Institut für
Technische Informatik und
Kommunikationsnetze

Sprechererkennung anhand des Prosodieverlaufs

Markus Schafroth
Michael Steiger

Semesterarbeit SA-2007-59

Herbstsemester 2007

Institut für Technische Informatik und Kommunikationsnetze

Gruppe für Sprachverarbeitung

Betreuer: Dr. B. Pfister und M. Gerber
Verantwortlicher: Prof. Dr. L. Thiele

Abstract

In der vorliegenden Arbeit werden auf prosodischen Merkmalen basierende Sprecherverifikations-Systeme betrachtet. Mehrere Verfahren zur Modellierung der aus den Sprachsignalen extrahierten prosodischen Merkmalen werden untereinander verglichen: Gauss'sche Mischmodelle angewendet auf den zeitlichen Verlauf der Grundfrequenz und der Energie, Gauss'sche Mischmodelle angewendet auf lokal ermittelte quantitative prosodische Grössen und N-Gramme zur qualitativen statistischen Erfassung des nach prosodischen Gesichtspunkten segmentierten Sprachsignals. Zusätzlich werden Systemkombinationen aller implementierten Systeme gebildet. Abschliessend wird mit einem cepstralen Referenzsystem verglichen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Das Gebiet der Sprachverarbeitung umfasst die zwei Kategorien Sprachanalyse und Sprachsynthese. Das Ziel der Analyse ist es, aus einem Sprachsignal die gewünschte Information zu extrahieren, beispielsweise einen gesprochenen Text schriftlich festzuhalten. Bei der Synthese soll dagegen ein schriftlich vorliegender Text in ein Sprachsignal übersetzt werden, das für den Menschen verständlich ist und möglichst natürlich klingen soll.

Innerhalb der Sprachanalyse lassen sich Spracherkennungssysteme und Sprechererkennungssysteme unterscheiden. Während bei Spracherkennungssystemen gesprochene Wörter und Sätze erkannt und identifiziert werden sollen, setzt sich die Sprechererkennung die Identifizierung oder Verifikation einer bestimmten sprechenden Person, des Sprechers, zum Ziel.

Dabei ist insbesondere der Unterschied zwischen der Identifizierung und der Verifikation eines Sprechers zu beachten. Bei der Identifizierung soll anhand einer Sprechprobe und einer Anzahl vorgegebener Sprecher entschieden werden, welche Person gesprochen hat. Bei der Verifikation steht einer Sprechprobe hingegen nur ein bestimmter Sprecher gegenüber; es soll entschieden werden, ob die Sprechprobe von diesem Sprecher stammt oder nicht.

1.1 Sprecherverifikation

Die vorliegende Arbeit beschränkt sich auf diese letzte Aufgabe: Sprecherverifikation. Anwendungsgebiete einer solchen Sprecherverifikation bestehen beispielsweise bei automatischen Zulassungssystemen aller Art sowie in der forensischen Analyse von Sprachdokumenten oder -aufzeichnungen.

Sprecherverifikation kann grundsätzlich in zwei Kategorien unterteilt werden. Bei der *textabhängigen* Verifikation muss in der Trainings- und der Anwendungsphase jeweils der gleiche Text gesprochen werden. Der Vergleich erfolgt hierbei oftmals auf Grund eines Pattern Matchings. Falls nicht davon ausgegangen werden kann, dass es sich beim Testsignal und dem oder den Referenzsignalen um denselben gesprochenen Text handelt, kommt die *textunabhängige* Sprecherverifikation zum Einsatz. Statistische Methoden können in diesem Fall für eine Entscheidung verwendet werden.

Sprecherverifikationssysteme basieren oft auf Merkmalen des Sprachsignals, wie sie auch bei der Spracherkennung verwendet werden. Im Unterschied zu diesen spektralen Merkmalen betrachten wir die prosodischen Eigenschaften bzw. Merkmale der Sprache. Diese umfassen unter anderem die Grundfrequenz und die Intensität bzw. deren zeitlicher Verlauf sowie die Sprechgeschwindigkeit. Wir versuchen also, anhand der Sprechmelodie, der Betonung sowie des Rhythmus einen Sprecher zu verifizieren. Verschiedene Ansätze zur Sprecherverifikation basierend auf prosodischen Merkmalen werden in der Literatur beschrieben. Drei davon wurden im Rahmen dieser Arbeit in MATLAB implementiert: Verifikation mittels Training der prosodischen Parameter mit Gauss'schen Mischmodellen (GMM) sowie die Modellierung des prosodisch segmentierten Sprachsignals mit N-Gramm Modellen. Beim dritten, dem sogenannten ProsVar Ansatz, erfolgt eine Kombination von Segmentierung und Modellierung mit Gauss'schen Mischmodellen, indem aus dem nach Stimmhaftigkeit segmentierten Sprachsignal zusätzliche Information gewonnen und mit einem GMM modelliert wird.

Nach der Implementation der drei Systeme erfolgt eine Evaluationsphase, in der verschiedene Systemparameter bestimmt werden sollen, sowie der Einfluss einer variierenden Anzahl von Trainingsdaten untersucht werden soll. Schliesslich werden die Systeme mit bis dahin nicht verwendeten Daten getestet. Die Performance der Einzelsysteme wie auch diejenige von Kombinationen mehrerer Systeme wird mit einem cepstralen System verglichen, welches in vergleichbaren Tests häufig als Referenzsystem herangezogen wird.

Die genaue, vom Institut formulierte Aufgabenstellung zur vorliegenden Arbeit ist in Anhang B zu finden.

1.2 Frühere und ähnliche Arbeiten

Nachfolgend werden einige ausgewählte Arbeiten zum Thema Sprecherverifikation basierend auf prosodischen Eigenschaften vorgestellt. Insbesondere von Interesse sind Artikel, welche die zu implementierenden Systeme beschreiben.

Modellierung der lokalen prosodischen Variationen Die Autoren von [2] beschreiben ein Verfahren, mit dem sprecherspezifische Information aus dem Verlauf der Grundfrequenz eines Sprachsignals gewonnen werden kann.

Dabei stehen im Gegensatz zu den meisten anderen Ansätzen weniger die Veränderung über längere Perioden als vielmehr die lokalen Variationen im Vordergrund. Dazu wird der Verlauf der Grundfrequenz in stimmhafte und stimmlose Abschnitte segmentiert und mit einem stückweise linearen Modell angenähert. Verschiedene Parameter der stimmhaften Segmente wie Median und Steigung der Grundfrequenz, deren Dauer, sowie die Dauer der stimmlosen Segmente, werden anschliessend als statistischer Fingerabdruck eines Sprechers weiterverwendet.

Modellierung von Trajektorien In [1] wird eine neue effiziente Möglichkeit zur Modellierung und Anwendung von prosodischen Eigenschaften für textunabhängige Sprecherverifikation vorgestellt.

Der Ansatz verwendet die Relation zwischen Grundfrequenz- und Energieverlauf um die Identität eines Sprechers zu charakterisieren. Die Idee besteht darin, dass die Dynamik der beiden Trajektorien gemeinsam verschiedene prosodische Gesten repräsentieren, die charakteristisch für einen spezifischen Sprecher sind. Zusätzlich werden damit auch weitere Eigenschaften wie z.B. Aufgeregtheit oder Monotonie

des Sprachsignals erfasst.

Konkret wird für jedes stimmhafte Frame ein vierdimensionaler Vektor erstellt, der die logarithmierten Grundfrequenz- und Energieverläufe, sowie deren Ableitungen enthält. Zur Modellierung wird ein einfaches Bigramm Modell der Symbolsequenzen verwendet. Ein Likelihood-Entscheider arbeitet mittels eines sprecherabhängigen Bigramm Modells sowie eines sprecherunabhängigen Modells, des sogenannten Universal Background Bigramm Model.

Bilden von N-Grammen In [3] wird vorgeschlagen, die zeitlichen Trajektorien der Grundfrequenz sowie der Energie für eine Segmentierung des Sprachsignals zu verwenden.

Die Kombination von Grundfrequenz, Intensität und Dauer charakterisiert aussergewöhnliche prosodische Gesten eines Sprechers. Daher wird das kontinuierliche Sprachsignal in eine Sequenz von diskreten Abschnitten unterteilt, welche das Signal hinsichtlich des zeitlichen Verlaufs von Grundfrequenz und Energie beschreiben. Diskrete Hidden Markov Modelle, Binäre Bäume und N-Gramme sind einige der Methoden, die zur Modellierung dieser Segmentsequenzen verwendet werden können. Ein Likelihood Entscheider kann anschliessend zur Erkennung eines Sprechers beigezogen werden, basierend auf einem sprecher- oder sprach*abhängigen* Modell sowie einem sprecher- oder sprach*unabhängigen* Modell, das auf Grund aller verfügbaren Daten erstellt wurde. Zusätzlich kann auch die Länge der Segmente berücksichtigt werden, womit eine genauere Charakterisierung des Sprechstils erreicht wird.

Auswertung von N-Grammen In [4] wird ein Likelihood-Entscheider für N-Gramm Statistiken beschrieben. Dabei wird auf Grund von UBM, Modell- und Testsprecher ein Änlichkeitswert berechnet, der anschliessend eine Schwellwertentscheidung ermöglicht.

1.3 Aufbau der Arbeit

Das nächste Kapitel gibt einen zusammenfassenden Überblick über die gesamte Arbeit. Anschliessend werden die folgenden Themen vertieft behandelt: Kapitel 3 erklärt die Grundprinzipien und Kapitel 4 die konkrete Umsetzung bzw. die Implementation der Sprecherverifikationssysteme. In Kapitel 5 werden die Systeme evaluiert und die Ergebnisse der Performancetests präsentiert, verglichen und diskutiert. Kapitel 6 enthält unsere Schlussfolgerungen und abschliessend geben wir in Kapitel 7 einige Anregungen für weiterführende Arbeiten auf dem Gebiet der prosodischen Sprecherverifikation.

2 Zusammenfassung

Bei der Sprecherverifikation wurden bisher vor allem spektrale Eigenschaften des Sprachsignals ausgewertet. In der Literatur finden sich verschiedene Ansätze, wie auch prosodische Merkmale der Sprache berücksichtigt werden können, um Sprecher erfolgreich zu unterscheiden. In dieser Arbeit soll ein solches Sprecherverifikations-System implementiert, getestet und ausgewertet werden. Dabei sollen die drei unterschiedlichen implementierten Ansätze sowohl untereinander als auch mit einem bestehenden auf cepstralen Merkmalen basierenden System verglichen werden. Die zu implementierenden Methoden sind

- Statistische Modellierung der prosodischen Merkmale mit GMM.
- Erfassen der prosodisch segmentierten Sprachsignale als N-Gramm Modelle.
- Statistische Modellierung von prosodischen Segmenteigenschaften mit GMM.

Die Auswertung der Systeme erfolgt auf Grund einer Auswahl von Sprachsignalen aus dem NIST Speech Recognition Evaluation (SRE) 2004 Korpus. Es stehen drei Sets zur Verfügung. Mit dem Evaluationsset werden die Systeme ausgewertet und einige Parameter optimiert. Das Testset dient den abschliessenden Performancetests der implementierten Systeme und besteht aus 48 weiblichen und 33 männlichen Sprechern. Ein Set wird für die Erstellung des Universal Background Model (UBM) verwendet.

Die Verarbeitung der Sprachsignal-Daten erfolgt in MATLAB, für die statistische Modellierung mit GMM wird das HTK Toolkit verwendet. Die Auswertung aller Performancetests zeigt für die Einzelsysteme vergleichbare Tendenzen wie in referenzierten Publikationen.

3 Design

Im Folgenden wird die Sprecherverifikation im Allgemeinen und das Design der implementierten Ansätze im Speziellen vorgestellt.

Für die auf prosodischen Merkmalen basierenden Sprecherverifikationssysteme wird eine Datenvorverarbeitung benötigt, da nicht das Sprachsignal selber, sondern davon abgeleitete, je nach System unterschiedliche Grössen betrachtet werden. Auf Grund dieser extrahierten Daten können dann verschiedene Systeme eine Verifikation durchführen und zu möglicherweise unterschiedlichen Entscheidungen gelangen.

Die drei implementierten Systeme sind der GMM-, N-Gramm- und der sogenannte ProsVar-Ansatz. Zusätzlich wird ein cepstrales Referenzsystem beschrieben. Schliesslich werden auch Ansätze zur Systemkombination diskutiert. Durch Kombination der Resultate der Einzelsysteme wird versucht, eine Verbesserung der Leistung gegenüber der Einzelsystemleistung zu erzielen.

Details zur Datenvorverarbeitung und zur Optimierung verschiedener Systemparameter sind in den Kapiteln 4 bzw. 5 enthalten.

3.1 Sprecherverifikation

Die grundsätzliche Funktionsweise der Sprecherverifikation mit den implementierten Systemen ist aus dem Schema in Abbildung 3.1 ersichtlich.

Ein Sprachsignal wird vorverarbeitet, indem daraus bestimmte prosodische Parameter oder Verläufe Y_{A_i} extrahiert werden. Y_{A_i} bezeichnet im Folgenden dieje-

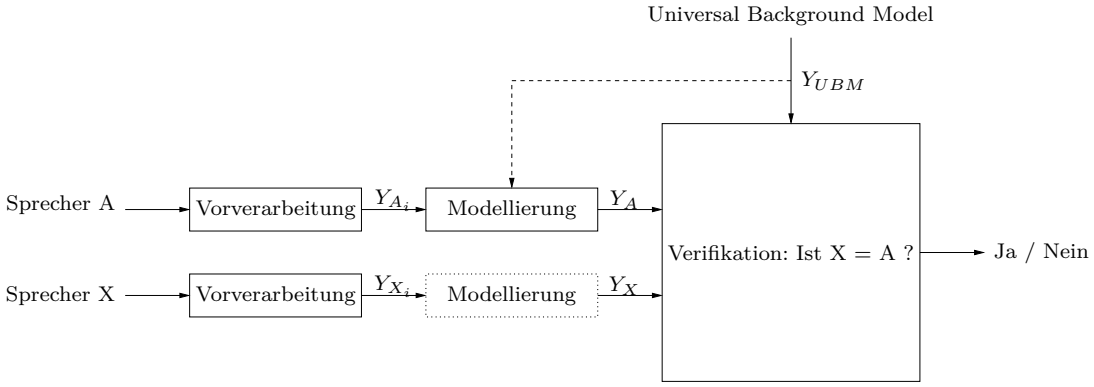


Abbildung 3.1: Funktionsweise der implementierten Sprecherverifikationssysteme.

nigen Daten, welche für die Modellbildung des Sprechers A verwendet werden. Im Weiteren werden dann nur noch diese Daten betrachtet und weiterverarbeitet.

Um mit den hier diskutierten Systemen erfolgreich eine Sprecherverifikation durchführen zu können, muss von Sprecher A eine grössere Menge von Sprachsignalen vorhanden sein. Daraus werden die für das System spezifischen Daten Y_{A_i} extrahiert und ein Modell Y_A für den Sprecher A erstellt. Soll nun verifiziert werden, ob es sich bei einem beliebigen Sprecher X um den Sprecher A handelt, müssen von einem Testsignal des Sprechers X die Parameter Y_{X_i} bestimmt werden, mit welchen das Modell Y_X gebildet werden kann. Durch den Vergleich der Modelle des Sprechers A, Y_A, sowie des Sprecher X, Y_X, kann dann entschieden werden, ob es sich bei den Sprechern A und X um dieselbe Person handelt oder nicht.

Da von einem Sprecher oft wenig Daten zur Verfügung stehen, wird ein sogenanntes Universal Background Model (UBM) Y_{UBM} zu Hilfe genommen. Dieses Modell enthält die statistische Beschreibung eines durchschnittlichen Sprechers gemäss dem angewandten System. Werden die Parameter des UBM gemäss denjenigen eines Sprechersignals adaptiert, d.h. in diesem Signal ersichtliche Trends der Verteilung werden qualitativ auf das UBM übertragen, kann ein Sprechermodell erzeugt werden, das demjenigen sehr ähnlich ist, das bei mehr verfügbaren Daten des Testsprechers entstanden wäre. Wie in Abbildung 3.1 durch eine gestrichelte Linie angedeutet, wird das UBM nicht bei allen Systemen bei der Modellbildung

beigezogen: Das N-Gramm System erstellt die Sprechermodelle gemäss [3] ausschliesslich auf Grund der vorhandenen Daten des jeweiligen Sprechers. Umgekehrt wird nur beim N-Gramm System eine Modellierung des zu verifizierenden Sprechers X durchgeführt (Y_X). Die anderen Systeme verwenden als Eingangsgrössen jeweils direkt die Parameter Y_{X_i}, welche analog zu den Y_{A_i} gebildet werden. Im Folgenden wird bei den Systemen GMM, ProsVar und CEP mit Modellbildung die Adaption des UBM mit Hilfe der Trainingsdaten gemeint.

3.2 Training

Die Grundlage für viele Sprecherverifikationen bilden das zu prüfende Testsignal und ein oder mehrere Referenzsignale, die zweifelsfrei einem Sprecher zugeordnet werden können. Je nach dem konkret eingesetzten Verifikationsverfahren ist vorgängig ein sogenanntes Training des Systems nötig. Das heisst, aus den vorhandenen, einem Sprecher zugeordneten Daten wird ein Modell für diesen spezifischen Sprecher generiert. In der Testphase wird das Testsignal anschliessend mit dem erstellten Sprechermodell verglichen. Es kann angenommen werden, dass die Verifikation bei mehr zur Verfügung stehenden Modelldaten zuverlässiger wird. In Kapitel 5 wird dieser Zusammenhang untersucht.

3.3 Systeme

Wir bilden drei verschiedene, auf prosodischen Merkmalen basierende Systeme: Die Modellierung von Grundfrequenz- und Energieverlauf als Gauss'sches Mischmodell (I), die Segmentierung und Beschreibung mit N-Grammen (II) sowie die Modellierung verschiedener Parameter des nach Stimmhaftigkeit segmentierten Sprachsignals als Gauss'sche Mischmodelle (III). Als Referenz dient ein auf GMM basierendes cepstrales System (IV).

Im Folgenden werden die Konzepte der ersten drei Systeme detailliert erläutert und das Referenzsystem kurz skizziert. Es soll aufgezeigt werden, wie die systemspezifischen Parameter Y_{A_i} bestimmt werden und wie der Entscheidungsprozess abläuft.

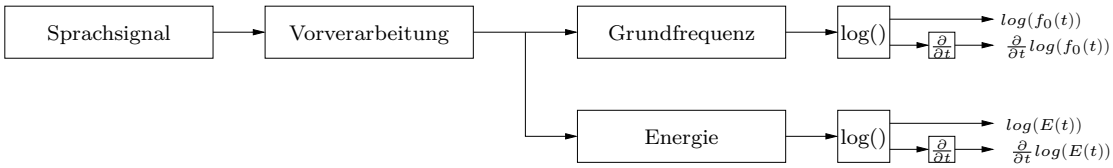


Abbildung 3.2: Extrahierung der Daten für das GMM System.

3.3.1 System I: GMM

Wie bereits erwähnt, sind für die auf prosodischen Merkmalen basierende Sprecherverifikation die prosodischen Eigenschaften Grundfrequenz und Intensität der Sprache zwei der wichtigsten Grössen. Insbesondere Mittelwert und Varianz werden als charakteristisch für einen Sprecher vermutet. Mit dem GMM Ansatz modellieren wir gemäss [1] den Grundfrequenz- und Intensitätsverlauf der Sprachmelodie eines Sprechers und trainieren damit ein Gauss'sches Mischmodell. Der Extrahierungsprozess ist in Abbildung 3.2 dargestellt.

Das Sprachsignal wird zuerst vorverarbeitet, indem längere Sprachpausen weggeschnitten werden. Anschliessend wird der Grundfrequenzverlauf extrahiert und der Energieverlauf bestimmt, wobei der Letztere durch das Proportionalitätsverhältnis die Intensität repräsentiert. Wie in [1] beschrieben, werden diese beiden Verläufe logarithmiert und abgeleitet.

Das Modell Y_A für den Sprecher A ergibt sich somit aus den Daten Y_{A_i} und dem UBM, wobei für Y_{A_i} gilt:

$$Y_{A_i} = \begin{cases} Y_{A_1} = log(f_0(kT)) \\ Y_{A_2} = \dfrac{\partial}{\partial t} log(f_0(t))|_{t=kT} \\ Y_{A_3} = E(kT) \\ Y_{A_4} = \dfrac{\partial}{\partial t} log(E(t))|_{t=kT} \end{cases}$$

Diese diskreten Kurven werden schliesslich mittels des Hidden Markov Toolkit (HTK)[1] als Gauss'sche Mischmodelle modelliert. Für die Entscheidungsfindung und die Anwendung des HTK Toolkit siehe Abschnitt 3.4 und 4.3.3. Umfassende Informationen zum HTK Toolkit gibt [5].

3.3.2 System II: N-Gramm

Allgemein wird unter einem N-Gramm eine Untersequenz der Länge N einer längeren Sequenz verstanden. Je nach der Anzahl c verschiedener Segmenttypen, aus welchen eine Sequenz bestehen kann und abhängig von der Länge N gibt es $k = c \cdot (c-1)^{N-1}$ mögliche verschiedene N-Gramme. Durch die Auswertung der relativen Auftretenshäufigkeiten der einzelnen N-Gramme in einer oder mehreren Sequenzen kann dann eine statistische Aussage über diese Daten gemacht werden.

Dieser Ansatz wird nun auf die Sprachsignale übertragen, mit dem Ziel, das Sprachsignal in eine auswertbare Sequenz von Klassen zu transformieren. Diese Sequenz wird dann gemäss obiger Anleitung ausgewertet um einen statistischen Fingerabdruck für den jeweiligen Sprecher zu modellieren.

Die verschiedenen Klassen sind definiert durch das tendenzielle Verhalten des Grundfrequenz- und Energieverlaufs innerhalb eines Signalabschnitts. Diese Segmentierung in eine Abfolge von Klassen für die Bildung der N-Gramme nach [3] ist in Abbildung 3.3 veranschaulicht. Für stimmhafte Signalabschnitte gibt es die insgesamt vier Klassen 2, 4, 6 und 8. Ausschlaggebend für die Zuteilung ist jeweils die Tendenz des Grundfrequenz- und Energieverlaufs. Die Segmentierung erfolgt durch paralleles Abtasten des Grundfrequenz- und Energieverlaufs. Das heisst, es muss eruiert werden, ob es sich bei einem Wendepunkt um ein lokales Minimum oder Maximum handelt. Bei jedem Wendepunkt einer der beiden Kurven ergibt sich automatisch auch ein Wechsel der Klasse.

Bei fallendem $f_0(t)$ und fallendem $E(t)$ wird das Segment der Klasse 2 zugeordnet, bei fallendem $f_0(t)$ aber steigendem $E(t)$ der Klasse 4. Segmente mit

[1] Siehe auch http://htk.eng.cam.ac.uk

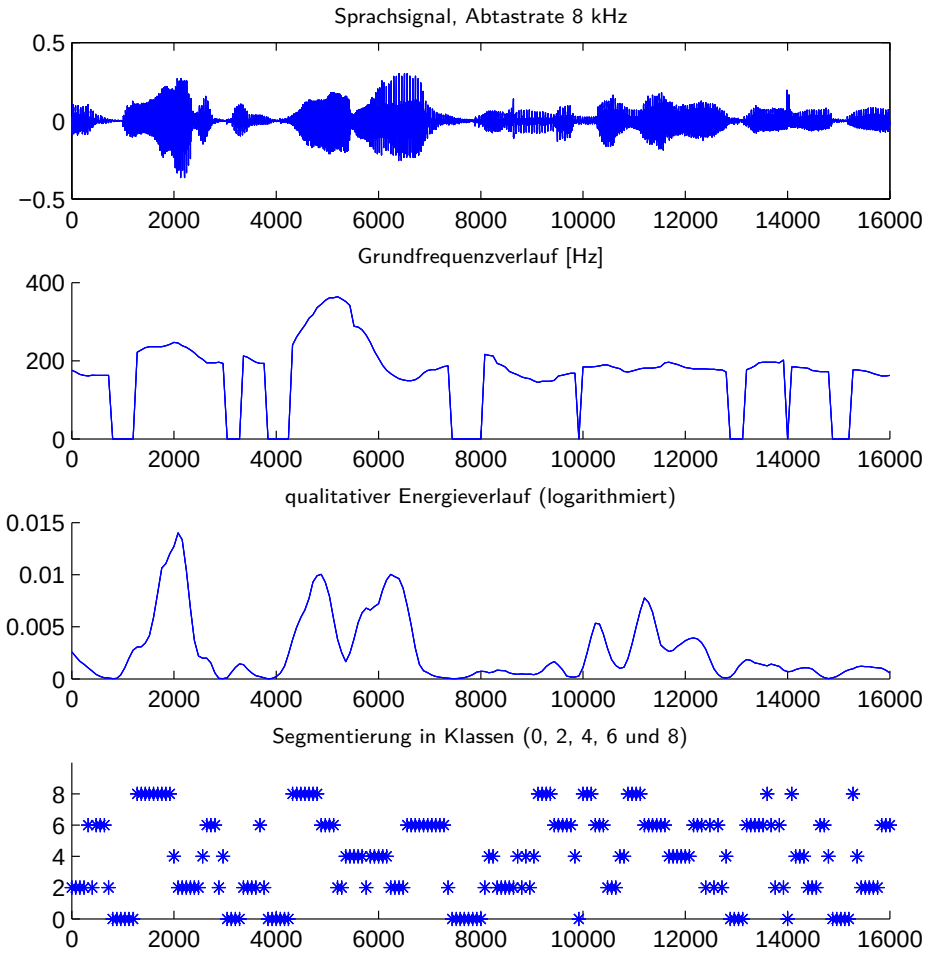


Abbildung 3.3: Segmentierung des Sprachsignals und Extraktion der Klassen: Der qualitative Verlauf von Grundfrequenz- und Energieverlauf des Sprachsignals bestimmt die Klasse, welcher das Segment zugeordnet wird.

steigendem $f_0(t)$ aber fallendem $E(t)$ sind in der Klasse 6 zusammengefasst und jene mit steigendem $f_0(t)$ und steigendem $E(t)$ in der Klasse 8. Sprachsegmente, in denen keine Grundfrequenz detektierbar ist, die also stimmlos sind, werden der Klasse 0 zugeteilt. In Tabelle 3.1 ist diese Zuordnung übersichtlich dargestellt.

Klasse	Zeitlicher Verlauf
0	stimmlos
2	fallendes $f_0(t)$ und fallendes $E(t)$
4	fallendes $f_0(t)$ und steigendes $E(t)$
6	steigendes $f_0(t)$ und fallendes $E(t)$
8	steigendes $f_0(t)$ und steigendes $E(t)$

Tabelle 3.1: Zuordnung der Segmenttypen zu den Klassen.

Somit wird das Sprachsignal auf Grund dessen Grundfrequenz- und Energieverlauf in eine diskrete Folge von Klassen unterteilt, die anfangs erwähnte Sequenz von Klassen. Optional kann auch die Länge der jeweiligen Klassensegmente mitberücksichtigt werden, also z.B. bei besonders langem Anstieg von Grundfrequenz und Energie. So gibt es für jeden Klassentypen die zwei Ausprägungen *long* und *normal*, womit sich die Anzahl möglicher Klassentypen auf zehn verdoppelt. Die zusätzlichen Klassen erhalten jeweils die dazwischenliegenden ungeraden Nummern zugeteilt, also 1, 3, 5, 7 und 9. Die Zuteilung, wann ein Klasse als *long* gilt, wird in Abschnitt 5.2.2 beschrieben.

Für die gewünschte Länge N der N-Gramme wird nun das Signal ausgewertet. Das heisst es wird gezählt, wie oft jedes mögliche N-Gramm in der aus dem Sprachsignal erstellten Abfolge von Klassen vorkommt. Die Parameter Y_{A_i} stellen im N-Gramm System also die Abfolge von Klassen für ein Sprachsignal dar. Das Modell Y_A wird für den Sprecher A aus allen verfügbaren Sprachsignalen erstellt und entspricht demnach der relativen Auftretenshäufigkeit aller N-Gramm Typen von Y_{A_i} für die gewünschte Länge N.

Bei der Verifikation eines Sprechers erfolgt der Vergleich des N-Gramm Fingerabdrucks des Sprechers A, Y_A, und des Sprechers X, Y_X, gemäss [4] nach folgender Formel:

$$Score = \frac{\sum_k N_{tokens}(k) \cdot log[\Lambda_A(k)/\Lambda_{BG}(k)]}{\sum_k N_{tokens}(k)}$$

wobei k den N-Gramm Typen spezifiziert und $N_{tokens}(k)$ die Anzahl N-Gramme vom Typ k im Testsignal repräsentiert. Λ_A und Λ_{BG} stehen für die relativen Auftretenshäufigkeiten des jeweiligen N-Gramms im Referenzmodell des Sprechers A bzw. im Background Model und sind definiert durch

$$\Lambda_i(k) = \frac{\#\text{ gezähltes Vorkommen des N-Gramms vom Typ } k \text{ im Modell } i}{\#\text{ total gezählte N-Gramme im Modell } i}$$

Λ_{BG} stellt hier einen Mittelwert für jedes N-Gramm über alle betrachteten Sprecher des entsprechenden Sets, aus welchem das UBM gebildet wird, dar. Anhand eines Schwellwertes (Threshold) für den berechneten Score entscheidet sich jeweils, ob die Sprecher identisch sind oder nicht.

3.3.3 System III: ProsVar

Im N-Gramm Ansatz wird das Sprachsignal auf Grund seines Grundfrequenz- und Energieverlaufs in Klassen unterteilt. Dabei geht ein grosser Teil an lokaler Information verloren. Der N-Gramm Ansatz berücksichtigt nur qualitative Komponenten; so wird zum Beispiel ein Anstieg der Grundfrequenz als solcher registriert, jedoch nicht dessen quantitative Ausprägung. Ebenso wird die Länge einer Klasse nur durch *long* oder *normal* beschrieben, innerhalb dieser Kategorien jedoch nicht weiter klassifiziert. In [2] wird ein Ansatz beschrieben, wie diese Informationen erfasst werden können.

Im Unterschied zum N-Gramm System wird nur der Grundfrequenzverlauf des Sprachsignals betrachtet und das Signal in stimmlose und stimmhafte Abschnitte segmentiert. Von diesen Segmenten wird nun einerseits die Länge erfasst, und andererseits werden die stimmhaften Segmente weiter quantitativ analysiert. Konkret werden der Median und die Steigung der Grundfrequenzverläufe der jeweiligen stimmhaften Abschnitte berechnet. Der Medianwert wird logarithmiert, die Steigung der Grundfrequenz innerhalb eines stimmhaften Segements wird mittels einer

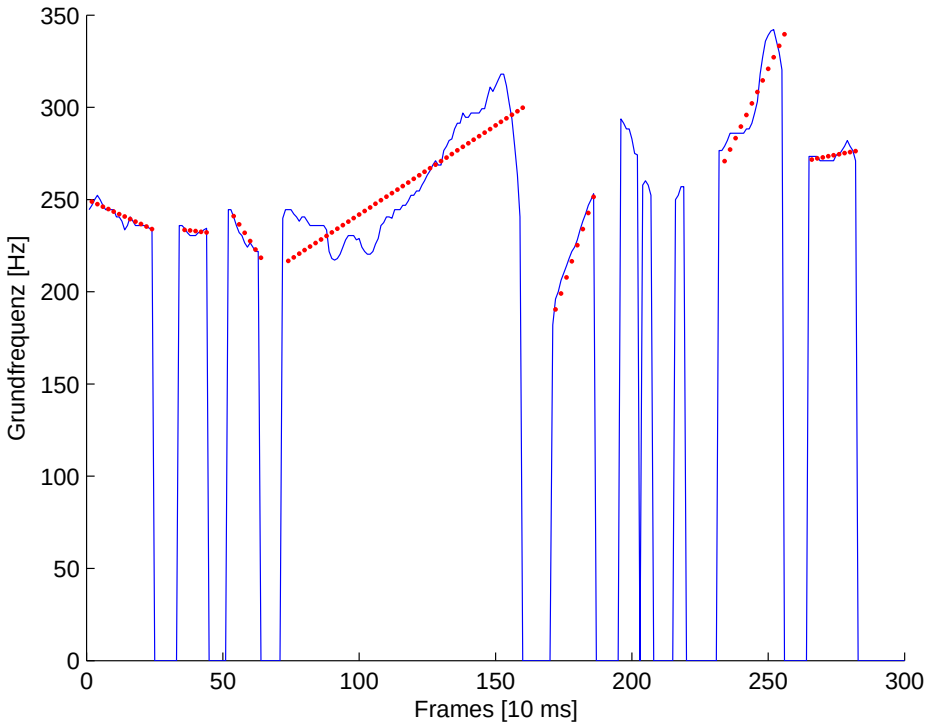


Abbildung 3.4: Die Segmentierung des Sprachsignals für das ProsVar System auf Grund stimmhafter und stimmloser Abschnitte. Der Grundfrequenzverlauf ist blau gezeichnet, die mittlere Steigung pro Segment rot.

linearen Mininum Mean Square Error (MMSE) Schätzung berechnet, wie dies in Abbildung 3.4 illustriert wird. Das Sprechermodell Y_A wird aus den Y_{A_i}, d.h. den extrahierten Daten und dem UBM gebildet, wobei für die Y_{A_i} gilt:

$$Y_{A_i} = \begin{cases} Y_{A_1} = t_{NS1}, t_{NS2}, t_{NS3}, ... \\ \\ Y_{A_2} = t_{S1}, t_{S2}, t_{S3}, ... \\ \\ Y_{A_3} = log(med_{S1}), log(med_{S2}), log(med_{S3}), ... \\ \\ Y_{A_4} = slope_{S1}, slope_{S2}, slope_{S3}, ... \end{cases}$$

S bezeichnet stimmhafte und NS stimmlose Segmente. Die Verifikationsentscheidung basiert auf dem HTK Toolkit, analog zum GMM System.

3.3.4 System IV: Cepstrales Vergleichssystem CEP

Das cepstrale Vergleichssystem führt wie in [6] beschrieben eine MFCC-Merkmalsextraktion für die Sprachsignale durch und modelliert die erhaltenen Merkmalsvektoren anschliessend mit einem Gauss'schen Mischmodell. MFCC steht für Mel Frequency Cepstral Coefficients, wobei Mel die Masseinheit für die wahrgenommene Tonhöhe ist. Über das ganze betrachtete Spektrum, das durch die halbe Abtastfrequenz nach oben beschränkt ist, wird dabei eine Anzahl Dreiecksfilter verteilt und auf das Kurzzeit-Fourierspektrum angewendet. Das so entstandene Mel Spektrum beschreibt nun gewissermassen ganze Frequenzkanäle des ursprünglichen Signals. Mittels Logarithmierung und Anwendung der diskreten Cosinustransformation auf das Mel Spektrum ergibt sich eine Glättung des Spektrums, wodurch unerwünschte hochfrequente Anteile wegfallen. Die Mel-Frequenz-Cepstrum-Koeffizienten können nun direkt aus der Energie an den Ausgängen der Filter berechnet werden. Weiterführende Beschreibungen zu MFCC finden sich in [7].

Zur statistischen Beschreibung werden die erhaltenen Merkmalsvektoren Y_{A_i} eines Sprechers A werden analog zum GMM System im HTK Toolkit mit Gauss'schen Mischmodellen modelliert.

3.4 Modellierung von GMM mit dem HTK Toolkit

Die Ansätze GMM, ProsVar und CEP benützen für die Entscheidungsfindung und Modellierung das bereits erwähnte HTK Toolkit. Dies ist eine vom Cambridge Universitiy Engineering Department entwickelte Software, welche primär dazu dient, Hidden Markov Modelle (HMM) zu erstellen. Das Gauss'sche Mischmodell, das bei diesen drei Ansätzen eingesetzt wird, ist ein Spezialfall eines HMM mit nur einem versteckten Zustand.

Mit dem HTK Tool kann nun das UBM in eine verwertbare Form transformiert werden. Dazu werden zuerst der globale Mittelwert und die Kovarianzen des Modells berechnet. Dann werden mittels Baum-Welch Algorithmus die Parameter des HMM berechnet.

Das Training eines Sprechers erfolgt auf ähnliche Weise. Schliesslich stehen das UBM und das an den Sprecher adaptierte UBM (trainiertes Modell eines bestimmten Sprechers) zur Verfügung. Nun wird mit einem Viterbi-Erkenner zuerst das Testsignal gegen das trainierte Modell getestet, was in einem $Score_{Target}$ resultiert. Anschliessend wird das Testsignal gegen das UBM getestet um die Individualität des Sprechers zu evaluieren; dies ergibt $Score_{UBM}$. Das Ergebnis des Tests ist die Differenz dieser beiden Scores: $Score = Score_{Target} - Score_{UBM}$. Anhand eines Schwellwertes für diesen Score entscheidet sich schliesslich, ob die Sprecher als identisch eingestuft werden.

3.5 Kombination der Systeme

Durch Kombination der drei Systeme GMM, N-Gramm und ProsVar wird versucht, eine insgesamt bessere Performance als diejenige der Einzelsysteme zu erzielen. Dazu werden die zwei Ansätze „Linearkombination“ und „Best-of-three“ verfolgt.

Lineare Kombination Die von den Einzelsystemen zurückgelieferten Masse für die Übereinstimmung zweier Sprecher, die sogenannten Scores, werden linear kombiniert. Ein Gewichtungsfaktor für jedes System wird bestimmt um den unterschiedlichen Zuverlässigkeiten der Systeme gerecht zu werden.

Best-of-three Da drei Systeme zur Verfügung stehen, kann ein einfacher Mehrheitsentscheid gefällt werden. Sind zwei oder drei Systeme für die gleiche Entscheidung, so wird diese getroffen.

3.6 Performanceauswertung der Systeme

Ein Sprecherverifikationssystem hat jeweils zwei Entscheidungsmöglichkeiten: Entweder werden zwei Sprecher als *identisch* oder als *nicht identisch* eingestuft. Daher gibt es die vier in Tabelle 3.2 dargestellten Entscheidungsszenarien:

Fall	Sprechermodell	Sprecher	Entscheidung	Bemerkung
1	X	X	Sprecher identisch	
2	X	Y	Sprecher identisch	Fehler 2. Art
3	X	X	Sprecher nicht identisch	Fehler 1. Art
4	X	Y	Sprecher nicht identisch	

Tabelle 3.2: Mögliche richtige und falsche Entscheidungen bei der Sprecherverifikation.

Richtige Entscheidungen werden in Fall 1 und Fall 4 gefällt, wo die Sprecher korrekterweise als identisch bzw. nicht identisch erkannt werden. Fall 2, bei dem ein Sprecher erfolgreich verfiziert wird, obwohl es sich um die falsche Person handelt, wird in der Wahrscheinlichkeitstheorie allgemein als Fehler 2. Art bezeichnet. Fall 3, bei dem ein Sprecher fälschlicherweise nicht verifziert werden kann, obwohl es sich um dieselbe Person handelt, wird als Fehler 1. Art bezeichnet. Diese beiden Fehler werden in der Praxis, je nach konkreter Anwendung, als unterschiedlich gravierend eingestuft. So dürfte es beispielsweise weniger schlimm sein, wenn der Kunde einer Bank nicht sofort als Inhaber eines Kontos verifiziert werden kann, als wenn eine Person zu Unrecht Zugriff auf fremde Daten erlangt. Hier würde der Fehler 2. Art als kritischer eingestuft als derjenige 1. Art.

Da ein System, welches die Fehler 2. Art effizient unterdrückt oftmals hohe Raten der Fehler 1. Art zur Folge hat und umgekehrt, stellt bei der Auswertung solcher Systeme die Equal Error Rate (EER) ein gutes Gesamtmass für die Performance dar. Über die EER lassen sich verschiedene Systeme aussagekräftig vergleichen. Sie bezeichnet die Fehlerrate im Schnittpunk der beiden Verteilungsfunktio-

nen $d_0 = D(x_0)$ und $d_1 = 1 - D(X1)$, wobei $D(x_0)$ die Verteilungsfunktion des Ähnlichkeitsmasses der als identisch und $D(x_1)$ diejenige der als nicht identisch einzustufenden Sprecher bezeichnet. Etwas einfacher formuliert wird die Entscheidungsschwelle für das Ähnlichkeitsmass zweier Sprecher angegeben, so dass die Wahrscheinlichkeit für Fehler 1. und 2. Art gleich hoch ist.

4 Implementation

Dieses Kapitel beschreibt, wie die drei auf prosodischen Eigenschaften basierenden Sprechererkennungssysteme GMM, N-Gramm und ProsVar sowie das cepstrale Vergleichssystem in MATLAB implementiert sind. Wichtig sind insbesondere auch die Datenverwaltung und die Vorverarbeitung der Eingangsgrössen für die jeweiligen Erkennungsysteme. Schliesslich wird noch auf die Entscheidungsfindung bei der Verifikation und die Kombination mehrerer Systeme eingegangen.

4.1 Datenverwaltung

Die zur Verfügung stehenden NIST Daten liegen als WAVE-Dateien vor, deren Namen mit vier Buchstaben codiert sind. In einer Textdatei werden diese Namen den korrekten Sprechern zugeordnet. Neben dem Geschlecht des Sprechers ist hier auch vermerkt, welche Daten für Training und welche für Tests bestimmt sind und welchem der drei Sets der Sprecher angehört. Je eines dieser Sets wird zur Bildung der UBM, für die Optimierung verschiedener Parameter und für die abschliessenden Tests verwendet.

Die Funktion „ImportFile.m“ generiert aus dieser Textdatei die drei Matrizen *V1*, *V2* und *V3* wie in Abbildung 4.1 dargestellt. Die Matrix *V1* enthält die zum Sprecher X korrespondierende Setnummer und das zutreffende Geschlecht. *V2* weist jedem Sprecher X Trainingsdaten zu, mit denen der Sprecher X modelliert werden kann. Die dem Sprecher X zugewiesenen Testdaten sind in *V3* angegeben.

Die verwendete Abbildung der Dateinamen auf Dateinummern (z.B. „taag“ auf „80324212“) ist umkehrbar und erleichtert die Handhabung in MATLAB erheblich, da auf die Verwendung von String Matrizen verzichtet werden kann.

V1

Set	Sex	ID
1	f	3550
2	m	5125
...	...	...

V2

ID	Datei (Training)	...
3550	20060610	...
5125	20100513	...
...	...	...

V3

ID	Datei (Test)	...
3550	24012504	...
5125	24011823	...
...	...	...

Abbildung 4.1: Die Matrizen *V1*, *V2* und *V3* enthalten die Kontextinformationen zu jeder Sprachdatei.

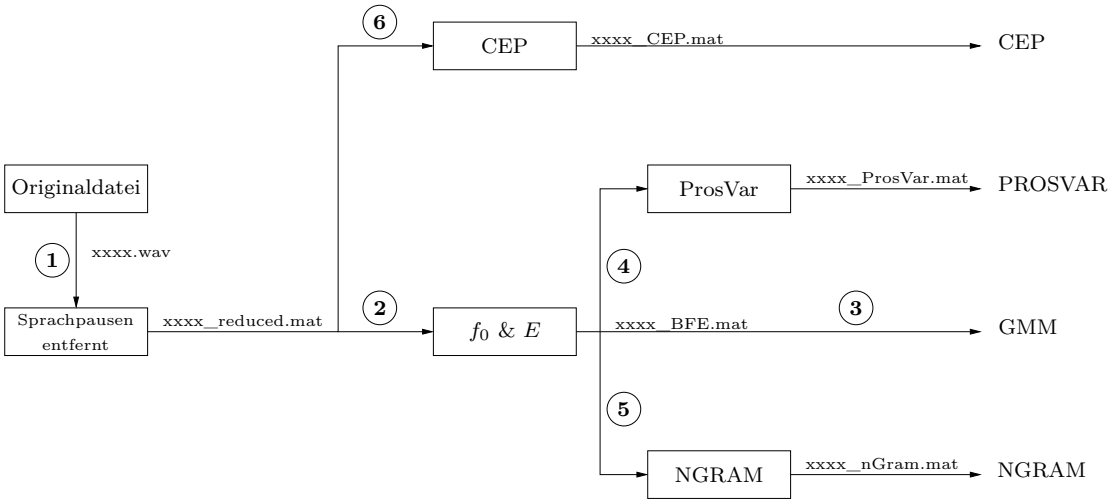


Abbildung 4.2: Vorverarbeitungsstufen für die verschiedenen Systeme.

4.2 Datenverarbeitung

Für die vier Systeme GMM, N-Gramm, ProsVar und CEP werden diverse Eingangsgrössen benötigt. In mehreren gestaffelten Schritten werden für jedes System vorgängig die nötigen Operationen ausgeführt. Abbildung 4.2 zeigt dies schematisch.

4.2.1 Entfernen der Sprachpausen (1)

Da die zur Verfügung gestellten NIST Daten einkanalig aufgezeichnete Telefongespräche enthalten (d.h. es ist nur einer der beiden Gesprächsteilnehmer zu hören), entfernt die Funktion „RemoveSpeechPauseNormal.m“ die Sprachpausen aus den Originaldateien. Es wird auf Grund einer minimalen Lautstärke zwischen Sprach-

und Pausesequenzen unterschieden, wobei Pausen so weggeschnitten werden, dass nur wenig relevante Sprachgeräusche verloren gehen. Dieser Schritt ist sinnvoll, da für die prosodische Analyse nur deutlich gesprochene und hörbare Sequenzen von Bedeutung sind. Der Postfix der generierten Dateien lautet „_reduced.mat“, der Inhalt entspricht dem mit 8 kHz gesampelten WAVE-Signal.

4.2.2 Bestimmen des Grundfrequenz- und Energieverlaufs (2)

Aus den „_reduced.mat“-Dateien werden nun der Grundfrequenz- und der Energieverlauf extrahiert. Die Funktion „getspecfundamentalfreq.m“ gibt aus einem Sprachabschnitt den Grundfrequenzverlauf zurück. Sie wurde uns vom Institut zur Verfügung gestellt. Aus Performancegründen wird die Funktion sequentiell ausgeführt, d.h. es werden jeweils Sprachabschnitte von ca. 100'000 Samples nacheinander verarbeitet. Als Parameter für die Kurzzeitanalyse werden 300 Samples als Fensterbreite (*winsize*) und 80 für die Fensterverschiebung (*winshift*) gesetzt.

Die resultierende Grundfrequenzkurve weist gegenüber dem Ursprungssignal eine um den Faktor 1/80 kleinere Zahl von Funktionswerten auf, da jeweils pro Fensterverschiebung (80 Samples) nur ein Grundfrequenzwert berechnet wird. Diese Stauchung muss beim Energieverlauf ebenfalls berücksichtigt werden, da eine gleiche Anzahl von Werten im Grundfrequenz- wie Energieverlauf wünschenswert ist. Die Funktion „GetEnergy.m“ wendet deshalb ein Meanfilter auf den Energieverlauf an und tastet anschliessend mit der entsprechenden Verschiebung (*winshift*) ab. Da wir nur am qualitativen Energieverlauf interessiert sind, wird in Übereinstimmung mit [3] noch dessen Logarithmus berechnet.

Das Detektieren der Grundfrequenz birgt gewisse Risiken. Die Gefahr, dass Frequenzwerte um das Doppelte zu hoch erkannt werden (Doubling), ist relativ gross. Das heisst, es wird statt der Grundfrequenz die erste Harmonische erkannt. Dieser Effekt tritt vor allem bei qualitativ schlechteren Telefongesprächen oft auf. Zwar wird als Abhilfe bereits die höchste zu detektierende Grundfrequenz auf einen fixen Wert gesetzt gesetzt (in dieser Implementation 600 Hz), trotzdem kommt das Doubling immer noch häufig vor. Die Funktion „BaseFreqcorrector.m“ nimmt sich diesem Problem an. Es wird zuerst der durchschnittliche Wert aller Grundfrequenz-

Samples in einer Gesprächssequenz berechnet (dabei sind stimmlose Teile ausgeschlossen). Die Grenze, oberhalb welcher die Grundfrequenz halbiert werden soll, wird auf einen vom Durchschnitt abgeleiteten Wert gesetzt, der im Kapitel 5 näher erläutert wird. Für stimmlose Abschnitte wird der Wert der Grundfrequenz auf Null gesetzt.

Die Ergebnisse dieser zweiten Vorverarbeitungsstufe, also der Grundfrequenz- und Energieverlauf, werden als Dateien mit dem Postfix „_BFE.mat“ gespeichert.

4.2.3 Bestimmen der GMM Eingangsgrössen (3)

Da gemäss [2] der Grundfrequenzverlauf bzw. dessen Logarithmus normalverteilt ist, wird für das GMM System der Logarithmus der Grundfrequenz betrachtet. Dasselbe gilt auch für den Energieverlauf. Schliesslich wird noch die zeitliche Ableitung der beiden Verläufe berechnet, womit dann die Eingangsgrössen für das GMM System bereitstehen.

4.2.4 Bestimmen der ProsVar Eingangsgrössen (4)

Das ProsVar System benötigt folgende Eingangswerte, welche auf dem vorher berechneten Grundfrequenzverlauf basieren: Median und Steigung der Grundfrequenz pro stimmhaftes Segment, Dauer der stimmhaften Segmente sowie die Dauer der stimmlosen Segmente. Das Signal bzw. dessen Grundfrequenzverlauf wird also in stimmhafte und stimmlose Elemente segmentiert. Für die stimmhaften Abschnitte wird die Steigung mit dem minimalen quadratischen Fehler (Minimum Mean Square Error, MMSE) in der Funktion „GetSegParameters.m“ linearisiert. Diese Funktion bestimmt zuerst eine Schätzung der Steigung auf Grund des ersten und letzten Wertes eines stimmhaften Segmentes und variiert dann die Steigung und den Biaswert, um so den kleinsten MSE zu finden. Zusätzlich wird der Median der Grundfrequenz pro Segment bestimmt. Möglicherweise nicht korrigiertes Doubling lässt sich als „Salt and Pepper Noise“ auffassen, d.h. als kurzzeitig starke Ausreisser der Grundfrequenz. Der Median liefert hier aussagekräftigere Resultate als beispielsweise der Mittelwert, da er unempfindlicher gegenüber dieser Art von Störungen ist. Schliesslich werden auch die Längen der stimmhaften und der

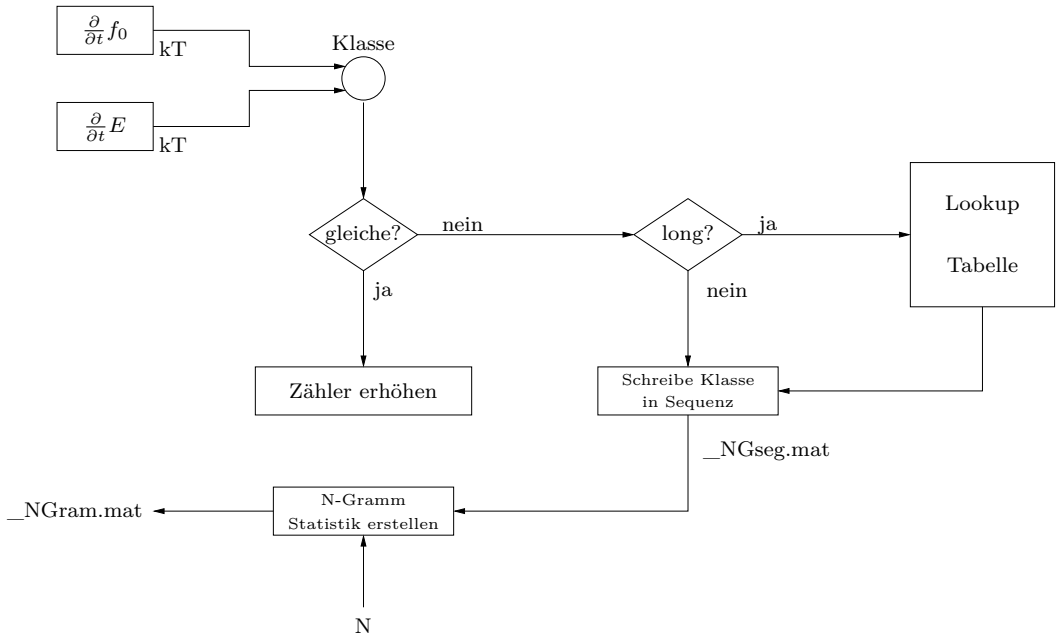


Abbildung 4.3: Berechnung der N-Gramme auf Grund des Grundfrequenz- und Energieverlaufs.

stimmlosen Segmente erfasst. Die vier Grössen werden für jedes mindestens $100ms$ lange Segment aufgezeichnet und als „_ProsVar.mat“-Datei gespeichert.

4.2.5 Berechnung der N-Gramme (5)

Die N-Gramme werden auf Grund des qualitativen Grundfrequenz- und Energieverlaufs berechnet, welcher gemäss 4.2.2 in „_BFE.mat“ gespeichert ist. Das genaue Vorgehen ist in Abbildung 4.3 illustriert.

Nach der Berechnung der Ableitungen von Grundfrequenz- und Energie wird entschieden, welcher Klasse das jeweilige Wertepaar $(\frac{\partial}{\partial t} f_0(t),\ \frac{\partial}{\partial t} E(t))$ angehört. Mögliche Kombinationen sind ++, –, +-, -+ und stimmlos, wobei „+“ einem positiven und „-“ einem negativen Ableitungswert entspricht. Die Klassenzuteilung erfolgt gemäss Tabelle 3.1. Falls die Klasse des aktuellen Wertepaars identisch ist mit derjenigen des vorangehenden, werden lediglich ein Zähler erhöht und die nächsten zwei Werte geladen. Es erfolgt kein neuer Eintrag in der Klassensequenz. Falls die zwei Wertepaare unterschiedlich klassifiziert werden, wird die Klasse des neuen Segments ebenfalls in die Klassensequenz eingetragen. Der Zähler dient da-

bei der Evaluierung der Länge der jeweiligen Segmente: Es wir überprüft, ob der letzte Eintrag das Prädikat *long* erhält oder nicht; die Minimallänge ist dabei abhängig von der Klasse. In einer separaten Tabelle (siehe Abschnitt 5.2.2) sind die Längen für jede Klasse definiert. Für die N-Gramme mit lediglich 5 Klassen wird der Zähler nicht benötigt.

Die Klassensequenzen werden als „_Ngseg.mat“ gespeichert. Dieser Zwischenschritt ist nötig, da die Länge N der N-Gramme dynamisch soll gewählt werden können. Die nun vorliegende Sequenz wird in N-Gramme unterteilt. Dazu bildet die Funktion „BuildNGRAMS.m“ auf Grund des Parameters N die N-Gramm Matrix und einen Vektor mit der jeweiligen Anzahl N-Gramme eines Typs k. Diese Auflistung wird als „_NGramX0.mat“ gespeichert. Die Erstellung der N-Gramm Matrix ist in 4.4 visualisiert für eine Klassensequenz der Länge 8, wobei $N = 3$ und $k = 24$ ist. Die vier Klassen werden *A, B, C* und *D* genannt. Im ersten Schritt werden also die ersten 3 Klassen (also das erste 3-Gramm [*A, B, C*]) in die erste Zeile der N-Gramm Matrix geschrieben, im zweiten Schritt die Klassen 2 bis 4 ([*B, C, B*]) in Zeile 2 und so weiter. Abschliessend wird im Statistikvektor gespeichert, welcher N-Gramm Typ wie oft in der Klassensequenz enthalten ist.

4.2.6 Berechnung der cepstralen Koeffizienten (6)

Um die auf prosodischen Eigenschaften basierenden Systeme aussagekräftig vergleichen zu können, wird ein cepstrales Referenzsystem benötigt. Dazu extrahieren wir aus dem pausenreduzierten Sprachsignal die Mel-Frequency Cepstral Coefficients (MFCC), die eine kompakte Darstellung des Spektrums erlauben. In unserem Fall wird eine Mel Filterbank mit 34 Filtern verwendet. Zur Extraktion der Merkmale wird das pausenreduzierte Sprachsignal in Analysefenster von $25ms$ Länge mit einer Überlappung von $10ms$ aufgeteilt. Unter Anwendung von 34 Dreiecksfiltern werden anschliessend mit der vom Institut zur Verfügung gestellten Funktion „mfcc.m“ die 12 ersten cepstralen Koeffizienten berechnet, wobei der nullte Koeffizient (die Energie) nicht betrachtet wird. Abschliessend werden der Mittelwert aller Koeffizienten komponentenweise abgezogen und diese mittelwertfreien Daten als „_coeff.mat“ gespeichert.

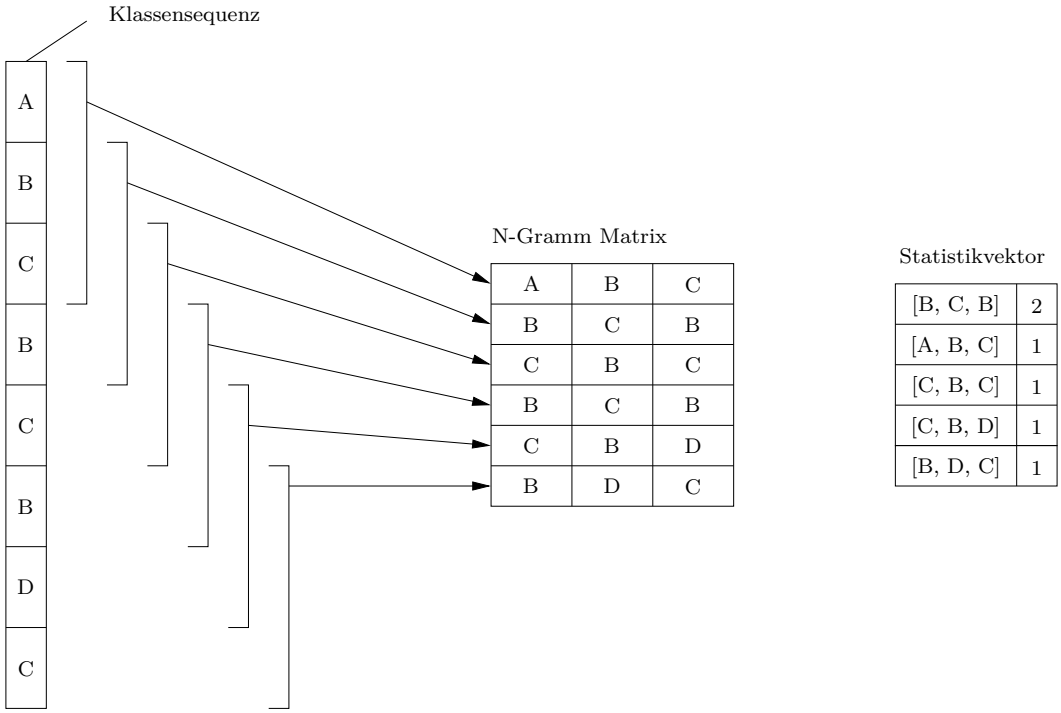


Abbildung 4.4: Erstellung der N-Gramm Matrix und des zugehörigen Statistikvektors.

4.3 Verifikation

Die Systeme GMM, CEP und ProsVar werden mit Hilfe des HTK Toolkit verarbeitet. Das N-Gramm System wird hingegen vollständig in MATLAB ausgewertet. Im Folgenden wird auf die Erstellung der jeweiligen UBM, der Sprechermodelle sowie auf den genauen Ablauf der Tests eingegangen.

4.3.1 Bildung des UBM und trainieren des Sprechermodells

Aus den vorgegebenen Datensets 0, 1 und 2 verwenden wir für die Erstellung des UBM das Set 0. Set 2 enthält die Daten für die Optimierung verschiedener Systemparameter und Set 1 diejenigen für die abschliessenden Performancetests.

N-Gramm

Das UBM für das N-Gramm System wird von der Funktion „CalculateUBMforN-GRAM.m“ erstellt. Hierzu werden alle Dateien aus Set 0 für das entsprechende N-Gramm ausgewertet und die Anzahl der entsprechenden N-Gramm Typen aufaddiert (je ein UBM für $N = 2, 3, 4, 5$). Somit entsteht je ein UBM für beide Geschlechter und jeden N-Gramm Typ. Alle UBM werden sowohl für normale wie auch für modifizierte N-Gramme erstellt (mit Attribut *long*).

Das Sprechermodell wird genau gleich gebildet, indem alle für den entsprechenden Sprecher zur Verfügung stehenden Daten bzw. die Auftretenshäufigkeiten der N-Gramm Typen aufaddiert werden.

GMM, ProsVar und CEP

Das UBM für alle im HTK Tool zu analysierenden Systeme (GMM, ProsVar und CEP) wird ähnlich gebildet. Eine Auswahl an Daten wird zusammengefasst und im HTK Tool als UBM modelliert. Die Modellierungszeit hängt stark von der Menge der zu analysierenden Daten ab, weshalb hier je nach System unterschiedlich vorgegangen werden muss. Aus Zeitgründen war es im Rahmen dieser Arbeit nicht möglich, für jedes System ein UBM mit allen verfügbaren Daten zu erstellen.

- GMM: Alle Dateien aus Set 0 werden berücksichtigt, aber jeweils nur ein Zwanzigstel der Daten pro Datei wird verwendet.
- ProsVar: Alle Dateien aus Set 0 werden vollständig verwendet.
- CEP: Alle Dateien aus Set 0 werden berücksichtigt, jedoch wird hier ein Fünfzigstel der Daten pro Datei verwendet.

4.3.2 Testablauf

Um die Sprecherverifikation durchzuführen werden jeweils ein passendes UBM, Modellierungsdaten eines bekannten Sprechers A und eine Testdatei des zu verifizierenden Sprechers X benötigt. Auf Grund dieser Information soll nun entschieden werden, ob es sich beim fraglichen Sprecher um die Person A handelt. Der Testablauf funktioniert nach der nachstehend in Pseudo-Code dargestellten Prozedur.

```
1   % Laden der Matrizen V1, V2, V3
2   V1 = [Set; Sex; SpeakerID]
3   V2 = [SpeakerID; TrainingsfileID1; TrainigsfileID2 ... ]
4   V3 = [SpeakerID; TestfileID1; TestfileID2 ... ]
5   SexID = S
6   SetID = N
7
8   for i = [SpeakerID1, SpeakerID2, ... ]
9    for j = [SpeakerID1, SpeakerID2, ... ]
10     for k = [TestfileID1 von j, TestfileID2 von j, ... ]
11       if(AND(Sex(j) == S, Sex(i) == S, SetID(j) == N, SetID(i) == N)
12          TRAINMODEL = TrainingsfileIDs von i
13          UBM = UBM(SexID)
14          TEST = k
15          Result(i,j,1) = D E C I D E(UBM, TRAINMODEL, TEST, Option)
16          if( i == j )
17           Result(i,j,2) = 1
18          else
19           Result(i,j,2) = 0
20          end
21        end
22      end
24    end
25  end
```

Zu Beginn werden alle Informationen zu den Sprachdateien wie in Kapitel 4.1 besprochen aus den drei Datenverwaltungsmatrizen geladen (Zeilen 2 bis 6). Anschliessend wird mit den Variablen *SexID* und *SetID* definiert, welches Geschlecht

und welches Datenset evaluiert werden soll.

Anschliessend wird jeder Sprecher gegen jeden Sprecher verifiziert (Zeilen 8 und 9). Da für jeden Sprecher in der Regel mehrere Testdateien vorhanden sind und mit mehreren Tests ein System noch besser analysiert werden kann, führen wir für alle zur Verfügung stehenden Testdateien einen Test durch (Zeile 10). Die Verifikation wird nur durchgeführt, falls die beiden Sprecher das vorgegebene Geschlecht haben und dem definierten Set angehören (Zeile 11). Nun wird das Modell von Sprecher i erstellt indem alle *FileID*s von Sprecher i zu *TRAINMODEL* zusammenfasst werden (Zeile 12). Das UBM wird nach Methode (GMM, N-Gramm, ProsVar, CEP) und Geschlecht gewählt (Zeile 13). Zeile 14 wählt die vom Sprecher j gegen den Sprecher i zu verifizierende Testdatei. Schliesslich wird eine Entscheidung getroffen (Zeile 15). Über den Parameter *Option* kann die Menge der Modellierungsdaten reduziert werden (siehe Kapitel 5). Um schliesslich das Ergebnis auswerten zu können, wird auf Grund der in den Matrizen *V2* und *V3* gespeicherten Informationen die getroffene Entscheidung auf ihre Richtigkeit überprüft.

4.3.3 Entscheidungsfindung

Der Funktion *D E C I D E* werden die Kennungen eines Modells, der Testdaten sowie eines UBM übergeben. Daraus berechnet diese ein Mass für die Übereinstimmung der Testdaten mit dem Modell, unter Berücksichtigung des entsprechenden UBM.

N-Gramm

Für das N-Gramm System wird die in Abschnitt 3.3.2 erläuterte Formel in der Funktion „VAREvaluateNGRAM.m“ angewandt und direkt ein Score ausgegeben.

GMM, ProsVar, CEP

Für die Systeme GMM, ProsVar und CEP werden dem HTK Tool die Testdaten, das UBM und das Sprechermodell (adaptiertes UBM) übergeben. Nun werden die Distanz der Testdaten zum Sprechermodell sowie diejenige der Testdaten zum UBM berechnet. Durch Subtraktion der Distanz UBM <-> Testdaten von der

Distanz Testdaten <-> Sprechermodell wird die gewünschte Gewichtung des statistisch durchschnittlichen UBM mit den Eigenheiten des Testsprechers erreicht. Schliesslich liefert das HTK Tool den gewünschten Wert für die Ähnlichkeit zwischen Test- und Modellsprecher.

4.3.4 Auswertung

Die in 4.3.2 gebildete dreidimensionale Matrix *Result* enthält im Element *Result(i,j,1)* den Score der Verifikation des Modells von Sprecher i gegenüber einem Testsignal von Sprecher j. Um die Systeme auswerten zu können, muss die Information, ob es sich dabei um den gleichen Sprecher handelt, vorhanden sein. Diese ist in *Result(i,j,2)* gespeichert. Der Eintrag *1* steht für gleiche, *0* für ungleiche Sprecher.

Die Matrix *Result* wird in zwei Vektoren zerlegt, wobei Vektor V_{ID} die Scores all jener Tests enthält, die als identische Sprecher verifiziert werden sollten. Vektor V_{NID} enthält alle übrigen Scores, d.h. diejenigen, welche von Tests unterschiedlicher Sprecher stammen.

Da das Verifikationssystems eine eindeutige Antwort liefern soll, muss der berechnete Score anhand eines Schwellwertes einer Entscheidung zugeordnet werden können. Der ideale Wert dieser Schwelle ist systemspezifisch und hängt zusätzlich von den dem System bekannten Sprechern ab, d.h. all jenen Sprechern, für die ein trainiertes Modell existiert. Der Schwellwert ist also für die Sprecher des Testsets optimiert.

Im Idealfall könnte man einen Schwellwert finden, welcher

a) kleiner ist als alle Scores in V_{ID} und

b) grösser als alle Scores in V_{NID}.

In der Realität sind die Bedingungen a) und b) aber unrealistisch bzw. unvereinbar. Setzt man den Wert gemäss a), entstehen angewandt auf V_{NID} Fehler der Art, dass nicht identische Sprecher als identisch erkannt werden (Fehler 2. Art).

Setzt man die Schwelle hingegen gemäss b), so werden identische Sprecher nicht mehr erkannt (Fehler 1. Art).

Die Equal Error Rate, also die Fehlerrate am Punkt da beide Fehler gleich wahrscheinlich sind, bietet einen Ansatz zum Vergleich verschiedener Systeme. Die Funktion "GetEERandTHRES.m"berechnet aus den beiden Vektoren V_{ID} und V_{NID} die EER und den dazugehörigen Schwellwert für die Entscheidung.

4.4 System-Kombination

Nachstehend werden die Linearkombination der Systeme sowie die Best-of-three Kombination erklärt.

4.4.1 Linearkombination der Systeme

Um die Resultate mehrerer Systeme linear kombinieren zu können, müssen diese vorgängig in eine einheitliche Form gebracht werden. Das heisst, die resultierenden Scores müssen mittelwertfrei und mit der gleichen Varianz vorliegen. Deshalb werden der entsprechende Schwellwert der Entscheidungsgrenze von allen Scores subtrahiert und zusätzlich die Scores durch ihre Varianz geteilt.

Die Funktion „fusion.m“ führt diese Operationen durch und bildet die Linearkombination der drei Systeme, indem die Scores je mit einem zu bestimmenden Faktor gewichtet aufsummiert werden.

4.4.2 Best-of-three

Dieses System greift direkt auf die von den Einzelsystemen getroffenen Entscheidungen zurück. Anhand dieser wird immer so entschieden, wie die Mehrheit der betrachteten Systeme individuell entschieden hat. Konkret wird für jeden Einzelsystem-Score, welcher grösser ist als die Entscheidungsschwelle, eine *1* notiert (d.h. erfolgreiche Verifizierung), ansonsten eine *0* (abgelehnte Verifizierung). Die acht möglichen Kombinationen führen auf die folgende Entscheidungstabelle 4.1. Für die

Spalte „Entscheidung“ gilt wiederum, dass die Ziffer *1* für eine erfolgreiche und die Ziffer *0* für eine nicht erfolgreiche Verifizierung steht.

System I	System II	System III	Entscheidung
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Tabelle 4.1: Entscheidungstabelle der Best-of-three Kombination.

Das Best-of-three System ist diskret in dem Sinne, dass jeweils nur eine eindeutige Antwort (verifiziert / nicht verifiziert) zurückgegeben wird, also kein Score berechnet wird wie bei der Linearkombination. Somit können nur die Wahrscheinlichkeiten der Fehler 1. und 2. Art als Mass für die Performance dieser Systemkombination berechnet werden.

5 Evaluation

Dieses Kapitel ist wie folgt gegliedert: Zuerst werden in Abschnitt 5.1 alle Vorverarbeitungs- und verschiedene Systemparameter bestimmt, in Abschnitt 5.2 wird evaluiert und anschliessend in Abschnitt 5.3 getestet. In Abschnitt 5.4 werden die verschiedenen Systemkombinationen untersucht.

Die Systeme werden, wenn nicht anders erwähnt, mit Set 2 evaluiert und es werden Parameter bestimmt. Anschliessend wird mit Set 1 ein Test mit den in Set 2 gefundenen Parametern durchgeführt und die Performance festgehalten. Männliche und weibliche Sprecher werden immer separat getestet. Abschliessend folgt eine Übersicht der Resultate.

Einige Beobachtungen während der frühen Evaluationsphase liessen darauf schliessen, dass die Daten von Set 1 und 2 nicht in jeder Hinsicht vergleichbaren Charakter haben. Aus diesem Grund sind einige der Evaluationen, welche eigentlich nicht als setspezifisch gälten, im Testset durchgeführt worden.

Für die Test- und alle Evaluationsserien werden jeweils pro Sprecher 2 Testsequenzen ausgewählt und jedes Sprechermodell gegen alle Testsequenzen verifiziert. Bei den 48 weiblichen Sprechern (96 Testsequenzen) werden insgesamt 4704 Verifikationsvorgänge benötigt, bei den 33 männlichen Sprechern (66 Testsequenzen) ergeben sich insgesamt 2178 Verifikationsvorgänge. Die Länge der Sprachdateien beträgt (nach der Entfernung der Sprachpausen gemäss Abschnitt 4.2.1) durchschnittlich ca. 1.6 Minuten.

Von Interesse ist auch die Systemperformance in Abhängigkeit von der Menge der verfügbaren Daten zur Sprechermodellbildung. Hierzu wird die Menge der Mo-

dellierungsdaten variiert. Es werden jeweils 1/3, 2/3 und alle Daten für die Bildung des Sprechermodells verwendet. Es sind pro Sprecher 3, manchmal 6 Sprachsequenzen für das Training vorhanden. Diese Evaluation wird in Set 2 durchgeführt, die Zusammenhänge sind aus den folgenden Unterkapiteln ersichtlich.

5.1 Grundlegende Parameter

Nachfolgend werden grundlegende Parameter für die Vorverarbeitung der Sprachsignale besprochen und bestimmt.

5.1.1 Energieschwelle zur Detektion der Sprachpausen

Bei prosodischen Analysen gilt das Interesse der gesprochenen Sprache und nicht kurzen und leisen Sprachfragmenten wie sie in den zur Verfügung gestellten Daten (Telefongespräche) relativ oft vorkommen. Dies können auch kurze akustische Äusserungen zur Meinungsübereinstimmung mit dem Gegenüber sein. Räuspern oder andere Nebengeräusche, Pausen und sehr leise Passagen werden deshalb aus den betrachteten Sprachsignalen weggeschnitten. In Abbildung 5.1 ist eine solche Pausenentfernung dargestellt. Die Schwelle ist bei 0.02 auf der linearen Skala von 0 bis 1 bzw. -78.24 dB auf der logarithmischen Skala angesetzt. In der stichprobenartigen akustischen Auswertung der verarbeiteten Daten erweist sich diese Schwelle als angemessen.

5.1.2 Grundfrequenzkorrektur

Durch visuelle Schätzung und gestützt auf [8] liegt der Anteil der um den Faktor 2 zu hoch extrahierten Grundfrequenzwerte bei ca. 10%. Der verwendete Grundfrequenzdetektor bestimmt die Grundfrequenz zwischen 0 und 600 Hz. Diese Werte werden mit der in 4.2.2 erwähnten Funktion „Basefreqcorrector.m“ auf Frequenzdoubling überprüft. Dabei wird der Mittelwert der Grundfrequenz berechnet und alle Frequenzwerte, welche oberhalb des 1.5-fachen dieses Mittelwertes liegen, werden halbiert. In Abbildung 5.2 ist ein Beispiel für diese Korrektur dargestellt.

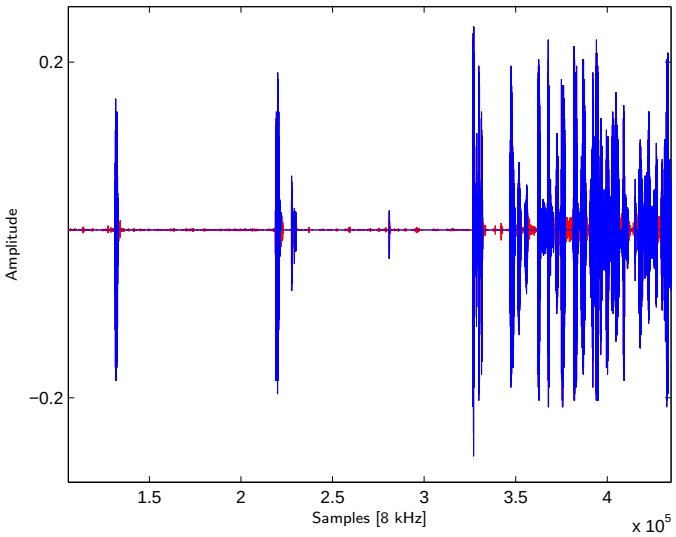


Abbildung 5.1: Entfernung der Sprachpausen: Das betrachtete Signal (blau) sowie die weggeschnittenen Komponenten (rot).

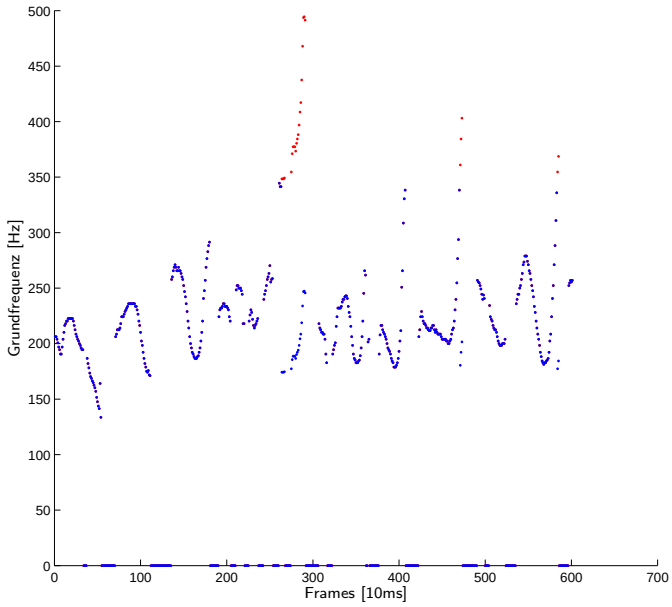


Abbildung 5.2: Korrektur des Grundfrequenzverlaufs: Der korrigierte Grundfrequenzverlauf (blau) sowie zugehörige gedoublete ursprünglichen Werte (rot).

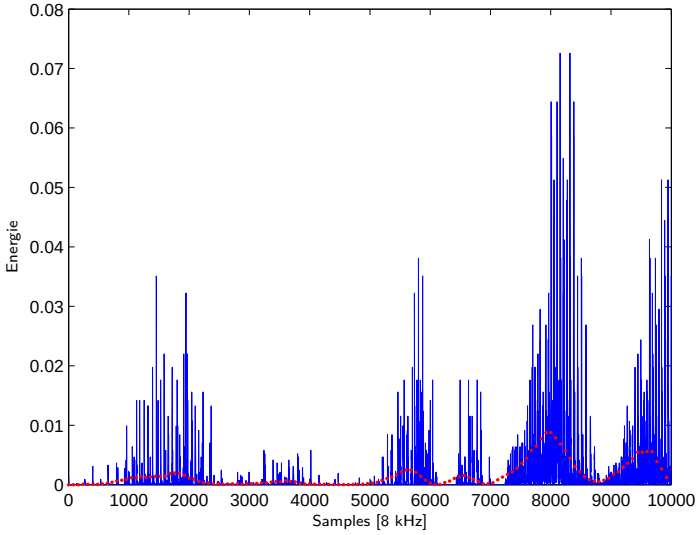


Abbildung 5.3: Filterung des Energieverlaufs: Der Energieverlauf (blau) sowie dessen Mean-Filterung (rot).

5.1.3 Filterung des Energieverlaufs

Da die Grundfrequenz jeweils über die Länge *winsize* und um *winshift* überlappend verschoben berechnet wird, muss die Anzahl der Energiepunkte ebenfalls reduziert werden. So wird der Energieverlauf wie in 5.3 gezeigt mit einem Meanfilter der Länge *winsize* gefaltet und mit *winshift* abgetastet.

5.2 Einzelsysteme - Evaluation

In diesem Abschnitt werden die optimalen Parameter der Systeme GMM, N-Gramm und ProsVar bestimmt.

5.2.1 GMM

Die Grundfrequenz bzw. ihr Logarithmus ist gemäss [2] gaussverteilt. Wir gehen davon aus, dass sich die Energie ähnlich verhält. In kleineren Evaluationstests (weniger Verfikiationsvorgänge, weniger Daten zur Modellbildung, kleineres UBM) auf Set 1 wurde die Anzahl Mischkomponenten variiert. So wurden mit 8 Mischkomponenten bei weiblichen Sprechern eine Equal Error Rate von 35.86% und mit 32

Mischkomponenten eine solche von 33.45% erreicht. Mit 1024 Mischkomponenten wurde keine weitere Steigerung erreicht, weshalb die Anzahl der Mischkomponenten auf 32 gesetzt wird. Diese Evaluationstests wurden mit einer Sprechermodellbildung erstellt, in der nur Grundfrequenz und Energie berücksichtigt wurden (ohne Ableitungen). Die genaue Konfiguration für die Mischkomponenten und die maximale Anzahl Iterationen lautet wie folgt:

- Mischkomponenten = [1 2 4 8 16 32]
- Max. Iterationen = [1 3 3 3 3 5]

Es werden vier UBM benötigt, da für beide Geschlechter die Modellbildung anhand von 1) Grundfrequenz- und Energieverlauf und 2) zusätzlich mit Berücksichtigung deren zeitlichen Ableitungen getestet wird. Die UBM werden aus einem Zwanzigstel der verfügbaren Daten erstellt. Die Modellierung dauert in der Grössenordnung von 24 Stunden, so dass der limitierende Faktor die Anzahl Daten im UBM für den Fall 2) ist. Um aber Vergleiche zwischen Fall 1) und 2) ziehen zu können ist es erforderlich, die gleiche Menge von Ausgangsdaten zu verwenden. Da im Fall 2) vier Komponenten berechnet werden, ergibt sich hier die doppelte Menge an Daten für die Erstellung des UBM. Die maximale Anzahl Iterationen konnte auf Grund der relativ schnellen Konvergenz sehr tief gehalten werden.

Die Abhängigkeit der Systemperformance zu den zur Sprechermodellierung verwendeten Daten wird mit den zu Beginn dieses Kapitels erstellten Szenarien getestet. Das Sprechermodell wird mit Daten aus Set 2 erstellt. In Tabelle 5.1 sind die Ergebnisse dargestellt.

Geschlecht	1/3 Daten	2/3 Daten	3/3 Daten
weiblich	35.3%	33.3%	33.8%
weiblich (mit Ableitungen)	32.7%	30.2%	31.7%
männlich	40.8%	38.0%	37.2%
männlich (mit Ableitungen)	38.2%	36.4%	34.0%

Tabelle 5.1: Evaluation des GMM Systems in Set 2: Für beide Systeme (mit und ohne Ableitungen) wird die Menge der Trainingsdaten variiert.

Die Zunahme der Daten für die Modellierung führt bei männlichen Sprechern zu besseren Leistungen. Auch bei weiblichen Sprechern ist dieser Zusammenhang ersichtlich. Die Einbindung der Ableitungen von Grundfrequenz- und Energieverlauf bringt eine signifikante Verbesserung der EER von 6.2% und 8.6% bei weiblichen bzw. männlichen Sprechern.

5.2.2 N-Gramm

Der N-Gramm Ansatz unterscheidet sich von den anderen Systemen darin, dass keine zeitintensiven Modelle berechnet werden müssen. Das N mit der besten N-Gramm-Performance soll bestimmt werden. Schliesslich soll der Ansatz mit den *long* Klassen untersucht werden. All diese Evaluationstests werden in Set 2 durchgeführt.

Bestimmen des optimalen Parameters N

Geschlecht	2-Gramm	3-Gramm	4-Gramm	5-Gramm
weiblich	31.8%	31.6%	29.6%	31.1%
männlich	34.1%	28.9%	28.7%	29.6%

Tabelle 5.2: Bestimmen des Parameters N für das N-Gramm System in Set 2.

Wie aus Tabelle 5.2 ersichtlich ist, ergeben die 4-Gramme für beide Geschlechter die besten Ergebnisse. Daher wird für die 4-Gramme nun die Anzahl der Trainingsdaten variiert. Die zusätzlichen Daten führen, wie Tabelle 5.3 zeigt, bei männlichen und weiblichen Sprechern zu besseren Leistungen.

Geschlecht	1/3 Daten	2/3 Daten	3/3 Daten
weiblich	32.4%	30.5%	29.6%
männlich	33.4%	29.2%	28.7%

Tabelle 5.3: Evaluation des N-Gramm Systems mit $N = 4$ in Set 2: Variieren der Trainingsdaten.

Die weniger grosse Steigerung bei den weiblichen Sprechern führen wir darauf zurück, dass sich männliche Sprecher durch ihre tendenziell eher monotonere

Sprachmelodie prosodisch weniger stark unterscheiden als weibliche Sprecher. Dadurch ist ihr Bedarf an Trainigsdaten höher.

Bestimmen des Parameters *long*

Die in [1] vorgeschlagene Unterteilung der Klassen in normale und solche mit Attribut *long* wird untersucht mit dem Ziel, eine möglichst optimale Schwelle für die Unterscheidung zu finden. Wir setzen diese Schwelle für alle Klassen und Geschlechter einzeln. Dazu betrachten wir gemäss [1] drei Ansätze, welche die jeweilige Schwellenlänge bestimmen. Ausgewertet werden diese Ansätze auf das UBM Set.

1. Median (M): Alle Segmente einer Klasse i, welche länger sind als der Median der Länge aller Segmente dieser Klasse, erhalten das Attribut *long*.
2. Median 66 (M66): Alle Segmente einer Klasse i, welche länger sind als die längsten 2/3 aller Segmente dieser Klasse, erhalten das Attribut *long*.
3. Kumulation 66 (K66): Alle Segmente einer Klasse i, welche länger sind als die kürzesten bis zu 2/3 der durchschnittlichen Länge aufsummierten Segmentlängen, erhalten das Attribut *long*.

Die Ergebnisse der Berechnung dieser Schwellwerte, also die Mindestlängen der Klassen für die Kennzeichnung mit dem Attribut *long*, sind in Tabelle 5.4 zusammengestellt.

Klasse	M (w)	M (m)	M66 (w)	M66 (m)	K66 (w)	K66 (m)
0	5	4	7	7	17	24
2	2	2	2	2	5	4
4	2	2	2	2	3	4
6	2	2	2	2	5	4
8	2	2	2	2	4	4

Tabelle 5.4: Bestimmen des Parameters *long* für das N-Gramm System mit $N = 4$.

Die Bestimmung der Schwellwerte wird in Set 2 vorgenommen. Tabelle 5.5 zeigt die Ergebnisse dieser drei Ansätze für weibliche Sprecher.

Long-Typ	2-Gramm	3-Gramm	4-Gramm	5-Gramm
Median	30.1%	30.7%	34.2%	35.0%
Median 66	30.9%	27.5%	31.8%	37.0%
Kumuliert 66	30.4%	31.8%	34.5%	35.7%

Tabelle 5.5: Bestimmung des Parameters *long* für das N-Gramm System mit $N = 2, 3, 4, 5$ für weibliche Sprecher.

Auffallend ist, dass einige Ergebnisse unter Verwendung des Attributs *long* schlechter ausfallen als die entsprechenden Tests ohne diese zusätzliche Kategoriesierung (vergleiche Tabelle 5.2). Da die Median 66 Methode mit 3-Grammen die besten Resultate unter Verwendung von *long* ergibt, wird sie für die weiteren Tests verwendet.

Schliesslich wird der Effekt der Variierung der Trainingsdaten mit der Median 66 Methode in Set 2 evaluiert. Diese in Tabelle 5.6 zusammengestellten Ergebnisse zeigen eine Verbesserung gegenüber den normalen 4-Grammen (weiblich: 29.6% männlich: 28.7%) um 7.0% bzw. 3.8%. Es zeigt sich, dass die Verfügbarkeit von Trainingsdaten einen grossen Einfluss auf die Performance hat. Damit sind alle Systemparamter für den N-Gramm Ansatz bestimmt.

Geschlecht	1/3 Daten	2/3 Daten	3/3 Daten
weiblich	33.8%	30.8%	27.5%
männlich	30.0%	29.5%	27.6%

Tabelle 5.6: Evaluation des N-Gramm Systems mit $N = 3$ und Median 66 (*long*) in Set 2: Variieren der Trainingsdaten.

5.2.3 ProsVar

Die Komponenten des ProsVar Systems sind gemäss [2] je gauss- oder exponentialverteilt. Daher werden tendenziell wenig Mischkomponenten benötigt. Da ein UBM aus möglichst vielen Daten, aber innerhalb einer vernünftigen Zeit erstellt werden sollte, wurde die Anzahl von Mischkomponenten auf 16 gesetzt. Die maximale Anzahl Iterationen konnte auch hier auf Grund der relativ schnellen Konvergenz tief gehalten werden.

- Mischkomponenten = [1 2 4 8 16]
- Max. Iterationen = [1 2 2 2 5]

Das Universal Background Modell für das ProsVar System wird im HTK Toolkit mit allen verfügbaren Daten erstellt. Dieser Evaluationstest wurde in Set 2 durchgeführt. In Tabelle 5.7 sind die Ergebnisse zusammengestellt.

Geschlecht	1/3 Daten	2/3 Daten	3/3 Daten
weiblich	30.4%	29.8%	30.3%
männlich	34.2%	33.4%	32.1%

Tabelle 5.7: Evaluation des ProsVar Systems in Set 2: Variieren der Trainingsdaten.

Es zeigt sich, dass zusätzliche Daten für die Modellierung bei männlichen Sprechern zu einer besseren Leistung führen. Bei weiblichen Sprechern ist dieser Zusammenhang hingegen kaum ersichtlich.

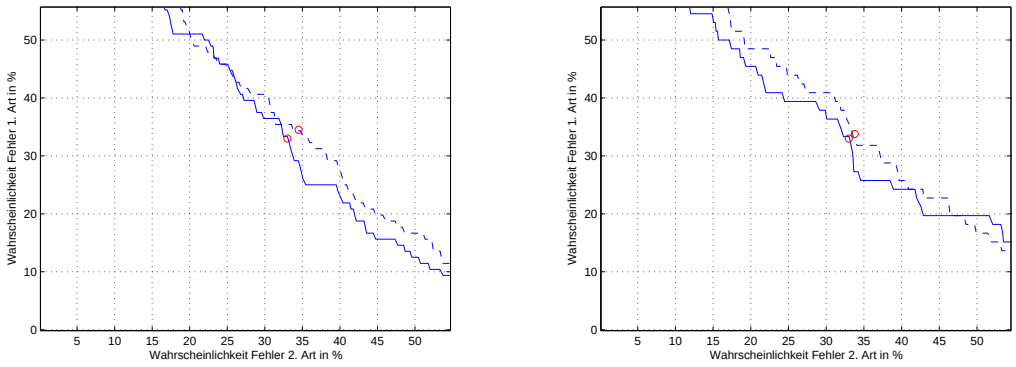
5.3 Einzelsysteme - Resultate

In diesem Abschnitt werden die Systeme mit den zuvor bestimmten Parametern getestet. Zusätzlich wird das cepstrale System als Referenz beigezogen.

5.3.1 GMM

Die vier UBM für das GMM System werden gemäss den in 5.2.1 festgelegten Mischkomponenten und maximalen Iterationen berechnet. In Abbildung 5.4(a) und 5.4(b) sind die Ergebnisse des GMM Tests in Set 1 dargestellt.

Auf der Ordinate ist die Wahrscheinlichkeit des Fehlers 1. Art und auf der Abszisse die Wahrscheinlichkeit des Fehlers 2. Art aufgetragen. Die durchgezogene Kurve ist diejenige des Systems mit 4 Komponenten (f_0, E, $\frac{\partial}{\partial t}f_0$, $\frac{\partial}{\partial t}E$) und die gestrichelte Linie diejenige des 2 Komponenten Systems (f_0, E). Die EER liegt für weibliche Sprecher bei 33.0% (für 2 Komponenten bei 34.5%) und für männliche



(a) (b)

Abbildung 5.4: Die Beziehung zwischen den Fehlerwahrscheinlichkeiten 1. und 2. Art beim GMM System für (a) weibliche und (b) männliche Sprecher. Durchgezogene Linie: 4 Komponenten, gestrichelte Linie: 2 Komponenten.

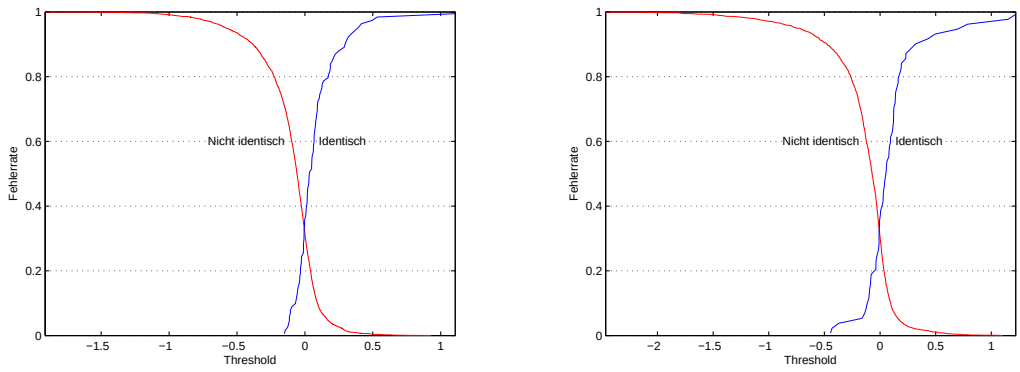
Sprecher bei 32.8% (für 2 Komponenten bei 33.7%). Die Systemverbesserung beträgt also bei weiblichen Sprechern 4.3% und bei männlichen Sprechern 2.7%.

Die Verteilungsfunktionen der Systeme mit 4 Komponenten sind in Abbildung 5.5(a) und 5.5(b) dargestellt.

5.3.2 N-Gramm

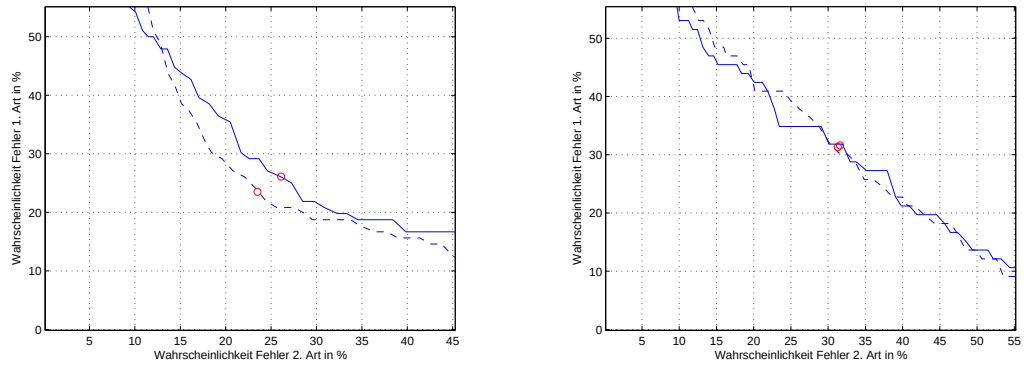
Die in Abschnitt 5.2.2 festgelegten Systemparamter werden auf Set 1 angewendet. In Abbildung 5.6(a) und 5.6(b) sind diese Ergebnisse dargestellt.

Die EER liegt für weibliche Sprecher bei 26.1% (ohne *long*: 23.6%) und für männliche Sprecher bei 31.6% (ohne *long*: 31.3%). Diese Ergebnisse sind insofern überraschend, als die N-Gramme mit *long* Segmenten nicht besser abschneiden als diejenigen ohne. Ein Erklärungsansatz bildet die starke Abweichung dieser Ergebnisse von Set 1 bereits beim normalen N-Gramm-Ansatz (weibliche Sprecher deutlich besser). Die Verteilungsfunktionen sind in Abbildung 5.7(a) und 5.7(b) dargestellt.



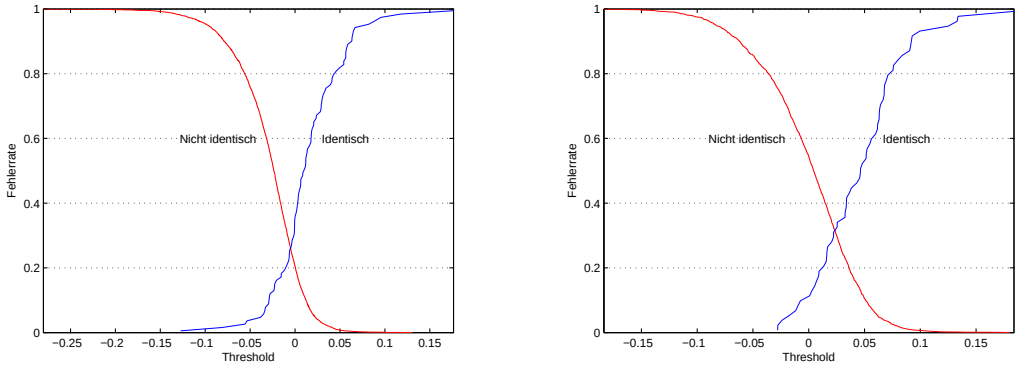
(a) (b)

Abbildung 5.5: Die Verteilungsfunktionen der Scores für die identischen bzw. nicht identischen Sprecher beim GMM System (4 Komponenten). (a) weibliche und (b) männliche Sprecher. Im Schnittpunkt beider Kurven ist die EER abzulesen.



(a) (b)

Abbildung 5.6: Die Beziehung zwischen den Fehlerwahrscheinlichkeiten 1. und 2. Art beim N-Gramm System für (a) weibliche und (b) männliche Sprecher. Durchgezogene Linie: *long*, gestrichelte Linie: normal.



(a) (b)

Abbildung 5.7: Die Verteilungsfunktionen der Scores für die identischen bzw. nicht identischen Sprecher beim N-Gramm System (*long*). (a) weibliche und (b) männliche Sprecher. Im Schnittpunkt beider Kurven ist die EER abzulesen.

5.3.3 ProsVar

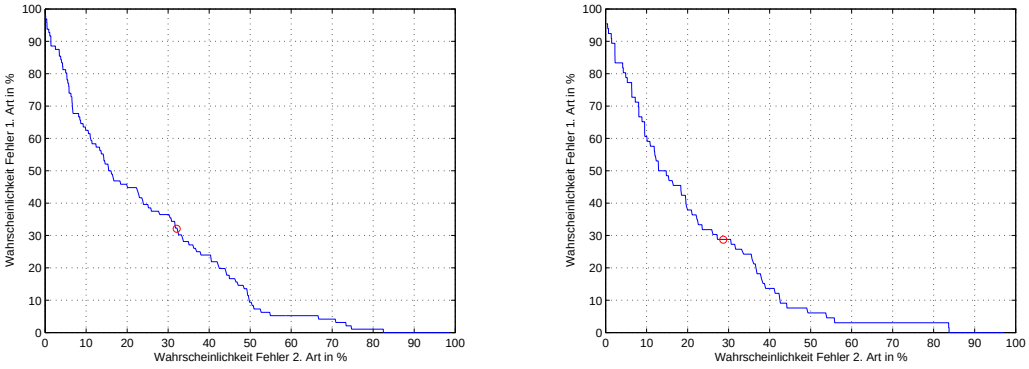
Für das ProsVar System werden mit den in Abschnitt 5.2.3 bestimmten Mischkomponenten und maximalen Iterationen die UBM berechnet und es wird in Set 1 getestet. In Abbildung 5.8(a) und 5.8(b) sind die Ergebnisse dargestellt. Die EER liegt für weibliche Sprecher bei 32.1% und für männliche Sprecher bei 28.7%. Die Verteilungsfunktionen sind in Abbildung 5.9(a) und 5.9(b) dargestellt.

5.3.4 Cepstrales Vergleichssystem

Die cepstralen Koeffizienten werden gemäss Abschnitt 4.2.6 extrahiert. Mit der gegebenen Anzahl Mischkomponenten und Iterationen wird das UBM aus einem Fünfzigstel der verfügbaren Daten gebildet. Um eine Systemkombinationsevaluation mit den prosodischen Systemen zu erlauben, wird der Test in Set 2 durchgeführt. Die HTK-Konfiguration wurde vom Institut folgendermassen vorgegeben:

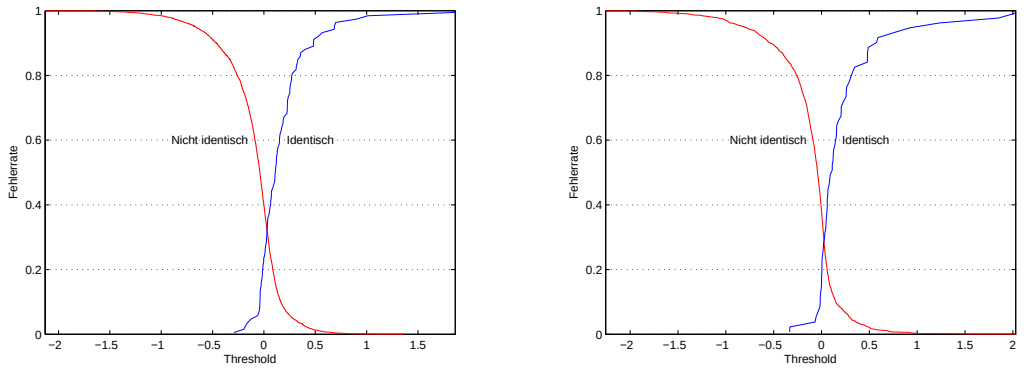
- Mischkomponenten = [1 2 4 8 16 32 64 128 256 512 1024]
- Max. Iterationen = [1 1 1 1 1 1 2 2 2 3 5]

Für männliche Sprecher erreicht das System eine EER von 19.7%, für weibliche 21.4%.



(a) (b)

Abbildung 5.8: Die Beziehung zwischen den Fehlerwahrscheinlichkeiten 1. und 2. Art beim ProsVar System für (a) weibliche und (b) männliche Sprecher.



(a) (b)

Abbildung 5.9: Die Verteilungsfunktionen der Scores für die identischen bzw. nicht identischen Sprecher beim ProsVar System. (a) weibliche und (b) männliche Sprecher. Im Schnittpunkt beider Kurven ist die EER abzulesen.

5.4 Systemkombinationen

Es werden eine Linearkombination und das in Kapitel 4 beschriebene Best-of-three System untersucht und getestet.

5.4.1 Linearkombination der prosodischen Systeme

Die drei Systeme GMM, N-Gramm und ProsVar stehen zur Verfügung. Für männliche und weibliche Sprecher sind die besten Systeme GMM (4 Komponenten), 3-Gramm (mit Median 66) und ProsVar.

Die Auswertung der möglichen Linearkombinationen in Set 2 liefert folgende Gewichtungsfaktoren:

System	GMM	N-Gramm	ProsVar	EER
weiblich	0.8	1	3	25.1%
männlich	0	1	1.28	26.9%

Tabelle 5.8: Bestimmung der Gewichtungsfaktoren (N-Gramm = 1) für die Linearkombinationen in Set 2.

Mit diesen bestimmten Linearkoeffizienten bildet man nun die Linearkombination für Set 1. Interessant ist der Fall der männlichen Sprecher: Offenbar bringt das GMM Sysstem nur redundante Daten und wird deshalb für die Linearkombination nicht benötigt.

Unerwarteterweise ergibt der *long* Ansatz für N-Gramme in Set 1 keine besseren Ergebnisse als das normal gebildete N-Gramm (siehe Bemerkung zu Beginn dieses Kapitels). Aus Konsistenzgründen führen wir die Linearkombination mit den schlechteren Ergebnissen, also dem *long* Ansatz durch.

Die Testergebnisse und der Vergleich mit den Einzelsystemen für Set 1 sind in Tabelle 5.9 dargestellt. Die berechnete prozentuale Verbesserung bezieht sich jeweils auf das beste Einzelsystem.

Geschlecht	Linearkombination	GMM	N-Gramm	ProsVar	Verbesserung
weiblich	24.5%	33.0%	26.1%	32.1%	6.1%
männlich	30.0%	32.8%	31.6%	28.7%	keine

Tabelle 5.9: Testergebnisse der Linearkombination und Vergleich mit den Einzelsystemen für Set 1.

Bei den weiblichen Sprechern erfolgt eine Verbesserung um 6.1%. Bei männlichen Sprechern ist keine Verbesserung zu beobachten. Prinzipiell könnte man die Linearkombination an das Set bzw. die Sprecher anpassen, um so eine bessere Leistung zu erzielen (siehe Kapitel 6).

5.4.2 Best-of-three

Dieser Ansatz benötigt keine Anpassung systemspezifischer Parameter, sondern wird direkt auf Set 1 angewendet. Er erbringt folgende Ergebnisse:

Geschlecht	Best-of-three	Best-of-three	GMM	N-Gramm	ProsVar
weiblich	1. Art: 27.1%	2. Art: 29.7%	33.0%	26.1%	32.1%
männlich	1. Art: 28.1%	2. Art: 27.6%	32.8%	31.6%	32.1%

Tabelle 5.10: Testergebnisse der Best-of-three Kombination und Vergleich mit den Einzelsystemen für Set 1.

Da dieses System diskret ist („ja“ oder „nein“ Entscheidung, kein Score wird ermittelt), kann keine EER berechnet werden, sondern nur die Wahrscheinlichkeit der Fehler 1. und 2. Art. Bei männlichen Sprechern liegen beide Wahrscheinlichkeiten deutlich unter den EER aller Einzelsysteme, die Strategie ist also besser. Bei weiblichen Sprechern ist dies nicht der Fall.

5.4.3 Linearkombination der prosodischen Systeme mit dem cepstralen System

Die Linearkombination der drei prosodischen Einzelsysteme kann mit dem cepstralen System kombiniert werden. Da aber in unserem Fall das cepstrale System nicht gezielt optimiert wurde (die Literatur lässt weit tiefere EERs vermuten),

wird dieser Kombinationsansatz nicht abschliessend getestet. Um das Potential einer solchen Kombination abschätzen zu können, wurden aber analog zu Abschnitt 5.4.1 die Linearkoeffizienten bestimmt, welche die beste Performance liefern. Das cepstrale System wird jeweils mit 1 gewichtet, bei männlichen Sprechern wird die beste Linearkombination mit 640 und bei weiblichen Sprechern mit 1160 gewichtet. Die Performance dieses Tests ist in Tabelle 5.11 aufgeführt und erscheint vielversprechend.

Geschlecht	LK CEP - Pros	CEP	LK pros	Verbesserung
weiblich	19.1%	21.4%	25.1%	10.7%
männlich	17.0%	19.7%	26.9%	13.7%

Tabelle 5.11: Testergebnisse der Linearkombination und Vergleich mit den Einzelsystemen für Set 2.

6 Fazit

Da die verfügbaren Daten für die Erstellung der UBM nicht für beide Geschlechter gleich ist, trotzdem aber möglichst viele Sprecher im UBM enthalten sein sollen, wurden die UBM für beide Geschlechter mit unterschiedlich vielen Daten erstellt. Dadurch lassen sich keine direkten Vergleiche zwischen Systemen für männliche und weibliche Sprecher machen. Die für die weiblichen Sprecher gebildeten UBM werden mit ca. 30% mehr Daten als diejenigen für die männlichen gebildet. Die Ergebnisse bzw. die EER der jeweiligen Systeme sind aber trotzdem nicht signifikant besser für die weiblichen Sprecher. Zusätzlich ergeben das Evaluations- (Set 2) und das Testset (Set 1) in dieser Hinsicht jeweils inverse Ergebnisse. Daraus schliessen wir, dass Unterschiede des UBM in dieser Grössenordnung auf die Systemperformance kaum Einfluss haben.

System	Kommentar	männlich	weiblich	Referenzwert
GMM	2 Komponenten	33.7%	34.5%	-
	4 Komponenten	32.8%	33.0%	16.3%
N-Gramm	4-Gramm	31.3%	23.6%	19.2%
	3-Gramm long	31.6%	26.1%	14.1%
ProsVar		28.7%	32.1%	-
Linearkombination		30.0%	24.5%	-
Best-of-three	Fehler 1. Art	28.1%	27.1%	-
	Fehler 2. Art	27.6%	29.7%	-
CEP	für Set 2	19.7%	21.4%	-
Linearkombination CEP	für Set 2	17.0%	19.1%	-

Tabelle 6.1: Übersicht über die besten erreichten Performancewerte aller Systeme für Set 1. Die Referenzwerte beziehen sich auf [1].

Vergleich GMM <–> N-Gramm Das N-Gramm System ergibt eine bessere EER als das GMM System. Allerdings wird das UBM für das GMM System aus

Zeitgründen mit nur einem Zwanzigstel der verfügbaren Daten erstellt, das N-Gramm System benützt hingegen alle Daten für die UBM Bildung. Die in [1] erreichten EER für beide Systeme sind deutlich besser. Das GMM wie auch das N-Gramm System wird dort aber mit mehr Trainingsdaten und grösserem UBM getestet.

Mit Blick auf die in Set 2 durchgeführte Evaluation der beiden Systeme ist der Nutzen einer Erhöhung der Anzahl Trainingsdaten ersichtlich (siehe Evaluationsprotokoll in Anhang A). Insbesondere beim N-Gramm System mit Parameter *long* ist der Einfluss von zusätzlichen Trainingsdaten klar ersichtlich.

Das N-Gramm System, welches im Testfall für normale N-Gramme bessere Ergebnisse liefert als die mit *long* Klassen gebildeten N-Gramme, kann weiter verbessert werden. So können die Parameter zur Bestimmung der Zuteilung in die *long* Klassen setspezifisch gebildet werden. Das System wird somit auf einen bestimmten Sprecherkreis angepasst. Da wir nicht über genügend Testdaten für eine solche Strategie verfügen, konnte sie nicht weiter evaluiert werden.

ProsVar System Da für das ProsVar System in [2] keine Performance angegeben ist, es wird nur eine Kombination mit einem cepstralen System getestet, kann hier kein Referenzwert angegeben werden. Die Performance unserer ProsVar Implementation liegt zwischen derjenigen des GMM und N-Gramm Systems.

Systemkombinationen Unsere beiden Ansätze (Linearkombination und Best-of-three) führen teilweise auf eine Verbesserung, nicht jedoch in derart erheblichem Ausmass wie sie in anderen Publikationen ([1], [3]) erreicht wird. Dort wird aber keine Aussage über die Art der eingesetzten Systemkombination gemacht. Es ist ein Ansatz mit neuralen Netzen zu vermuten.

In dieser Arbeit wird die Linearkombination aus dem ProsVar und dem N-Gramm System gebildet, das GMM System wird für männliche Sprecher gar nicht und für weibliche schwächer gewichtet hinzugezogen. Dass sich das ProsVar und

das N-Gramm System gut ergänzen ist insofern nicht überraschend, als das ProsVar System die im N-Gramm System nicht berücksichtigten quantitativen Komponenten (siehe Abschnitt 3.3.3) erfasst. Es ist somit eine gute Ergänzung der beiden Systeme N-Gramm und ProsVar zu beobachten.

Der Best-of-three Ansatz erscheint vielversprechend, insbesondere im Hinblick auf die Einbindung zusätzlicher Systeme mit vergleichbaren EER.

CEP System - Vergleich Das CEP System erreicht eine wesentlich bessere Leistung als die beste Systemkombination der übrigen implementierten prosodischen Systeme. Prosodische Systeme können daher nicht als Ersatz oder Alternative zu cepstralen Systemen betrachtet werden, jedoch können sie dank der Berücksichtigung komplementärer Aspekte des Sprachsignals in der Kombination mit cepstralen Systemen eine Steigerung der Gesamtperformance ermöglichen. Dies wurde in Abschnitt 5.4.3 ansatzweise gezeigt.

7 Ausblick

Dieses Kapitel zeigt einige Möglichkeiten für weiterführende Untersuchungen basierend auf den in dieser Arbeit implementierten Systemen auf.

Mehr Trainingsdaten Entsprechende Tests haben gezeigt, dass die Systemperformance aller untersuchten Systeme stark von der Menge der zur Verfügung gestellten Trainingsdaten abhängt. Es wäre daher interessant, die Sprechermodelle mit entsprechend grösseren Datenmengen zu erstellen. Insbesondere wäre es von Interesse zu untersuchen, welcher Art der Zusammenhang zwischen Anzahl Trainingsdaten und erreichter Performance ist. Ebenfalls könnte die benötigte Rechenzeit in Abhängigkeit mit den berücksichtigten Daten erfasst werden. Da eine gewisse Abflachung der Performance-Kurve mit zunehmenden Modelldaten zu erwarten ist, könnte mit dem entsprechenden Wissen ein geeigneter Kompromiss zwischen Fehlerwahrscheinlichkeit und Rechenzeit gefunden werden.

UBM mit mehr / angepassten Daten In eine ähnliche Richtung zielt die Erhöhung der Datenmenge für die UBM Erstellung. Der Zusatznutzen von generell grösseren UBM-Datensets dürfte sich allerdings Grenzen halten. Hingegen könnte mit einer Abstimmung der UBM-Daten auf die konkret zu verifizierenden Sprechergruppen eine Steigerung der Performance einhergehen.

Datenvorverarbeitung Da es sich bei den untersuchten Sprachsignalen um Telefongespräche handelt, variiert die Sprachqualität stark. Mit geeigneten Vorverarbeitungsmethoden könnten beispielsweise Störgeräusche beseitigt und Verzerrungen der Übertragungsleitung weggefiltert werden.

Datenextraktion Eine gute Extraktion der Grundfrequenz ist die Grundlage für die hier betrachteten Systeme. Daher könnte ein exakterer Algorithmus zur Vermeidung von Doubling entwickelt werden. Im Weiteren könnten auch bestimmte Charakteristika des Grundfrequenzverlaufs (beispielsweise An- und Abklingen) erkannt werden und in die Verifikation mit einfliessen.

Systemkombinationen Das Gebiet der Systemkombinationen bietet viele Optimierungsmöglichkeiten. So könnten verschiedene Kombinationsarten basierend auf Mustererkennung, beispielsweise Neuronale Netze, zur Anwendung kommen.

Literaturverzeichnis

[1] D. Reynolds J. Godfrey A. Adami, R. Mihaescu. Modeling prosodic dynamics for speaker recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4:788–791, 2003.

[2] L. Heck M. Weintraub K. Soenmez, E. Shriberg. Modeling dynamic prosodic variation for speaker verification. *Proceedings of the International Conference on Spoken Language Processing*, 7:3189 – 3192, 1998.

[3] H. Hermansky A. Adami. Segmentation of speech for speaker and language recognition. *Eurospeech, Geneva*, pages 841–844, 2003.

[4] G. Doddington. Speaker recognition based on idiolectal differences between speakers. *Proceedings of Eurospeech*, pages 2521–2524, 2001.

[5] J. Odell D. Ollason V. Valtchev P. Woodland S. Young, D. Kershaw. *The HTK Book*. 1999.

[6] C. Fredouille G. Gravier I. Magrin-Chagnolleau S. Meignier T. Merlin J. Ortega-Garcia D. Petrovska-Delacrétaz D. Reynolds F. Bimbot, J. Bonastre. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430 – 451, 2004.

[7] R. Beutler B. Pfister. *Skript zur Vorlesung Sprachverarbeitung I.* Institut für technische Informatik und Kommunikationsnetze, ETH Zürich, 2007.

[8] M. Weintraub E. Shriberg K. Soenmez, L. Heck. A lognormal tied mixture model of pitch for prosody-based speaker recognition. *Proceedings of Eurospeech*, 3:1391 – 1394, 1997.

Anhang

A Evaluationsprotokolle

Evaluation (Set 2)

System I (GMM)	männliche Sprecher			weibliche Sprecher		
	1	2	3	1	2	3
2 Komponenten	40.8	38.0	37.2	35.3	33.3	33.8
4 Komponenten	38.2	36.4	34.0	32.7	30.2	31.7
System II (N-Gramm) Normal						
2-Gramm	35.1	34.3	34.1	35.1	32.2	31.8
3-Gramm	34.1	31.5	28.9	34.7	31.3	31.6
4-Gramm	33.4	29.2	28.7	32.4	30.5	29.6
5-Gramm	34.3	32.2	29.6	34.4	31.6	31.1
"long median"						
2-Gramm	30.7	28.7	31.2	36.8	33.5	30.1
3-Gramm	30.2	27.2	27.2	34.3	30.0	30.7
4-Gramm	36.6	29.9	31.1	34.9	34.0	34.2
5-Gramm	35.7	32.8	32.4	36.9	34.4	35.0
"long 66-median"						
2-Gramm	33.1	31.8	30.7	35.1	31.5	30.9
3-Gramm	30.0	29.5	27.6	33.8	30.8	27.5
4-Gramm	32.1	28.6	28.0	34.3	32.6	31.8
5-Gramm	34.4	29.7	29.9	38.4	35.9	37.0
"long 66-kumulativ"						
2-Gramm	34.8	34.4	34.7	32.3	31.6	30.4
3-Gramm	36.2	34.3	32.1	32.5	31.8	31.8
4-Gramm	36.3	35.1	32.7	34.7	34.3	34.5
5-Gramm	39.3	38.2	37.0	38.8	40.0	35.7
System III (ProsVar)						
Resultat	34.2	33.4	32.1	30.4	29.8	30.3
System IV (CEP)						
Resultat			19.7			21.4
System Fusion						
LK alle 3			26.9			25.1
LK mit CEP			17.0			19.1
Best-Of-Three	W'keit Fehler 1. Art		31.1	W'keit Fehler 1. Art		27.1
	W'keit Fehler 2. Art		29.1	W'keit Fehler 2. Art		28.7

Test (Set 1)

System I (GMM)	männliche Sprecher			weibliche Sprecher		
	1	2	3	1	2	3
2 Komponenten			33.7			34.5
4 Komponenten			32.8			33.0
System II (N-Gramm) Normal						
2-Gramm						
3-Gramm						
4-Gramm			31.3			23.6
5-Gramm						
"long median"						
2-Gramm						
3-Gramm						
4-Gramm						
5-Gramm						
"long 66-median"						
2-Gramm						
3-Gramm			31.6			26.1
4-Gramm						
5-Gramm						
"long 66-kumulativ"						
2-Gramm						
3-Gramm						
4-Gramm						
5-Gramm						
System III (ProsVar)						
Resultat			28.7			32.1
System IV (CEP)						
Resultat						
System Fusion						
LK alle 3			30.0			24.5
LK mit CEP						
Best-Of-Three	W'keit Fehler 1. Art		28.1	W'keit Fehler 1. Art		27.1
	W'keit Fehler 2. Art		27.6	W'keit Fehler 2. Art		29.7

B Aufgabenstellung

Institut für
Technische Informatik und
Kommunikationsnetze

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo

Herbstsemester 2007
(SA-2007-59)

Semesterarbeitsaufgabenstellung

für

Herrn Markus Schafroth und Herrn Michael Steiger

Betreuer: M. Gerber ETZ D97.4
Stellvertreter: Dr. B. Pfister ETZ D97.6

Ausgabe: 24. September 2007
Abgabe: 21. Dezember 2007

Sprechererkennung anhand des Prosodieverlaufs

Einleitung

Ein Sprachsignal enthält nicht nur Informationen über den gesprochenen Text, sondern auch über die Stimme des Sprechers. Dies nutzt die Sprecher-Verifikation aus, die beispielsweise für automatische Zulassungssysteme gebraucht wird, in denen die Benutzer anhand der Stimme identifiziert werden.

Grob wird zwischen der Text abhängigen und der Text unabhängigen Sprecher-Verifikation unterschieden. In der Text abhängigen Sprecher-Verifikation wird davon ausgegangen, dass in den zwei Sprachsignalen, die verglichen werden müssen, dasselbe gesprochen wurde. In diesem Fall wird oft Pattern Matching verwendet. In der Text unabhängigen Sprecher-Verifikation wird davon ausgegangen, dass in den zwei zu vergleichenden Sprachsignalen voneinander unterschiedlicher Text gesprochen wurde. Oft wird in diesem Fall mit statistischen Methoden (z.B. Gauss'schen Mischmodellen, GMMs) gearbeitet. Eine Zusammenfassung insbesondere der Text unabhängigen Sprecher-Verifikation findet sich z.B. in [1].

Meistens werden für die Sprecher-Verifikation aus dem Sprachsignal dieselben Merkma-

le wie für die Spracherkennung extrahiert. Diese Merkmale repräsentieren den groben spektralen Verlauf. Sie vernachlässigen die prosodischen Eigenschaften der Sprache wie z.B. Grundfrequenz, Sprechgeschwindigkeit oder Intensität mehrheitlich. Untersuchungen haben jedoch gezeigt, dass sich Sprecher auch anhand dieser prosodischen Merkmale unterscheiden lassen.

In der Literatur sind einige Ansätze beschrieben, wie Sprecherverifikation basierend auf der Prosodie gemacht werden können. Im Folgenden werden einige Ansätze erwähnt:

1. GMMs werden für prosodische Merkmale statt für die sonst üblichen spektralen Merkmale trainiert. Ein solcher Ansatz wird in [2] als Vergleichssystem verwendet.
2. Einige Ansätze (z.B [3] oder [4]) beruhen auf der Segmentierung der Sprachsignale in Silben. Diese Aufteilung in Silben wird in [3] aufgrund von rein prosodischen Merkmalen gemacht. Andere Ansätze verwenden dazu zusätzlich Transkriptionen des Sprachsignals (z.B. [4]). Die prosodischen Merkmale der gefundenen Silben werden mit Hilfe von N-Gram Modellen (siehe [5]) modelliert.

Aufgabenstellung

In dieser Arbeit geht es darum, ein Sprecherverifikations-System zu implementieren, welches prosodische Merkmale zur Unterscheidung von Sprechern verwendet. Es wird empfohlen wie folgt vorzugehen:

- Das für die Arbeit nötige Wissen soll erarbeitet werden. Dazu sind insbesondere die Vorlesungsskripte ([6] und [5]) zu empfehlen.
- In der Literatur soll nach Arbeiten gesucht werden, welche Sprecherverifikation aufgrund von prosodischen Merkmalen machen. Als Ausgangspunkt werden die Publikationen [3] und [4] empfohlen.
- Es werden vom Institut einige Werkzeuge zur Untersuchung von Sprachsignalen zur Verfügung gestellt. Insbesondere ist ein Algorithmus zur Detektion der Grundfrequenz F_0 vorhanden. Um einen Einblick in die prosodischen Merkmalen zu erhalten, soll mit diesen Werkzeugen experimentiert werden.
- Zwei Methoden Spracherkennung mit Prosodie sollen implementiert werden:
  - Statistische Modellierung der prosodischen Merkmale mit GMMs.
  - Bilden von N-Gram Modellen wie in [3] beschrieben.
- Die implementierten Methoden sollen getestet und sowohl untereinander als auch mit einem auf cepstralen Merkmalen basierenden System verglichen werden. Vom Institut werden Testdaten zur Verfügung gestellt.

Die ausgeführten Arbeiten und die erhaltenen Resultate sind in einem Bericht zu dokumentieren (siehe dazu [7]), der in gedruckter und in elektronischer Form abzugeben ist. Zusätzlich sind im Rahmen eines Kolloquiums zwei Präsentationen vorgesehen: etwa

drei Wochen nach Beginn soll der Arbeitsplan und am Ende der Arbeit die Resultate vorgestellt werden. Die Termine werden später bekannt gegeben.

Literaturverzeichnis

[1] F. Bimbot and J.-F. Bonastre et al. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, pages 430–451, April 2004.

[2] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey. Modeling prosodic dynamics for speaker recognition. In *ICASSP 2003*, 2003.

[3] A. G. Adami and H. Hermansky. Segmentation of speech for speaker and language recognition. In *Proc. of the Eurospeech 2003*, pages 841–844, 2003.

[4] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3–4):455–472, July 2005.

[5] B. Pfister, R. Beutler und C. Traber. *Sprachverarbeitung II.* Vorlesungsskript für das Sommersemester 2006, Departement ITET, ETH Zürich, 2006.

[6] B. Pfister und R. Beutler. *Sprachverarbeitung I.* Vorlesungsskript für das Wintersemester 2005/2006, Departement ITET, ETH Zürich, 2005.

[7] B. Pfister. *Richtlinien für das Verfassen des Berichtes zu einer Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.
(http://www.tik.ee.ethz.ch/~spr/SADA/richtlinien_bericht.pdf).

[8] B. Pfister. *Hinweise für die Präsentation der Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.
(http://www.tik.ee.ethz.ch/~spr/SADA/hinweise_praesentation.pdf).

Zürich, den 13. September 2007 Prof. Dr. L. Thiele

C Software CD