier die durch das UASR-System generierten Merkmaltypen pfv_30, sfv_24 und lfv_36.

E Vergleich mit UASR-System

Das Training der Gaussian Mixture Models sowie eine einfache Klassifikation nach (3.1.23) wurden, damit ein Vergleich der Resultate des Prototyps stattfinden konnte, in das UASR-System implementiert. Es fand anschließend ein punktueller Vergleich mit den Merkmaltypen *pfv_30*,*sfv_24* und *lfv_36* unter den Modellordnungen $I = 1\backslash 2\backslash 4$ statt. Als Basis des Vergleichs diente dabei der Test 1.2.

Erkennungsraten des Prototyps:

Modelordnung $\rightarrow$ Merkmaltyp $\downarrow$	1	2	4
pfv_30	$(77{,}92 \pm 5{,}37)\%$	$(82{,}50 \pm 4{,}92)\%$	$(82{,}92 \pm 4{,}87)\%$
sfv_24	$(84{,}58 \pm 4{,}67)\%$	$(85{,}83 \pm 4{,}51)\%$	$(81{,}25 \pm 5{,}05)\%$
lfv_36	$(83{,}33 \pm 4{,}82)\%$	$(89{,}17 \pm 4{,}02)\%$	$(90{,}00 \pm 3{,}88)\%$

Tabelle E.1 : Vergleich der Erkennungsraten des UASR-Systems und des Prototyps

Erkennungsraten des UASR-Systems:

Modelordnung $\rightarrow$ Merkmaltyp $\downarrow$	1	2	4
pfv_30	$(77{,}9 \pm 5{,}4)\%$ HMM 0_0	$(85{,}4 \pm 4{,}6)\%$ HMM 1_3	$(83{,}8 \pm 4{,}8)\%$ HMM 2_0
sfv_24	$(84{,}6 \pm 4{,}7)\%$ HMM 0_0	$(85{,}0 \pm 4{,}6)\%$ HMM 1_1	$(85{,}0 \pm 4{,}6)\%$ HMM 2_5
lfv_36	$(83{,}3 \pm 4{,}8)\%$ HMM 0_0	$(83{,}8 \pm 4{,}8)\%$ HMM 1_3	$(90{,}4 \pm 3{,}8)\%$ HMM 2_10

Tabelle E.2 : Vergleich der Erkennungsraten des UASR-Systems und des Prototyps

Aus den beiden Tabellen ist ersichtlich, dass beide Klassifikatoren bei einer Modellordnung von $I = 1$ die selben Erkennungsraten besitzen. Die Variation bei den Modellordnungen $I = 2$ und $I = 4$ ist mit den unterschiedlich verwendeten Cluster Verfahren beim EM-Algorithmus zu erklären. Dementsprechend kann davon ausgegangen werden, dass die Berechnungen innerhalb des Prototyps unter diesen Bedingungen korrekt sind.

F Spracheingaben - Transliterationen

F.1 Gruppe -nicht autorisierte Person-

Abkürzungen der nicht autorisierten Personen:
APO, CME, CMU, EVO, JLA, JLE, JPU, JSC, JST, KLE, LHA, LSC, MUB, NKA, PGY, RKE, RSC, THA, TRI

Anzahl der Wiederholungen	**Spracheingabe**
5	Autorisierung Wilke Paula Theodor
5	Autorisierung Fellendorf Berta Nordpol
5	Autorisierung Baschinski Paula Anton
5	Autorisierung Kaiser Wilhelm Heinrich
5	Autorisierung Förster Dora Übermut
5	Autorisierung Kuhnke Viktor Kaufmann
5	Autorisierung Ahlers Julius Cäsar
5	Autorisierung Aschenbach Gustav Emil
5	Autorisierung Ewald Gustav Martha
5	Autorisierung Geßler Cäsar Otto
5	Autorisierung Klug Dora Berta
5	Autorisierung Birth Ludwig Anton

F.2 Gruppe -autorisierte Person-

Proband - CEW

Eigene Passphrase: Autorisierung Ewald Gustav Martha
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (Termin v003):

authentifizierung ewald gustav martha
autorisierung gustav martha
autorisierung ewald martha gustav
ich befinde mich im sprachlabor
das sind testaufnahmen fuer ne bachelorarbeit
autir autorisierung ewald gustav martha
autorierung ewald gustav martha
heute ist mittwoch
es ist beergfeeest
auuutoooriiisiiieeeruuung eeeewaaald guuustaaav maaarthaaa
autorisierung ewald gustav martha
der bildschirm ist viereckig
computer geh an ich autorisiere mich mit der passphrase ewald gustav martha
autorisierung gustav ewald martha
es ist gleich sechzehn uhr
die aufnahmen werden ueber eine externe soundkarte erstellt
es ist wahr ich bin der handballer
ich mache aufnahmen mit einem rode mikrofon
autorsierung ewald gustav marrrtha
das ist die letzte testaufnahme

Ungültige Phrasen (Termin v004):

ich bin im sprachlabor
heut ist montag
momentan befinden sich hier vier personen
der bildschirm ist viereckig
peter hat keine ahnung
lehrstuhl kommunikationstechnik
ich befinde mich im lg drei a
das sind aufnahmen ohne phrase
momentan mach ich sprachaufnahmen
auf dem rechner ist windows installiert
wir brauchen noch zehn aufnahmen
heute fahre ich nach berlin
heute beginnt das probestudium
ich bin momentan viertes semester
eigendlich bin ich achtes semester
ich mache aufnahmen mit einem rode mikrofon
das programm ist schlecht geschrieben
hier sind sehr viele rechner
gott sei dank gibts kilmaanlagen
das ist die letzte testaufnahme jeay

Proband - DAS

Eigene Passphrase: Autorisierung Aschenbach Gustav Emil
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

aschen
authentifizierung aschenbach gustav emil
autorisierung asch gu emil
dienstag
computer autorisierung gustav emil
authentifizierung aschenbach gustav emil
autorirung aschenbach gustav emil
autorisierung aschenbach gustav emil
autorisierung aschenbach gustav emil
ich befinde mich im sprachlabor
autorisierung aschenbach emil gustav
autorisierung gustav emil
es ist halb zehn das heisst wir koennen noch was
autooorisierung aschenbach
autorisierung gustav emil aschenbach
computer autorisierung
autorisierung aschenbach gustav emil
der raum ist schallisoliert
autorisierung aaaschenbaaach gustav emil
das ist die letzte testaufnahme

Ungültige Phrasen (v004):

heute ist mittwoch
ich befinde mich im sprachlabor wieeder
dass ist das lehrgebaeude lg drei a
meine passphrase geht kein was an
autorisierung sprachlabor
computer ich autorisiere mich mit einer passphrase
die aufnahmen finden ueber ein java programm statt
ich nehme mit einem rode mikrofon auf
ich befinde mich ausserhalb der sprachkabine
es ist neun uhr dreiundvierzig
du kommst hier net rein
es befindet sich jetzt ein kalender im sprachlabor
eigendlich muss ich noch sechszehn minuten aufnahmen machen
abeer wir sind schneller
der vorhang ist blau
das sind gut ausgesteuerte aufnahmen
beim naechstenmal bin ich auch wieeder dabei
ich hoffe das war das letztemal
jedesmal muss ich so ein komischen zettl unterschreibn
das ist die letzte testaufnahme

Proband - IFE

Eigene Passphrase: Autorisierung Fellendorf Berta Nordpol
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung fellendorf berta nordpol
mhm nochmal kurz
autorisierung fellendorf nordpol berta
autorisierung berta nordpol
autorirung fellendorf berta nordpol
audorisierung fellendorf berta nordpol
ich befinde mich im sprachlabor
heute ist mittwoch
computer geh an ich autorisiere mich mit der passphrase fellendorf berta nordpol
meine arbeitswoche endete gestern
das sind sprachaufnahmen fuer eine bachelorarbeit
ich autorisiere mich mit einer passphrase
autorisi fellen aeh
autorisierung fellendorf berta
autorisierung fellendorf berta nordpol genau
auuutoooriiisiiieruuung felleeendooorf beeertaaa nooordpoool
autorisierung fellendorf berta nordpol
noch drei aufnahmen brauchen wir
nordpol berta fellendorf autorisierung
das ist die letzte testaufnahme

Ungültige Phrasen (v004):

autorisierung ohne korrekte passphrase
es ist zwoelf uhr achtzehn
ich befinde mich im sprachlabor
die vorhaenge sind dunkel blau
autorisierung fellendorf computer geh an
ich war grad beim blutplasma spenden
heute ist mittwoch
ich war fuenf minuten zu spaet
draussen ist schlechtes wetter wie beim letzten mal
ich war schon drei ma hier
das sprachlabor ist schallisoliert
autorisierung fellen dorf berta
die aufnahmen finden ueber ein java programm statt
wir brauchen noch sieben aufnahmen
die klimaanlage ist nicht an
morgen spielt deutschland gegen die usa
ich muss arbeiten am donnerstag
die aufnahmen finden mit einem rode mikrofon statt
hoffentlich war dass das letztemal das ich testaufnahmen mache
das ist die letzte aufnahme

Proband - IKL

Eigene Passphrase: Autorisierung Klug Dora Berta
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung klug dora berta
computer geh an ich autorisiere mich mit der passphrase
computer geh an ich autorisiere mich mit der passphrase klug dora berta
autorisierung klug berta dora
autorirung klug dora berta
heute ist freitag gleich ist frei
ich befinde mich im sprachlabor
autorisierung klak dora berta
autorisierung klug dora berta
linux ist besser als windows
auuutoooriiisiiieruuung kluuug beeertaaa
ich mach neue sprachaufnahmen
computer geh an
ingesamt gibt es zwoelf autorisierte sprecher
autorisier
autorisierunk kluk berta
au u die haste bei mir zuhauf
morgen geh ich feiern
ich moechte den grossen fernseher bitte mit nach hause nehmen
das ist die letzte passphrase

Ungültige Phrasen (v004):

ich befinde mich im sprachlabor
heute ist mittwoch
es ist sechzehn uhr dreizn
die vorhaenge sind blau
das wetter heute ist sehr schlecht
das sind aufnahmen fuer den lehrstuhl kommunikationstechnik
autorisierung klug
linux ist besser als windows
ich befinde mich ausserhalb der sprachkabine
die testaufnahmen werden mit einem java programm durchgefuehrt
morgen spielt deutschland gegen die usa
das sind aufnahmen fuer eine bachelorarbeit
alle angaben wie immer ohne gewaehr
und jetzt die lotto zahlen
gleich kommt der bus
ich mach zum dritten ma sprachaufnahmn
naechsten dienstag ist fakultaetgrillen
noch drei aufnahmen
das ist eine autorisierung ohne korrekte passphrase
das ist jetzt die letzte aufnahme

Proband - JAH

Eigene Passphrase: Autorisierung Ahlers Julius Caesar
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung ahlers julius caesar
auuutorisierung aaahlers juuulius caeaeaesar
autorisierung ahlers julius caese
autorisierung ahlers julius caese
computer geh an ich autorisiere mich mit der passphrase ahlers julius caesar
ich befinde mich im sprachlabor
das sind sprachaufnahmen fuer eine bachelorarbeit
autorisierung julius caesar
autorirung ahlers julius caesar
autorisierung ahlers caesar julius
der bildschirm ist rund
das sind neue sprachaufnahmen
auuutoooriiisiiieruuung aaahleeers juuuliuuus caeaeaesaaar
autorisierung julius caesar ahlers
heute ist mittwoch
ich war keine sieben minuten zu spaet
mir faellt nix mehr ein
ich bin demnaechst masterstudent
diese aufnahme ist eine wiederholung
das ist die letzte sprachaufnahme

Ungültige Phrasen (v004):

heute ist mittwoch
ich befinde mich im sprachlabor
die vorhaenge sind rot
der bildschirm ist viereckig
ich war keine fuenfzehn minuten zu spaet
ich befinde mich im lehrgebaeude drei a
ich mache testaufnahmen
wir haben hier eine klimaanlage
das sind aufnahmen ohne gueltige passphrase
wir brauchen noch ein paar aufnahmen
ich mache zum vierten mal testaufnahmen
es ist zehn uhr zweiundzwanzig
hier sind viele bildschirme
ich befinde mich ausserhalb der sprachkabine
der grossrechner war ueber nacht an
es befindet sich jetzt ein kalender im labor
wir machen aufnahmen ueber ein java programm
es ist zehn uhr dreiundzwanzig
knoppas ist vorbei
das ist die letzte testaufnahme

Proband - LFO

Eigene Passphrase: Autorisierung Foerster Dora Uebermut
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

aaalleee
authentifizierung foerster dora uebermut
aurorisierung foerster dora uebermut
autorisierung foerster dora uebamud
ich befinde mich im sprachlabor
autorisierung fjoerstda dora uebermut
linux ist besser als alles
auuutoooriiisiiieruuung foeoeoersteeer doooraaa ueueuebeeermuuut
audorisierun foerster dora uebermut
dies sind sprachaufnahmen fuer eine bachelorarbeit
autorisierunk foerster dora uebermut
autorisie foers dor ueber mu
der bildschirm ist viereckig
fleckel es ist ein oktobaer
aeh foerster dora uebermuts
computer authentifizierung foerster dora uebermuuut
brrr hmm
computer bitte authentifiziere mich jetzt meine name ist foerster und meine passphrase dora uebermut
heute ist mittwoch es ist nicht so schoenes wetter
dies ist meine letzte sprachaufnahme

Ungültige Phrasen (v004):

heute ist donnerstag es ist neun uhr zwoelf viel zu frueh
ich befinde mich im eh sprach
ich befinde befinde mich im sprachlabor im lehrgebaeude lg eins c ne
die vorhaenge sind dunkelblau und aufgezogen und befinden sich an der rechten seite des sprachlabors
die notausgaenge befinden sich rechts und vorne links und sind mit blauem filzstift makiert
herzlich willkommen liebe sportfreunde willkommen hier zum tunier der kegelvereine gruen weiss erfurt und rot weiss essen
dies sind testaufnahmen fuer eine bachelorarbeit
autorisierung foerster
computer geh an ich autorisiere mich mit einer falschen passphrase
die aufnahmen finden ueber ein java programm statt
diese set von sprachaufnahmen ist bereits das dritte von meiner person
ich befinde mich ausserhalb der sprachkabine
linux ist besser als alles andere vielleicht ausser bsd
heute steht noch an gruppenarbeit dass koennte tatsaechlich eine weile dauern
wir brauchen noch fuenf testaufnahmen
ich habe auch an diesem lehrstuhl bachelorarbeit geschrieben
gleich gibt es noch eine lustige kleine fragestunde mit einem komilitonen
der bildschrim hat vier ecken vier ecken hat der bildschirm
die ist die letzte testaufnahme

Proband - MBI

Eigene Passphrase: Autorisierung Birth Ludwig Anton
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

autorirung birth ludwig anton
autorisierung b ludwig anton
autorisierung birth le anton
autorisierung birth ludwig an ton
computer birth ludwig anton
computer autorisierung birth ludwig anton
computer autorisierung birth ludwig
heute ist montaaag
autorisierung birth ludwig berta
autorisierung birth luuudwig anton
autorisierung birth anton
ich befinde mich im sprachlabor
authentifizierung birth ludwig anton
autorisierung birth ludwig anton
authentification birth ludwig anton
authentifizierung birth ludwig aaanton
autorisierung birth ludwig anton
autorisierung birth ludwig anton
das ist die vorletzte sprachaufnahme
autorisierung ludwig anton

Ungültige Phrasen (v004):

ich befinde mich im sprachlabor
das ist lehrgebaeude lg drei a
das sind aufnahmen fuer den lehrstuhl kommunikationstechnik
ich befinde mich an brandenburgischen technischen universitaet cottbus senftenberg
die vorhaenge sind marinen blau
die aufnahmen finden ueber ein beschissenes java programm statt
die aufnahmen finden mit einem rode mikrofon statt
osx ist besser als linux und windows
ich befinde mich ausserhalb der sprachkabine
jetzt sind se doof wei die au
der bildschirm ist schwarz
ich kann mich nicht anmelden
autorisierung birth
sesam oeffne dich
der industrierechner laeuft
licht ist an
alle geloeteten kabel funktionieren
es ist vierzehn uhr vierundzwanzig
morgen spielt deutschland gegen die usa
das ist die letzte aufnahme

Proband - MKA

Eigene Passphrase: Autorisierung Kaiser Wilhelm Heinrich
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

autorisierung kaiser wilhelm hei
autorisierung kai wilhelm heinrich
computer autorisierung kaiser wilhelm heinrich
auto kaiser wilhelm heinrich
heute ist montag
auto kai wi hei
auto kaiser wilhelm hei
autoriri auto nee
autorisierun mariu ach
autorisierung kaiser heinrich wilhelm
Ich bin im sprachlabor
autorisierung wilhelm heinrich kaiser
autorisierung wilhelm kaiser heinrich
authentifizierung kaiser wilhelm heinrich
autorisierung heinrich kaisa wilhelm
autorisierung kaiser wilhelm heinrich
autosierung kaiser wilhelm heinrich
heute ist ein schoener tag
autorisierung kaaaiiiser wilheeelm heinrich
autorisierung letzte kaiser aufnahme

Ungültige Phrasen (v004):

heut ist mittwoch
die vorhaenge sind dunk dunkel
jemand hat den vorhang aufgezogen
es ist dreizehn uhr fuenfzich
ich befinde mich im sprachlabor
das ist im lehrgebaeude lg drei a
die aufnahmen laufen ueber ein java programm
gleich kommt der naechste proband
heute stehn noch zwei probanden an
autorisierung kaiser
autorisierung computer geh an
heute abend fahr ich nach berlin
der bildschirm ist viereckig
tj hat ein neues handy
die aufnahmen passieren mit einem rode mikrofon
das ist keine korrekte passphrase
ich mache jetzt zum dritten mal sprachaufnahmen
hoffentlich das letztemal
wir brauchen noch zwei aufnahmen
das ist die letzte aufnahme

Proband - PGE

Eigene Passphrase: Autorisierung Gessler Caesar Otto
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung gessler caesar otto
heute ist samstag
authentifizierung gessler julius caesar
das war eine falsche passphrase
authentifizierung computer geh an
authentifizierung gessler caesar
authentifizierung gessler caesar ot
authentifizierung gessler
authentifizierung gessler sprachlabor
authentifizierung gessler caesarrr otto
authenifizierung gesslerrr caesarrr otto
autorisieruung gessleerrr caesaarrr otto
autorisierung gessler caesar ot
ich versuche den spracherkenner zu ueberlisten
authentifizierung ist egal
authentifizierung computer gessler
heute nachmittag arbeite ich weiter
noch drei aufnahmen
autorisierung gessler caesar ottooo
letzte aufnahme fuer samstag

Ungültige Phrasen (v004):

ich befinde mich im sprachlabor
autorisierung ohne gueltige passphrase
dass ist ein laengerer test fuer die testdaten ohne passphrase
heute ist donnerstag wir haben momentan schoenes wetter
autorisierung gessler
neun mal neun ist einundachtich einundachtzich durch neun ist neun
das sind sprachaufnahmen fuer eine bachelorarbeit am lehrstuhl kommunikationstechnik
die aufnahmen werden mit einem java programm durchgefuehrt
ich befinde mich ausserhalb der sprachkabine die vorhaenge sind blau und das muster grau
heute abend kommt deutschland gegen die usa ich hoffe deutschland gewinnt
momentan befinden sich im sprachlabor zwei personen
wir haben es zwoelf uhr zweiundreissich
ich habe noch genau ein monat zeit fuer meine bachelorarbeit
die aufnahmen werden ueber ein rode mikrofon getaetigt
dass ist insgesamt das vierte testset was ich aufnehme
wir brauchen noch fuenf aufnahmen
die klimaanlage kuehlt die geraete und den raum
autorisierung ohne gueltige passphrase computer geh an
ich befinde mich am lehrstuhl kommunikationstechnik
dass ist die letzte testaufnahme

Proband - SBA

Eigene Passphrase: Autorisierung Baschinski Paula Anton
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung baschinski paula anton
authen
autorisierung baschinki paula anton
autorisierung paula anton
computer an ich autorisiere mich mit der passphrase baschinski paula anton
heute ist der siebte fuenfte zweitausendfuenfzehn
ich befinde mich im sprachlabor des lehrstuhls kommunikationstechnik
meine autorisierungspassfrage besteht aus zwei woertern
autorirung baschinski paula anton
autorisieren baschinsk paul anton
ich nehme an den testaufnahmen teil
autorisieruuung baschinski paula anton
es regnet schon wieder
die aufnahmen finden mit einer fokusrite scarlet
autorisierung weiss nicht paula weiss nicht
das ist ein selbstgeschriebenes aufnahmeprogramm
eine person wartet schon draussen
das aufnahmeprogramm ist mit java geschrieben
auuuutoooriiiisiiieruuung baaaschiiinskiii pauuulaaa aaantooon
um veert

Ungültige Phrasen (v004):

ich befinde mich im sprachlabor
ich mach schon zum dritten ma testaufnahmen
joschi kam heute fuenfzehn minuten zu spaet
das ist das akademische viertel
es ist ziemlich heiss hier drin
ich bin ein bisschen erkaeltet
ich bin stark erkaeltet
es ist elf uhr zweiundvierzich
nach mir kommen heute noch vier weitere probanden
das sind die testaufnahmen v null null vier
autorisierung computer geh an
draussen ist schlechtes wetter
wie beim letztenmal
wir haben jetzt ein kalender im sprachlabor
am donnerstag schau ich fussball
ich befinde mich ausserhalb der sprachkabine
die vorhaenge sind blau
ich mache aufnahmen mit einem rode mikrofon
das ist die vorletzte aufnahme
endlich schluss

Proband - SWI

Eigene Passphrase: Autorisierung Wilke Paula Theodor
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung wilke paula theodor
autorirung wilke paula theodor
autorisierung paula theodor
autorisierung wilke theodor paula
heute ist freitag
auuutoooriiisiiieruuung wiiilke pauuulaaa theooodooor
computer geh am ich autorisiere mich mit der passphrase autorisierung wilke paula theodor
ich befinde mich im sprachlabor
computer wilke paula theodor
theodor paula wilke autorisierung
autorisierung wiiilkeee paula theodor
ich mache sprachaufnahmen
es ist vierzehn uhr acht
autori wil pau theo
das wetter ist nicht schoen
computer paula theodor wilke
das ist schon das dritte mal das ich sprachaufnahmen mache
autorisierung juhnke paula theodor
das ist die vorletzte sprachaufnahme
letzte sprachaufnahme endlich geschafft

Ungültige Phrasen (v004):

ich befinde mich im sprachlabor
heute ist mittwoch
es ist zehn uhr fuenumvierzich
das sprachlabor ist schallisoliert
autorisierung ich auto mh
autorisierung computer geh an
autorisierung wilke computer geh an
ich autorisiere mich mit einer nicht korrekten passphrase
die vorhaenge sind blau
ich mach zum vierten ma testaufnahmen
achso das soll ich sagen
wir brauchen noch neun aufnahmen ohne korrekte passphrase
der bildschirm ist viereckich
hoffentlich muss ich zum letztenma testaufnahmen machen
autorisierung wilke paula
auf dem grossen bildschirm wuerde ich gerne mal fussball sehen
nur noch ein paar testaufnahmen
theodor paula wilke autorisierung
wilke paula autorisierung
das ist die letzte testaufnahme

Proband - TKU

Eigene Passphrase: Autorisierung Kuhnke Viktor Kaufmann
Anzahl an Wiederholungen der eigenen Passphrase: 30 (v003), 30 (v004)

Ungültige Phrasen (v003):

authentifizierung
authentifizierung kuhnke viktor kaufmann
autorisierung kuhn viktor kaufmann
i love u vita ger mann
ich mach sprachaufnahmen fuer eine bachelorarbeit
ich befinde mich im sprachlabor des kommun
computer an ich autorisiere mich mit kuhnke viktor kaufmann
wir haben zwoelf autorisierte sprecher
autorisierung kuhnke kaufmann viktor
auuutoooriiisiiieruuung kuuuhnkeee viiiktooor kauuufmaaann
autorirung kuhnke viktor kaufmann
autorisierung kuhnke
autorirung kuhnke viktor kaufmann
der bildschirm ist viereckig
linux ist besser als windows
heute ist mittwoch
autorisierung lehrstuhl kommunikationstechnik
autorisierung computer geh an
wir testen hier den spracherkennus kennungs
das ist die letzte sprachaufnahme

Ungültige Phrasen (v004):

autorisierung kuhnke
ich befinde mich im sprachlabor l lehrgebaeude lg drei a
ich mache testaufnahmen fuer eine bachelorarbeit
ich mache bereits zum dritten mal testaufnahmen
die vorhaenge sind blau und das wetter ist schlecht
die aufnahmen werden mit einem java programm aufgefue durchgefuehrt
es ist dreizehn uhr neun ich war gerade mittagessen
ich befinde mich fuer die testaufnahmen ausserhalb der sprachkabi sprachkabi
heute ist donnerstag der sechundzwanzigste sechste
die schweiz hat gestern drei zu null gegen honduras gewonnen
ich hatte heute frei im gegensatz zu dir
naechste woche ist fakultaetsgrillen
linux ist besser als windows
aufm grossen plasmafernseher wuerde ich gern gerne mal fussball gucken
wir brauchen noch sechs aufnahmen
in diesem raum ist eine klimaanlage installiert
dass ist eine laengere aufnahme damit eine groessere sekundenlaenge existiert
ich autorisiere mich ohne korrekte passphrase computer geh an
die autorisierung wird fehlschlagen
das ist die letzte testaufnahme

Literaturverzeichnis

[Bro] Brookes, Mike: *VOICEBOX: Speech Processing Toolbox for Matlab.* `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

[Dev] Devleker, Kirthi: *Matlab Signal Processing Toolbox.* `http://www.mathworks.de/products/signal/`

[DG64] D.M. Green, J.A.Swets: Signal Detection Theory and Psychophysics. (1964). ISBN 0–471–32420–5

[Fel12] Fellbaum, Prof. Dr.-Ing. K.: *Sprachverarbeitung und Sprachübertragung.* 2012. – ISBN 978–3–642–31503–9

[Fli95] Fliegner, Dr. L.: *Textabhängige Sprecherverifizierung unter Berücksichtigung der Endpunktdetektion.* Wissenschaft und Technik Verlag, 1995. – ISBN 3–928943–40–5

[Foe14] Foerster, Leonard: Rückweisung ungrammatischer Spracherkenner-Eingaben. In: *Lehrstuhl Kommunikationstechnik - Brandenburgische Technische Universität* (2014), S. 19

[Har01] Hardt, Dr.-Ing. D.: *Textabhängige und phonetisch-basierte Sprecherverifizierung für den Einsatz in der Telekommunikation.* 2001. – ISBN 3–89685–362–7

[Her10] Hertlein, Dr.-Ing.Heinz R.: *Fusion von Klassifikationssystemen für die automatische Sprechererkennung.* 2010. – ISBN 978–3–8325–2525–5

[Hof98] Hoffmann, Prof. Dr.-Ing. R.: *Signalanalyse und -erkennung.* Springer-Verlag, 1998. – ISBN 3–540–63443–6

[Mil07] Mildner, Dr.-Ing. V.: *Signalverarbeitungskonzepte zur robusten Sprechererkennung.* 2007. – ISBN 978–3–8322–6504–5

[Nab] Nabney, Ian: *Netlab Toolbox for Matlab.* `http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab`

[Roh03] Rohdenburg, Thomas: *Klassifikation von Audio-Signalen.* 2003

[Ros02] Rose, Philip: *Forensic Speaker Identification.* 2002. – ISBN 0–415–27182–7

[RR75] Reynolds, D.A. ; Rose, R.C.: Robust text-independent speaker identification using Gaussain mixture speaker Models. In: *IEEE Transactions on ASSP* (1975), Nr. 3, S. 72–83

[RW99] Roberts, William J. ; Willmore, Jonathan P.: Automatic speaker recognition using Gaussian Mixture Models. In: *In Information, Decision and Control, IDC* (1999), S. 465–470

[Tal95] Talamazzini, Schukat: *Automatische Spracherkennung - Grundlagen, statistische Modelle und effiziente Algorithmen.* Springer-Vieweg, 1995

[Wen04] Wendemuth, Andreas: *Grundlagen der digitalen Signalverarbeitung.* 2004. – ISBN 3–540–21885–8

[Wol72] Wolf, J. J.: Efficient Acoustic Parameters for Speaker Recognition. In: *The Journal of the Acoustical Society of America* (1972), Nr. 2, S. 2044–2056

[Wol11] Wolff, Prof. Dr.-Ing. M.: *Akustische Mustererkennung.* 2011. – ISBN 978–3–942710–14–5