Universität Karlsruhe
Institut für Logik, Komplexität
und Deduktionssysteme
Am Fasanengarten 5
D–76128 Karlsruhe

Klassifizierung und Erkennung von Sprachsegmenten

Diplomarbeit von

Jan-Constantin Buckow

Betreuer:

Prof. Dr. Alex Waibel
Dipl.-Ing. Martin Westphal

angefertigt am

Computer Science Department
Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.

buckow@cs.cmu.edu

buckow@ira.uka.de

März-Oktober 1996

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig erstellt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 31. Okt. 96
Jan Buckow

Inhaltsverzeichnis

Zusammenfassung

Die zunehmende Leistungsfähigkeit von Spracherkennungssytemen hat dazu geführt, daß es heutzutage möglich ist, Anwendungen für diese Systeme zu entwickeln, die vor wenigen Jahren noch undenkbar gewesen wären.

Ein Problem, das erst seit kurzem untersucht wird, ist Spracherkennung auf großen inhomogenen Audioaufnahmen. Ein Beispiel hierfür sind Radiosendungen. Im Rahmen dieser Diplomarbeit werden Möglichkeiten untersucht, wie mit den verschiedenen Schwierigkeiten, die in diesem Zusammenhang auftreten, umgegangen werden kann.

Es wird ein Ansatz vorgestellt, mit dem Audioaufnahmen in Segmente unterteilt werden können. Die Segmentierung wird in Abhängigkeit von den jeweils vorliegenden akustischen Bedingungen vorgenommen. In diesem Zusammenhang wird untersucht, welche akustischen Bedingungen zuverlässig automatisch unterschieden werden können. Neben einer intuitiv getroffenen Entscheidung, was unterschieden werden soll, wird eine Vorgehensweise beschrieben, wie rein aufgrund der statistischen Eigenschaften des Audiosignals eine Einteilung vorgenommen werden kann.

Herkömmliche Spracherkennungssyteme sind sehr empfindlich gegenüber veränderten akustischen Bedingungen. Die Inhomogenität von Radiosendungen stellt daher ein großes Problem in Bezug auf Spracherkennung dar. Diesem Problem wird begegnet, indem mehrere Spracherkenner verwendet werden, die für die verschiedenen akustischen Bedingungen in Radiosendungen angepaßt wurden. Für die einzelnen Segmente wird durch Klassifikation bestimmt, welcher der spezialisierten Spracherkenner am besten geeignet ist. Für welche akustischen Bedingungen Spracherkenner angepaßt werden, wird automatisch entschieden, indem –wie schon bei der Segmentierung– ähnliche akustische Bedingungen in Klassen zusammengefaßt werden.

Kapitel 1

Einleitung

Automatische Spracherkennungssysteme *(ASES)* sind mittlerweile schon in vielen Bereichen des täglichen Lebens vorzufinden. Automatische Reisezugauskunft, Kontoführung per Telephon und automatische Diktiersysteme gehören zu den bekanntesten Beispielen. Diesen Systemen ist gemeinsam, daß Sprache als Medium für Dateneingabe verwendet wird. In solchen Fällen ist Sprache aus vielen Gründen verhältnismässig einfach automatisch zu erkennen. Oft sind z.B. nur einzelne Worte zu erkennen, ist das zu verwendende Vokabular beschränkt, bleiben die akustischen Bedingungen unverändert, sind die Sprachsegmente kurz und Anfang und Ende der Segmente bekannt.

Durch die Verwendung von Sprache wird es in vielen Fällen ermöglicht, monotone und lästige Aufgaben, die bisher von Menschen verrichtet werden mußten, von Computern erledigen zu lassen. In manchen Fällen ist Sprache als Eingabemedium notwendig, weil die Hände aufgrund einer Körperbehinderung, oder weil diese anderweitig gebraucht werden, nicht zur Verfügung stehen. Oft ist es aber auch einfach praktischer, Eingaben mittels Sprache statt über eine Tastatur zu machen.

Ein anderes Einsatzgebiet automatischer Spracherkennungssysteme, das viele neue Probleme mit sich bringt, ist Spracherkennung auf Audiodaten, die gar nicht in erster Linie für den Datenaustausch mit dem Computer gedacht sind; hierzu zählen insbesondere Audiodaten von Radio- und Fernsehsendungen. Für Spracherkennung auf solchen Daten gibt es viele praktische Anwendungen, wie z.B. die automatische Erzeugung von Untertiteln für Hörgeschädigte oder das automatische Erstellen von Datenbasen, die es ermöglichen, Radio- und Fernsehbeiträge zu einem bestimmten Schlagwort zu finden.

Gegenüber den Fällen, in denen ASES zur Dateneingabe verwendet werden, zeichnet sich Spracherkennung auf Audiodaten von Radio- und Fernsehsendungen dadurch aus, daß naturgemäß keine Einschränkungen bezüglich des Wortschatzes und der Sprechgeschwindigkeit gemacht werden können. In Radio- und Fernsehsendungen ist weiterhin nahezu alles vorhanden, was Spracherkennung erschweren kann, wie z.B. Hintergrundmusik, Telephongespräche, Hintergrundgeräusche, häufige Wechsel der akustischen Gegebenheiten, der Sprecher und der Eigenschaften des Übertragungskanals. Hinzu

kommt, daß in den meisten Fällen nur die Anfangs- und Endzeiten von Radio- und Fernsehsendungen bekannt sind. Spracherkennung auf Audiosegmenten der Länge ganzer Sendungen ist aber unter anderem aufgrund des hohen Speicherbedarfs selten machbar und auch nicht sinnvoll.

Die unterschiedlichen in einer Sendung vorkommenden Einflüsse, denen das Sprachsignal unterliegt (z.B. Bandbegrenzung des Sprachsignals durch Übertragung über einen Telephonkanal), machen es sinnvoll, speziell auf die jeweils vorliegenden Einflüsse zugeschnittene Verfahren bei der Spracherkennung einzusetzen.

In der vorliegenden Diplomarbeit werden Möglichkeiten untersucht, Audiodaten in Segmente zu unterteilen, die für eine anschließende Spracherkennung geeignet sind *(Segmentierung)*. Wegen der Empfindlichkeit von Spracherkennungssystemen gegenüber Änderungen der akustischen Bedingungen werden Segmente gebildet, innerhalb derer die Einflüsse auf das Sprachsignal möglichst konstant bleiben. Hierfür werden die in den Audiodaten vorliegenden Einflüsse bestimmten Klassen zugeordnet *(Klassifizierung)*. Die Einteilung von akustischen Bedingungen in Klassen wird automatisch und von Hand vorgenommen *(Klassenbildung)*. Abhängig von der ermittelten Klassenzugehörigkeit eines Segments kann einer von mehreren speziell angepaßten Spracherkennern ausgewählt werden.

Im Kapitel 2 wird erst ein Überblick über den Spracherzeugungsprozeß und Spracherkennung beim Menschen gegeben. Anschließend wird der grundlegende Aufbau automatischer Spracherkennungssysteme erläutert und das *Janus Recognition Toolkit (JRTk)* vorgestellt, mit dem es möglich ist, solche Systeme zu entwickeln, und das in der vorliegenden Arbeit verwendet wurde.

Grundlage für alle im Rahmen dieser Diplomarbeit durchgeführten Untersuchungen bildeten die Daten, die den Teilnehmern der sogenannten ARPA HUB-4 Evaluation 1995 zur Verfügung gestellt wurden. Diese Daten werden im Kapitel 3 beschrieben.

Auf das Problem der Segmentierung wird in Kapitel 4 eingegangen. Da das Zerteilen von großen Audioaufnahmen eine zentrale Aufgabe im gegebenen Zusammenhang darstellt, wurde der hierfür gewählte Ansatz ausgiebig untersucht. Die entsprechenden Versuche und Versuchsauswertungen werden in Kapitel 5 beschrieben.

Die Anpassung von Spracherkennern an bestimmte akustische Gegebenheiten, die Bildung von Klassen von akustischen Bedingungen für die Entwicklung spezialisierter Spracherkenner und die Auswahl des geeigneten Spracherkenners für ein Segment werden in Kapitel 6 behandelt. In Kapitel 7 werden Spracherkennungsversuche beschrieben und anhand dieser der Einfluß automatischer Segmentierung und Klassifizierung sowie der Verwendung mehrerer Spezialerkenner auf die Wortfehlerrate untersucht.

Abschließend wird in Kapitel 8 das entwickelte System mit anderen Systemen verglichen und bewertet. Außerdem werden Möglichkeiten zur Weiterentwicklung aufgezeigt.

Kapitel 2

Spracherzeugung und Spracherkennung

Spracherzeugung und Spracherkennung ist für den Menschen ein ganz natürlicher Vorgang. In den ersten Lebensjahren wird Sprechen und Verstehen der Muttersprache in aller Regel mühelos erlernt. Es wäre daher wünschenswert, wenn Sprache als Medium für die Kommunikation von Mensch und Maschine voll zur Verfügung stünde. Das ist heute noch nicht der Fall, aber in den letzten 30 Jahren wurden große Fortschritte in dieser Richtung gemacht.

In diesem Kapitel werden Spracherzeugung und -wahrnehmung beim Menschen, soweit es für die späteren Ausführungen von Bedeutung ist, und der bislang erfolgreichste Ansatz zur maschinellen Spracherkennung beschrieben. Eine ausführlichere Darstellung von Spracherzeugung und -wahrnehmung ist in [31] zu finden, von maschineller Spracherkennung in [34, 35, 43]. Anschließend wird das *JRTk* vorgestellt, mit dem es möglich ist, solche Spracherkennungssysteme aufzubauen.

2.1 Spracherzeugung und -wahrnehmung

Sprache ist eine der wichtigsten Formen zwischenmenschlicher Kommunikation. Eine Information, die mittels Sprache übertragen werden soll, muß vom Menschen in ein Sprachsignal umgewandelt werden. Hierbei durchläuft die Information verschiedene Codierungsstufen. Erst wird die Information als eine Folge von Worten codiert; die einzelnen Worte wiederum entsprechen einer Folge von Lauten, und diese Laute werden schließlich erzeugt durch eine Folge von Nervenimpulsen, die Vokaltrakt, Lunge und Stimmbänder steuern.

Ein Zuhörer muß, um im Sprachsignal enthaltene Informationen zu extrahieren, entsprechende Decodierungsschritte ausführen. Ausgehend vom Sprachsignal muß die zugehörige Folge von Lauten ermittelt, hieraus die gesagten Worte bestimmt und aus diesen die Information wiedergewonnen werden.

Einige der Prozesse, die beim Menschen ablaufen, um diese Codierung und Decodierung vorzunehmen, sind bereits recht ausführlich erforscht worden, andere hingegen sind noch weitgehend ungeklärt.

Für maschinelle Spracherkennung ist, was die Spracherzeugung beim Menschen anbelangt, vor allem der Zusammenhang von Phonem und Sprachsignal wichtig.

2.1.1 Das Sprachsignal

Jede Sprache verfügt über eine Anzahl kleinster unterscheidbarer linguistischer Einheiten, den **Phonemen**. Davon gibt es in einer Sprache ungefähr 50. Die Menge der Phoneme einer Sprache ist allerdings nicht eindeutig festgelegt. Der Mensch kann mit Hilfe des Vokaltrakts unbegrenzt viele verschiedene Laute erzeugen. Nur eine Teilmenge dieser Laute hat eine Bedeutung innerhalb einer Sprache, d.h. es ist nicht für alle Laute möglich, diese einem Phonem zuzuordnen. Die Sprachlaute werden oft auch **Phon** *(engl.: phone)* genannt.

Das Prinzip der Erzeugung von Sprachlauten im Vokaltrakt ist in allen Sprachen weitgehend gleich. Die folgenden Beschreibungen beziehen sich in erster Linie auf die englische und deutsche Sprache, gelten aber auch für die meisten anderen Sprachen.

Sprache wird in Form von Änderungen des Luftdrucks übermittelt. Der Mensch erzeugt diese Druckänderungen, indem er in der Lunge durch Muskelkontraktion einen Luftstrom hervorruft. Durch diesen Luftstrom können die Stimmbänder zum Vibrieren gebracht werden, oder es kann durch eine Verengung an einer Stelle im Vokaltrakt zu Turbulenzen im Luftstrom kommen. Weiterhin kann durch völligen Verschluß, Aufbauen eines Drucks und nachfolgendes plötzliches Öffnen des Vokaltrakts ein Luftstoß verursacht werden. Vibration, Turbulenz oder Luftstoß regen den Vokaltrakt, der sich akustisch wie eine Röhre verhält, zur Resonanz an. Abhängig von der Art der Anregung unterscheidet man die im folgenden beschriebenen drei Klassen von Lauten.

Stimmhafte Laute

Durch den Grad der Muskelkontraktion der Stimmbänder kann der Mensch deren Flexibilität und die Größe der Öffnung, die für den von der Lunge kommenden Luftstrom zur Verfügung steht, bestimmen. Wird die Flexibilität entsprechend gewählt, kann der Luftstrom eine Vibration der Stimmbänder verursachen. In gewissen Grenzen kann durch Anspannen oder Entspannen auch die Frequenz der Vibration bestimmt werden.

Laute, bei denen die Stimmbänder sich in dieser Weise periodisch öffnen und schließen, werden **stimmhafte Laute** genannt. Die Häufigkeit, mit der der Luftstrom periodisch unterbrochen wird, die Grundfrequenz des Sprachsignals bei stimmhaften Lauten, wird meistens mit **F0** bezeichnet. **F0**

ist die physikalische Größe, die als Stimmlage wahrgenommen wird.

Die Frequenz **F0** ist auch von der Länge der Stimmbänder abhängig. Die unterschiedliche Stimmlage von Männern, Frauen und Kindern ist durch diese unterschiedliche Länge zu erklären.

Nicht stimmhafte Laute

Neben den stimmhaften Lauten, gibt es noch die **nicht stimmhaften Laute**, bei denen durch eine Verengung im Vokaltrakt eine Turbulenz erzeugt wird.

Plosivlaute

Ein Luftstoß, der durch einmaliges Schließen, Aufbauen eines Drucks und anschließendes plötzliches Öffnen an einer Stelle im Vokaltrakt hervorgerufen wird, erzeugt einen sogenannten **Plosivlaut**.

Vokaltrakt und Artikulatoren

Der Vokaltrakt stellt eine Röhre mit elastischen, feuchten und warmen Wänden dar, deren variable Querschnittsfläche durch die Position von Zunge, Gaumen, Stimmbändern, Lippen, Kiefern und Zähnen, den sogenannten **Artikulatoren**, verändert werden kann (siehe Abbildung 2.1).

Die periodischen Luftstöße im Falle von stimmhaften Lauten, die turbulente Strömung im Falle von nicht stimmhaften Lauten und auch ein einmaliger Luftstoß im Falle von Plosivlauten regen den Vokaltrakt zur Resonanz an, d.h. einige der Frequenzen, die in der den Vokaltrakt anregenden Druckänderung enthalten sind, und deren Oberschwingungen werden verstärkt andere gedämpft.

a) Vokaltrakt

b) Vokaltrakt als Röhrensystem mit variabler Querschnittsfläche

c) vereinfachte Darstellung des Vokaltrakts als Röhrensystem

Abbildung 2.1: a) Schematische Darstellung des Vokaltrakts als Teil des menschlichen Kopfes und b) als Röhrensystem mit variabler Querschnittsfläche sowie c) idealisiert als System aus zwei Röhren.

Die Form des Vokaltrakts bestimmt dessen Resonanzverhalten. Da diese Form durch die Artikulatoren variiert werden kann, ist es dem Menschen möglich, gezielt das Resonanzverhalten des Vokaltrakts zu verändern und so unterschiedliche Laute hervorzubringen.

Die Resonanzfrequenzen des Vokaltrakts werden auch **Formanten** genannt und oft mit **F1, F2, F3, ...** bezeichnet, wobei die **F***i* aufsteigend geordnet sind. Die Formanten sind für die Spracherkennung sehr wichtig. In [24] wird z.B. ein Verfahren beschrieben, durch das basierend auf Messungen von **F0, F1, F2** und **F3** über einer Testmenge von Aufnahmen englischer Vokale der Vokal mit 93%-iger Genauigkeit bestimmt werden kann. Das entspricht in etwa der Genauigkeit, mit der Menschen Vokale in einem Konsonant-Vokal-Konsonant *(KVK)* Kontext erkennen können.

Eigenschaften des Sprachsignals

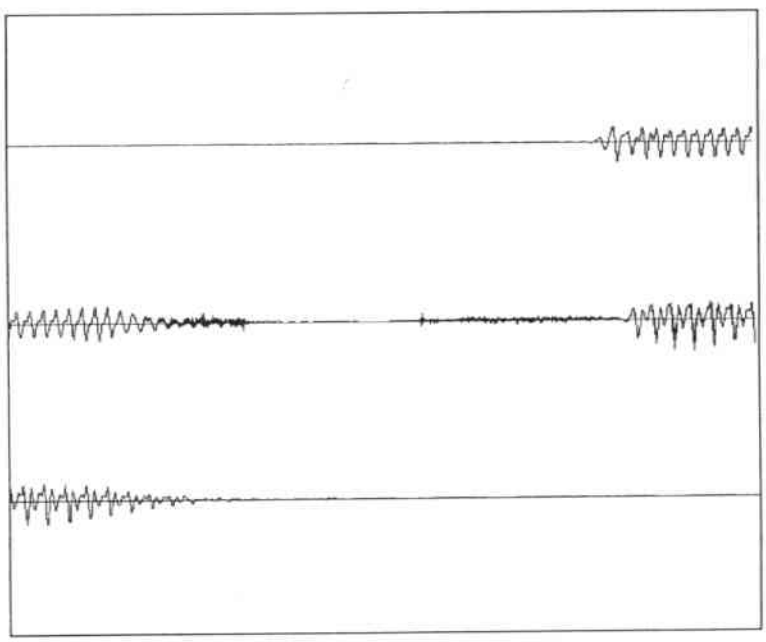

Abbildung 2.2: Das Sprachsignal des englischen Satzes „He was cute" gesprochen von einem männlichen Sprecher; eine Zeile entspricht 300ms.

In Abbildung 2.2 ist ein typisches Beispiel für ein Sprachsignal abgebildet. Dargestellt ist das Signal für den englischen Satz „He was cute". An dieser Abbildung sind viele generelle Eigenschaften von Sprachsignalen zu erkennen.

- Es sind Abschnitte auszumachen, innerhalb derer die Eigenschaften des Signals weitgehend konstant bleiben. Diese Abschnitte lassen sich grob in drei Klassen einteilen:

  quasi periodisch Eine Grundschwingung ist deutlich zu erkennen, und die einzelnen Perioden des Signals weisen viele Gemeinsamkeiten auf, stimmen aber nicht völlig überein.

nicht periodisch Perioden sind nicht auszumachen. Das Signal ist schwächer.

Stille Die Amplitude des Signals ist sehr klein.

- Es ist zwar möglich, z.B. Unterschiede zwischen den einzelnen periodischen Segmenten zu erkennen, aber eine Phonemzuordnung mit bloßem Auge ist nahezu unmöglich.

- Die Dauer der Abschnitte variiert stark, liegt hier aber im Mittel bei 80ms bis 100ms, je nachdem wieviele Segmente man unterscheidet.

Die Anregungsfunktion im Falle stimmhafter Laute ist eine Folge von Luftstößen mit dem zeitlichen Abstand von $\frac{1}{F0}$. Das führt dazu, daß im Sprachsignal die Energie bei ganzzahligen Vielfachen der Grundfrequenz $F0$ konzentriert ist.

Im Falle nicht stimmhafter Laute ist die Anregungsfunktion weitgehend zufällig und energieärmer als bei stimmhaften Lauten. Ein gutes Modell für die Anregungsfunktion ist nach [31] ein Signal mit flachem Spektrum und einer normalverteilten Amplitude.

Durch die genaue Form des Vokaltrakts wird bestimmt, welche Frequenzen verstärkt und welche gedämpft werden, d.h. wo die Formanten liegen. Ausgehend vom Energiespektrum des Sprachsignals ist es also möglich, die Lage der Formanten zu bestimmen, was wiederum Rückschlüsse auf die Position der Artikulatoren zuläßt.

Es folgt eine Auflistung weiterer Eigenschaften des Sprachsignals, die für automatische Spracherkennung (siehe Abschnitt 2.2 und Kapitel 6) sowie die Segmentierung und Klassifizierung von Audiosegmenten (siehe Kapitel 4) wichtig sind.

- Bedingt durch die Länge, Form und Beschaffenheit des Vokaltrakts sind im Sprachsignal in erster Linie Frequenzen bis 8kHz enthalten. Die Energie des Sprachsignals für stimmhafte Laute ist vor allem im Bereich der ersten drei Formanten konzentriert. Für einen männlichen Sprecher liegen diese Formanten im Durchschnitt bei 500Hz, 1500Hz und 2500Hz, schwanken aber beträchtlich in Abhängigkeit von der Position der Artikulatoren.

  Die Amplitude des Sprachsignals bei nicht stimmhaften Lauten ist weitaus geringer als bei stimmhaften bedingt durch den Enerieverlust an der Stelle, an der der Vokaltrakt verengt ist. Die Energie ist in Bereichen über 2500Hz konzentriert, was daran liegt, daß vor allem Frequenzen in der Nähe der Resonanzfrequenzen des kurzen Vokaltraktstücks von der Verengung bis zu den Lippen verstärkt werden. Die Bandbreiten der Resonanzfrequenzen im Falle nicht stimmhafter Laute sind viel größer als im Falle stimmhafter.

Während stimmhafte und nicht stimmhafte Laute innerhalb eines Zeitabschnitts nahezu konstante Signaleigenschaften aufweisen (geringe Variation des Spektrums), sind Plosivlaute in erster Linie anhand von Änderungen im Spektrum charakterisierbar.

- Die Länge des Vokaltrakts variiert von Sprecher zu Sprecher. Die durchschnittlichen Vokaltraktlänge von Männern liegt bei 17cm, die von Frauen bei 13cm. Das führt zu einer Verschiebung der Resonanzfrequenzen.

- Das Sprachsignal enthält viel Redundanz. Menschen sind z.B. in der Lage synthetisch erzeugte Vokallaute zu identifizieren, wenn $\mathbf{F}i$ für $i \geq 3$ fest gelassen, und nur die ersten zwei Formanten variiert werden.

- Durchschnittswerte für **F0** sind nach [31] 132Hz bei Männern und 223Hz bei Frauen. Beim normalen Sprechen schwankt **F0** um ca. eine Oktave.

- Die Übergänge des Sprachsignals sind nicht abrupt, sondern kontinuierlich.

- Die Dauer von Sprachlauten beträgt im Durchschnitt etwa 80ms, variiert allerdings stark. Die durchschnittliche Dauer eines Diphtongs beträgt nach [31] z.B. 180ms, die von Vokalen 130ms, die von Konsonanten 70ms.

- Das Sprachsignal variiert für ein Phonem sehr stark, wenn es in unterschiedlichen Phonemkontexten oder unterschiedlich schnell ausgesprochen wird.

  Das liegt daran, daß die Artikulatoren eine Folge von Bewegungen ausführen müssen, um eine gewünschte Phonemfolge auszudrücken. Je nach Phonemkontext wird also die Zielposition der Artikulatoren für ein bestimmtes Phonem aus anderen Ausgangspositionen angesteuert. Hinzu kommt, daß die einzelnen Artikulatoren sich nur unterschiedlich schnell bewegen können. Hierdurch wird das Sprachsignal des aktuellen Phonems stark durch das vorhergehende Phonem beeinflußt.

  Es gibt auch den umgekehrten Fall, daß sich bestimmte Artikulatoren, die für das folgende Phonem eine andere Stellung einnehmen müssen, im Voraus verändern, wenn nicht die für die Erkennung des aktuellen Phonems nötigen Eigenschaften des Sprachsignals dadurch verändert werden. Im ersten Fall spricht man von vorwärtsgerichteter **Koartikulation**, im zweiten Fall von rückwärtsgerichteter *(engl.: forward/backward coarticulation)*.

- Einen großen Einfluß auf das Sprachsignal hat die jeweils vorliegende Gesprächssituation. Gelesene Sprache unterscheidet sich stark von

spontaner Sprache, z.B. in Bezug auf die Länge der Pausen und Phoneme sowie die Betonung. Weiterhin kommen in Spontansprache oft Dehn- und Störlaute vor, werden häufiger Füllworte und umgangssprachliche Redewendungen benutzt. Zudem sind spontan gesprochene Sätze oftmals nicht grammatikalisch korrekt; es werden z.B. Sätze abgebrochen und neu begonnen oder falsch beendet.

Außerdem können sich Betonung, Wortauswahl, Sprechgeschwindigkeit und Länge der Pausen in Abhängigkeit davon unterscheiden, ob man mit Erwachsenen, Kindern, Personen, die die Sprache nicht vollständig beherrschen, oder Computern spricht.

2.1.2 Hören und Verstehen

Die Vorgänge, die beim Hören und Verstehen von Sprache beim Menschen ablaufen, sind nur zu einem kleinen Teil bekannt. Welche Verarbeitung das Sprachsignal erfährt, bis es schließlich in Form von Nervenimpulsen an das Gehirn weitergeleitet wird, hingegen ist weitgehend geklärt.

Das Gehör des Menschen

Das menschliche Ohr ist dreigeteilt. Der erste Teil von der Ohrmuschel zum Trommelfell dient unter anderem dem Schutz des Ohres, dem Lokalisieren von Geräuschen und verstärkt als Röhrenstück einer Länge von ca. 2.7cm und einer Breite von ca. 0.7mm die Frequenzen zwischen 3kHz und 5kHz.

Das Mittelohr transformiert Schwingungen des Trommelfells über eine Konstruktion aus drei kleinen Knochen in mechanische Energie am *ovalen Fenster* des Innenohres. Außerdem wird durch die Knochenkonstruktion und die Muskeln des Mittelohrs das empfindliche innerne Ohr innerhalb gewisser Grenzen vor Schäden durch zu laute Geräusche bewahrt.

Das innere Ohr ist eine mit Flüssigkeit gefüllte schneckenförmige Röhre, die durch zwei Membrane in drei Kammern geteilt wird. Bewegungen am ovalen Fenster des Innenohrs führen dazu, daß die Flüssigkeit im Innenohr zwischen den Kammern hin- und herströmt. Das versetzt die Membrane in Bewegung. Auf einer der Membrane befinden sich s.g. Haarzellen, die durch diese Bewegung gebeugt werden. Die Beugung der Haarzellen löst Nervenimpulse an den Enden des Hörnervs, mit dem die Haarzellen verbunden sind, aus.

Verarbeitung des Sprachsignals

Die Membran, mit der die Haarzellen verbunden sind, verändert ihre Beschaffenheit und Breite entlang ihres Verlaufs. Jede Stelle entlang der Membran hat eine bestimmte *charakteristische Frequenz (CF)*. Wird die Membran durch diese Frequenz angeregt, so vibriert sie an dieser Stelle maximal.

Die Aktivität der mit den Haarzellen verbundenen Nervenzellen in Reaktion auf bestimmte Geräusche kann gemessen werde. Auf diese Weise ist es möglich zu untersuchen, wie das Sprachsignal im Ohr des Menschen vorverarbeitet wird, bis es in Form von Nervenimpulsen zum Gehirn gelangt. Einige in diesem Zusammenhang wichtige Informationen sind im folgenden aufgelistet.

- Die mit den Haarzellen verbundenen Nervenzellen senden Inpulse (Abweichungen vom elektrischen Ruhepotential) mit einer Dauer von etwa 0.5-1ms zum Gehirn. Bei Stille werden ca. 10-50 Impulse pro Sekunde in zufälligem Abstand voneinander gesendet. In Anwesenheit von Geräuschen nimmt die Anzahl der Impulse zu.

- Die Aktivität der Nervenzellen des Hörnervs spiegelt durch die Verknüpfung der Nervenzellen mit den Haarzellen einer bestimmten CF die Anwesenheit oder Abwesenheit einer bestimmten Frequenz in einem Geräusch wieder.

- Die Membran, auf der sich die Haarzellen befinden, verhält sich bei Anregung durch eine bestimmte Frequenz CF_i wie ein Bandpaßfilter, wobei das Verhältnis von Bandbreite BB_i und CF_i nahezu konstant ist, d.h. es gilt $\frac{BB_1}{CF_1} \approx \frac{BB_2}{CF_2}$ (siehe Abbildung 2.3). Das bedeutet, daß die Frequenzauflösung im Innenohr für kleine Frequenzen am besten ist und stetig abnimmt.

  Weiterhin verhält sich die Position d_i auf der Membran zu der zugehörigen CF_i ungefähr logarithmisch (siehe Abbildung 2.3).

- Es ist ausgiebig untersucht worden, wie Amplitude, Änderung der Amplitude, Änderungen der anwesenden Frequenzen, Pausen zwischen akustischen Reizen, Dauer akustischer Reize und Kombinationen verschiedener akustischer Reize das Aktivitätsmuster der Nervenzellen beeinflussen (siehe z.B. [31, Kapitel 4]).

- Wie der Mensch schließlich die Aktivitätsmuster der Nervenzellen im Gehirn auswertet und so an die im Sprachsignal enthaltenen Informationen gelangt, ist weitgehend unbekannt.

  Welche akustischen Eigenschaften des Sprachsignals für die Spracherkennung beim Menschen höchstens von Bedeutung sein können, kann durch Versuche ermittelt werden, in denen bestimmte Eigenschaften von akustischen Signalen verändert und andere fest gelassen werden. So können dann die gerade noch wahrnehmbaren Unterschiede festgestellt werden.

- Für das Verstehen von Sprache ist in erster Linie der Frequenzbereich 200-5600Hz von Bedeutung. Das entspricht auch dem Bereich, für den das Gehör des Menschen besonders empfindlich ist und der den Hauptanteil der Energie des Sprachsignals enthält.

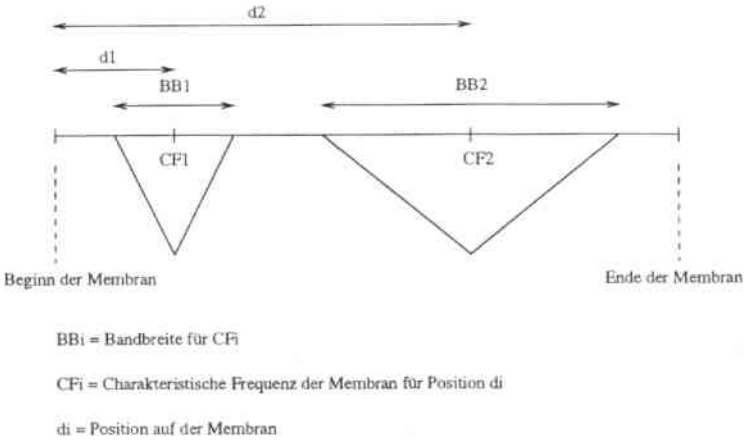

Abbildung 2.3: Schematische Darstellung der Zusammenhänge von charakteristischer Frequenz CF_i und der zugehörigen Bandbreite BB_i in Abhängigkeit von der Position d_i entlang der Membran

2.1.3 Einflüsse auf das Sprachsignal

Das im Vokaltrakt erzeugte Sprachsignal wird, bevor es das Ohr des Zuhörers erreicht, in verschiedener Weise verändert.

- Dem Sprachsignal werden Hintergrundgeräusche überlagert. Diese Hintergrundgeräusche können zufälliger Natur sein (Rauschen), strukturiert und regelmäßig (Maschinengeräusche), einmalig (Türklappen) und auch harmonisch (Musik). Solche Überlagerungen sind mit dem Sprachsignal unkorreliert.

- Weiterhin wird das Sprachsignal selbst vom Boden und ggf. den Wänden reflektiert und so sich selbst überlagert. Diese Beiträge sind mit dem Sprachsignal korreliert.

- Wird das Signal noch durch ein Mikrophon in ein elektrisches Signal umgewandelt, übertragen, evtl. auf Band aufgenommen oder in sonstiger Weise behandelt (z.B. durch Dämpfen oder Verstärken bestimmter Frequenzbereiche in einer Stereoanlage) und anschließend wieder über einen Lautsprecher zurück in eine Druckänderung transformiert, so wirken zusätzlichen Einflüsse auf das Sprachsignal, die mit diesem korreliert und unkorreliert sein können.

Die Abfolge von Transformationen, die das Signal erfährt, wird **Übertragungskanal** genannt. Viele der mit dem Sprachsignal korrelierten Überlagerungen können durch lineare Filter modelliert werden. Hierzu zählen z.B. Echoeffekte und einige der durch Mikrophone bewirkten Transformationen.

Verschiedene Übertragungskanäle erschweren die Rückgewinnung der reinen Sprachinformation und somit die Spracherkennung sehr.

2.2 Maschinelle Spracherkennung

In den letzten 30 Jahren wurden viele verschiedene Ansätze zur maschinellen Spracherkennung untersucht. In [43] werden vier Schulen der Spracherkennung unterschieden, und zwar der *auf Schablonen basierende Ansatz*, der *auf Wissensbasen aufbauende Ansatz*, der *stochastische Ansatz* und der *neuronale Ansatz*. Eine andere Einteilung wird in [35] vorgenommen. Hier wird zwischen dem *akustisch-phonetischen Ansatz*, *dem auf Mustererkennung basierenden Anzatz* und dem *Künstliche Intelligenz Ansatz* unterschieden.

Eine eindeutige Trennung der Ansätze ist allerdings oft nicht möglich, insbesondere bei hybriden Systemen, die z.B. künstliche Neuronale Netze für die Klassifizierung von Merkmalsvektoren in einem ansonsten auf stochastischen Modellen basierenden System einsetzen. Die heutzutage erfolgreichsten Systeme sind in erster Linie der Klasse der stochastischen Systeme, bzw. der auf Mustererkennung basierenden Klasse von Systemen zuzuordenen. Im folgenden wird diese Art von Systemen näher beschrieben.

2.2.1 Genereller Aufbau

Während sich Spracherkennungssysteme in den Details stark unterscheiden, ist der grundlegende Aufbau recht ähnlich. Um das Sprachsignal für den Computer bearbeitbar zu machen, wird es erst abgetastet. Wie man an Abbildung 2.2 sehen kann, ist das reine Sprachsignal nicht leicht zu interpretieren. Daher wird das Sprachsignal vorverarbeitet, wobei die zu verarbeitende Datenmenge reduziert und die für die Spracherkennung wichtigen Merkmale extrahiert werden.

Anschließend wird das vorverarbeitete Sprachsignal segmentiert und die einzelnen Segmente klassifiziert; diese Schritte erfolgen bei stochastischen Systemen oft implizit (siehe weiter unten). Die Begriffe Segmentierung und Klassifizierung in diesem Zusammenhang sind nicht zu verwechseln mit dem Zerteilen langer Audioaufnahmen in Stücke mit gleichbleibenden akustischen Bedingungen, wie sie später gebraucht werden. Aus der Folge von klassifizierten Segmenten wird dann unter Berücksichtigung von Randbedingungen und Zusatzwissen die Wortfolge ermittelt, die am besten auf das abgetastete Sprachsignal paßt.

Die einzelnen Systeme unterscheiden sich nun in folgenden Bereichen, auf die weiter unten im Detail eingegangen wird.

Vorverarbeitung Welche Merkmale aus dem Sprachsignal extrahiert werden, hängt z.B. vom Einsatzgebiet des Spracherkenners ab. Ziel der Vorverarbeitung kann es sein, die Einflüsse verschiedener Mikrophone oder einer bestimmten Art von Störungen zu eliminieren.

Segmentierung und Klassifizierung Dieser Schritt dient dazu, den zusammenhängenden Strom von Audiodaten entsprechend den Sprach-

ereignissen, die diese Audiodaten hervorgerufen haben, zu zerteilen und die Sprachereignisse zu identifizieren.

Bei dem *auf Mustererkennung basierenden Ansatz* vergleicht man die extrahierten Merkmale des zu testenden Sprachsignals mit erlernten Referenzmustern. Dieser Vergleich wird in zwei Stufen durchgeführt; lokal, um Ähnlichkeiten des zu testenden Sprachsignals mit den Referenzmustern festzustellen *(pattern classsification)*, und global, um der Tatsache gerecht zu werden, daß die Dauer einzelner Sprachereignisse (z.B. Dauer von Phonemen, Dauer von Phonemübergängen, Länge von Pausen) stark variieren kann *(time alignment)*. Die Segmentierung erfolgt bei diesem Spracherkennungsansatz implizit bei dem globalen Anpassen der Referenzmuster an die zu testende Folge von Merkmalsvektoren. Die Klassifizierung der einzelnen Merkmalsvektoren erfolgt bei dem lokalen Vergleich. Hierauf wird im Abschnitt 2.2.3 näher eingegangen.

Meistens wird nicht nur das am besten passende Muster, bzw. das am besten passende Phonem, berücksichtigt, sondern die n besten. Das Ergebnis dieses Segmentierungs- und Klassifizierungsschrittes ist daher meistens ein gerichteter Graph (in diesem Zusammenhang in der Literatur oft Gitter; *engl: lattice*), wobei jeder Pfad durch den Graph einer Hypothese darüber, was gesagt wurde, entspricht. Die Knoten des Graphs können im Falle des auf Musterkennung basierenden Ansatzes beliebige Einheiten repräsentieren (z.B. Worte, Silben, Phoneme, Subphoneme).

Suche/Decodierung In diesem Schritt wird derjenige Pfad durch den gerichteten Graph ausgewählt, der unter Berücksichtigung von Randbedingungen und Zusatzwisssen am besten paßt, wobei die Entscheidung, was *am besten paßt*, aufgrund von Abstandsmaßen getroffen wird. Randbedingungen und Zusatzwissen können in diesem Zusammenhang sein, daß eine Phonemfolge ein sinnvolles Wort ergeben muß oder eine Wortfolge einen syntaktisch korrekten Satz. Eine Art von Randbedingung stellen s.g. *Sprachmodelle* dar, die kurz in Abschnitt 2.2.4 beschrieben werden.

Randbedingungen und Zusatzwissen können an dieser Stelle in verschiedener Form eingebracht werden [37].

2.2.2 Vorverarbeitung

Abgesehen von der Vorverarbeitung, die das Sprachsignal im Ohr des Menschen erfährt, bis es schließlich in Form von Nervenimpulsen ans Gehirn weitergeleitet wird, ist nicht bekannt, wie der Mensch an die im Sprachsignal enthaltenen Informationen gelangt und welche Merkmale des Sprachsignals für diese Decodierung wichtig sind.

Allgemein sollte die Vorverarbeitung möglichst viele Informationen über das Sprachsignal erhalten, die für die Spracherkennung wichtig sind, diese Informationen in einer Form darstellen, die eine möglichst einfache Interpretation erlaubt, und die für die Spracherkennung nicht notwendigen oder sogar störenden Anteile im Signal entfernen. Es folgen einige Beispiele dafür, was die Aufgabe der Vorverarbeitung sein kann.

- Der Frequenzbereich des Sprachsignals von 200Hz bis 5600Hz ist für das Verstehen von Sprache von größter Bedeutung. Um diese Informationen zu erhalten, muß das Sprachsignal mit mindestens 11200Hz abgetastet werden (siehe [30]). Stellt man jeden Abtastwert als 8Bit Wert in einem Rechner dar, so ist das eine recht große Datenmenge (ca. 10KByte/s). Wenn man hingegen bedenkt, daß ein Mensch pro Sekunde etwa 8 Phoneme ausspricht, wobei die Menge der Phoneme sich in 6 Bits codieren läßt (ca. 50 Phoneme; $50 < 2^6$), so wird deutlich, daß im Sprachsignal viel Redundanz enthalten ist. Die Vorverarbeitung dient dazu, diese Datenmenge zu reduzieren.

- An Abbildung 2.2 ist kaum ablesbar, was gesagt wurde. Der Übergang in den Frequenzbereich aber macht es schon möglich, für viele Laute mit bloßem Auge das gesagte Phonem mit recht großer Sicherheit zu bestimmen.

  In Abbildung 2.4 ist oben der Logarithmus des Energiespektrums[1] zum Zeitpunkt 0.61s für das Sprachsignal in Abbildung 2.2 dargestellt. Es sind deutlich die ersten drei Formanten erkennbar. Durch einfaches Ablesen der Formantenwerte und Nachschauen in einer Tabelle läßt sich ermitteln, daß es sich bei dem gesagten Laut vermutlich um ein /U/ handelt. Dennoch sind in diesem Spektrum mehr Informationen enthalten, als eigentlich nötig wären, um zu dieser Hypothese zu gelangen.

  Ein weiterer Vorverarbeitungsschritt kann sein, das Spektrum ähnlich der Vorverarbeitung, die das Sprachsignal im Innenohr des Menschen erfährt, in einer Anzahl von sich überlappenden Frequenzbändern zusammenzufassen, wobei die Frequenzbänder entsprechend der Frequenzauflösung des menschlichen Gehörs gewählt werden. Ein Beispiel hierfür ist die s.g. *Melscale-Filterung* (siehe Abbildung 2.4 unten).

- Weitere Aufgaben der Vorverarbeitungen können sein

  1. Sprechernormalisierung, d.h. die aus dem Sprachsignal extrahierten Merkmale werden in einer Weise verändert, die der Tatsache gerecht wird, daß der Vokaltrakt von Person zu Person verschieden ist (siehe z.B. [20]),

[1] Das Energiespektrum wurde über einem Zeitfenster der Länge 16ms berechnet.

2. Beseitungung von additiven Störungen im Frequenzbereich; Normalisierung bzgl. der Eigenschaften des Übertragungskanals (siehe z.B. [21, 46]),
3. Normalisierung bzgl. der akustischen Umgebung (siehe z.B. [22]).

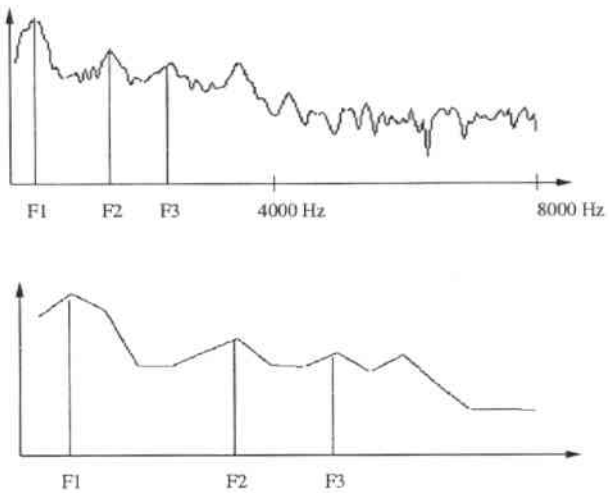

Abbildung 2.4: Der Logarithmus des Energiespektrums des weiter oben abgebildeten Sprachsignals des Satzes „He was cute" zum Zeitpunkt 0.61s berechnet über einem Fenster von 16ms (oben) sowie dasselbe Spektrum nach Melscale-Filterung (unten)

Das Sprachsignal $s(t)$ wird also erst durch Abtastung in ein Folge von Abtastwerten $(s_j)_{j=1}^m$ umgewandelt. In aller Regel wählt man das Abtastinterval fest und kleiner als $\frac{1}{11200\text{Hz}}$, um keine Informationen über die für die Spracherkennung wichtigen Frequenzen im Bereich von 200-5600Hz zu verlieren (siehe z.B. [30, 36]); d.h. $s_j = s(jT_a)$ für $j = 1, 2, 3, \ldots, m$, wobei T_a das Abtastinterval ist.

Da das Sprachsignal sich, wie in Abschnitt 2.1.1 beschrieben, nicht abrupt ändert, sondern über kurze Zeitintervalle nahezu konstante Eigenschaften aufweist, analysiert man das Sprachsignal innerhalb entsprechend gewählter Zeitfenster.

Diese Analyse liefert eine Folge von Merkmalsvektoren $(\boldsymbol{x}_i)_{i=1}^n$, die, wie oben beschrieben, für die Spracherkennung wichtige Informationen enthalten, und möglichst robust gegenüber Störeinflüssen sein sollten (für eine detailiertere Erörterung dieser Vorverarbeitungsschritte siehe z.B. [36]).

2.2.3 Akustische Modellierung

Der Vorverarbeitungsschritt liefert eine Folge von Merkmalsvektoren $(\boldsymbol{x}_i)_{i=1}^n = (\boldsymbol{x}_n, \ldots, \boldsymbol{x}_2, \boldsymbol{x}_1)$. Ausgehend von dieser Folge von Merkmalsvektoren wird versucht, die Wortfolge $(w_j)_{j=1}^m = (w_m, \ldots, w_2, w_1)$ zu bestimmen,

die mit größter Wahrscheinlichkeit gesagt wurde. Es wird also (w_j^*) gesucht mit

$$P((w_j^*)|(\boldsymbol{x}_i)) = \max_{(w_j)}\{P((w_j)|(\boldsymbol{x}_i))\} \tag{2.1}$$

Die Bayes-Regel für bedingte Wahrscheinlichkeiten angewendet auf den Ausdruck, über den maximiert wird, ergibt

$$P((w_j)|(\boldsymbol{x}_i)) = \frac{p((\boldsymbol{x}_i)|(w_j))P((w_j))}{p((\boldsymbol{x}_i))} \tag{2.2}$$

Hierbei nennt man $P((w_j)|(\boldsymbol{x}_i))$ die *A-posteriori-Wahrscheinlichkeit* für die Wortfolge (w_j) gegeben die Folge von Merkmalsvektoren $(\boldsymbol{x}_i)$, $p((\boldsymbol{x}_i)|(w_j))$ die *klassenbedingte Wahrscheinlichkeitsdichte* von $(\boldsymbol{x}_i)$ gegeben (w_j) und $P((w_j))$ die *A-priori-Wahrscheinlichkeit* der Wortfolge (w_j). $p((\boldsymbol{x}_i))$ ist die Wahrscheinlichkeitsdichte dafür, daß die Vektorfolge $(\boldsymbol{x}_i)$ beobachtet wird.

Als Aufgabe eines Spracherkenners kann es angesehen werden, die A-posteriori-Wahrscheinlichkeit durch die richtige Wahl von (w_j) in (2.2) zu maximieren. $p((\boldsymbol{x}_i))$ ist im Sinne dieser Maximierungsaufgabe unwesentlich, da dieser Wert von (w_j) unabhängig, also konstant für alle Wortfolgen ist. Es müssen daher nur die klassenbedingte Wahrscheinlichkeitsdichte $p((\boldsymbol{x}_i)|(w_j))$ und die a-priori Wahrscheinlichkeit $P((w_j))$ für Folgen (w_j) bestimmt werden.

In stochastischen Systemen wird das explizit gemacht, indem man dem Spracherzeugungsprozeß Modelle zugrunde legt, durch die $p((\boldsymbol{x}_i)|(w_j))$ und $P((w_j))$ approximiert werden. Die Modellierung von $p((\boldsymbol{x}_i)|(w_j))$ wird *akustische Modellierung* genannt, da man die Erzeugung der Merkmalsvektoren des Sprachsignals bei gegebener Wortfolge modelliert. Die Modellierung von $P((w_j))$ wird entsprechend *Sprachmodellierung* genannt, da man versucht, die Wahrscheinlichkeiten für Wortfolgen in einer Sprache mathematisch zu erfassen.

In diesem Abschnitt wird auf die akustische Modellierung eingegangen, im nächsten auf Sprachmodelle.

Klassifikatoren

Beim auf Mustererkennung basierenden Ansatz werden Klassifikatoren trainiert, mit denen für einen zu klassifizierenden Merkmalsvektor bestimmt wird, welcher Klasse dieser zuzuordnen ist. Die in dem Zusammenhang bei Spracherkennung am häufigsten verwendeten Klassifikatoren sind künstliche Neuronale Netze und auf Mixturen multivariater Normalverteilungen basierende Klassifikatoren.

Im allgemeinen trainiert man Klassifikatoren, indem die Parameter jedes Klassifikators so eingestellt werden, daß auf einer Trainingsmenge der Fehler der Klassifikatoren minimiert oder die klassenbedingte Wahrscheinlichkeit einem Modell gemäß maximiert wird (siehe z.B. [4, 2]). Im Falle von Neuronalen Netzen werden die Gewichte der einzelnen Neuronen eingestellt, im

Falle von Mixturen multivariater Gaußverteilungen werden die Mittelwertvektoren, Kovarianzmatrizen und Mixturgewichte entsprechend angepaßt (siehe hierzu [12, 2, 42, 6]).

Hidden Markov Models (HMM)

Die Einordnung einzelner Merkmalsvektoren wird in auf Mustererkennung basierenden Systemen durch Klassifikatoren vorgenommen. Die zeitliche Zuordnung *(time alignment)* der gesamten Folge von Merkmalsvektoren erfolgt entweder durch Methoden der Dynamischen Programmierung oder Hidden Markov Models.

Die zur Zeit erfolgreichsten Systeme zur automatischen Spracherkennung basieren auf *Hidden Markov Models*. Die für Spracherkennung verwendete Ausprägung dieser Modelle besteht aus

1. einer Anzahl N von Zuständen q_i,
2. einer Wahrscheinlichkeitsverteilung, die angibt, wie wahrscheinlich ein Übergang von Zustand q_i in q_j ist,
3. einer Wahrscheinlichkeitsverteilung, die angibt, wie wahrscheinlich es ist, daß das HMM im Zustand q_i die Ausgabe $\boldsymbol{x}_i$ liefert, die s.g. *Emissionswahrscheinlichkeitsverteilung*, sowie
4. der Wahrscheinlichkeit π_i, daß sich das HMM zum Zeitpunkt $t = 0$ im Zustand i befindet.

Zur Modellierung der Emissionswahrscheinlichkeitsverteilung werden in ASES meistens Mixturen multivariater Gaußverteilungen oder Neuronale Netze verwendet. Will man an dieser Stelle Neuronale Netze verwenden, so ist es für die wahrscheinlichkeitstheoretische Deutung notwendig, die Ausgaben der Neuronalen Netze in eine Form zu bringen, die als klassenbedingte Wahrscheinlichkeitsdichte gedeutet werden kann. Das ist aber ohne weiteres möglich (siehe [2]).

Generell kann ein HMM ein beliebig verknüpfter gerichteter Graph sein. Da HMMs in Spracherkennungssystemen die Aufgabe haben, die zeitliche Abfolge von Sprachereignissen zu modellieren, also eine nichtlineare Stauchung oder Streckung der Zeitachse vorzunehmen, sind nur s.g. Links-Rechts HMMs sinnvoll (siehe [35, Kapitel 6]). In Abbildung 2.5 ist ein HMM dargestellt, wie es für die Modellierung eines Phonems in einem Spracherkenner verwendet werden könnte.

Das Phonem wird hier durch zwei Subphoneme modelliert, die wiederum durch eine Anzahl von Zuständen modelliert werden. Aufgrund der starken Kontextabhängigkeit von Phonemen werden in Spracherkennungssystemen meistens Polyphone, d.h. Modelle für Phoneme im Kontext mehrerer anderer Phoneme, verwendet. Triphone modellieren z.B. Phoneme mit rechtem und linkem Kontext.

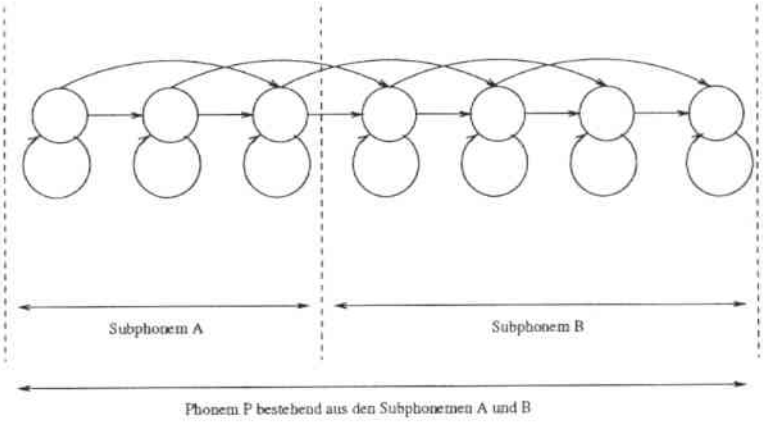

Abbildung 2.5: Links-Rechts HMM, wie es für Spracherkennung verwendet werden könnte

Das in Abbildung 2.5 dargestellte HMM kann also beispielsweise den Diphtong /e^y/ mit rechtem Kontext /b/ und linkem Kontext /t/ modellieren. Die Entscheidung, dieses Phonem durch zwei Subphoneme zu modellieren, kann getroffen worden sein, weil ein Diphtong von einer Vokaltraktkonfiguration eines Vokals zu einer Vokaltraktkonfiguration eines anderen Vokals übergeht. Die Anzahl der Zustände für die Subphoneme kann gewählt worden sein, um die verschiedenen Phasen der Bewegung, die die Artikulatoren ausführen müssen, zu modellieren. Die verschiedenen Zustandsübergänge (Übergang in denselben, den nächsten oder den übernächsten Zustand) werden der Tatsache gerecht, daß Phoneme unterschiedlich ausgesprochen werden können und daher manche Zustände mehrmals, nur einmal oder gar nicht durchlaufen werden müssen.

Will man HMMs für die Spracherkennung verwenden, muß man folgende drei Probleme bewältigen (λ bezeichnet im folgenden die Menge der Parameter eines HMMs, $(\boldsymbol{x}_i)$ eine Folge von Merkmalsvektoren und (Q_i) eine Folge von Zuständen).

Auswertung Man muß die Wahrscheinlichkeitsdichte dafür bestimmen können, daß eine Folge $(\boldsymbol{x}_i)$ von Merkmalsvektoren durch ein HMM mit λ erzeugt wird. Diesen Wert $p((\boldsymbol{x}_i)|\lambda)$ braucht man, um für zwei HMMs bei gegebenem Sprachsignal entscheiden zu können, für welches HMM das beobachtete Signal wahrscheinlicher ist.

Decodierung Man muß zu gegebenem $(\boldsymbol{x}_i)$ die Folge (Q_i) bestimmen können, die für ein λ optimal ist. Was in diesem Zusammenhang optimal bedeutet, wird durch ein Optimalitätskriterium festgelegt. Das an dieser Stelle am häufigsten verwendete Kriterium ist die Maximierung der Wahrscheinlichkeit $P((Q_i)|(\boldsymbol{x}_i), \lambda)$ für den gesamten Pfad. Die Zustandsfolge entspricht einer Folge von Sprachereignissen und stellt somit die Decodierung der Folge von Merkmalsvektoren für ein λ dar.

Training Man muß die Parameter λ eines HMMs über einer Trainingsmenge $(\boldsymbol{x}_{\text{Training},i})$ so bestimmen, daß das HMM möglichst gut zur Trainingsmenge paßt. Oder anders ausgedrückt, man muß die Parameter so einstellen, daß die Wahrscheinlichkeit dafür, daß das Modell die tatsächlich beobachtete Merkmalsvektorfolge hervorbringt, maximiert wird. Man sucht also die Menge von Parametern λ^* für die gilt $\lambda^* = \arg\max_{\lambda}\{p((\boldsymbol{x}_{\text{Training},i})|\lambda)\}$.

Für die Lösung dieser Probleme gibt es effiziente Algorithmen; für das Auswertungsproblem den *Forward-Algorithmus* und den *Backward-Algorithmus*, für die Bestimmung der besten Zustandsfolge den *Viterbi-Algorithmus* und für das Trainingsproblem den *Baum-Welch-Algorithmus* (siehe [34, 35]).

Da es in der Größenordnung von 50^n verschiedene n-Phone gibt, ist klar, daß schon ab Triphonen eine sehr große Anzahl von Parametern bestimmt werden muß. Das verursacht einerseits einen sehr hohen Rechenzeit- und Speicherplatzbedarf und erfordert andererseits sehr große Trainingsmengen, um die Parameter zuverlässig zu bestimmen. Weiterhin ist es aufgrund der Ähnlichkeit vieler Kontexte auch gar nicht unbedingt nötig, diese gesondert zu modellieren.

Ein Schritt, der daher in den meisten Systemen dieser Art durchgeführt wird, ist ein s.g. *context clustering*. Hierbei wird eine Menge von Kontextklassen gebildet, innerhalb derer die Verteilungen für die Berechnung der lokalen Wahrscheinlichkeiten sich so ähnlich sind, daß man wenig an Genauigkeit verliert, wenn man all diese Kontexte durch nur eine Verteilung modelliert (siehe z.B. [49, 29, 47]).

Werden in einem System Mixturen multivariater Gaußverteilungen verwendet, so kann die Anzahl der zu bestimmenden Parameter außerdem reduziert werden, indem für bestimmte Mengen von Modellen dieselben Mittelwertvektoren und Kovarianzmatrizen benutzt und nur die Mixturgewichte angepaßt werden. Wählt man z.B. für alle Modelle, die ein bestimmtes Phonem im Kontext anderer Phoneme modellieren, dieselben Mittelwertvektoren und Kovarianzmatrizen, so spricht man von einem phonetisch gebundenen halb-kontinuierlichen HMM (*phonetically tied semi-continuous HMM*).

Verzichtet man auf diese Form der Parameterreduktion und verwendet für alle Modelle eigene Verteilungen, so sprich man von einem kontinuierlichen HMM (*(fully) continuous HMM*). Bildet man durch Vektor-Quantisierung die Folge $(\boldsymbol{x}_i)$ auf eine Folge von diskreten Werten (O_i) ab und modelliert die Emissionswahrscheinlichkeit durch diskrete Verteilungen, so spricht man von einem diskreten HMM (*discrete HMM*).

2.2.4 Sprachmodelle

Wie in Abschnitt 2.2.3 erwähnt, ist für Spracherkennung neben der klassenbedingten Wahrscheinlichkeitsdichte $p((\boldsymbol{x}_i)|(w_j))$ die A-priori-Wahrscheinlichkeit einer Wortfolge $P((w_j))$ wichtig. Es gilt

$$
\begin{aligned}
P((w_j)) &= P((w_n, w_{n-1}, \ldots, w_1)) \\
&= \prod_{i=1}^{n} P(w_i|(w_{i-1}, w_{i-2}, \ldots, w_1))
\end{aligned} \tag{2.3}
$$

Die Wahrscheinlichkeit in Gleichung 2.3 für alle möglichen Wortfolgen zuverlässig zu schätzen, ist aufgrund der Anzahl möglicher Wortfolgen unmöglich. Bei einem Vokabular von 10000 Worten gäbe es beispielsweise schon 10^{12} mögliche Folgen bestehend aus 3 Worten, für die man die Wahrscheinlichkeit bestimmen müßte. Ein Ansatz ist daher, Wortfolgen in Äquivalenzklassen zusammenzufassen.

$$
\begin{aligned}
P((w_j)) &= P((w_n, w_{n-1}, \ldots, w_1)) \\
&\approx \prod_{i=1}^{n} P(w_i|Q((w_{i-1}, w_{i-2}, \ldots, w_1)))
\end{aligned} \tag{2.4}
$$

Hierbei bezeichnet $Q((w_{i-1}, w_{i-2}, \ldots, w_1))$ die Äquivalenzklasse, in die die Wortfolge $(w_{i-1}, w_{i-2}, \ldots, w_1)$ fällt.

Diese Äquivalenzklassen können z.B. basierend auf Grammatiken gebildet werden. Die zur Zeit erfolgreichsten Sprachmodelle sind allerdings alle Erweiterungen einfacher sogenannter Trigramm-Sprachmodelle. Hierbei approximiert man die Wahrscheinlichkeit für eine Wortfolge, indem man für jedes Wort nur die beiden vorausgehenden Worte betrachtet, also

$$
\begin{aligned}
P((w_j)) &= P((w_n, w_{n-1}, \ldots, w_1)) \\
&\approx \prod_{i=1}^{n} P(w_i|(w_{i-1}, w_{i-2}))
\end{aligned} \tag{2.5}
$$

Die $P(w_i|(w_{i-1}, w_{i-2}))$ werden mittels der Häufigkeiten der Trigramme (w_i, w_{i-1}, w_{i-2}) über einer Trainingsmenge bestimmt. Da bei der großen Anzahl möglicher Trigramme die einzelnen Trigramme selbst in sehr großen Trainingstexten selten oder nie gesehen werden, wird $P(w_i|(w_{i-1}, w_{i-2}))$ mit Hilfe der Trigramm-, Bigramm- und Unigramm-Häufigkeiten interpoliert. Hierbei sind Bigramme Wortpaare und Unigramme einzelne Worte.

2.2.5 Suche/Decodierung

Die Suche hat die Aufgabe, aus allen Wortfolgen diejenige auszusuchen, deren A-posteriori-Wahrscheinlichkeit am größten ist. Aufgrund der großen Anzahl zu betrachtender Wortfolgen ist es ein sehr großer Rechenaufwand, diese Wahrscheinlichkeiten für alle denkbaren Wortkombinationen zu berechnen.

Um dennoch das Sprachsignal mit vertretbarem Aufwand in eine Wortfolge decodieren zu können, muß die Suche den Suchraum reduzieren. Suchstrategien im allgemeinen werden in [26] beschrieben. In [48] wird das Problem der Suche im Zusammenhang mit großen Wortlisten (siehe Abschnitt 7.1.1) und Spracherkennung in Echtzeit eingegangen.

Eine typische Maßnahme zur Beschleunigung der Suche, die aber auch Fehler verursachen kann, ist, in einem Suchschritt beim Übergang von Merkmalsvektor i zum Merkmalsvektore $i+1$ nicht alle bis zu diesem Punkt berechneten Pfade durch ein HMM weiterzuverfolgen, sondern nur die n besten. Eine weitere Möglichkeit ist, bei jedem Suchschritt nur die Pfade weiterzuverfolgen, deren Gesamtwahrscheinlichkeit um nicht mehr als Δ vom wahrscheinlichsten Pfad abweicht. Δ und n sind hier Parameter der Suche, die so eingestellt werden sollten, daß die Suche ausreichend schnell ist, aber nicht zu viele Fehler verursacht werden.

2.3 Das *Janus Recognition Toolkit (JRTk)*

JRTk ist eine Entwicklungsumgebung für HMM-basierte Spracherkennungssysteme, die an der Universität Karlsruhe und der Carnegie Mellon University in Pittsburgh (USA) entwickelt wurde. Bei Evaluationen im Jahr 1996 gehörten die Systeme, die mit dem JRTk erstellt wurden, zu den erfolgreichsten[2].

Das JRTk stellt eine Erweiterung von Tcl/Tk dar (siehe [45, 32]). Tcl ist eine Skriptsprache, die es erlaubt, den Befehlssatz durch selbstdefinierte Befehle zu vergrößern. In dieser Form sind die JRTk Befehle in Tcl/Tk integriert.

Mit dem JRTk können sehr flexibel Spracherkennungssysteme, wie sie im letzten Abschnitt beschrieben wurden, aufgebaut werden. Es stehen Befehle zur Verfügung, mit denen Objekte erzeugt werden können. Diese Objekte wiederum verfügen über Parameter, Unterobjekte, lokale Datenstrukturen und Methoden.

Für alle Teile eines auf HMMs basierenden Spracherkenners existieren Objekttypen. Es gibt zum Beispiel Objekttypen zur Modellierung von

- HMM Zuständen,
- Zustandsübergängen eines HMMs,
- HMM Topologien,
- Emissionswahrscheinlichkeitsverteilungen

[2]Mit großem Abstand bestes System bei der VERBMOBIL Evaluation 1996; unter den besten Systemen bei der SWITCHBOARD Evaluation 1996 mit geringem Abstand vom besten System.

und viele mehr. Ein Spracherkennungssystem kann aufgebaut werden, indem Objekte und Unterobjekte erzeugt, die Parameter der Objekte eingestellt und durch die Methoden die Datenstrukturen der Objekte manipuliert werden.

Segmentierer, Klassifizierer und Spracherkenner, die in den folgenden Kapiteln beschrieben werden, sind auf diese Weise mit JRTk erzeugt worden.

Kapitel 3

Die verwendeten Daten

Für die vorliegende Diplomarbeit wurden die Daten verwendet, die für die s.g. *HUB-4* Evaluation 1995 der *Advanced Research Project Agency (ARPA)* bereitgestellt wurden. Im Rahmen dieser Evaluation ging es darum, ein Spracherkennungssystem zu entwickeln, das die Sprachanteile von *Marketplace*-Nachrichtensendungen transkribiert. An der Evaluation haben viele der zur Zeit auf dem Gebiet automatischer Spracherkennung erfolgreichsten Forschungsgruppen teilgenommen (siehe [10, 9, 8, 15, 44, 13]).

3.1 Die *Marketplace*-Nachrichtensendungen

Marketplace ist eine Nachrichtensendung, die von *Public Radio International (PRI)*, früher *American Public Radio (APR)*, produziert wird. Ausgestrahlt werden diese Sendungen in den gesamten USA und von öffentlichen Radiosendern in der ganzen Welt. In den ca. 30 Minuten langen Sendungen werden sehr unterschiedliche Themen angeschnitten mit einem Schwerpunkt auf Wirtschaftsnachrichten (siehe [33]).

Neben reinen Musiksegmenten, verlesenen Nachrichten und Korrespondentenberichten kommen noch Telephoninterviews, Berichte von Ereignissen vor Ort, Abschnitte mit Spontansprache und Abschnitte mit Musik im Hintergrund vor. Außer dem Hauptansager, David Brancaccio, sprechen noch weitere geschulte Sprecher sowie Sprecher mit Dialekten und verschiedenen ausländischen Akzenten.

3.2 Audiodaten und Transkriptionen

Für die Entwicklung der Systeme wurden allen Teilnehmern der Evaluation folgende Datenmengen zur Verfügung gestellt:

Trainingsdaten 10 Marketplace-Sendungen abgetastet mit 16kHz im WAV-Dateiformat.

Entwicklungstestdaten 6 Marketplace-Sendungen im WAV-Format mit zugehörigen Transkriptionen im Sgml-Format. Durch eine Auswahl von Stücken aus diesen 6 Sendungen werden zwei Mengen definiert, die in Art, Dauer und Zusammensetzung in etwa der Evaluationsmenge entsprechen.

Evaluationstestdaten 5 Marketplace-Sendungen sowie die Angabe, welche Ausschnitte der Sendungen im Rahmen der Evaluation für die Auswertung der Systeme verwendet werden sollen. Entsprechend dieser Vorgabe soll eine Nachrichtensendung ganz, von zwei Sendungen nur das erste Drittel ohne die ersten Minuten[1] und von zwei anderen Sendungen nur etwa das letzte Drittel ohne die letzten Minuten verwendet werden.

Auf den Evaluationstestdaten wurden schließlich im Rahmen der Evaluation die Systeme der Teilnehmer ausgewertet. Die Teilnehmer mußten selbst die Trainings- und Entwicklungstestdaten in einer Form aufbereiten, wie sie für die Entwicklung eines Spracherkennungssystems benötigt werden.

Für die vorliegende Diplomarbeit konnten die bereits zum Teil aufgearbeiteten Trainingsdaten zu fünf der insgesamt 10 Radiosendungen in der Trainingsmenge von der *CMU Robust Speech Recognition Group*[2] verwendet werden. Zum Teil aufgearbeitet heißt in diesem Fall, daß die Transkriptionen der Trainingsdaten mit Zeitmarken und einem aus vier Feldern bestehenden Buchstaben-Code versehen waren.

Folgendes Beispiel soll verdeutlichen, in welcher Form die Daten vorlagen.

```
....
4.477 7.062 A F - M   FROM LOS ANGELES THIS IS MARKETPLACE
7.062 18.649 - - - M
18.649 23.487 B M - M AS EXPECTED THE JAPANESE STOCK MARKET
          PLUMMETED MONDAY AS THE GOVERNMENT THERE TEETERS
....
```

Eine Zeile stellt ein Segment dar. Die ersten zwei Zahlen sind Anfangs- und Endzeit des Segments in der Audiodatei. Die folgenden vier Felder haben die Bedeutung:

Sprecher-Identifikation: Ein Buchstabe, der den Sprecher identifiziert.

Geschlecht des Sprechers: M, F, G[3] oder - für männlich, weiblich, nicht erkennbar oder kein Sprecher im Vordergrund.

[1] In diesen ersten Minuten findet in der Regel nur die Begrüßung durch den Ansager statt.

[2] `http://www.cs.cmu.edu/afs/cs.cmu.edu/user/robust/www/home.html`

[3] G wurde für die Markierung der Trainingsdaten nicht verwendet. Offenbar war immer klar, welchen Geschlechts der Sprecher war. Daher wird diese Markierung in den folgenden Betrachtungen nicht weiter berücksichtigt.

Kanaleigenschaften: **T**, **O** oder - als Angabe für die Eigenschaften des Übertragungskanals. Hierbei steht **T** für Telephon, **O** für starke Umgebungseinflüsse wie Echo oder Verzerrung und ähnliches[4] und - für Aufnahmen in Studioqualität.

Art der Hintergrundgeräusche: Eine Buchstabenkombination aus **M**, **S** und **O**, die die vorliegenden Hintergrundgeräusche in dem Segment beschreibt, oder -, falls keine Hintergrundgeräusche vorliegen. **M** steht hierbei für Musik, **S** für leise Sprache weniger Sprecher im Hintergrund, **O** für unverständliche Sprache vieler Sprecher im Hintergrund oder Lärm.

Die letzten Felder in einer Zeile stellen die Transkription des in dem Segment Gesagten ohne Unterscheidung von Klein- und Großschrift, Interpunktion, Sonderzeichen oder ähnliches dar. Bei reinen Musiksegmenten sind diese Felder leer.

Für die Diplomarbeit wurden nur die Evaluationstestmenge sowie die 5 Sendungen der Trainingsmenge verwendet, für die Transkriptionen mit Zeitmarken und Markierungen zur Verfügung standen. Wird in den folgenden Ausführungen die Trainingsmenge erwähnt, so sind nur die 5 Sendungen mit Zeitmarken gemeint, falls keine weiteren Angaben gemacht werden.

[4]Das ist die Bedeutung dieses Buchstaben-Codes, wie sie von der *CMU Robust Speech Recognition Group* angegeben wurde; im Original `strong environmental characteristics such as reverberation, distortion, etc...`

Kapitel 4

Segmentierung und Klassenbildung

Wie schon in der Einleitung erwähnt, gibt es verschiedene Gründe dafür, Audiodaten aus Radio- und Fernsehsendungen zu segmentieren, bevor man versucht, auf diesen Daten Sprache zu erkennen. Einige der wichtigsten Gründe sind:

1. Kleine Stücke sind leichter zu bewältigen. Z.B. sind keine besonderen Maßnahmen für eine geschickte Speicherverwaltung nötig, sondern man kann die gesamten Audiodaten eines Segments und die für die Spracherkennung notwendigen komplexen Datenstrukturen insgesamt im Speicher halten. Bei ganzen Radiosendungen hingegen ist das meistens nicht möglich.

2. Liegt eine Radiosendung zerteilt in Segmente vor, kann Spracherkennung auf den einzelnen Segmenten parallel erfolgen und so die Erkennung insgesamt beschleunigt werden.

3. Radiosendungen enthalten oft Segmente, in denen überhaupt keine Sprache, sondern nur Musik vorkommt. Versucht man mit einem herkömmlichen Spracherkenner auf solchen Segmenten Sprache zu erkennen, wird das Ergebnis in aller Regel eine sinnlose Folge von Worten sein. Ein Herausschneiden solcher Segmente ist also sinnvoll.

4. Wird durch eine korrekte Segmentgrenze z.B. ein Satzbeginn oder -ende angezeigt, so kann diese Information vom Sprachmodell verwendet werden. Ist eine solche Grenze nicht vorhanden, kann das zu einer falschen Bewertung der Wahrscheinlichkeit einer Wortfolge über Satzgrenzen hinweg führen. Korrekte Grenzen beseitigen auch die Möglichkeit, daß Erkennungsfehler an dieser Stelle durch falsches *time alignment* (siehe Abschnitt 2.2.3) gemacht werden.

5. Der wichtigste Grund für eine Segmentierung ist aber, daß die akustischen Modelle (siehe Abschnitt 2.2.3) in aller Regel sehr empfindlich

sind gegenüber veränderten akustischen Bedingungen. Hintergrundgeräusche, andere Übertragungskanäle, Hintergrundmusik und ähnliches bewirken, daß die extrahierten Merkmale nicht gut zu den über der Trainingsmenge berechneten akustischen Modellen passen; die Erkennungsleistung sinkt dramatisch.

Vorverarbeitungsmethoden, die robust sind gegenüber allen denkbaren Störungen, aber die für Spracherkennung wichtigen Informationen erhalten, sind nicht bekannt. Akustische Modelle für alle Sprachlaute in Anwesenheit der verschiedenen Störungen zu trainieren und in einem Spracherkenner zu vereinigen, macht einerseits diesen einen Erkenner sehr komplex, und führt andererseits dazu, daß die Verwechselbarkeit der einzelnen akustischen Modelle zunimmt, was wiederum zu einer geringeren Erkennungsleistung führt.

Ein Zerteilen der Radiosendungen in Segmente, innerhalb derer die akustischen Bedingungen nahezu konstant bleiben, ermöglicht es, die einzelnen Segmente spezialisierten Spracherkennern zuzuführen. Für ein Segment kann der Erkenner ausgewählt werden, der am besten zu den Einflüsse paßt, denen das Sprachsignal innerhalb des Segments unterliegt.

In diesem Kapitel wird der im Rahmen der vorliegenden Diplomarbeit gewählte Ansatz zur Segmentierung erläutert. Dieser Ansatz basiert auf einer Einteilung von akustischen Bedingungen in Klassen und der Zuordnung von Audiodaten zu diesen Klassen. Es wurden Klasseneinteilungen automatisch und von Hand erzeugt, entsprechende Segmentierer entwickelt und auf einer Testmenge ausgewertet.

4.1 Segmentierung durch Klassifikatoren

Wird eine große inhomogene Audioaufnahme in Segmente zerteilt, auf denen Sprache erkannt werden soll, so sind für die erzeugten Segmente folgende Eigenschaften wünschenswert:

1. Keine Segmentgrenzen in einem Wort, da in diesem Fall das Wort nicht mehr richtig erkannt werden kann.

2. Keine Segmentgrenzen, die eine syntaktische Einheit zerteilen, da in diesem Fall durch das Sprachmodell nicht mehr die korrekte Wahrscheinlichkeit für die Wortfolge ermittelt werden kann.

3. Keine Segmente, in denen das Sprachsignal verschiedenen Einflüssen unterliegt, die sich in unterschiedlicher Weise auf die extrahierten Merkmale auswirken, da in diesem Fall die akustischen Modelle eines Spracherkenners nicht mehr genau zu den akustischen Bedingungen im Segment passen können.

4. Erkennen von Segmenten die reine Musik enthalten.

Manche der Forderungen lassen sich bei normalen Radiosendungen oft nicht gleichzeitig erfüllen. Wenn zum Beispiel Musik langsam ein- oder ausgeblendet wird, widersprechen sich die Forderungen 1 und 3. In diesem Zusammenhang sollte die Segmentierungsentscheidung getroffen werden, die den kleinsten Fehler des gesamten Systems, d.h. die kleinste Wortfehlerrate bei der nachfolgenden Spracherkennung, verursacht.

Es gibt zwei prinzipiell verschiedene Ansätze für die Segmentierung. Wenn man schon weiß, welche Klassen von Einflüssen auf das Sprachsignal man unterscheiden will, können Segmente gebildet werden, indem man mit Hilfe von Klassifikatoren eine Zuordnung der Audiodaten zu den verschiedenen Klassen vornimmt und dort Segmentgrenzen setzt, wo sich diese Zuordnung ändert. Ein anderer Ansatz ist, daß man mit Hilfe eines Maßes, das die Änderung der akustischen Bedingungen angibt, dort Segmentgrenzen setzt, wo eine ausreichend deutliche Änderung angezeigt wird. Im Rahmen dieser Diplomarbeit wurde der erste Ansatz gewählt.

4.1.1 Klassenbildung für die Segmentierung

Beim Entwurf des Segmentierers haben folgende Überlegungen und Eigenschaften der Trainingsdaten eine Rolle gespielt:

1. Die Markierungen der Transkriptionen bzgl. Geschlecht des Sprechers, Kanaleigenschaft und Hintergrund ergeben auf der Trainingsmenge 34 verschiedene akustische Bedingungen, die im folgenden **Basisklassen** genannt werden. Die Gesamtdauer der Segmente der einzelnen Basisklassen in den Trainingssendungen ist sehr unterschiedlich. Zwei der Klassen kommen z.B. weniger als 1.5s lang vor, während eine andere Klasse etwa 2841s lang vorkommt, mehr als dreimal so lang wie die Gesamtdauer von Segmenten der nächst häufigsten Basisklasse (siehe Anhang A).

2. Unterschiedliche akustische Bedingungen spiegeln sich in den statistischen Eigenschaften der Merkmalsvektoren, die aus dem Audiosignal extrahiert werden, wieder. Um eine zuverlässige Aussage über die statistischen Eigenschaften der Merkmalsvektoren für eine Klasse von akustischen Bedingungen machen zu können, muß eine ausreichende Anzahl von Merkmalsvektoren dieser Klasse zur Verfügung stehen.

3. Da für einige Basisklassen sehr wenig Trainingsdaten vorliegen, müssen Basisklassen zu neuen Klassen zusammengefaßt werden, damit zuverlässige Aussagen über deren statistische Eigenschaften gemacht werden können.

4. Verschiedene Klasseneinteilungen basierend auf den vorgegebenen Basisklassen scheinen intuitiv sinnvoll: Zum Beispiel ist sicher aufgrund

der bekannten Einflüsse des Telephonkanals (siehe z.B. [25]) eine Klasse für Sprache über Telephonverbindungen sinnvoll. Ebenso scheint eine eigene Klasse für Sprache mit Hintergrundmusik sinnvoll zu sein. Wie ist aber z.B. Vordergrundsprache mit Hintergrundsprache einzuordnen? Ist Vordergrundsprache mit Hintergrundsprache von reiner Vordergrundsprache zu unterscheiden?

Zwei Vorgehensweisen wurden untersucht. Erst wurde eine Klasseneinteilung gewählt, die intuitiv sinnvoll erscheint. Dann wurden, wie später im Detail erklärt, automatisch verschiedene Klasseneinteilungen erzeugt.

4.1.2 Ein Modell für Klassenzugehörigkeit

Als Modell für Klassenzugehörigkeit wurde folgende Wahl getroffen:

- Die einzelnen Merkmalsvektoren einer Klasse werden zufällig erzeugt und sind unabhängig voneinander.
- Als Wahrscheinlichkeitsverteilung für die Erzeugung der Merkmalsvektoren einer Klasse wird eine Mixtur von multivariaten Gaußverteilungen angenommen.
- Um zu verhindern, daß zu kleine Segmente entstehen, wird der Erzeugung der Merkmalsvektoren ein Links-Rechts HMM aus 15 Zuständen zugrunde gelegt (siehe Abschnitt 2.2.3 und Abbildung 4.1).

Zwei vereinfachende (und nicht korrekte) Annahmen, die allerdings im Zusammenhang mit Modellbildung häufig gemacht werden, sind die Unabhängigkeit der einzelnen Merkmalsvektoren sowie die den Hidden Markov Modellen zugrunde liegende Annahme, daß der nächste Zustand und die aktuelle Beobachtung nur vom aktuellen Zustand abhängen. Es hat sich in vielen Bereichen gezeigt, daß trotz dieser vereinfachenden Annahmen das Modell meistens sehr gut mit den wahren Verhältnissen übereinstimmt.

Mixturen multivariater Normalverteilungen

Die Wahrscheinlichkeitsdichtefunktion einer Mixtur von multivariaten Normalverteilungen hat folgende Form:

$$p(\boldsymbol{x}) = \sum_{i=1}^{n} P(M_i) n(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4.1}$$

$$n(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)} \tag{4.2}$$

Hierbei ist $n(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ die Dichtefunktion einer Normalverteilung mit Mittelwertsvektor $\boldsymbol{\mu}_i$ und Kovarianzmatrix $\boldsymbol{\Sigma}_i$. $P(M_i)$ gibt die Wahrscheinlichkeit an, mit der ein Merkmalsvektor von der i-ten Komponente M_i der Mixtur erzeugt wird.

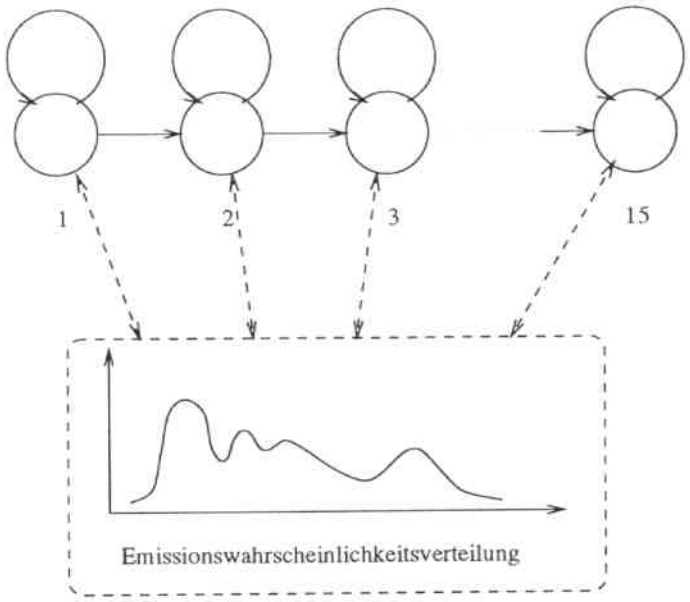

Abbildung 4.1: Für den Segmentierer verwendetes Links-Rechts HMM mit 15 Zuständen. Die einzelnen Zustände verwenden dieselbe Emissionswahrscheinlichkeitsverteilung, da sie ausschließlich dazu dienen, eine Mindestlänge für Segmente zu erzwingen.

Durch Mixturen multivariater Gaußverteilungen läßt sich jede kontinuierliche Wahrscheinlichkeitsdichtefunktion mit beliebiger Genauigkeit approximieren, vorausgesetzt natürlich man verwendet ausreichend viele Komponenten und wählt die Parameter korrekt (siehe [2]).

Die Wahrscheinlichkeitsverteilungen haben die Aufgabe, die in den Segmenten der zugehörigen Klasse vorkommenden Laute zu modellieren. Aufgrund der vielen verschiedenen Laute ist die Anzahl der Mixturkomponenten ausreichend groß zu wählen. Mit der Anzahl der Komponenten wächst allerdings auch die Zahl der zu bestimmenden Paramter und damit der Bedarf an Trainingsdaten. Da die Segmente vor allem Sprachlaute enthalten und diese Laute wiederum Ausprägungen der ungefähr 50 im Englischen vorkommenden Phoneme sind, wurde die Zahl von Komponenten auf 256 gesetzt. Diese Wahl erlaubt eine feine Modellierung der zu erwartenden Laute, ist aber noch mit den zur Verfügung stehenden Daten trainierbar, vorausgesetzt die Zahl der zu unterscheidenden Klassen wird nicht zu groß und die Verteilung der Trainingsdaten auf die Klassen ist nicht zu ungleichmäßig.

Die Topologie des Hidden Markov Modells

Für die Segmentierung werden alle 10ms Merkmalsvektoren aus den Audiodaten über einem Zeitfenster der Größe von 16ms extrahiert. Diese Wahl der Fenstergröße und die Häufigkeit der Berechnung von Merkmalsvektoren orientiert sich an der Geschwindigkeit, mit der sich das Sprachsignal ändert, und der Dauer, während der das Sprachsignal nahezu konstante Signaleigen-

schaften aufweist.

Durch die in Abbildung 4.1 dargestellte HMM Topologie wird verhindert, daß kurze Schwankungen im Audiosignal sehr kleine Segmente erzeugen. Es wird daher für alle Zustände dieselbe Emissionswahrscheinlichkeitsverteilung verwendet. Durch die gewählte Topologie und die Berechnung von Merkmalsvektoren alle 10ms werden Segmente von mindestens 150ms Länge erzeugt. Damit ist das kürzeste in der Trainingsmenge vorkommende Segment noch modellierbar.

4.1.3 Segmentierung unter Verwendung des Modells

Unter Verwendung des eben beschriebenen Modells kann folgendermaßen segmentiert werden. Ein HMM für eine gesamte Nachrichtensendung wird aus HMMs für einzelne Klassen erzeugt. Die Wahrscheinlichkeiten $P(\mathcal{K}_n|(\mathcal{K}_i)_{i=1}^{n-1})$ dafür, daß eine Klasse $\mathcal{K}_n$ in einer Radiosendung nach der Folge von Klassen $(\mathcal{K}_i)_{i=1}^{n-1}$ kommt, entsprechen einem Sprachmodell (siehe Abschnitt 2.2.4) und werden im Segmentierer durch ein einfaches Bigramm-Sprachmodell approximiert. Welche HMM-Sequenz und damit welche Folge von Klassen am besten zu gegebenen Audiodaten paßt, kann durch das Suchobjekt des *JRTk* ermittelt werden (siehe Abschnitt 2.2.5 und 2.3). Durch die Zuordnung der Audiodaten zu HMM-Zuständen erhält man Beginn- und Endzeit der den Klassen zugeordneten Audiosegmente und somit eine Segmentierung entsprechend der Klassenzugehörigkeit.

Um die Parameter der Modelle für die einzelnen Klassen zu bestimmen, können die Trainingsverfahren des *JRTk* verwendet werden, da das gewählte Modell für Klassenzugehörigkeit der akustischen Modellierung im Falle von Spracherkennung sehr ähnlich ist.

4.1.4 Klassifizierung von Segmenten

Um für ein Audiosegment eine Klassenzuordnung zu treffen, kann folgendermaßen vorgegangen werden.

Für jede Klasse $\mathcal{K}$ wird für die Merkmalsvektorfolge $(\boldsymbol{x}_i)_{i=1}^n$ des Segments die Wahrscheinlichkeitsdichte $p((\boldsymbol{x}_i)_{i=1}^n|\lambda_\mathcal{K})$ berechnet. Dieser Wert ist ein Maß dafür, wie wahrscheinlich es ist, daß die Folge $(\boldsymbol{x}_i)_{i=1}^n$ vom akustischen Modell der Klasse $\mathcal{K}$ erzeugt wurde. Mit $\lambda_\mathcal{K}$ wird wieder wie in Abschnitt 2.2.3 die Menge der Parameter des Modells bezeichnet. Diejenige Klasse $\mathcal{K}$, für die das Produkt $p((\boldsymbol{x}_i)_{i=1}^n|\lambda_\mathcal{K})P(\mathcal{K})$ maximal ist, wird dem Segment als Klassenzuordnung zugewiesen. Diese Vorgehensweise minimiert die Wahrscheinlichkeit für eine Fehlklassifikation (siehe [4]).

$P(\mathcal{K})$ ist hierbei die A-priori-Wahrscheinlichkeit der Klasse $\mathcal{K}$. Da die Wahrscheinlichkeit dafür, daß eine Klasse von akustischen Bedingungen an einer bestimmten Stelle in einer Radiosendung auftaucht, vom akustischen Kontext abhängt[1], müßte diese Wahrscheinlichkeit unter Berücksichtigung

[1] Es gibt z.B. kein Vorkommen von Sprache über Telephon direkt nach reiner Musik

der vorausgehenden Segmente berechnet werden. Hier wird jedoch diese Wahrscheinlichkeit nur durch die Häufigkeit der Klasse $\mathcal{K}$ in der Trainingsmenge abgeschätzt. Stimmt die so gefundene Klassenzuordnung mit der von Menschen vorgenommenen überein, so zählt das als korrekte Klassifikation.

4.2 Von Hand gewählte Klasseneinteilung

Für den gewählten Segmentierungsansatz müssen Parameter für Mixturen multivariater Gaußverteilungen, die Klassen von akustischen Bedingungen modellieren, bestimmt werden. Für welche Klassen, und damit für welche Kombinationen von Basisklassen, eigene Modelle erstellt werden, wird durch eine Klasseneinteilung festgelegt.

Eine erste Einteilung der Basisklassen in Klassen, die in diesem Zusammenhang unterschieden werden sollen, wurde von Hand vorgenommen. Die Einteilung wurde aufgrund von Überlegungen, welche akustischen Bedingungen die Merkmalsvektoren in unterschiedlicher Weise beeinflussen, getroffen. Außerdem wurde in Betracht gezogen, wieviele Trainingsbeispiele für die einzelnen Klassen in der Trainingsmenge vorkommen.

Die Klasseneinteilung wurde aus den Basisklassen, wie sie in Kapitel 3 beschrieben werden, erzeugt. Hierbei wurden nur die drei Felder für das Geschlecht des Sprechers, den Übertragungskanal und den Hintergrund berücksichtigt. Das Feld mit der Sprecheridentifikation wurde unberücksichtigt gelassen, weil außer dem Hauptansager keiner der Sprecher so oft vorkam, daß eine Unterscheidung für möglich und sinnvoll gehalten wurde.

4.2.1 Darstellung von Klasseneinteilungen

In den folgenden Erörterungen wird eine abkürzende Schreibweise verwendet, um anzugeben, welche Basisklassen zu neuen Klassen zusammengefaßt wurden. Es werden drei Felder getrennt durch einen Bindestrich dargestellt. Die Felder entsprechen den im Abschnitt 3.2 beschriebenen. Die hier verwendeten Buchstabencodes unterscheiden sich jedoch von den dort verwendeten, weil hier Mengen von Basisklassen dargestellt, Hintergrundsprache und Lärm nicht unterschieden und Abkürzungen für entsprechende deutsche Worte[2] verwendet werden.

Sprecher: -, **M**, **F**, **MF** und $*$ für *kein Sprecher*, *ein männlicher Sprecher*, *ein weiblicher Sprecher*, *ein Sprecher beliebigen Geschlechts* und *keine Einschränkung*.

Kanaleigenschaften: **T**, **K**, **V** oder $*$ für *Telephonkanal*, *klarer Kanal*, *verzerrter Kanal* oder *keine Einschränkung*.

über einen klaren Kanal in der Trainingsmenge.

[2] Mann, Frau, Telephonkanal, verzerrter Kanal, klarer Kanal, Musik, Lärm

Hintergrundgeräusche: -, **M**, **L**, **M(L)**, **ML** oder $*$ für *keine Hintergrundgeräusche*, *reine Hintergrundmusik*, *Hintergrundgeräusche ohne Musik*, *Hintergrundgeräusche bestehend aus Musik und (nicht unbedingt) anderen Hintergrundgeräuschen*, *Hintergrundgeräusche bestehend aus Musik und (unbedingt) anderen Geräuschen* oder *keine Einschränkung*.

Dieser Konvention entsprechend bezeichnet zum Beispiel *MF-*-–* die Menge aller Basisklassen, in denen jemand spricht, keine Hintergrundgeräusche vorhanden sind und ein beliebiger Kanal vorliegt. Die von Hand vorgenommene Klasseneinteilung ist gemäß dieser Konvention in folgender Tabelle aufgeführt.

Klasseneinteilung 1: 9 von Hand gewählte Klassen	
Neue Klasse	Menge von Basisklassen
H1	**-K-–*
H2	**-K-L*
H3	*MF-K-M(L)*
H4	*–-K-M(L)*
H5	**-V-–*
H6	**-V-L*
H7	**-V-M(L)*
H8	**-T-–*
H9	**-T-L* **-T-M(L)*

Eine andere Form der Darstellung einer Klasseneinteilung wurde in Abbildung 4.2 gewählt. Der Wurzelknoten des Baums repräsentiert hier die Menge aller Basisklassen, jeder andere Knoten bezeichnet eine Teilmenge davon. Welche Basisklassen diese Teilmenge umfaßt, kann ermittelt werden, indem man dem Pfad vom Wurzelknoten zum aktuellen Knoten folgt. Jede Verbindung von zwei Knoten im Baum stellt eine Auswahl aus der Menge der Basisklassen, die der übergeordnete Knoten enthält, dar.

Diese Auswahl kann durch eine Frage oder Mengenbeschreibung dargestellt werden. Z.B. kann aus der Menge aller Basisklassen die Menge derjenigen Klassen, in denen ein klarer Übertragungskanal vorliegt, durch die Frage „*Kanal = klar ?*“ ausgewählt werden. Eine äquivalente Darstellung durch eine Mengenbeschreibung entsprechend der oben angegebenen Konvention ist **-K-**. In der Abbildung sind der Anschaulichkeit halber Fragen an den Ästen des Baums angegeben. Nachfolgend wird nur noch die kürzere Darstellung durch Mengenbeschreibungen verwendet.

Die in Abbildung 4.2 dargestellte Klasseneinteilung entspricht der in der Tabelle angegebenen. Der Teilbaum, dessen Blätter die Basisklassen mit Telephonkanal enthalten, wurde nicht weiter unterteilt in Klassen, in denen Hintergrundmusik vorliegt oder nicht, da in der Trainingsmenge kein Beispiel für Musik über Telephonleitungen vorkommt. Die Basisklassen, in

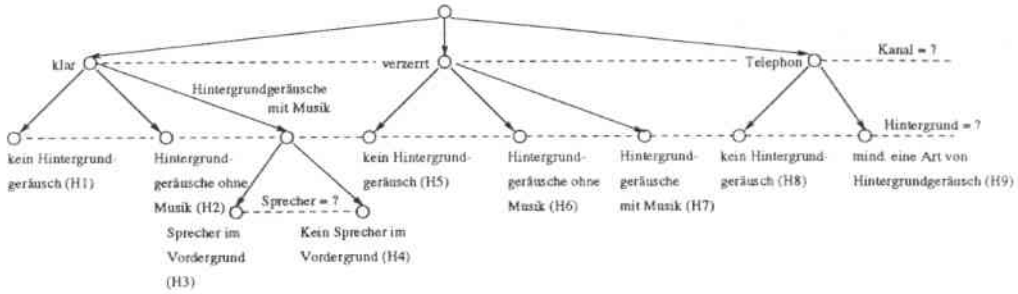

Abbildung 4.2: Darstellung einer Einteilung der Basisklassen in übergeordnete Klassen in Form eines Baums

denen ein klarer Kanal und Musik vorliegt, wurden weiter bzgl. Sprecher/kein Sprecher unterschieden, da viele reine Musiksegmente mit einer Gesamtdauer von ca. 9 Minuten in der Trainingsmenge vorkommen und es, wie weiter oben begründet, für die Spracherkennung wichtig ist, reine Musik von Sprache zu trennen.

4.2.2 Gründe für die gewählte Klasseneinteilung

Die obige Klasseneinteilung wurde in der Annahme vorgenommen, daß sich die Effekte verschiedener Übertragungskanäle und Hintergrundgeräusche in unterschiedlicher Weise auf das Sprachsignal und die extrahierten Merkmale auswirken und somit die Klassen gut voneinander unterschieden werden können.
Diese Annahmen wurden durch folgende Überlegungen motiviert:

- Eine gute Unterscheidbarkeit des Telephonkanals von den anderen Kanälen wurde erwartet, da bei Übertragung über einen Telephonkanal neben anderen Effekten nur Frequenzen im Bereich von ungefähr 300Hz bis 3400Hz übertragen werden, was am Kurzzeitspektrum erkennbar ist (siehe z.B. [25, 41]).

- Eine Trennbarkeit des verzerrten und des klaren Kanals wurde angenommen, da der Einfluß verschiedener Übertragungskanäle bekanntermaßen im Spektrum sichtbar ist (siehe z.B. [35, Abschnitt 5.7] und [22, 46]).

- Die Trennbarkeit von Musik und Sprache ist schon ausführlich untersucht worden. In [40] wird z.B. ein Verfahren beschrieben, das sogar ohne Verwendung einer Analyse des Kurzzeitspektrums, nur basierend auf Nulldurchgangsrate und Signalenergie, eine Trennung von Sprache und Musik mit einer Genauigkeit von 98% erreicht.

- Da sich die Struktur von Musik im Zeit- und Frequenzbereich sehr von der Struktur von Sprache und zufälligen Geräuschen unterscheidet (in Bezug auf Bandbreite, Energieverteilung im Spektrum, Tonalität; siehe

[11, 40, 41]) wurde erhofft, daß sich diese Unterschiede auch für Hintergrundgeräusche und Hintergrundsprache gegenüber Hintergrundmusik im Kurzzeitspektrum, in Signalenergie und Nulldurchgangsrate bemerkbar machen.

- Die sehr charakteristischen Eigenschaften von Sprache (Wechsel von stimmhaften und nichtstimmhaften Lauten mit den einhergehenden Nulldurchgangsraten- und Signalenergieänderungen, sowie die typische Energieverteilung im Spektrum [40]) sollten eine Trennung von Segmenten, die Sprache enthalten, und solchen ohne Sprache ermöglichen.

- Verunreinigte Sprache und nicht verunreinigte Sprache haben z.B. verschiedene durchschnittliche Energiespektren (siehe [41]). Weitere Effekte von Verunreinigungen, die anhand des Kurzzeitspektrums, der Nulldurchgangsrate und der Signalenergie nachweisbar sein müßten, sind auch in [35, Abschnitt 5.7] beschrieben.

4.3 Automatische Erzeugung von Klasseneinteilungen

Bei einer mehr oder weniger intuitiv gewählten Klasseneinteilung besteht die Gefahr, daß Zusammenhänge übersehen oder falsch eingeschätzt werden. Das kann zu nicht optimalen Lösungen führen. Um die im letzten Abschnitt beschriebene Klasseneinteilung in dieser Hinsicht zu bewerten, wurden die Basisklassen zusätzlich automatisch zu neuen Klassen zusammengefaßt und die Güte der automatisch gefundenen Klasseneinteilungen mit der von Hand gewählten verglichen.

4.3.1 Divisive hierarchische Klassenbildung

Die Klasseneinteilungen sollen verwendet werden, um Klassifikatoren zu trainieren, und mit Hilfe der Klassifikatoren soll segmentiert werden. Wenn man die im Sinne der Segmentierung beste Klasseneinteilung finden will, muß man die Güte von Segmentierungen bewerten können, d.h. man braucht eine Gütefunktion. Im gegebenen Fall kann eine Gütefunktion für eine Segmentierung über dem Anteil der richtig bestimmten und der Anzahl der falsch ermittelten Segmentgrenzen sowie über dem Anteil der korrekt klassifizierten Merkmalsvektoren und der mittleren Segmentlänge definiert werden (siehe Abschnitt 5.2.1). Die Maximierung einer solchen Gütefunktion über einer Testmenge erfordert einen sehr hohen Rechenaufwand. Es müßten für alle möglichen Klasseneinteilungen Klassifikatoren trainiert, eine Segmentierung durchgeführt und ausgewertet werden.

Um den unvertretbar hohen Aufwand einer solchen Maximierung zu umgehen, wurde hier, wie das in solchen Fällen oft gemacht wird [4], ein heuristischer Ansatz gewählt. Beim auf Klassifizierung basierenden Ansatz zur

Segmentierung werden Segmentgrenzen dann gut gefunden werden, wenn die Klassifikatoren die Merkmalsvektoren zuverlässig den jeweiligen Klassen zuordnen können. Das wird dann der Fall sein, wenn die Basisklassen so zu Klassen zusammengefaßt werden, daß die Merkmalsvektoren in einer Klasse sehr *ähnlich* sind und die Merkmalsvektoren der Klassen untereinander sehr *verschieden* sind. *Ähnlich* und *verschieden* bezieht sich hier und im folgenden auf die statistischen Eigenschaften der Mengen von Merkmalsvektoren.

Ein Problem unüberwachten Lernens

Mengen von Basisklassen mit ähnlichen Merkmalsvektoren zu bilden und so neue Klassen zu erzeugen, die sich stark unterscheiden, stellt eine Form des unüberwachten Lernens dar (für eine allgemeine Erörterung siehe [4]). Um Ähnlichkeit bewerten zu können, wurde ein Abstandsmaß definiert.

Unter Verwendung dieses Maßes wurde durch einen iterativen Prozeß die Menge aller Basisklassen schrittweise zerteilt. In jedem Schritt wurde die Aufspaltung vorgenommen, durch die das Abstandsmaß maximiert wurde. Diese Vorgehensweise ist motiviert durch die Überlegung, daß, wenn man in jedem Schritt die bezüglich des Abstandsmaßes unterschiedlichsten akustischen Bedingungen voneinander trennt, schließlich eine Klassenaufteilung entsteht, bei der die einzelnen Klassen sich sehr unterscheiden. Die Maximierung einer Gütefunktion kann jedoch nicht garantiert werden.

Der Algorithmus

Für die automatische Klassenbildung wurde ein divisiver hierarchischer Ansatz (siehe [4]) verwendet, wie er im folgenden beschrieben ist.

1. Erst werden, dem im Abschnitt 4.1.2 beschriebenen Modell entsprechend, Mittelwertvektoren $\boldsymbol{\mu}_i$ und Kovarianzmatrizen $\boldsymbol{\Sigma}_i$ für die Komponenten M_i einer Mixtur von Normalverteilungen über einer Trainingsmenge bestimmt (ohne Unterscheidung der Basisklassen $\mathcal{B}_j$).

2. Dann werden die Mixturgewichte $P_{\mathcal{B}_j}(M_i)$ für die Basisklassen bezüglich der im vorigen Schritt berechneten Mixturkomponenten ermittelt sowie die Anzahl der Trainingsbeispiele $N_{\mathcal{B}_j}$, die auf die einzelnen Basisklassen entfallen, bestimmt.

3. Aus den Mixturgewichten $P_{\mathcal{B}_j}(M_i)$ und der Anzahl der Trainingsbeispiele $N_{\mathcal{B}_j}$ können aus den Mixturen $p_{\mathcal{B}_j}(\boldsymbol{x})$ für die einzelnen $\mathcal{B}_j$ die Mixturen $p_{\mathcal{B}'}(\boldsymbol{x})$ für Klassen $\mathcal{B}' := \{\mathcal{B}_{j_1}\mathcal{B}_{j_2}\ldots\mathcal{B}_{j_k}\}$ bestehend aus mehreren Basisklassen ermittelt werden.

4. Es wird wie in Punkt 3 für die Menge bestehend aus allen Basisklassen die Mixtur $p_{\text{Gesamt}}(\boldsymbol{x})$ berechnet. Dann wird die Menge aller Basisklassen durch eine Frage, wie in Abschnitt 4.2.1 beschrieben, in zwei Teilmengen $\mathcal{B}_{\text{Ja}}$ und $\mathcal{B}_{\text{Nein}}$ geteilt und die Mixturen $p_{\mathcal{B}_{\text{Ja}}}(\boldsymbol{x})$ und $p_{\mathcal{B}_{\text{Nein}}}(\boldsymbol{x})$ bestimmt.

5. Es kann nun bewertet werden, wieviel mehr Information in einem System, das die Trainingsmenge durch $p_{\mathcal{B}_{\text{Ja}}}(\boldsymbol{x})$ und $p_{\mathcal{B}_{\text{Nein}}}(\boldsymbol{x})$ statt nur durch $p_{\text{Gesamt}}(\boldsymbol{x})$ modelliert, enthalten ist. Über diesen Zugewinn an Information ist das Abstandsmaß $D(p_{\mathcal{B}}, p_{\mathcal{B}_{\text{Ja}}}, p_{\mathcal{B}_{\text{Nein}}})$, das weiter unten beschrieben wird, definiert.

6. Die Schritte 3 bis 5 werden für jede Frage q aus einer Menge von Fragen Q durchgeführt. Die Frage, die das Abstandsmaß maximiert, wird verwendet, um die Menge von Basisklassen aufzuteilen.

7. Nach dem Aufspalten der Gesamtmenge von Basisklassen hat man zwei Teilmengen. Für beide Teilmengen kann wieder für alle Fragen q aus der Menge Q der Informationsgewinn durch eine Aufspaltung ermittelt werden. Die Frage, die den größten Informationsgewinn bewirkt, wird gewählt und die entsprechende Aufspaltung vorgenommen. In dieser Weise kann man weiter verfahren. Ein Abbruchkriterium für diesen iterativen Vorgang kann eine gewünschte Anzahl von Teilmengen sein oder die Tatsache, daß der Informationsgewinn unter eine Schwelle sinkt.

Die so erzeugte Klasseneinteilung kann man, wie in Abschnitt 4.2.1 beschrieben, durch einen Baum darstellen. Da die Teilmengen bei dem angegebenen Algorithmus immer in zwei Mengen aufgespaltet werden, entsteht ein Binärbaum.

In Schritt 5 des Algorithmus' betrachtet man nur solche Aufteilungen, bei denen für beide neu gebildeten Modelle ausreichend viele Trainingsbeispiele existieren, d.h. man prüft $N_{\mathcal{B}_{\text{Ja}}} \geq N_{\text{minimal}}$ und $N_{\mathcal{B}_{\text{Nein}}} \geq N_{\text{minimal}}$.

Die Menge von Fragen Q hat die Funktion, die Menge aller möglichen Klasseneinteilungen, die pro Schritt geprüft werden müssen, einzuschränken. Wenn man man sich überlegt, daß es $2^{33} - 1$ Möglichkeiten gibt, die 34 Basisklassen in zwei Mengen aufzuteilen[3], wird klar, daß der Rechenaufwand für das Prüfen all dieser Klasseneinteilungen sehr groß ist. Weiterhin sind manche Klasseneinteilungen auch überhaupt nicht sinnvoll; eine Klasse bestehend aus M-$K-$ und $-K$-M wird sicher nie gebildet werden, da anhand des Spektrums sofort mit bloßem Auge erkennbar ist, daß sich diese Basisklassen sehr unterscheiden.

Der in diesem Abschnitt beschriebene Mechanismus zur Klassenbildung wird häufig im Zusammenhang mit dem *Clustering* von akustischen Modellen für Phoneme im Kontext mehrerer anderer Phoneme in ASES verwendet. Hier werden z.B. Fragen gestellt wie „*Ist das vorhergehende Phonem ein Vokal?*". Eine Beschreibung des Algorithmus' in diesem Zusammenhang ist in [49, 29, 47] zu finden.

[3]Das kann durch vollständige Induktion gezeigt werden.

Das Abstandsmaß

Beim *Clustering*-Prozeß werden im ersten Schritt Mittelwertvektoren und Kovarianzmatrizen über den Trainingsdaten aller Basisklassen berechnet. Für jede Mixtur, die im Verlaufe des Algorithmus' betrachtet wird, werden diese Mixturkomponenten verwendet und nur die Mixturgewichte neu bestimmt. Die Modelle $p_{\mathcal{B}}(\boldsymbol{x})$ stellen daher gewissermaßen diskrete Verteilungen über den Mixturkomponenten dar. Die Komponenten nehmen die Rolle eines durch Vektor-Quantisierung erzeugten Codebuchs ein.

Die Zunahme an Entropie, wenn das Modell $p_{\mathcal{B}}$ durch eine Frage in $p_{\mathcal{B}_{\text{Ja}}}$ und $p_{\mathcal{B}_{\text{Nein}}}$ aufgespaltet wird, ist gegeben durch

$$
\begin{aligned}
D(p_{\mathcal{B}}, p_{\mathcal{B}_{\text{Ja}}}, p_{\mathcal{B}_{\text{Nein}}}) &= N_{\mathcal{B}_{\text{Ja}}} H(p_{\mathcal{B}_{\text{Ja}}}) + N_{\mathcal{B}_{\text{Nein}}} H(p_{\mathcal{B}_{\text{Nein}}}) - N_{\mathcal{B}} H(p_{\mathcal{B}}) \quad (4.3)\\
\text{mit} \qquad H(p_{\mathcal{B}}) &= \sum_{i=1}^{M} P_{\mathcal{B}}(M_i) \log P_{\mathcal{B}}(M_i)\\
H(p_{\mathcal{B}_{\text{Ja}}}) &= \sum_{i=1}^{M} P_{\mathcal{B}_{\text{Ja}}}(M_i) \log P_{\mathcal{B}_{\text{Ja}}}(M_i)\\
H(p_{\mathcal{B}_{\text{Nein}}}) &= \sum_{i=1}^{M} P_{\mathcal{B}_{\text{Nein}}}(M_i) \log P_{\mathcal{B}_{\text{Nein}}}(M_i)
\end{aligned}
$$

M bezeichnet hier die Anzahl von Mixturkomponenten, $N_{\mathcal{B}}$, $N_{\mathcal{B}_{\text{Nein}}}$ und $N_{\mathcal{B}_{\text{Ja}}}$ die Anzahl von Trainingsbeispielen, die auf die jeweiligen Mengen entfallen, und $H(.)$ die Entropie. Die Zunahme an Entropie kann als Zugewinn an Information interpretiert werden (siehe [2]).

Bei dieser Sichtweise ist die Motivation für die Wahl von $D(.,.,.)$ als Abstandsmaß in Gleichung 4.3 folgende. Die Mixturen $p_{\mathcal{B}}(\boldsymbol{x})$ für die Teilmengen $\mathcal{B}$ repräsentieren die statistischen Eigenschaften der Merkmalsvektoren dieser Mengen. Wenn durch Aufspalten einer Teilmenge ein großer Informationsgewinn erzielt wird, ist das gleichbedeutend damit, daß sich die neuen Verteilungen stark unterscheiden, was damit wiederum bedeutet, daß die statistischen Eigenschaften der Merkmalsvektoren der neu gebildeten Teilmengen sehr unterschiedlich sind. Wählt man also in jedem Schritt diejenige Aufteilung einer Menge von Basisklassen, die den größten Informationsgewinn bewirkt, so bildet man auf Basis der Fragen Q die Teilmengen, deren statistische Eigenschaften sich besonders stark unterscheiden. Das ist genau die zu Beginn des Abschnitts formulierte Zielsetzung der Klassenbildung.

4.3.2 Die erzeugten Klasseneinteilungen

Die Menge von Fragen Q ist bei dem in Abschnitt 4.3.1 beschriebenen Algorithmus natürlich von großer Bedeutung. Enthält Q die Frage, die die beste[4]

[4] *beste* im Sinne des gewählten Abstandsmaßes

Aufteilung einer Menge bewirken würde, nicht, so kann der Algorithmus diese Klasseneinteilung auch nicht finden.

Für die Entwicklung des Segmentierers wurde die Trainingsmenge geteilt. Vier der fünf Nachrichtensendungen aus der Trainingsmenge wurden zum Trainieren der akustischen Modelle verwendet, und auf der verbleibenden Nachrichtensendung wurde der Segmentierer ausgewertet. Für die Segmentierung der Sendungen der Evaluationstestmenge wurden die akustischen Modelle über allen fünf Sendungen neu bestimmt.

Für die vier Sendungen und auch für die gesamte Trainingsmenge wurden Klasseneinteilungen automatisch erzeugt, um den Einfluß der Größe der verwendeten Trainingsmenge auf den Klassenbildungsprozeß zu untersuchen.

Die verwendeten Mengen von Fragen

Die beiden für die automatische Klassenbildung verwendeten Fragenmengen sind im Anhang B aufgelistet. Diese Mengen entsprechen sich bis auf die Tatsache, daß in dem einen Fall eine Unterscheidung des Geschlechts des Sprechers erlaubt wird und im anderen Fall nicht.

Für die Segmentierung ist eine Unterscheidung des Sprechergeschlechts sinnvoll, da ein Sprecherwechsel eine sinnvolle Segmentgrenze darstellt (in der Regel fällt ein Sprecherwechsel mit einer syntaktischen Grenze zusammen; siehe hierzu auch die für eine Segmentierung geforderten Eigenschaften in Abschnitt 4.1). Die zweite Menge von Fragen wurde für das Training von spezialisierten Spracherkennern verwendet.

Die automatisch gefundene Klasseneinteilungen über 4 Trainingssendungen

In Abbildung 4.3 ist der Binärbaum dargestellt, den der Algorithmus zur Klassenbildung mit Fragen der Menge 1 aufgebaut hat. Es sind an den Knoten jeweils die Fragen angegeben, die zur Aufteilung verwendet wurden. Weiterhin ist angegeben, in welcher Reichenfolge die Fragen gestellt wurden.

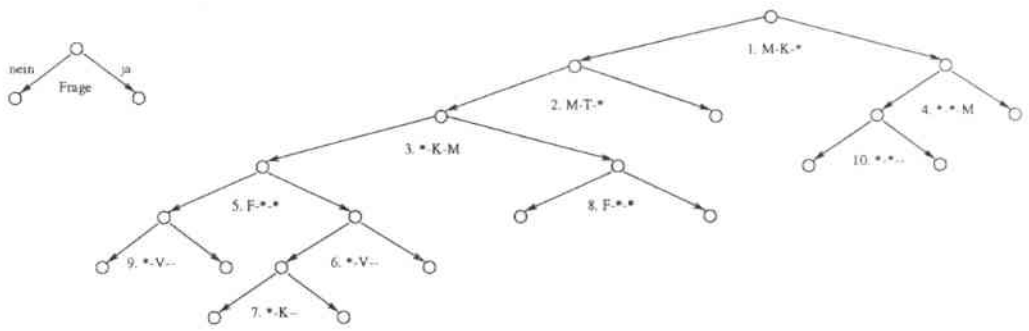

Abbildung 4.3: Darstellung der automatisch gefundenen Klasseneinteilung bei Unterscheidung des Sprechergeschlechts. Die akustischen Modelle wurden über 4 der 5 Trainingssendungen berechnet. An jedem Knoten ist die Frage angegeben, die die Basisklassen des Knotens aufgespaltet hat.

In der Abbildung und in den folgenden Erörterungen wird wieder die in Abschnitt 4.2.1 beschriebene abkürzende Schreibweise für Fragen, bzw. die äquivalenten Mengenbeschreibungen verwendet.

Es sind in obiger Abbildung nur 10 Aufspaltungen von Knoten dargestellt, obwohl weitere Aufspaltungen berechnet wurden. Ab der 11. Frage hängt die Gestalt des Baums davon ab, ob man eine Trennung von Teilmengen zuläßt, für die weniger als 2000 Trainingsbeispiele vorliegen. 2000 Merkmalsvektoren entsprechen 20s an Trainingsdaten. Für die akustischen Modelle wurden Mixturen mit 256 Komponenten verwendet. Trainiert man eine solche Mixtur mit weniger als 2000 Merkmalsvektoren, so werden im Durchschnitt noch nicht einmal 8 Vektoren zum Bestimmen der Parameter einer Komponente verwendet. Weniger als 8 Vektoren sind jedoch nicht genug, um Mixturgewicht, Kovarianzmatrix und Mittelwertvektor zuverlässig zu schätzen. Der oben abgebildete Teilbaum stellt somit nur die Klassenaufteilungen dar, die unabhänig davon vorgenommen werden, ob man beim in Abschnitt 4.3.1 beschriebenen Prozeß $N_{\text{minimal}} \geq 2000$ setzt oder nicht.

*M-K-** als Frage zum Aufspalten des ersten Knotens bedeutet, daß die Menge aller Basisklassen aufgeteilt wurde in die Basisklassen, in denen ein männlicher Sprecher über einen klaren Kanal bei beliebigen Hintergrundgeräuschen spricht, und alle übrigen. Daß diese Frage als erste gewählt wurde, bedeutet, daß durch diese Trennung ein maximaler Informationsgewinn erzielt wurde. Als zweite Frage wurde *M-T-** für all die Segmente gestellt, für die die erste Frage mit *Nein* beantwortet wurde. Diese Frage bezieht sich auf männliche Sprecher über einen Telephonkanal. Die dritte Frage **-K-M* teilt wiederum die Menge von Basisklassen auf, für die die ersten zwei Fragen mit *Nein* beantwortet wurden.
Zur Interpretation dieser automatisch gefundenen Lösung:

- Es fällt auf, daß der Kanal für männliche Sprecher offenbar von größerer Bedeutung ist als bei weiblichen, da eine Trennung des Kanals bei männlichen Sprechern in den ersten beiden Fragen erfolgt, bei weiblichen hingegen erst in Frage 6 und 7.

- Die Unterschiede der Sprache von männlichen und weiblichen Sprechern scheint sich deutlich in den statistischen Eigenschaften der Merkmalsvektoren bemerkbar zu machen, da eine recht gute Trennung in den ersten Fragen erfolgt (abhängig vom Kanal allerdings).

- Abwesenheit von Hintergrundgeräuschen (Fragen 6, 7, 9 und 10) und Anwesenheit von reiner Hintergrundmusik (Fragen 3 und 4) haben einen großen Einfluß auf die Merkmalsvektoren.

- Die Basisklassen, die übrig bleiben, wenn man den *Nein*-Pfaden im Baum folgt, weisen meistens Mischungen mehrerer Hintergrundgeräusche oder einen verzerrten Kanal auf.

Die automatisch gefundene Klasseneinteilung über 5 Trainingssendungen

Abbildung 4.4 zeigt die Klasseneinteilung, die über allen 5 Nachrichtensendungen erzeugt wurde. Man sieht deutliche Unterschiede zur Abbildung 4.3.

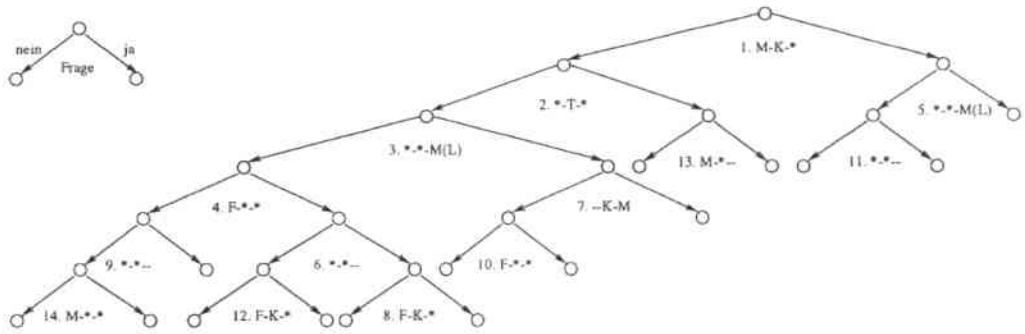

Abbildung 4.4: Darstellung der automatischen gefundenen Klasseneinteilung bei Unterscheidung des Sprechergeschlechts. Die akustischen Modelle wurden unter Verwendung aller 5 Trainingssendungen berechnet. An jedem Knoten ist die Frage angegeben, die die Basisklassen des Knotens aufgespaltet hat.

Es fallen an diesem Baum folgende Dinge auf:

- Die Einflüsse des Telephonkanals scheinen sehr stark zu sein (2. Frage).
- Basisklassen mit Hintergrundgeräuschen, die Musik enthalten, werden sehr bald abgespaltet (Fragen 3 und 5).
- Während ein klarer Kanal für männliche Sprecher von großem Einfluß ist (1. Frage), wird eine Trennung bzgl. des Kanals bei weiblichen Sprechern erst an 8. und 12. Stelle durchgeführt.
- Die statistische Eigenschaften der Merkmalsvektoren werden durch das Geschlecht des Sprechers offenbar deutlich beeinflußt (Frage 4 und indirekt Frage 1). Eine Trennung erfolgt recht bald und recht vollständig.

Vergleich der automatischen Klasseneinteilungen

Bei all den Unterschieden zwischen den Klasseneinteilungen über 4 und 5 Trainingssendungen fallen auch viele Gemeinsamkeiten auf. Teilbäume verschieben sich und Fragen verändern sich geringfügig, doch nicht grundlegend. Folgende Entwicklungen werden deutlich:

- Die Unterscheidbarkeit bzgl. des verzerrten und klaren Kanals wird schlechter (Fragen 6 und 9 in beiden Abbildungen).
- Im rechten Teilbaum der beiden Abbildungen (männlicher Sprecher und klarer Kanal) verändert sich nicht viel; das kann damit zusammenhängen, daß aufgrund der großen Anzahl von Trainingsbeispielen die akustischen Modelle schon recht gut trainiert sind.

- Der Einfluß von Hintergrundgeräuschen, die Musik enthalten, gewinnt an Bedeutung (Frage 3 in Abbildung 4.4); die Trennbarkeit von reiner Hintergrundmusik und Hintergrundgeräuschen, die Musik enthalten nimmt jedoch ab (Fragen 3 und 4 in Abbildung 4.3 gegenüber Fragen 3 und 5 in Abbildung 4.4).

Insgesamt deutet die Instabilität der Aufteilung darauf hin, daß die Menge an Trainingsdaten noch nicht ausreicht, um zuverlässige Aussagen darüber zu machen, welche akustischen Bedingungen sich ähnlich sind. Ein weiterer Grund vor allem für die sich verändernde Bedeutung des verzerrten Kanals kann die Art und Weise sein, wie die Markierungen und Zeitmarken auf der Trainingsmenge vergeben wurden (siehe Kapitel 3 und 5.1.3).

Kapitel 5

Segmentierungsversuche

Mit den von Hand und automatisch erzeugten Klasseneinteilungen, die im letzten Kapitel beschrieben wurden, wurden folgende Versuche durchgeführt.

1. Klassifizieren von vorgegebenen Segmenten.
2. Segmentieren einer gesamten Radiosendung.

Die Versuche zu Punkt 1 wurden gemacht, da der gewählte Segmentierungsansatz auf der Klassifizierung einzelner Merkmalsvektoren basiert. Die Klassifizierungsfähigkeit der Klassifikatoren kann bewertet werden, indem man über einer Testmenge prüft, ob die automatisch gefundene Klassenzuordnung mit der von Menschen vorgenommenen übereinstimmt.

Letztendlich soll mit den Klassifikatoren segmentiert werden. Die Versuche zu Punkt 2 dienten dazu, die Güte der Klassifikatoren in dieser Hinsicht auszuwerten.

5.1 Klassifizierung vorgegebener Segmente

Ziel der Versuche auf vorgegebenen Segmenten war, eine geeignete Vorverarbeitung auszusuchen und ein Gefühl für die Unterscheidbarkeit der für die Segmentierung gewählten Klassen zu bekommen. Außerdem konnten erste Vergleiche der automatisch gefundenen mit der intuitiv gewählten Klasseneinteilung vorgenommen werden.

5.1.1 Vorverarbeitung und Training der Klassifikatoren

Wie schon in Abschnitt 4.1.1 erläutert, basiert der gewählte Segmentierungsansatz darauf, daß Einflüsse, die sich in unterschiedlicher Weise auf die statistischen Eigenschaften der für die Spracherkennung extrahierten Merkmale auswirken, durch Klassifikatoren erkannt werden.

Eine Wahl der Vorverarbeitung wäre bei dieser Zielsetzung, genau die Merkmale aus dem Audiosignal zu extrahieren, die auch für die Spracherkennung verwendet werden. Bei der Spracherkennung versucht man jedoch, den Einfluß verschiener Übertragungskanäle und Störungen des Sprachsignals durch geeignete Maßnahmen zu reduzieren.

Für die Segmentierungsentscheidung bzgl. des verzerrten und klaren Kanals ist es aber z.B. wünschenswert, die Einflüsse der verschiedenen Kanäle in den Merkmalsvektoren zu erhalten, um so die Unterscheidbarkeit zu erleichtern. Ein Vorverarbeitungsschritt, der in den später verwendeten Spracherkennungssystemen verwendet wird, ist die Subtraktion des Mittelwertvektors $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{mel}_i$ der *Melscale-Spektralkoeffizienten* $\boldsymbol{mel}_i$ (siehe weiter unten) eines Segments mit n Merkmalsvektoren von den einzelnen $\boldsymbol{mel}_i$, um additive Störungen im logarithmierten Frequenzbereich zu beseitigen, die von verschiedenen Kanälen verursacht werden. Diesen Schritt sollte man also bei der Vorverarbeitung für den Segmentierer nicht durchführen.

Kurzzeitspektrum, Nulldurchgangsrate und Signalenergie sind für die Spracherkennung von großer Bedeutung. Weiterhin sind nicht nur die Signaleigenschaften während der quasi-stationären Abschnitte des Sprachsignals wichtig, sondern vor allem bei Plosivlauten auch die Veränderungen des Signals (siehe Abschnitt 2.1.1).

Daher wurden folgende Merkmale des Sprachsignals in den anschließend beschriebenen Vorverarbeitungen eingesetzt (der Index t bezieht sich immer auf das Zeitfenster zum Zeitpunkt t):

Melscale-Spektralkoeffizienten $\boldsymbol{mel}_t$: Die Signalenergie wird in 16 sich überlappenden Frequenzbändern zusammengefaßt und logarithmiert; die Frequenzbänder sind entspechend der Melscale gewählt (siehe Abschnitt 2.2.2 und [35]). Die im Vektor $\boldsymbol{mel}_t$ zusammengefaßten 16 Werte werden *Melscale-Spektralkoeffizienten* genannt.

Delta-Koeffizienten $\Delta_T(t)$: Die Definition der Delta-Koeffizienten ist $\Delta_T(t) := \boldsymbol{mel}_{t+T} - \boldsymbol{mel}_{t-T}$.

Nulldurchgangsrate z_t: Die Anzahl z_t der Änderungen des Vorzeichens des Sprachsignals heißt Nulldurchgangsrate.

Signalenergie p_t: Die Gesamtenergie des Sprachsignals im Zeitfenster wird mit p_t bezeichnet.

Die Extraktion der Merkmale aus dem mit 16kHz abgetasten Signal erfolgte alle 10ms über Zeitfenstern der Länge 16ms. Es wurden folgende drei Vorverarbeitungen untersucht ($\boldsymbol{v}_t$ bezeichnet jeweils das Ergebnis der Vorverarbeitung).

Vorverarbeitung 1:

$\boldsymbol{v}_t := \boldsymbol{mel}_t$

Vorverarbeitung 2:

$$\boldsymbol{v}_t := L_2(\boldsymbol{mel}_{t-40ms}, \boldsymbol{mel}_{t-20ms}, \boldsymbol{mel}_t, \boldsymbol{mel}_{t+20ms}, \boldsymbol{mel}_{t+40ms}, p_t, z_t)^T$$

Vorverarbeitung 3:

$$\boldsymbol{v}_t := L_3(\boldsymbol{mel}_t, \boldsymbol{\Delta}_{20ms}(t), \boldsymbol{\Delta}_{40ms}(t), \boldsymbol{\Delta}_{80ms}(t), p_t, z_t)^T$$

L_2 und L_3 bezeichnen zwei lineare Transformationen. Die Transformationen wurden so berechnet, daß die $\boldsymbol{v}_t$ 20-dimensional sind und für die verschiedenen Klassen über einer Trainingsmenge möglichst gut getrennt werden (siehe hierzu auch [4, 23]). Das Trainieren von akustischen Modellen erfolgt in drei Schritten:

Berechnen einer linearen Transformation Dieser Schritt wurde nur bei Vorverarbeitung 2 und 3 durchgeführt (siehe oben).

Clustering Die Audiosegmente der einzelnen Klassen werden vorverarbeitet (gegebenenfalls unter Verwendung der im letzten Schritt berechneten linearen Transformation). Durch ein Clustering Verfahren (siehe [4]) werden für jede der Klassen aus den extrahierten Merkmalsvektoren 256 Punktwolken im Merkmalsraum gebildet. Die Mittelwertvektoren $\boldsymbol{\mu}_i$ $(i = 1, 2, \ldots, 256)$ der Mixturen von Normalverteilungen werden auf die verschiedenen Mittelpunktsvektoren der Punktwolken gesetzt.

Iteratives Anpassen der akustischen Modelle Ausgehend von den im letzten Schritt bestimmten Mittelwertvektoren werden ähnlich wie im Falle von Spracherkennungssystemen (siehe Abschnitt 2.2.3) die Parameter der akustischen Modelle angepaßt. Da im gegebenen Fall durch die Zeitmarken für die einzelnen Segmente der Trainingsmenge eine eindeutige Zuordnung von Trainingsdaten zu den Zuständen der HMMs gegeben ist, war kein Training wie im Falle von Spracherkennungssystemen notwendig[1], sondern es mußten lediglich die Parameter der Modelle entsprechend der Zuordnung angepaßt werden.

5.1.2 Von Hand gewählte Klasseneinteilung

Unter Verwendung der drei oben beschriebenen Vorverarbeitungsmethoden wurden Klassifikatoren für die von Hand gewählten Klassen über 4 der 5 Sendungen in der Trainingsmenge trainiert und die Segmente der fünften Sendung klassifiziert. Die Ergebnisse nach 1, 2, 3 und 4 Trainingsiterationen sind in Abbildung 5.1 dargestellt.

[1] Bei Spracherkennung ist in der Regel keine eindeutige Zuordnung von Trainingsbeispiel zu HMM Zustand gegeben. Es müssen daher entweder mittels Baum-Welch Training die akustischen Modelle für alle möglichen Pfade durch das HMM unter Berücksichtigung der Wahrscheinlichkeit für diese Pfade angepaßt werden, oder es kann durch den Viterbi-Algorithmus der beste Pfad durch das HMM bestimmt und so eine Abbildung der Trainingsbeispiele auf HMM Zustände erzeugt werden, die dann für das Anpassen der Parameter verwendet werden kann (Viterbi-Training).

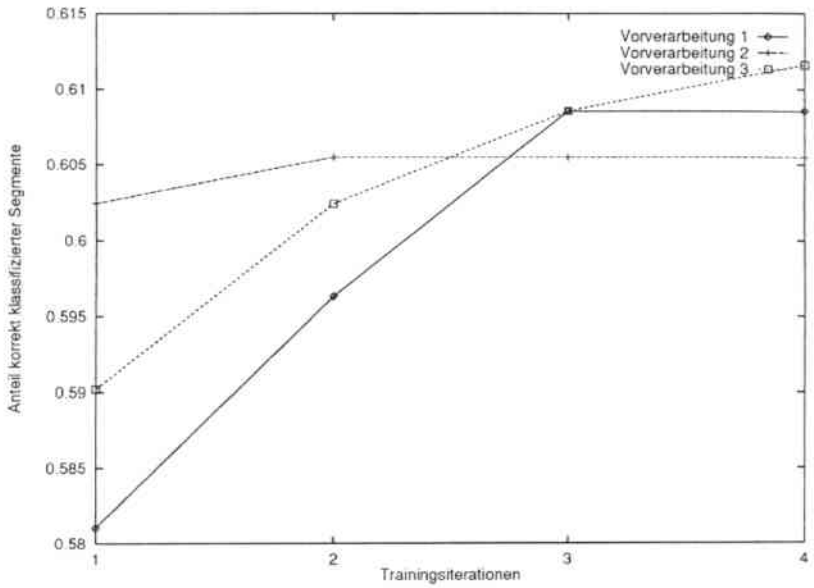

Abbildung 5.1: Anteil der durch die Klassifikatoren des Segmentierers korrekt klassifizierten Segmente für drei verschiedene Vorverarbeitungsmethoden nach 1, 2, 3 und 4 Trainingsiterationen.

5.1.3 Analyse der Versuche mit von Hand gewählter Klasseneinteilung

Auf den ersten Blick sehen die Klassifikationsergebnisse recht schlecht aus. Über ein Drittel aller Klassifikationen ist falsch. Außerdem deutet die Tatsache, daß die weitaus komplexere Vorverarbeitung 3 nur unwesentlich besser ist als Vorverarbeitung 1, darauf hin, daß Nulldurchgangsrate, Signalenergie und Kontextinformationen[2] kaum zusätzliche Information gegenüber dem Kurzzeitspektrum bieten, was eine Unterscheidung der Klassen anbelangt.

Konfusionsmatrix der 9 von Hand gewählten Klassen

		Segmentmarkierung								
Klassifikation		H1	H2	H3	H4	H5	H6	H7	H8	H9
H1	$*$-*K*-—	112	2	9	0	76	0	0	0	0
H2	$*$-*K*-*L*	1	0	24	0	3	0	0	0	0
H3	*MF*-*K*-*M(L)*	0	0	28	1	0	0	0	0	0
H4	—-*K*-*M(L)*	0	0	0	15	0	0	0	0	0
H5	$*$-*V*-—	0	0	0	1	19	0	0	0	0
H6	$*$-*V*-*L*	0	0	1	0	1	0	1	0	0
H7	$*$-*V*-*M(L)*	0	0	0	0	0	0	3	0	0
H8	$*$-*T*-—	1	0	0	0	0	1	0	23	5
H9	$*$-*T*-{*L*,*M(L)*}	0	0	0	0	0	0	0	0	0

[2]Diese Kontextinformationen sind durch die Delta-Koeffizienten in Vorverarbeitung 3 und die benachbarten Melscale-Spektralkoeffizienten in Vorverarbeitung 2 gegeben.

Wenn man sich anschaut, wie die Fehlklassifikationen zustande kommen, wird der Grund für dieses Ergebnis deutlich. In der oben dargestellten Konfusionsmatrix ist in Zeile i und Spalte j angegeben, wie oft Klasse Hi erkannt wurde, wenn ein Segment der Klasse Hj vorlag. Ist $i = j$, so handelt es sich um eine korrekte Klassifikation, andernfalls um eine Verwechslung der korrekten Klasse Hj mit Hi. Es sind die Ergebnisse für Vorverarbeitung 3 nach 4 Trainingsiterationen dargestellt, da diese Vorverarbeitung die besten Klassifikationsergebnisse erbracht hat. Die *Hi* ($i = 1, 2, \ldots 9$) bezeichnen hierbei die in Abschnitt 4.2.1 beschriebenen von Hand gewählten Klassen.
Am häufigsten werden drei Klassenpaare verwechselt. Wie kommt es zu den Verwechslungen?

H1 für H3 Die Verwechlsung der Klassen *MF-K-M(L)* und $*$-*K*-- ist dadurch zu erklären, daß die vorgegebenen Segmentgrenzen aufgrund syntaktischer Gesichtspunkte gesetzt wurden. Kommt in einem Segment Musik vor, wurde im Feld des Buchstabencodes, das den Hintergrund beschreibt, ein *M* eingetragen (siehe Kapitel 3), selbst wenn das Segment zum Teil aus klarer Sprache ohne Musik besteht. Weiterhin ist die Hintergrundmusik in einigen Fällen selbst für Menschen nur bei sehr aufmerksamen Hinhören wahrnehmbar.

H2 für H3 Für die Verwechslung von *MF-K-M(L)* und $*$-*K-L* trifft eben genannter Grund genauso zu. Oft liegt nur in einem Teil eines Segments Musik vor; da aber das Stück mit Hintergrundmusik keine vollständige syntaktische Einheit umfaßt, wurde es nicht als eigenes Segment markiert.

H1 für H5 Am häufigsten wurde $*$-*V*-- mit $*$-*K*-- verwechselt. Diese Verwechslung rührt daher, daß die Markierung *V* für den Kanal in den fünf Nachrichtensendungen für sehr unterschiedliche Kanaleigenschaften vergeben wurde. In diesem Zusammenhang muß erklärt werden, wie die Markierung der Sendungen vorgenommen wurden: Fünf verschiedene Personen der *CMU Robust Speech Recognition Group* haben je eine Nachrichtensendung transkribiert. Die Transkription wurde entsprechend einer kurzen Beschreibung der zu verwendenden Markierungen (wie in Kapitel 3 wiedergegeben) vorgenommen. Das hat dazu geführt, daß sehr unterschiedliche Kanalcharakteristiken mit *V* markiert wurden; von fast ungestörtem Kanal bis zu einem Kanal, der Sprache auch für Menschen nahezu unverständlich macht, ist alles vorzufinden.

Berücksichtigt man die eben genannten Gründe, so ist das Klassifizierungsergebnis weit besser zu bewerten, als die ca. 61.2% vermuten lassen. Für ein Segment, das mit *V* markiert wurde, aber einen nahezu klaren Kanal enthält, werden die extrahierten Merkmalsvektoren eher mit dem Modell für einen klaren Kanal übereinstimmen. Genauso ist zu erklären, wie es zu den anderen Verwechslungen kommt: Liegt nur sehr leise oder nur teilweise Musik

vor, so werden die Merkmalsvektoren mehr mit den Modellen für Klassen ohne Musik übereinstimmen als mit denen für Klassen mit Musik.

Für die Segmentierung bedeuten diese Verwechslungen, daß eine Trennung von klarem und verzerrten Kanal unter Umständen nicht sinnvoll ist. Hat man zwei leicht miteinander verwechselbare Klassen, so besteht die Gefahr, daß kleine Schwankungen in der Qualität der Audiodaten zu fehlerhaften Segmentgrenzen führen können, durch die dann unter Umständen in einem Spracherkennungssystem weit mehr Worterkennungsfehler verursacht werden, als eine fehlende Segmentgrenze bewirken würde.

Von den oben aufgezählten Verwechslungspaaren abgesehen, können, wie in Abschnitt 4.2.2 vermutet, die Klassen recht gut auseinander gehalten werden. Telephonkanal, reine Musik und klare Sprache werden fast vollständig getrennt. Zwischen diesen drei Klassen treten insgesamt 4 Fehlklassifikationen auf. Das entspricht bei 327 Segmenten einer korrekten Klassifikation in mehr als 98% der Fälle.

5.1.4 Automatisch gefundene Klasseneinteilung

In folgender Tabelle sind 9 automatisch erzeugte Klassen A1-A9 und zugehörige Mengenbeschreibungen aufgelistet.

Automatisch gefundene Klassen A1-A9			
	Mengenbeschreibungen für die Klassen A1-A9		
A1	*M-K-M*		
A2	*M-K-–*	*M-K-L*	*M-K-ML*
A3	*M-T-**		
A4	*F-K-M*		
A5	*–-K-M*		
A6	*F-V-–*		
A7	*F-K-–*		
A8	*F-V-L* *F-K-L* *F-T-**	*F-V-M(L)* *F-K-ML*	
A9	*M-V-** *–-V-** *–-T-** *–-K-–*	*–-K-L*	*–-K-ML*

Diese 9 Klassen wurden dem Baum in Abbildung 4.3 gemäß erzeugt. Es wurden nur 9 Klassen erzeugt, da die Gestalt des Baums in der Abbildung, wie in Abschnitt 4.3.2 erklärt, höchstens bis zur 10. Frage als nach statistischen Gesichtspunkten stabil betrachtet werden kann. Die Aufspaltungen der Klassen wurden mit Hilfe des in Abschnitt 4.3.1 beschriebenen Abstandsmaßes vorgenommen. Es wurde in jedem Schritt die Aufteilung gewählt, die das Abstandsmaß maximierte. Dieser Abstand fiel mit der Anzahl von Aufspaltungen schnell ab, war bis zur Frage 8 jedoch noch deutlicher größer

als bei Frage 11. Daher wurden für den Segmentierer Klassifikatoren für die Klassen A1-A9 trainiert, die durch die ersten 8 Aufspaltungen erzeugt wurden.

Die Mengenbeschreibungen können durch Konjunktion der positiv und der negierten negativ beantworteten Fragen, die zur Bildung der Klassen geführt haben, hergeleitet werden. Wenn man die Menge aller Basisklassen z.B. erst durch die Frage *M-K-** einschränkt und diese Einschränkung dann verfeinert durch **-*-M*, so wählt man insgesamt diejenigen Basisklassen aus, für die *M-K-M* zutrifft.

Die Klassifikatoren für diese Klassen wurden in derselben Weise trainiert, wie es weiter oben für die von Hand gewählten Klassen beschrieben ist. Auch hier wurde die Vorverarbeitung 3 verwendet, da diese für die von Hand vorgenommene Klasseneinteilung die besten Resultate erbracht hat.

Die Klassifikationsergebnisse für die automatisch gefundene Klasseneinteilung sind in folgender Matrix angegeben.

Konfusionsmatrix der automatisch gewählten Klassen A1-A9

Klassifikation	Segmentmarkierung								
	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1	2	0	0	6	0	0	0	0	0
A2	4	41	0	0	0	0	0	0	48
A3	0	0	23	0	0	0	0	5	3
A4	0	0	0	16	2	0	0	0	0
A5	0	0	0	0	14	0	0	0	0
A6	0	0	0	0	0	0	0	0	0
A7	0	0	0	6	0	29	66	0	2
A8	0	0	0	22	0	0	0	0	0
A9	4	0	0	0	1	2	0	0	27

Es werden 58.5% der Segmente korrekt klassifiziert.

5.1.5 Analyse der Versuche mit automatisch erzeugter Klasseneinteilung

Die Fehlklassifikationen im Fall der automatisch gefundenen Klassen werden wie auch im Fall der von Hand gewählten Klassen hauptsächlich durch drei Klassenpaare verursacht.

A8 für A4 Ganz offenbar ist *F-K-M* (A4) nicht zuverlässig von den vielen anderen Basisklassen mit weiblichen Sprechern und Hintergrundgeräuschen, die in A8 enthalten sind, zu trennen. Die für diese Aufspaltung verantwortliche Kombination der Fragen 3 und 8 in Abbildung 4.3 ist in Abbildung 4.4 nicht mehr vorhanden. Daß sich dieser Teilbaum beim Übergang zu dem auf 5 Trainingssendungen basierenden Baum entsprechend verändert, ist ein Zeichen dafür, daß diese Aufspaltung statistisch nicht ausreichend abgesichert war.

A7 für A6 Hier werden die Basisklassen *F-K--* und *F-V--* miteinander verwechselt. Das entspricht der Verwechselbarkeit von H1 und H5 bei den von Hand gewählten Klassen für weibliche Sprecher.

A2 für A9 Klasse A2 enthält Basisklassen mit männlichen Sprechern über einen klaren Kanal, Klasse A9 enthält hauptsächlich Basisklassen mit männlichen Sprechern über einen Kanal, der mit *V* markiert wurde. Damit entsprechen solche Fehlklassifikationen denen von H1 für H5 eingeschränkt auf männliche Sprecher.

Die weiteren Verwechslungen A2 mit A1 und A7 mit A4 beruhen wieder auf dem Erkennen von Sprache ohne Musik, wenn tatsächlich Sprache mit Musik vorlag (wie H1 mit H3). Die Fehlklassifikationen A9 für A1 kommen durch die schlechte Unterscheidbarkeit von klarem und verzerrtem Kanal zustande und sind deshalb vergleichbar mit H1 und H5.

Die Summe all dieser Verwechslungen ist ungefähr so groß wie im Falle der von Hand gewählten Klassen. Es kommen allerdings noch die Verwechslungspaare A1 für A4 und A3 für A8 hinzu; hier wird Sprache männlicher Sprecher mit Hintergrundmusik bzw. über einen Telephonkanal für Sprache weiblicher Sprecher mit Hintergrundmusik bzw. über einen Telephonkanal gehalten. Es sieht also auf den ersten Blick so aus, als wäre durch automatische Klassenbildung nichts gewonnen, sondern etwas verloren worden. Da sich die Klasseneinteilungen jedoch sehr voneinander unterscheiden (siehe Anhang A) und vor allem die Verteilung der Daten auf die Klassen A1-A9 und H1-H9 sehr unterschiedlich ist, können die Ergebnisse nicht direkt miteinander verglichen werden.

5.1.6 Vergleich der Klasseneinteilungen

In folgender Tabelle ist die Verteilung der Test- und Trainingsdaten auf die einzelnen Klassen der Klasseneinteilungen A1-A9 und H1-H9 angegeben. Es wird deutlich, daß die Daten viel homogener auf die automatischen erzeugten Klassen verteilt sind. Das ist eine Folge des gewählten Abstandsmaßes, da die Anzahl der Trainingsbeispiele $N_{\mathcal{B}}$, die für eine Klasse $\mathcal{B}$ existieren, beim Aufspalten der Mixtur $p_{\mathcal{B}}(\boldsymbol{x})$ in $p_{\mathcal{B}_{\text{Ja}}}(\boldsymbol{x})$ und $p_{\mathcal{B}_{\text{Nein}}}(\boldsymbol{x})$ in das Abstandsmaß mit eingehen. Das ist auch sinnvoll, weil man aufgrund der geringen statistischen Signifikanz wenig an zusätzlicher Information gewinnt, wenn man eine Klasse, für die sehr wenige Trainingsbeispiele vorliegen, von einer Klasse mit vielen Trainingsbeispielen abtrennt.

Diese homogene Verteilung der Daten auf die Klassen ist für die Segmentierung wünschenswert, da die Aufgabe des Segmentierens ist, kleine Stücke zu erzeugen. Ist die Verteilung der Daten sehr inhomogen, ist es wahrscheinlicher, daß verschiedene akustische Bedingungen in den Klassen, denen viele Daten zugeordnet sind, zusammengefaßt werden. Das bedeutet, daß Segmentübergänge übersehen und dadurch große Stücke erzeugt werden.

Trainingsdaten pro Klasse					
Klasse	Dauer	# Seg.	Klasse	Dauer	# Seg.
H1	3136.43	520	A1	754.332	147
H2	461.329	83	A2	2775.31	455
H3	1062.14	198	A3	762.836	98
H4	376.828	61	A4	293.377	47
H5	689.712	90	A5	357.233	58
H6	332.135	46	A6	241.695	29
H7	61.185	9	A7	454.836	82
H8	731.449	98	A8	476.151	78
H9	46.909	5	A9	782.344	116
Testdaten pro Klasse					
Klasse	Dauer	# Seg.	Klasse	Dauer	# Seg.
H1	553.095	114	A1	47.916	10
H2	6.409	2	A2	181.587	45
H3	290.933	62	A3	164.877	23
H4	118.919	17	A4	237.268	50
H5	554.713	99	A5	118.919	17
H6	2.504	1	A6	185.706	31
H7	11.486	4	A7	374.582	66
H8	156.665	23	A8	21.048	5
H9	29.26	5	A9	392.081	80

Was die Güte der Segmentierung anbelangt, sagen die obigen Versuche noch nicht sehr viel aus. Die Klassifizierungsfähigkeit der Klassifikatoren A1-A9 und H1-H9 unterscheidet sich kaum. Es kann jedoch vermutet werden, daß bei Segmentierung durch A1-A9 zusätzliche Segmentgrenzen aufgrund der Unterscheidung von männlichen und weiblichen Sprechern gefunden werden. Wenn es sich bei den Verwechlsungen zwischen Mann und Frau um konsistente Fehler handelt, d.h. wenn nur Frauen, deren Sprache ähnliche Merkmale wie die von Männern aufweist, als männliche Sprecher klassifiziert werden, so sind diese Fehlklassifikationen im Sinne der Segmentierung kein Problem. Es ist nur sicherzustellen, daß nicht aufgrund starker Verwechselbarkeit der Modelle für Mann und Frau fehlerhafte Segmentgrenzen verursacht werden. Das ist in den Versuchen zur Segmentierung einer gesamten Radiosendung zu untersuchen.

5.2 Segmentierung einer Radiosendung

Nachdem in den Versuchen auf vorgegebenen Segmenten die Güte der Klassifikatoren untersucht worden war, sollte durch die in diesem Abschnitt beschriebenen Experimente, die Eignung der Klasseneinteilungen in Bezug auf die Segmentierung ermittelt werden.

5.2.1 Bewertung einer Segmentierung

Eine ganze Nachrichtensendung wurde, wie in Abschnitt 4.1.2 beschrieben, unter Verwendung des *JRTk* Suchobjekts segmentiert. Hierfür wurde wieder die Nachrichtensendung aus der Trainingsmenge verwendet, die nicht für das Training der Klassifikatoren benutzt worden war.

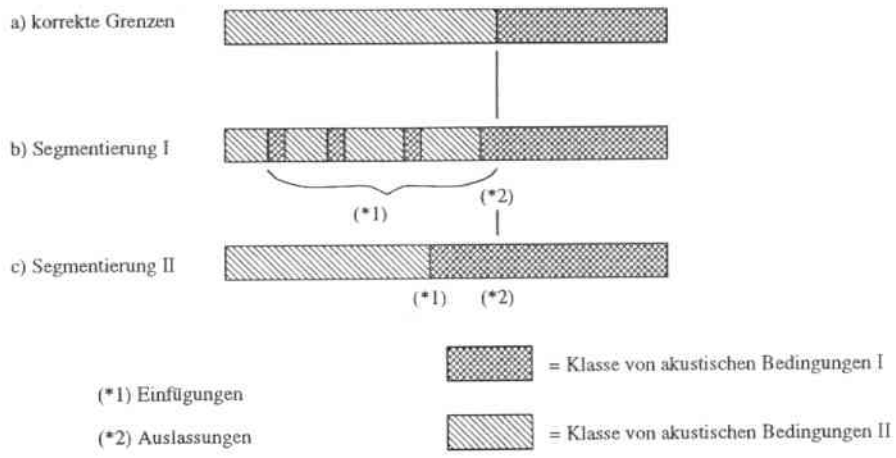

Abbildung 5.2: Beispiele für Segmentierungsfehler

Um zu entscheiden, durch welche von zwei Klasseneinteilungen die bessere Segmentierung erzeugt wird, muß die Güte einer Segmentierung bewertet werden können. Folgende Kriterien wurden für die Bewertung betrachtet:

Korrekte Klassifizierung von Merkmalsvektoren (KKM): Bei der Segmentierung einer gesamten Nachrichtensendung bekommt man für jeden einzelnen Merkmalsvektor eine Zuordnung zu einer Klasse. Diese Zuordnung ist verhältnismäßig unabhängig von den benachbarten Merkmalsvektoren. Die einzige Abhängigkeit wird durch die geforderte Mindestlänge eines Segments geschaffen. Im Falle vorgegebener Segmentgrenzen war das anders; dort wurde die Klassenzuordnung eines Vektors durch alle Merkmalsvektoren des Segments bestimmt.

Eine Anforderung an eine gute Segmentierung ist, daß die Klassifikation der Merkmalsvektoren für möglichst viele Vektoren korrekt sein muß. Der Prozentsatz der korrekt klassifizierten Merkmalsvektoren wird von nun an mit KKM abgekürzt. Dieses Maß gibt nur indirekt die Güte der Segmentgrenzen an. Es ist im Extremfall möglich, daß alle Segmentgrenzen korrekt bestimmt werden und dennoch kein einziger Merkmalsvektor richtig klassifiziert wird.

Anzahl falsch gesetzter Segmentgrenzen (ES): In Abbildung 5.2 sind verschiedene Segmentierungen eines Audiostücks angegeben. Unterschiedlich schraffierte Flächen stellen unterschiedliche akustische Bedingungen dar. Der Anteil der fehlerhaft klassifizierten Stücke ist in den Segmentierungen I und II gleich groß. Dennoch ist Segmentierung

II weitaus besser zu bewerten als Segmentierung I: Viele fehlerhafte Grenzen innerhalb eines Segments machen es unmöglich, Worte, die diese Grenzen überschneiden, korrekt zu erkennen. Weiterhin wird nicht nur die Erkennung dieser Worte beeinträchtigt, sondern auch die der angrenzenden Worte, da sich ein Worterkennungsfehler durch das Sprachmodell auf benachbarte Worte auswirkt. Außerdem wird beim *time alignment* versucht, die im Sinne des akustischen Modells wahrscheinlichste Wortfolge in einem Segment zu bestimmen. Durch eine falsche Segmentgrenze kann eine nicht korrekte Wortfolge wahrscheinlicher werden als die korrekte, selbst wenn bei richtig gesetzten Segmentgrenzen, die korrekte Wortfolge erkannt worden wäre. Dieses Beispiel macht deutlich, daß die Anzahl der falsch gesetzten Segmentgrenzen minimiert werden muß. Dieser Wert wird im folgenden Einfügungen genannt und mit ES (Eingefügte Segmentgrenzen) abgekürzt.

Anteil nicht gefundener Segmentgrenzen (AS): Ziel der Segmentierung ist es, kleine Segmente mit gleichbleibenden akustischen Bedingungen zu erzeugen. Werden Übergänge zwischen verschiedenen akustischen Bedingungen übersehen, können einerseits die akustischen Modelle von auf bestimmte akustische Bedingungen angepaßten Spracherkennern nicht mehr genau zu den Bedingungen im Segment passen. Andererseits entstehen längere Segmente. Der Anteil der nicht gefundenen Segmentgrenzen an den tatsächlich vorhandenen ist also zu minimieren. Dieser Anteil wird mit AS (Ausgelassenen Segmentgrenzen) abgekürzt.

Durchschnittliche Segmentlänge (MS): Wie im letzten Punkt erwähnt, ist ein Ziel der Segmentierung, kleine Segmente zu erzeugen. Ist die Klasseneinteilung, die zum Segmentieren verwendet wird, schlecht gewählt und umfassen z.B. einige Klassen viele verschiedene akustische Bedingungen, so können zwischen diesen akustischen Bedingungen keine Übergänge erkannt werden. Die Folge ist, daß lange uneinheitliche Segmente entstehen. Die Güte einer Klasseneinteilung in Bezug auf die Segmentierung spiegelt sich also auch in einer kleinen mittleren Segmentlänge (MS) wieder. Hier muß man allerdings darauf achten, daß gleichzeitig die Zahl der Einfügungen niedrig ist, da es durch eine große Zahl von Einfügungen natürlich auch möglich ist eine kleine mittlere Segmentlänge zu erzielen.

Diese Kriterien sind nicht unabhängig voneinander. Zum Beispiel wird KKM klein sein, wenn viele korrekte Segmentgrenzen nicht erkannt werden, d.h. AS groß ist. Um anhand dieser Kriterien die für die Segmentierung am besten geeignete Klasseneinteilung zu finden, muß

- KKM maximiert,
- AS und ES mininiert und

- MS unter Berücksichtigung von ES möglichst klein gehalten werden.

Die einzelnen Kriterien haben unterschiedliche Wichtigkeit in Bezug auf die Güte der Segmentierung. Werden z.B. viele fehlerhafte Segmentgrenzen gesetzt, ist also ES groß, so ist das sehr wahrscheinlich weitaus schädlicher für die anschließende Spracherkennung als die Auslassung von Grenzen zwischen akustisch ähnlichen Segmenten. Wenn beispielsweise *MF-K-–* sehr ähnlich zu *MF-V-–* ist, so werden die akustischen Modelle eines Spracherkenners, die mit Daten der Klasse *MF-K-–* trainiert wurden, verhältnismäßig gut zu beiden Klassen passen. Wird eine Grenze zwischen diesen Klassen nicht gesetzt, so führt das unter Umständen zu etwas schlechteren Worterkennungsraten, weil die akustischen Modelle des Spracherkenners nicht ganz genau zu beiden Klassen passen, und dazu, daß große Segmente entstehen. Werden aber aufgrund leichter Verwechselbarkeit der akustischen Modelle des Segmentierers viele fehlerhafte Segmentgrenzen eingefügt, so hat dies sicherlich negativere Auswirkungen auf die Worterkennungsrate bedingt durch die weiter oben beschriebenen Konsequenzen von Einfügungen auf Sprachmodell und *time alignment*. Für die Segmentierung sind ES und AS von besonderer Wichtigkeit.

5.2.2 Berechnung der Gütemaße

Wie KKM und MS zu berechnen sind, ist offensichtlich. KKM kann bestimmt werden, indem der zeitliche Anteil der korrekt klassifizierten Abschnitte an der Gesamtdauer der segmentierten Audiodaten ermittelt wird. MS ist einfach das arithmetische Mittel der Länge aller erzeugten Segmente.

Für die Bestimmung von AS und ES muß festgelegt werden, welche zeitliche Abweichung einer gefundenen von einer vorgegebenen Grenze man bei der Entscheidung zuläßt, ob es sich noch um eine korrekt bestimmte Grenze handelt oder nicht. Da im gegebenen Fall die Segmentgrenzen manuell gesetzt wurden und die Qualität dieser Grenzen leider sehr schlecht ist, mußte eine ausreichend große Toleranz zugelassen werden. Es kommt bei den Markierungen vor, daß die von Hand gesetzten Segmentgrenzen bis zu mehrere Worte von den tatsächlich vorhandenen abweichen.

Aufgrund dieser Tatsache wurde bei der Berechnung von AS und ES folgendermaßen vorgegangen: Wurde eine Segmentgrenze gefunden, die um mehr als 0.5s von allen vorgegebenen abwich, so wurde angenommen, daß zwischen dieser und den vorgegebenen Grenzen kein Zusammenhang besteht, also eine Einfügung vorliegt. Als Einfügungen wurden auch alle überschüssigen Grenzen gewertet, wenn mehrere gefundene Segmentgrenzen auf eine vorgegebene entfielen[3]. Wurde für eine vorgegebene Segmentgrenze keine Grenze in einer Umgebung von höchstens 0.5s gefunden, so wurde dies als Auslassung gewertet.

[3]Dieser Fall trat jedoch nie ein.

In der Praxis ist es so, daß die durch die Klassifikatoren gefundenen Segmentgrenzen viel genauer mit den tatsächlichen Änderung der akustischen Bedingungen übereinstimmen als die von Menschen gesetzten. Die wenigen Beispiele, in denen beim Testen des Segmentierers eine vorgegebene Grenze und eine automatisch gefundene um mehr als 0.5s abwichen, obwohl sie zusammengehörten[4], waren dadurch bedingt, daß die von Menschen gesetzten Zeitmarken so ungenau waren.

5.2.3 Ergebnisse der Segmentierung

Insgesamt sind auf der Sendung, die zum Testen des Segmentierers verwendet wurde, 327 Segmente vorgegeben. Bei vielen der 326 Segmentübergänge ändern sich weder der Sprecher noch die akustischen Bedingungen. Diese Segmentgrenzen unterteilen nur lange Abschnitte, zeigen also keine Änderung der Gegebenheiten in den Audiodaten an.

Betrachtet man für die 327 Segmente die Zuordnung zu den Klassen H1-H9 und A1-A9, so gibt es sowohl für die von Hand gewählten als auch für die automatisch bestimmten Klassen nur 94 unterscheidbare Abschnitte.

Sprecherwechsel bei gleichbleibenden akustischen Bedingungen können durch dem gewählten Segmentierungsansatz nicht erkannt werden, es sei denn man trainiert Klassifikatoren für die einzelnen Sprecher und die verschiedenen Kombinationen aus Hintergrundgeräuschen und Kanaleigenschaften, wofür aber im gegebenen Fall nicht ausreichend viele Trainingsbeispiele vorliegen.

Die Segmentierungen, die durch die Klassifikatoren A1-A9 und H1-H9 erzeugt wurden, sind bezüglich der Maße KKM, AS, ES und MS (siehe oben) bewertet worden. Die Ergebnisse sind in folgender Tabelle angegeben.

Bewertung der Segmentierungsergebnisse					
	KKM	AS	ES	MS	# Segmente
H1-H9	61.58%	35 (37%)	44	16.85s	94
A1-A9	60.94%	26 (27%)	37	16.21s	94

Die Segmentierung unter Verwendung von Klassifikatoren für A1-A9 ist in Bezug auf AS und ES deutlich besser als die für H1-H9. Der Anteil der korrekt klassifizierten Merkmalsvektoren KKM ist im Fall H1-H9 geringfügig größer, MS hingegen für A1-A9 geringfügig kleiner.

5.3 Bewertung der Versuche

Die Klassifizierung von vorgegebenen Segmenten und die Segmentierung einer Radiosendung haben gezeigt, daß die automatisch erzeugten Klassen A1-A9 bessere Eigenschaften haben als die Klassen H1-H9.

[4] Das kann durch Anhören überprüft werden.

Die Verteilung der Daten auf A1-A9 ist homogener, und es werden geringfügig kürzere Segmente erzeugt. Die Klassifikationsfähigkeit ist rein prozentual gesehen für H1-H9 geringfügig besser als für A1-A9, was sich allerdings mit der inhomogeneren Verteilung der Daten auf die Klassen erklären läßt.

Bezüglich der entscheidenden Kriterien AS und ES ist die automatisch gefundene Klasseneinteilung der von Hand erstellten jedoch deutlich überlegen.

5.4 Segmentieren der Evaluationstestmenge

Die in den letzten Abschnitten beschriebenen Untersuchungen haben ergeben, daß der Ansatz, Klassen von akustischen Bedingungen automatisch zu bilden und basierend auf diesen Klassen zu segmentieren, bessere Segmentierungsergebnisse erbringt als ein Segmentierer unter Verwendung intuitiv gewählter Klassen.

Der Übergang von Abbildung 4.3 zu Abbildung 4.4 hat aber auch deutlich gemacht, daß 4 Nachrichtensendungen noch nicht erlauben, im statistischen Sinn stabile Aussagen über die Merkmalsvektoren für die verschiedenen akustischen Bedingungen zu machen. Um alle zur Verfügung stehenden Trainingsdaten für das endgültige System zur Segmentierung zu verwenden, wurden Klassen A1'-A9' über allen 5 Nachrichtensendungen der Trainingsmenge bestimmt und entsprechende Klassifikatoren trainiert. Hierbei wurde ganz genauso vorgegangen, wie in den letzten Abschnitten für A1-A9 beschrieben.

In folgender Tabelle sind die über allen 5 Trainingssendungen automatisch gebildeten Klassen A1'-A9' angegeben.

Automatisch gefundene Klassen A1'-A9'		
	Mengenbeschreibungen für die Klassen A1'-A9'	
A1'	*M-K-M(L)*	
A2'	*M-K-–*	*M-K-L*
A3'	*∗-T-∗*	
A4'	*–-K-M*	
A5'	*M-V-M(L)* *F-V-M(L)* *F-K-M(L)* *–-K-ML* *–-V-M(L)*	
A6'	*F-K-–*	
A7'	*F-V-–*	
A8'	*F-V-L* *F-K-L*	
A9'	*M-V-–* *–-V-–* *–-K-–*	*M-V-L* *–-V-L* *–-K-L*

5.4.1 Versuche auf der Trainingsmenge

Für das Training der Klassifikatoren für A1'-A9' wurden alle verfügbaren Daten mit Markierungen bezüglich Sprechergeschlecht, Kanaleigenschaften und Hintergrundgeräuschen verwendet, um möglichst stabile Modelle für diese Klassen zu erhalten. Dadurch standen allerdings keine Daten mehr zum Testen dieser Modelle zur Verfügung.

Auf der Evaluationstestmenge wird nur zwischen Segmenten, bei denen der volle Frequenzbereich bis 8000Hz (volle Bandbreite), und solchen, bei denen nur ein reduzierter Frequenzbereich verwendet wird (reduzierte Bandbreite), sowie dem Vorhandensein bzw. Nichtvorhandensein von Musik unterschieden. Auf dieser Menge kann also der Segmentierer nicht bezüglich Klassifizierungs- und Segmentierungsfähigkeit getestet werden.

Um trotzdem einen Eindruck von der Qualität der Klassifikatoren zu bekommen, wurde der Segmentierer auf denselben Sendungen, die auch zum Training verwendet wurden, getestet. Die in folgender Tabelle dargestellten Ergebnisse sind dadurch natürlich nicht auf die Segmentierungsfähigkeit im allgemeinen Fall übertragbar.

Bewertung der Segmentierungsergebnisse					
	KKM	AS	ES	MS	# Segmente
A1'-A9'	93.11%	95 (23%)	121	18.34s	413

Die Anzahl der Auslassungen und Einfügungen ist hoch, obwohl auf derselben Menge getestet wurde, auf der auch die Klassifikatoren trainiert wurden.

In folgender Tabelle ist aufgeführt, welche falschen Klassen wie häufig in welche der korrekten Klassen eingefügt wurden.

Einfügungen von fehlerhaften Klassen									
Eingefügte	Segmentmarkierung								
Klassen	A1'	A2'	A3'	A4'	A5'	A6'	A7'	A8'	A9'
A1'	0	8	0	0	1	0	2	1	28
A2'	16	0	3	1	1	0	1	3	0
A3'	3	2	0	0	0	0	0	0	0
A4'	0	3	0	0	0	1	0	6	0
A5'	0	1	0	0	0	0	0	0	0
A6'	0	7	1	1	0	0	0	3	0
A7'	0	0	0	0	0	0	0	4	3
A8'	0	0	2	0	0	0	3	0	2
A9'	12	0	0	0	0	0	0	2	0

Die drei Klassen A1', A2' und A9' sind für über 53% der Einfügungen verantwortlich. In den Segmenten dieser Klassen kommen vor allem

männliche Sprecher über einen klaren oder verzerrten Kanal mit Hintergrundgeräuschen, die teilweise Musik einschließen, vor (ausgenommen ist lediglich die Menge *M-V-M(L)*). Weiterhin werden durch Klasse A9' noch Segmente ohne Sprecher mit einem klaren oder verzerrten Kanal mit Hintergrundgeräuschen ohne Musik oder ganz ohne Hintergrundgeräusche erfaßt; die Gesamtlänge solcher Segmente beträgt jedoch insgesamt nur ca. 30s (siehe Anhang A).

5.4.2 Zweistufige Segmentierung

Grund für einen Großteil der Einfügungen bei den Versuchen auf der Trainingsmenge ist das Ein- und Ausblenden von Musik, das mehrmals pro Sendung vorkommt. In solchen Fällen wird fast mit Sicherheit ein Wort, auf jeden Fall aber ein Satz durch eine Segmentgrenze zerschnitten. Ein Lösungsansatz für dieses Problem ist folgender. Durch Zusammenlegen der Klassen A1', A2' und A9' werden die fehlerhaften Segmentgrenzen nicht mehr gesetzt. Auf der anderen Seite sind dadurch aber auch korrekte Segmentübergänge zwischen diesen Klassen nicht mehr erkennbar. Durch einen zweiten Segmentierungsschritt, der basierend auf Stille Segmentgrenzen setzt, können jedoch wieder einige der durch das Zusammenlegen nicht mehr erkennbaren Grenzen wiedergefunden werden.

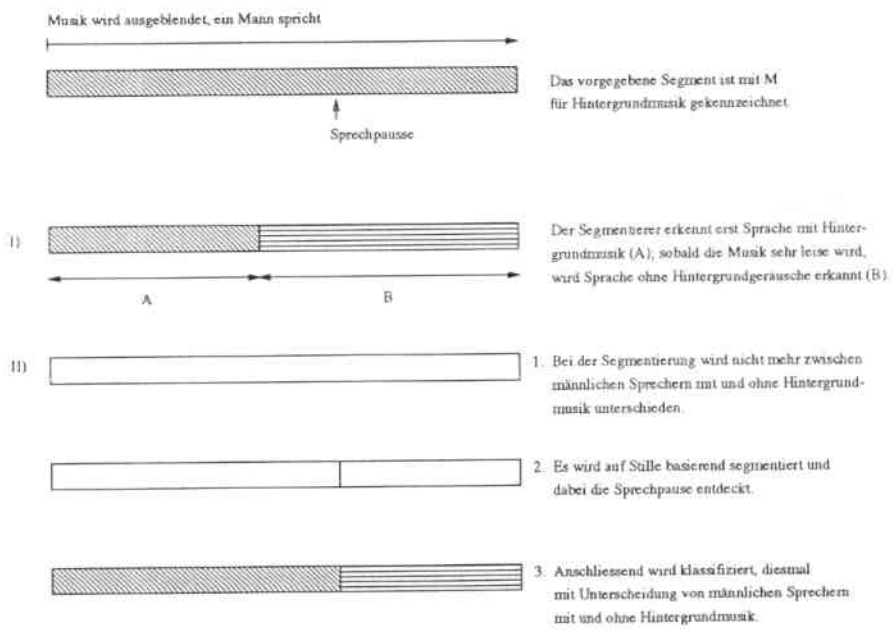

Abbildung 5.3: Gegenüberstellung von zwei Segmentierungsstrategien. I) Segmentierung mit Unterscheidung männlicher Sprecher mit und ohne Hintergrundmusik in einem Schritt. II) Segmentierung in zwei Schritten mit anschließender Klassifizierung.

Die nach diesem zweiten Segmentierungsschritt erzeugten Segmente können anschließend durch eine erneute Klassifizierung der jeweils korrekten

akustischen Klasse zugeordnet werden. Bei der erneuten Klassifizierung ist es sinnvoll, wieder zwischen den Klassen A1', A2' und A9' zu unterscheiden, da durch die auf Stille basierende Segmentierung neue Segmentgrenzen eingefügt wurden. Dieser Prozeß ist in Abbildung 5.3 dargestellt. Die gefundene Segmentierung entspricht zwar nicht der vorgegebenen, wird aber zu einer Minimierung der Worterkennungsfehler führen.

In folgender Tabelle sind KKM, AS, ES und MS nach Zusammenlegen der Klassen A1', A2' und A9' wiedergegeben.

Ergebnisse nach Klassenzusammenlegung				
Zusammengelegte Klassen	KKM	AS	ES	MS
A1', A2' und A9'	97.86%	55 (16.5%)	61	25.2889s

Durch diese Zusammenlegung wurde einerseits KKM erhöht und ES vermindert, andererseits ist es aber nach Zusammenlegung nicht mehr möglich Übergänge zwischen den Klassen A1', A2' und A9' zu entdecken. Die Anzahl der erkennbaren Übergänge nimmt ab.

Von den 412 Segmentübergängen, die zwischen den Klassen A1'-A9' existieren, sind nach dem zweistufigen Segmentieren nur 284 gefunden worden, nicht mehr 317 wie bei Unterscheidung aller neun Klassen. Damit beträgt die Zahl der Auslassungen insgesamt tatsächlich 128 und ist gegenüber der einstufigen Segmentierung also um 33 gestiegen.

Aufgrund der oben diskutierten Gründen für die Einfügungen wurde die Entscheidung getroffen, die Klassen A1', A2' und A9' im System zur Segmentierung der Evalutationsmenge nicht zu unterscheiden. Statt dessen wurden durch einen zweiten auf Stille basierenden Segmentierungsschritt kleinere Segmente erzeugt und später erneut klassifiziert.

5.4.3 Das Gesamtsystem zur Segmentierung

Die Evaluationstestmenge wurde mit Hilfe von Klassifikatoren für die zusammengelegten Klassen A1', A2' und A9' sowie die Klassen A3', A4', ..., A8' segmentiert. Nach dieser Segmentierung lagen zum Teil immer noch ziemlich große Stücke vor. Das größte von diesen Stücken war 6 Minuten lang.

Um diese lange Segmente weiter zu unterteilen, wurde ein Verfahren verwendet, das basierend auf Stille Segmentgrenzen setzt. Dieses Verfahren wurde für ein anderes Spracherkennungssystem entwickelt[5]. Es ist gut dafür geeignet, Sprechpausen zu erkennen, die von Menschen gemacht werden, um die syntaktische Struktur eines Satzes hervorzuheben (siehe [31, Kapitel 3]).

Obwohl zwischen den Klassen A1', A2' und A9' bei der Segmentierung nicht unterschieden wurde, ist es sinnvoll die Audiostücke, die durch den zweiten Segmentierungsschritt erzeugt wurden, bzgl. dieser Klassen zu unterscheiden (siehe Abbildung 5.3) .

[5]Mit dem auf Stille basierenden System wurden bei der SWITCHBOARD Evaluation 1996 lange Abschnitte mit Sprache über Telephonverbindungen zerteilt.

Kapitel 6

Verwendung von Spezialerkennern

Durch das im letzten Kapitel beschriebene Verfahren zur Segmentierung kann eine gesamte Radiosendung in kleine Stücke zerteilt werden, innerhalb derer jeweils eine bestimmte Klasse von akustischen Bedingungen vorliegt. Das ermöglicht, die einzelnen Segmente Spracherkennern zuzuführen, die an diese Klassen von Einflüssen auf das Sprachsignal speziell angepaßt wurden.

Im Zusammenhang mit der Verwendung von solchen speziell angepaßten Spracherkennern sind folgende Fragen zu klären:

1. Für welche akustischen Bedingungen sollen eigens Spracherkenner trainiert, bzw angepaßt werden?

2. Wie werden die Erkenner an die verschiedenen akustischen Gegebenheiten anpaßt? An dieser Stelle muß die sehr geringe Menge an Trainingsdaten berücksichtigt werden.

3. Wie wählt man für ein Segment den geeigneten Erkenner aus?

Auf diese Punkte wird in den folgenden Abschnitten eingegangen.

6.1 Spezialerkenner für verschiedene Einflüsse auf das Sprachsignal

Durch den Einsatz von mehreren Spezialerkennern soll insgesamt die Anzahl der Worterkennungsfehler minimiert werden. Je besser die akustischen Modelle eines Spracherkenners zu den akustischen Bedingungen in einem Audiosegment passen, je genauer die Vorverarbeitung auf die zu erwartenden Einflüsse, denen das Sprachsignal in einem Segment unterliegt, abgestimmt ist und je besser insgesamt die Parameter eines Spracherkenners auf die Gegebenheiten in einem Segment zugeschnitten sind, um so besser wird das Erkennungsergebnis sein.

Verfügt man über eine ausreichend große Menge von Trainingsdaten, so kann es sinnvoll sein, viele verschiedene speziell angepaßte Spracherkenner zu verwenden. Für Marketplace-Radiosendungen wären zum Beispiel folgende Spezialerkenner sinnvoll:

- Ein Erkenner für den Hauptansager. David Brancaccio spricht in den Sendungen bei weitem am häufigsten. Da sprecherabhänigige Spracherkennungssysteme in aller Regel eine bessere Erkennungsleistung bieten als sprecherunabhängige, würde die Verwendung eines solchen Erkenners die Wortfehlerrate für den Ansager sicher verringern.

- Ein Erkenner für Sprache über Telephonverbindungen. Für Telephonsprache bietet es sich an, eine andere Vorverarbeitung zu verwenden, die den Besonderheiten des Telephonkanals gerecht wird. Außerdem unterscheidet sich in den Radiosendungen Sprache über Telephon sehr von z.B. der Sprache des Ansagers beim Verlesen der Nachrichten. Daher könnte man ein speziell darauf zugeschnittenes Sprachmodell benutzen.

- Ein Erkenner für Sprache mit Hintergrundmusik. Menschen sind in der Lage, Sprache von Musik zu trennen. Die Erkennungsleistung von Spracherkennungssystemen sinkt jedoch dramatisch in Anwesenheit von Hintergrundmusik. Möglichkeiten, mit diesem Problem umzugehen, werden in [10] genannt.

- Jeweils Erkenner für weibliche und männliche Sprecher. Die Sprache von Männern und Frauen unterscheidet sich beträchtlich. Geschlechtsabhängige Spracherkenner können dieser Tatsache gerecht werden. Einige der Unterschiede können auch durch eine geeignete Vorverarbeitung kompensiert werden.

- Erkenner für Sprecher mit verschiedenen ausländischen Akzenten. Ausländische Sprecher sprechen oft Phoneme nicht korrekt aus, da die für die Artikulation eines Phonems nötigen Vokaltraktkonfigurationen oder Bewegungen von Artikulatoren in der Muttersprache nicht vorkommen. Diese Art von Fehlern können bis zu einem gewissen Grad ausgeglichen werden (siehe [28]).

Im vorliegenden Fall ist die Menge der zur Verfügung stehenden Trainingsdaten sehr gering. Daher ist es notwendig, Erkenner für sinnvolle Kombinationen von akustischen Bedingungen anzupassen. Die Entwicklung von mehreren Spracherkennern unter Berücksichtigung aller sinnvollen Anpassungen an die jeweils zu erwartenden akustischen Bedingungen ist sehr zeitaufwendig. Aus diesem Grund konnte im Rahmen dieser Diplomarbeit nur eine kleine Anzahl von Spezialerkennern verwendet, und nicht alle für die Entwicklung dieser Erkenner denkbaren Möglichkeiten der Anpassung ausgeschöpft werden. Das ist aber auch nicht nötig, um die Einflüsse automatischer Segmentierung und Klassifizierung sowie der Verwendung mehrerer Spracherkenner zu untersuchen.

Werden mehr Trainingsdaten verfügbar, so ist es jederzeit möglich, die einzelnen Erkenner zu verbessern, das System durch weitere Erkenner zu verfeinern und gegebenenfalls den Segmentierer anzupassen (um z.B. Segmentgrenzen zwischen Sprechern mit und ohne Akzent zu erkennen).

6.1.1 Erkenner für Klassen von akustischen Bedingungen

Damit ein Spracherkenner gute Erkennungsergebnisse liefern kann, ist es notwendig, daß die akustischen Modelle des Erkenners gut zu den Merkmalen passen, die aus den Audiodaten extrahiert werden, auf denen Sprache erkannt werden soll. Je größer also die Übereinstimmung der statistischen Eigenschaften der aus den Test- und Trainingsdaten extrahierten Merkmalsvektoren ist, um so besser wird das Erkennungsergebnis sein.

Es ist daher sinnvoll, solche Audiodaten für das Training eines Spezialerkenners zusammenzufassen, für die die extrahierten Merkmalsvektoren möglichst ähnliche statistische Eigenschaften haben, und verschiedene Spezialerkenner für Klassen von Einflüssen zu entwickeln, die sich sehr unterschiedlich auf die extrahierten Merkmale auswirken.

Die eben formulierten Anforderungen an die akustischen Bedingungen, für die eigens Erkenner trainiert werden sollen, entsprechen genau den in Abschnitt 4.3.1 geforderten Eigenschaften für Segmentierungsklassen. Um diesen Anforderungen gemäß Klassen von akustischen Bedingungen zu erzeugen, kann wieder die dort beschriebene Vorgehensweise gewählt werden.

6.1.2 Automatische Erzeugung einer Klasseneinteilung

Die Einteilung der akustischen Bedingungen in Klassen für das Trainieren von Spezialerkennern wurde durch den in Abschnitt 4.3.1 beschriebenen Ansatz vorgenommen.

Im Zusammenhang mit der automatischen Klassenbildung für die Segmentierung hat sich gezeigt, daß sich die Sprache von Männern und Frauen stark unterscheidet. Es wäre also sinnvoll, eigene Erkenner für männliche und weibliche Sprecher einzusetzen. Als Ausgangspunkt für die Entwicklung der Spezialerkenner wurde jedoch, wie weiter unten erläutert wird, ein geschlechtsunabhängiges Spracherkennungssystem verwendet. Die akustischen Modelle dieses Systems wurden mit Sprache von Männern und Frauen trainiert und spiegeln daher die statistischen Eigenschaften der Sprache von Sprechern beiderlei Geschlechts wieder.

Aus diesem Grund wurde bei der Klassenbildung für die Entwicklung von spezialisierten Spracherkennern keine Unterscheidung der Sprache von Männern und Frauen gemacht. Das wurde erreicht, indem bei dem in Abschnitt 4.3.1 beschriebenen Klassenbildungsprozeß keine Fragen erlaubt wurden, die eine Trennung von männlichen und weiblichen Sprechern zulassen.

Diese Fragen sind in Anhang B angegeben. Der unter Verwendung dieser Fragen erzeugte Baum ist in Abbildung 6.1 dargestellt.

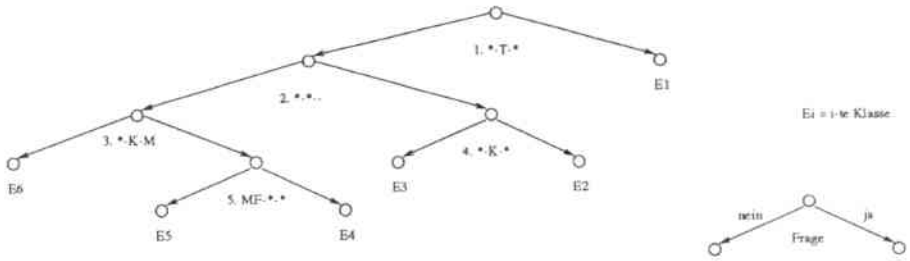

Abbildung 6.1: Klasseneinteilung E1-E6, die mit Hilfe von Fragen, die keine Unterscheidung männlicher und weiblicher Sprecher erlauben, automatisch erzeugt wurde. Für die Bildung der Klasseneinteilung wurden akustische Modelle über allen 5 Sendungen der Trainingsmenge berechnet. Die zur Aufspaltung verwendeten Fragen sind am jeweiligen Knoten angegeben und entsprechen der in Abschnitt 4.2.1 beschriebenen Konvention

Die automatisch erzeugte Klasseneinteilung legt nahe, spezielle Erkenner für folgende Klassen zu entwickeln:

- Sprache über Telephonverbindungen (E1),
- Sprache über einen klaren Kanal ohne Hintergrundgeräusche (E2),
- Sprache über einen verzerrten Kanal ohne Hintergrundgeräusche (E3),
- Sprache über einen klaren Kanal mit reiner Hintergrundmusik(E4),
- alle übrigen Klassen, die Sprache enthalten; also Sprache über klaren und verzerrten Kanal mit Hintergrundgeräuschen mit und ohne Hintergrundmusik (E6).

Segmente, die nur Musik über einen klaren Kanal enthalten (E5), werden schon im ersten Segmentierungsschritt (siehe Abschnitt 5.4) markiert und entfernt. Abschnitte ohne Sprache, die in den anderen Klassen enthalten sind, werden nicht gesondert behandelt, da solche Abschnitte in den Trainingssendungen insgesamt nur sehr selten vorkommen und zum Großteil beim zweiten Segmentierungsschritt entfernt werden. Verbleibende Abschnitte ohne Sprache sollten auf die akustischen Modelle für Stille der einzelnen Erkenner abgebildet werden.

In Hinblick auf die geringe Menge an Trainingsdaten wurden nicht für alle fünf der oben aufgezählten Klassen Erkenner entwickelt, sondern nur für diejenigen Klassen, die bei Aufspaltung durch die ersten zwei Fragen in Abbildung 6.1 entstehen, d.h. für

- Sprache über Telephonverbindungen (E1),
- Sprache über einen klaren oder verzerrten Kanal ohne Hintergrundgeräusche (E2 und E3),

- alle übrigen Klassen, die Sprache enthalten; also Sprache über einen klaren oder verzerrten Kanal mit Hintergrundgeräuschen (E4 und E6).

6.2 Training von Spezialerkennern

In den zur Zeit leistungsfähigsten Spracherkennungssytemen werden Phoneme im Kontext mehrerer anderer Phoneme, sogenannte *Polyphone*, durch eigene Modelle repräsentiert. Üblich sind hier Triphone; in neueren Systemen werden jedoch auch größere Kontexte verwendet. Diese Modelle bestehen meistens selbst aus mehreren HMM Zuständen. In einem solchen System sind aufgrund der großen Anzahl verschiedener Phonemkontexte sehr viele Parameter zu bestimmen. Selbst in sehr großen Trainingsmengen kommen viele der Phonemkontexte selten oder nie vor. Um dennoch Modelle zu erhalten, für die eine ausreichende Anzahl von Trainingsbeispielen existieren, müssen geeigente Maßnahmen getroffen werden. Einige Möglichkeiten werden in [29] beschrieben. Es können z.B. HMM Zustände, deren Emissionswahrscheinlichkeitsverteilungen ähnlich sind, Parameter gemeinsam benutzen, oder Parameter von Modellen, für die wenig Trainingsdaten vorliegen, können mit den Parametern ähnlicher, aber besser trainierter Modelle interpoliert werden.

6.2.1 Das WSJ System als Ausgangspunkt

Im vorliegenden Fall ist es aufgrund der sehr kleinen Trainingsmenge nahezu unmöglich, einen (und schon gar nicht mehrere) Spracherkenner vollständig durch Viterbi-Training oder den Baum-Welch-Algorithmus zu trainieren. Eine Strategie in dieser Situation ist, ein existierendes Spracherkennungssystem zu verwenden, das verhältnismäßig gut zu den vorliegenden Gegebenheiten paßt, und dieses Spracherkennungssystem durch Verfahren zu verändern, bei denen die akustischen Modelle trotz der wenigen Trainingsdaten zuverlässig und gleichmäßig in größere Übereinstimmung mit den vorliegenden akustischen Bedingungen gebracht werden.

Da in Marketplace-Sendungen vor allem geschulte Sprecher vorkommen, große Teile der Sendungen gelesene Sprache enthalten und der Schwerpunkt der Sendungen auf Wirtschaftsnachrichten liegt, sind gelesene Wirtschaftsnachrichten der in Marketplace-Sendungen vorkommenden Sprache ähnlich. In den zurückliegenden 5 Jahren wurden Spracherkenner oft auf der sogenannten *Wall Street Journal (WSJ)* Domäne miteinander verglichen. Hierbei handelt es sich um gelesene Zeitungsartikel aus dem *Wall Street Journal*. Der WSJ Trainingskorpus ist sehr groß, und daher kann ein Spracherkennungssystem mit vielen freien Parametern unter Verwendung dieser Daten zuverlässig trainiert werden.

Aus diesen Gründen wurde das WSJ Spracherkennungssystem des Lehrstuhls von Prof. Waibel an der Universität Karlsruhe [38] als Ausgangspunkt

für die Entwicklung der spezialisierten Marketplace-Spracherkenner verwendet.

6.2.2 Beschreibung des WSJ Systems

Als Basis für die Entwicklung von auf bestimmte akustische Klassen spezialisierten Spracherkennern wurde das neueste[1] WSJ System genommen. Hierbei handelt es sich um ein System mit folgenden Eigenschaften:

- Es werden Polyphone modelliert mit einem Kontext von bis zu drei vorausgehenden und nachfolgenden Phonemen, sogenannte *Septphone*.
- Phonemkontexte über Wortgrenzen hinweg werden berücksichtigt. Hierdurch wird man Koartikulationseffekten zwischen angrenzenden Worten gerecht. Das erhöht die Anzahl der vorkommenden Polyphone gegenüber dem Fall, in dem nur Phonemkontexte innerhalb eines Worts berücksichtigt werden, beträchtlich.
- Polyphonmodelle und das Modell für Störgeräusche, auf das beim Training Geräusche wie z.B. Papierrascheln abgebildet werden, bestehen aus drei Zuständen (Beginn-, Mittel- und Endzustand). Das Modell für Stille besteht nur aus einem Zustand.
- Der Prozeß zur Bestimmung der Parameter der Polyphonmodelle des Erkenners verlief in folgenden Schritten:

  1. Jeweils für die drei Zustände von 45 kontextunabhängigen Phonemmodellen wurde eine Mixtur von Normalverteilungen mit 16 Komponenten berechnet. Das ergab insgesamt 135 Mixturen.

     Von nun an wird der Kürze halber, wie in diesem Zusammenhang üblich, für die Mittelwertvektoren und Kovarianzmatrizen einer Mixtur der Begriff Codebuch verwendet. Im Zusammenhang mit Vektor-Quantisierung wird durch ein Codebuch eine Abbildung von Vektoren auf diskrete Werte festgelegt. Bei Mixturen zur Modellierung von Merkmalsvektoren treten gewissermaßen die Kovarianzmatrizen und Mittelwertvektoren an die Stelle der Codebücher. Mit Verteilungen über einem Codebuch sind die Mixturgewichte für die einzelnen Mixturkomponenten gemeint.
  2. Für die Modelle der verschiedenen in der Trainingsmenge vorkommenden Polyphone wurde für jeden Zustand eine Verteilung über dem Codebuch des entsprechenden kontextunabhängigen Phonemzustands berechnet. Das ergab insgesamt ca. 574000 Verteilungen über diesen 135 Codebüchern.

[1] im August 1996

3. Durch ein hierarchisches *Clustering*-Verfahren wurden solche Phonemkontexte zu Klassen zusammengefaßt, deren Verteilungen sich in Bezug auf ein Abstandsmaß ähnlich waren. Das Verfahren, das hierbei verwendet wurde, entspricht dem in Abschnitt 4.3.1 beschriebenen: Aus den Verteilungen der verschiedenen Phonemkontexte über einem Codebuch wurde eine gemeinsame Verteilung berechnet. Dann wurde anhand einer Menge von sich auf Phonemkontexte beziehenden Fragen diese Verteilung iterativ so aufgespaltet, daß in jedem Schritt ein Abstandsmaß maximiert wurde.

   Durch die Abfolge von Fragen, die beim iterativen Prozeß zum Aufspalten verwendet wurden, wurde ein Entscheidungsbaum erzeugt, der beim Training dazu dient, für einen Phonemkontext dasjenige Modell auszusuchen, das durch die Trainingsbeispiele angepaßt werden soll. Beim Test kann derselbe Entscheidungsbaum verwendet werden, um für während des Trainings nicht gesehene Phonemkontexte das geeignete Modell auszusuchen. Für eine detailierte Erörterung dieser Vorgehensweise siehe [47, 29, 49].

   Das Ergebnis dieses Schrittes war ein Entscheidungsbaum mit 4000 Blättern, die Klassen von Phonemkontexten repräsentieren.

4. Für diese Klassen von Phonemkontexten wurden in einem abschließenden Schritt Mixturen berechnet.

- Als Trainingsmenge für die akustischen Modelle wurde die sogenannte SI-284 Teilmenge des Wall Street Journal Korpus verwendet. Diese umfaßt die SI-84 Teilmenge des WSJ0 Korpus sowie 200 weitere Sprecher aus dem WSJ1 Korpus. Insgesamt sind das 45000 Sätze gesprochen von 141 Frauen und 143 Männern. Die akustischen Modelle wurden mit Daten von männlichen und weiblichen Sprechern trainiert. Das verwendete WSJ System ist also geschlechtsunabhängig.

- Die Übergangswahrscheinlichkeiten der HMMs haben einen festen Wert und wurden beim Training nicht angepaßt. Weiterhin findet keine Modellierung der Dauer eines Lauts (*duration modeling*) statt.

- Als Vorverarbeitung werden im Zeitfenster i 16 Melscale-Spektralkoeffizienten $\boldsymbol{mel}_i$ berechnet (siehe Abschnitte 2.2.2 und 5.1.2). Für jedes Segment wird der Mittelwert $\overline{\boldsymbol{mel}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{mel}_i$ dieser Koeffizienten ermittelt und von den einzelnen $\boldsymbol{mel}_i$ abgezogen. Dieser Schritt dient dazu, additive Störanteile im logarithmierten Frequenzbereich zu entfernen. Weiterhin wird eine Normalisierung der Amplitude durchgeführt. Die so entstehenden Vektoren $\boldsymbol{w}_i$ werden jeweils mit den 3 vorausgehenden und den 3 nachfolgenden ($\boldsymbol{w}_{i-3}, \boldsymbol{w}_{i-2}, \boldsymbol{w}_{i-1}, \boldsymbol{w}_{i+1}, \boldsymbol{w}_{i+2}$ und $\boldsymbol{w}_{i+3}$) zu einem 112 dimensionalen Vektor zusammengefaßt und,

wie in Abschnitt 5.1.1 beschrieben, durch eine lineare Transformation auf einen 48 dimensionalen Vektor abgebildet.

6.2.3 Training durch Adaption

Bei den Trainingverfahren für HMMs (Baum-Welch-Algorithmus oder Viterbi-Training; siehe Abschnitt 2.2.3) werden die Parameter eines Modells durch diejenigen Merkmalsvektoren der Trainingsmenge angepaßt, die bei einer Trainingsiteration auf das Modell entfallen. Aufgrund der wenigen Trainingsdaten würden im gegebenen Fall viele der Modelle des WSJ Systems nur unter Verwendung sehr weniger Trainingsbeispiele, also sehr unzuverlässig, oder in manchen Fällen gar nicht verändert.

Die akustischen Modelle würden sehr ungleichmäßig trainiert. In einem Fall wie diesem bietet es sich an, Trainingsdaten nicht zum Anpassen nur eines bestimmten Modells zu verwenden, sondern zum Berechnen einer Transformation für mehrere oder alle Modelle. Die Transformation wird hierbei so bestimmt, daß die Modelle des Erkenners in größere Übereinstimmung mit den extrahierten Merkmalsvektoren der Trainingsmenge gebracht werden. Diese Vorgehensweise wird Training durch Adaption genannt und z.B. in [28, 17, 18, 19] beschrieben.

Durch ein solches Training wurden die akustischen Modelle des WSJ Systems an die akustischen Bedingungen in den Marketplace-Trainingssendungen angepaßt. Adaption kann auf viele verschiedene Arten durchgeführt werden. Im Rahmen der Diplomarbeit wurde der sogenannte MLLR Ansatz verwendet (siehe [17, 18, 19]).

Bei der in *JRTk* implementierten Ausprägung von MLLR wird in einem ersten Schritt ein Baum erzeugt, der eine Klassenhierarchie für Klassen von akustischen Modellen darstellt. Liegen sehr wenige Trainingsdaten vor, kann nur eine grobe Transformation für die Gesamtmenge aller im Wurzelknoten des Baums enthaltenen Modelle berechnet werden. Was in diesem Zusammenhang *sehr wenig* bedeutet, wird durch einen Parameter der Adaption festgelegt. Vom Wurzelknoten ausgehend werden Teilmengen von der Gesamtmenge der Modelle abgespaltet. Liegen ausreichend viele Trainingsdaten für eine solche Teilmenge von Modellen vor, wird diese durch eine eigenen Transformation adaptiert. Je mehr Trainingsdaten und je tiefer in der Klassenhierarchie noch ausreichend viele Trainingsbeispiele für einzelne Klassen vorliegen, desto feiner wird die Adaption sein.

Durch die Adaption selbst werden nur die Mittelwertvektoren $\boldsymbol{\mu}_i$ der Mixturen verändert. Es wird eine Transformation der Form $\hat{\boldsymbol{\mu}}_i := A\boldsymbol{\mu}_i + \boldsymbol{a}$ berechnet. A und $\boldsymbol{a}$ werden dabei so bestimmt, daß die Wahrscheinlichkeit für die Trainingsbeispiele unter Verwendung der neuen Mittelwertvektoren $\hat{\boldsymbol{\mu}}_i$ maximiert wird.

Die Details der Berechnung der sogenannten Regressionsklassen in der Klassenhierarchie und der Adaptionstransformation sind in [17, 18, 19] beschrieben. Der große Vorteil dieser Art der Adaption ist, daß abhängig von

der Menge der zur Verfügung stehenden Trainingsdaten der Feinheitsgrad der Adaption bestimmt wird.

6.3 Auswahl der Spezialerkenner

Hat man für bestimmte Klassen von akustischen Bedingungen Spezialerkenner entwickelt, so muß man für ein Segment ermitteln, welcher dieser Klassen die akustischen Bedingungen in dem Segment am meisten entsprechen. Gesucht ist also eine Klassenzuordnung des Segments.

Da die Segmentierung durch eine Klassenzuordnung der Merkmalsvektoren vorgenommen wurde, kann diese Zuordnung verwendet werden, um den geeigneten Spracherkenner für ein Segment auszusuchen, falls eine Abbildung der Segmentierungsklassen auf diejenigen Klassen möglich ist, für die spezielle Spracherkenner existieren. Im gegebenen Fall ist das nicht immer möglich, da, wie Abschnitt 5.4 erläutert, keine Unterscheidung der dort verwendeten Klassen A1', A2' und A9' zugelassen wurde. Dadurch ist keine Unterscheidung der Sprache von Männern über einen klaren Kanal mit und ohne Hintergrundgeräusche und der Sprache von Männern über einen verzerrten Kanal ohne Hintergrundgeräusche und mit Hintergrundgeräuschen ohne Musik möglich.

Da nach dem ersten Segmentierungsschritt durch eine zweite auf Stille basierende Segmentierung weitere Grenzen gesetzt werden, ist es sinnvoll, Segmente, die der aus A1', A2' und A9' bestehenden Klasse von akustischen Bedingungen angehören, erneut zu klassifizieren und hierbei entsprechend den akustischen Klassen der Spracherkenner zu unterscheiden. Das Trainieren der Klassifikatoren kann wieder wie in Abschnitt 5.1.1 erfolgen, die Klassfizierung von Segmenten wie in Abschnitt 4.1.4.

Ein anderer Ansatz zur Klassfizierung ist, einfach alle spezialisierten Erkenner auf den Segmenten zu verwenden und das Erkennungsergebnis auszuwählen, für das die A-posteriori-Wahrscheinlichkeit (siehe Abschnitt 2.2.3) am größten ist. Die Maximierung dieser Wahrscheinlichkeit ist das in Abschnitt 2.2.3 formulierte Ziel eines Spracherkenners. Aufgrund der sehr langen Laufzeiten der Spracherkenner wurde im vorliegenden Fall davon abgesehen, so vorzugehen.

Kapitel 7

Spracherkennungsversuche

Durch die hier beschriebenen Versuche sollten die Einflüsse von automatischer Segmentierung und Klassifizierung sowie der Verwendung mehrerer spezialisierter Spracherkenner auf die Wortfehlerrate des gesamten Systems zur Transkription von Sprache auf Marketplace-Nachrichtensendungen untersucht werden.

7.1 Systeme und Testbedingungen

Um den Einfluß der einzelnen Komponenten des Systems auf die Wortfehlerrate analysieren zu können, wurden insgesamt 7 Versuche mit unterschiedlichen akustischen Modellen gemacht.

7.1.1 Systemkomponenten

Bei der späteren Beschreibung der Versuchsbedingungen werden die folgenden Abkürzungen und Begriffe verwendet.

Sprachmodell BN92-95: Ein Sprachmodell hat die Aufgabe, die Wahrscheinlichkeiten von Wortfolgen zu modellieren. Das in den später beschriebenen Versuchen verwendete Sprachmodell wurde über Transkriptionen von amerikanischen Nachrichtensendungen erstellt und paßt daher gut zu der Testdomäne.

Die für die Erstellung des Sprachmodells benutzten Transkriptionen werden von *Primary Source Media*[1] kommerziell auf CDs vertrieben. Auf den verwendeten CDs befinden sich Texte zu Sendungen der Sender ABC, CNN, NPR und PBS aus dem Zeitraum von 1992 bis 1995, die insgesamt 165 Millionen Worte umfassen. Dieser Korpus an Daten wird von nun an mit BN92-95 bezeichnet.

Die bereits vorverarbeiteten Textdaten wurden dankenswerterweise von Kristie Seymore und Roni Rosenfeld an der Carnegie Mellon University

[1] früher *Research Publications Inc.*

zur Verfügung gestellt. Diese Textdaten enthielten noch eine Anzahl von Fehlern, die in einer späteren Version nicht mehr vorhanden waren. Aus Zeitgründen konnten nicht mehr die bereinigten Daten verwendet werden. Bei einer Fortführung des Projekts sollte ein Vergleich stattfinden, welche Auswirkungen die Fehler hatten.

Für diese Diplomarbeit wurde ein Trigramm-Sprachmodell (siehe Abschnitt 2.2.4) unter Verwendung des in [14] beschriebenen *back-off* Verfahrens, nicht-linearer Interpolation und sogenanntem *absoluten Discounting* erstellt (siehe hierzu [27]). Im BN92-95 Korpus kamen weit über hunderttausend Worte vor, viele von diesen sehr selten. Zur Berechnung des Sprachmodells wurden die häufigsten 50000 Worte dieses Korpus' verwendet und ergänzt durch diejenigen Worte aus den 10 Marketplace-Trainingssendungen, die in den 50000 Worten nicht enthalten waren.

Wortliste: In Spracherkennungssystemen, in denen Phoneme modelliert werden, muß eine Phonemtranskription existieren, damit ein Wort erkannt werden kann. Nur so kann im Spracherkenner ein Modell für ein zu erkennendes Wort zusammengesetzt werden. Es ist also eine Auswahl zu treffen, welche Worte während eines Erkennungslaufs zur Verfügung gestellt werden. Diese Auswahl wird im folgenden Wortliste genannt.

Die Zusammensetzung und Größe dieser Liste beeinflußt die Erkennungsleistung eines Spracherkennungssystems stark. Kommen Worte in den Audiodaten vor, die nicht in der Wortliste eingetragen sind, so wird für diese zwangsläufig jeweils mindestens ein Erkennungsfehler begangen. Trägt man auf der anderen Seite sehr viele Worte in die Wortliste ein, so nimmt die Anzahl der Verwechslungsmöglichkeiten zu und der Suchraum wird größer, was wiederum bedeutet, daß die Erkennung länger dauert und unter Umständen die Erkennungsleistung sinkt.

Es ist also wichtig, eine geeignete Größe für die Wortliste zu wählen und die Auswahl von Worten so zu treffen, daß möglichst die in der Testdomäne vorkommenden Worte gut abgedeckt werden

Wortliste BN92-95-20k mit Varianten: Zur Anpassung eines Spracherkennungssystems wählt man im allgemeinen eine geeignete Größe für die Wortliste anhand von Tests über einer Entwicklungstestmenge aus. Das war aus Zeitgründen im gegebenen Fall leider nicht möglich. Es wurden daher die häufigsten 20000 Worte des BN92-95 Korpus' sowie in den 10 Marketplace-Trainingssendungen zusätzlich vorkommende Worte[2] (ca. 440), verwendet.

Wie oben beschrieben, ist die Aufgabe der Wortliste und der zugehörigen Phonemtranskriptionen, den Aufbau von HMMs für die zu erken-

[2] zu einem Großteil Eigennamen und Begriffe des aktuellen Zeitgeschehens aus der Zeit der Ausstrahlung der Sendungen der Trainingsmenge

nenden Worte zu ermöglichen. Viele Worte werden auf verschiedene Weise ausgesprochen, d.h. es gibt nicht nur eine, sondern mehrere mögliche Phonemtranskriptionen. Um dieser Tatsache gerecht zu werden, werden in Spracherkennungssytemen meistens verschiedene Transkriptionen für solche Aussprachevarianten eines Wortes verwendet. Einschließlich dieser Varianten umfaßt die in den später beschriebenen Versuchen verwendete Wortliste 26419 Worte.

akustische Modelle I (AM-WSJ): Die akustischen Modelle des WSJ Systems, das als Ausgangspunkt für die Entwicklung der spezialisierten Spracherkenner verwendet wurde, werden mit AM-WSJ bezeichnet.

akustische Modelle II (AM-1): Für einen der in den Versuchen verwendeten Spracherkenner wurden die akustischen Modelle des WSJ Systems durch Adaption (siehe Abschnitt 6.2.3) auf allen Sprache enthaltenden Segmenten der Trainingssendungen ohne Unterscheidung der akustischen Bedingungen angepaßt. Die so erzeugten akustischen Modelle werden im folgenden AM-1 genannt.

akustische Modelle III (AM-3-VORGEG): Auf den Evaluationstestdaten sind Segmentgrenzen vorgegeben. Die Segmente sind mit Markierungen bezüglich des Vorhandenseins von Musik und der im Segment verwendeten Bandbreite des Audiosignals versehen. Teilt man die Segmente der Trainingsmenge diesen Markierungen entsprechend ein, so erhält man die Klassen *Sprache mit Hintergrundmusik und voller Bandbreite (HM)*, *Sprache mit reduzierter Bandbreite (RB)* und *Sprache mit voller Bandbreite ohne Hintergrundmusik (VB)*. Sprache mit reduzierter Bandbreite und Hintergrundmusik kommt in den Trainingsdaten nicht vor.

Diese Unterteilung der Daten in Klassen unterscheidet sich von der Klasseneinteilung, die in Abschnitt 6.1.2 automatisch durch Aufspaltung anhand von 2 Fragen erzeugt wird.

Um überprüfen zu können, ob die automatisch erzeugte Klasseneinteilung für die Entwicklung von Spezialerkennern besser geeignet ist als die auf der Evaluationsmenge vorgegebene, wurden die akustischen Modelle des WSJ Systems für die Klassen VB, RB, HM mittels Adaption trainiert. Das ergab drei Mengen von akustischen Modellen. Diese Mengen werden im folgenden mit AM-3-VORGEG bezeichnet.

akustische Modelle IV (AM-3-AUTO): Durch automatische Klassenbildung ohne Unterscheidung männlicher und weiblicher Sprecher wurden bei Aufspaltung durch 2 Fragen drei Klassen von akustischen Bedingungen erzeugt. Wie in Abschnitt 6.1.1 und 6.1.2 begründet, wird erwartet, daß diese Klasseneinteilung am besten für die Entwicklung von drei Spezialerkennern geeignet ist. Gemäß dieser Klasseneinteilung wurden die akustischen Modelle des WSJ Systems adaptiert. Das

Ergebnis waren wieder drei Mengen von akustischen Modellen, die AM-3-AUTO genannt werden.

7.1.2 Testbedingungen

Zur Bewertung der einzelnen Komponenten des gesamten Systems sind folgende Fragen von Interesse:

1. Wie gut sind die untrainierten akustischen Modelle des WSJ System für die Marketplace-Domäne geeignet?

2. Wie sehr verbessert sich die Erkennungsleisung, wenn die akustischen Modelle des WSJ Systems mit allen Audiodaten der Marketplace-Trainingssendungen adaptiert werden, man jedoch nicht zwischen den verschiedenen akustischen Bedingungen unterscheidet?

3. Bringt die Verwendung von drei spezialisierten Erkennern statt eines einzigen einen Gewinn?

4. Ist es besser, die vorgegebenen Markierungen der Audiodaten für die Auswahl von spezialisierten Erkennern zu verwenden, oder sollte man automatisch für Segmente die Klassenzugehörigkeit bestimmen?

5. Welche Klasseneinteilung in drei Klassen ist besser für die Entwicklung von Spezialerkennern geeignet: die auf der Evaluationsmenge vorgegebene oder die automatisch erzeugte?

6. Welchen Einfluß hat die automatische Segmentierung auf das Erkennungsergebnis?

Um diese Fragen zu beantworten, wurde eine Anzahl von Versuchen zusammengestellt. Einige der Versuche konnten aus Zeitgründen nicht über der gesamten Evaluationsmenge durchgeführt werden. Um dennoch verhältnismäßig repräsentative Ergebnisse zu bekommen, wurden die auf der Evaluationsmenge vorgegebenen Segmente in eine zufällige Reihenfolge gebracht und die ersten 86 der 202 Segmente ausgewählt. Die so gebildete Testmenge (EVAL/3) entspricht vom Umfang her mehr als einem Drittel der Evaluationsmenge (EVAL), und die Zeitanteile der Klassen VB, RB und HM in EVAL/3 und EVAL sind in etwa gleich groß.
Die durchgeführten Versuche unterscheiden sich voneinander durch

- die akustischen Modelle, die für die Erkennung benutzt werden (AM-WSJ, AM-1, AM-3-VORGEG oder AM-3-AUTO),

- die Testmenge (EVAL/3 oder EVAL),

- automatische Klassifizierung der Segmente oder Verwendung der vorgegebenen Markierungen (autom. oder vorgeg.),

- automatische Segmentierung oder Verwendung der vorgegebenen Segmentgrenzen (autom. oder vorgeg.).

Gemeinsam ist in allen Versuchen die Wortliste BN92-95-20k mit Varianten, das Sprachmodell BN92-95 sowie die 4000 Klassen von Phonemkontexten, die durch eigene Mixturen modelliert werden.

In folgender Tabelle sind die Versuche den Unterscheidungsmerkmalen gemäß beschrieben und die Versuchsergebnisse angegeben. Ein - steht in der Tabelle für Einträge, die keine Bedeutung für den Versuch haben. Die Versuchsergebnisse sind als *Wortfehlerrate (WF)* angegeben; die WF ist die Summe der Anteile von Substitutionen, Einfügungen und Auslassungen an den insgesamt gesagten Worten.

Spracherkennungsexperimente					
Versuch	akustische Modelle	Test-menge	Klassen-zuordnung	Segmen-tierung	WF
I	AM-WSJ	EVAL/3	-	vorgeg.	46.1 %
II	AM-1	EVAL/3	-	vorgeg.	42.6 %
III	AM-3-VORGEG	EVAL/3	vorgeg.	vorgeg.	42.3 %
IV	AM-3-VORGEG	EVAL/3	autom.	vorgeg.	40.6 %
V	AM-3-VORGEG	EVAL	autom.	vorgeg.	42.5 %
VI	AM-3-VORGEG	EVAL	autom.	autom.	42.0 %
VII	AM-3-AUTO	EVAL	autom.	autom.	41.9 %

7.2 Versuchsauswertung

In folgender Tabelle sind die Versuchsergebnisse der Versuche I bis VII noch einmal als Wortfehlerrate (WF) dargestellt, diesmal aufgeschlüsselt nach dem Vorhandensein von Musik und der verwendeten Bandbreite.

Spracherkennungsexperimente					
Versuch	Gesamt	Hintergrundmusik		Bandbreite	
		Nein	Ja	voll	reduziert
I	46.1%	41.6%	73.4%	43.3%	56.9 %
II	42.6%	39.1%	64.1%	39.6%	53.8 %
III	42.3%	39.5%	59.9%	39.4%	53.4 %
IV	40.6%	37.6%	58.9%	39.5%	44.7 %
V	42.5%	39.8%	61.2%	37.4%	57.8 %
VI	42.0%	39.5%	59.8%	36.7%	57.7 %
VII	41.9%	39.3%	59.9%	36.7%	57.5 %

Versuche I und II Die deutlichste Verbesserung insgesamt wird durch die Adaption der akustischen Modelle an die Verhältnisse in den Marketplace-Trainingssendungen erzielt.

Versuche II, III und IV Die Verwendung von drei Erkennern bei vorgegebenen Segmentmarkierungen bringt gegenüber nur einem adaptierten Erkenner kaum eine Steigerung der Erkennungsleistung (Versuche II und III). Eine ziemlich deutliche Verringerung der Wortfehlerrate wird jedoch durch automatische Klassifizierung der Segmente erreicht (Versuche III und IV). Offenbar sind die von Menschen vorgegebenen Markierungen ungeeignet, um den passenden Erkenner für ein Segment auszusuchen. Das ist erstaunlich, da ja nur drei Klassen mit jeweils recht eindeutigen Merkmalen unterschieden werden. Deshalb wäre eigentlich zu vermuten, daß Menschen dieselbe Klassenzuordnung treffen, die auch automatisch durch Klassifikatoren vorgenommen wird. Der Grund muß sein, daß Menschen die Klassenzuordnung nicht aufgrund der auch für die Spracherkennung verwendeten Merkmale vornehmen.

In folgender Tabelle ist angegeben, welche Segmente von den vorgegebenen Markierungen abweichend klassifiziert wurden. RM bezeichnet Segmente die reine Musik enthalten. Solche sind unter den auf der Evaluationsmenge vorgegebenen Segmenten nicht enthalten. Ein Klassifizierer für diese Klasse wurde dennoch verwendet, um eine mögliche Verwechselbarkeit von RM und den anderen Klassen zu prüfen.

Konfusionsmatrix				
automatisch gefunden	Von Menschen vorgegeben			
	VB	RB	HM	RM
VB	47	6	10	0
RB	0	6	0	0
HM	1	0	16	0
RM	0	0	0	0

Aus den aufgeschlüsselten Spracherkennungsergebnissen geht hervor, daß die Verbesserung von Versuch IV gegenüber Versuch III vor allem durch Segmente verursacht wurde, die mit Markierungen für Sprache über Telephonleitungen versehen sind. Es wurde also offenbar in den 6 Fällen, in denen Telephonsegmente durch akustische Modelle für die Klasse VB erkannt wurden, deutlich mehr Worte korrekt erkannt.

Versuche IV und V Versuch V enthält Versuch IV als Teilversuch. Man sieht, daß die Teilmenge EVAL/3 offenbar nicht ganz repräsentativ für die gesamte Evaluationstestmenge EVAL ist.

Versuch V und VI Bei der Entwicklung des Segmentierers wurde davon ausgegangen, daß Segmentierungsfehler bei der Spracherkennung Fehler verursachen. Erstaunlicherweise wurde im gegebenen Fall durch automatische Segmentierung nicht nur keine Verschlechterung der Wortfehlerrate bewirkt, sondern sogar eine geringfügige Verbesserung.

Die in obiger Tabelle angegebene Aufschlüsselung der Spracherkennungsergebnisse hat ergeben, daß Versuch V in erster Linie für Sprache mit Hintergrundmusik schlechtere Worterkennungsraten erzielt hat als der vergleichbare Versuch VI. Eine Analyse der Erkennungsergebnisse hat gezeigt, daß bei lauter Hintergrundmusik viele Auslassungsfehler gemacht werden. Bei lauter Musik passen die akustischen Modelle des Erkenners schlecht zu den Audiodaten. Die Verwechselbarkeit von Worten nimmt zu, da die akustischen Modelle für die verschiedenen Worte gleichermaßen schlecht zu den Audiodaten passen. Aufgrund der großen Verwechselbarkeit und der Tatsache, daß durch das Sprachmodell Wortübergänge bestraft werden, kommt es zu den Auslassungsfehlern. In einem längeren Segment, das ein solches mit Musik verunreinigtes Segment als Teil enthält, kann der Kontext dazu führen, daß zusätzliche Worte in dem Teilsegment korrekt erkannt werden.

Versuch VI und VII Spezialerkenner entsprechend der automatisch gefundenen Klasseneinteilung bringen gegenüber Spezialerkennern, die gemäß der auf der Evaluationsmenge vorgegebenen Klasseneinteilung trainiert wurden, nur eine unwesentliche Verbesserung der Wortfehlerrate (Versuche VI und VII). Das ist sicher auf die kleine Trainingsmenge und die Tatsache zurückzuführen, daß nur wenig Entwicklungsaufwand in die einzelnen Erkenner gesteckt werden konnte.

Es kann daher nicht mit Bestimmtheit gesagt werden, daß die automatisch gefundene Klasseneinteilung besser für das Training von Spezialerkenner geeignet ist als die auf der Evaluationsmenge vorgegebene.

Kapitel 8

Zusammenfassung und Bewertung

Die Aufgabe, ein Spracherkennungssystem für Marketplace-Nachrichtensendungen zu entwickeln, war Gegenstand der sogenannten ARPA HUB-4 Evaluation 1995. Das erlaubt es, das im Rahmen dieser Diplomarbeit entwickelte System mit den Systemen anderer Forschungsgruppen zu vergleichen.

Sprache über Telephonverbindungen, mit Hintergrundmusik oder über verschiedene Übertragungskanäle zu erkennen, wird schon seit einiger Zeit untersucht. Was im Rahmen der HUB-4 Evaluation 1995 neu hinzukam, war das Problem der Inhomogenität der Audiodaten. Daher ist die Segmentierung und die Klassifizierung der Segmente in diesem Zusammenhang von besonderer Bedeutung.

In diesem Kapitel werden die Systeme der Teilnehmer der Evaluation kurz vorgestellt und mit dem für diese Diplomarbeit gewählten Ansatz verglichen. Hierbei wird ein besonderes Gewicht auf die Segmentierungskomponente gelegt. Anschließend wird der gewählte Ansatz bewertet und mögliche Erweiterungen des Systems diskutiert.

8.1 Die Systeme der HUB-4 Evaluation 1995

Im Rahmen der ARPA HUB-4 Evaluation 1995 wurden Systeme von IBM [10, 9, 8], Dragon [44], BBN [15] und der CMU Robust Speech Recognition Group [13] getestet. Bei der Entwicklung dieser Systeme wurden die Schwerpunkte sehr unterschiedlich gesetzt. Auch die Art, wie und auf welchen Daten die Systeme in der Entwicklungsphase ausgewertet wurden, unterscheidet sich stark. Daher ist es nicht einfach, Vergleiche zwischen den Komponenten der verschiedenen Systeme anzustellen.

Während der Evaluation waren die Testbedingungen jedoch für alle gleich: Die Sprachanteile der auf der Evaluationstestmenge vorgegebenen Abschnitte mußten ohne Verwendung manuell gesetzter Segmentgrenzen oder Markierungen voll automatisch transkribiert werden. Die Ergebnisse sind als Wortfehlerrate in der nachfolgenden Tabelle angegeben. Das Ergebnis des Sys-

tems, das im Rahmen der Diplomarbeit mit *JRTk* erstellt und ebenfalls unter Evaluationsbedingungen getestet wurde, ist in einer zusätzlichen Spalte angegeben.

Da das im Rahmen dieser Diplomarbeit entwickelte System ebenso an der Carnegie Mellon University (CMU) entstanden ist, wird zur Abgrenzung für das System der CMU Robust Speech Recognition Group die Bezeichnung *CMU (SPHINX)* verwendet[1].

Evaluationsergebnisse					**Diplomarbeit**
	IBM	Dragon	BBN	CMU (SPHINX)	JRTk
WF	27 %	41.4 %	42 %	41 %	41.9 %

Im folgenden werden die einzelnen Systeme grob beschrieben.

IBM: Segmentierung

- basierend auf Klassifikatoren
- Trennung von Klassen schrittweise durch je zwei Klassifikatoren; einen für die abzuspaltende Klasse und einen für all diejenigen Klassen, die nach den vorausgehenden Abspaltungsschritten übrig geblieben sind.
- eigene Vorverarbeitungen für verschiedene Abspaltungsschritte
- Abspaltungen: 1. reine Musik, 2. Telephonsprache, 3. Sprache verunreinigt durch Musik oder Lärm. Übrig bleiben Segmente mit klarer Sprache. Diese werden durch Klassifikatoren weiter unterteilt in klare Sprache von insgesamt 9 Sprechern, für die eine Menge an Trainingsdaten vorliegt, die zur Adaption ausreicht, und klare Sprache aller Sprecher, für die keine Adaption möglich ist.
- Verwendung eines HMMs zum Erzwingen einer Mindestlänge der Segmente
- Um die Funktionsweise zu verdeutlichen: Im ersten Schritt Verwendung eines Klassifikators für reine Musik und eines Klassifikators für alle übrigen akustischen Bedingungen. Entfernen aller Segmente, die reine Musik enthalten. Auf den verbleibenden Segmenten im zweiten Schritt Verwendung eines Klassifikators für Sprache über einen Telephonkanal und eines Klassifikators für alle akustischen Bedingungen außer reiner Musik und Telephonsprache. Entfernen aller Segmente, die Telephonsprache enthalten. Entsprechend wird weiter verfahren.

[1] SPHINX ist das Spracherkennungssystem, das von der Robust Speech Recognition Group verwendet wird

Sprachmodell

- erzeugt aus Sprachmodellen, die über 3 Korpora berechnet wurden
- die drei Korpora: 1. Zeitungsnachrichten über einen längeren Zeitraum mit Schwerpunkt auf Wirtschaftsnachrichten, 2. Transkriptionen von Nachrichtensendungen über einen längeren Zeitraum sowie 3. Zeitungsnachrichten und Nachrichtensendungen aus der Zeit um die Evaluation, um Aktualtität zu gewährleisten

Wortliste

- ca. 64000 Worte aus den drei Korpora, über denen das Sprachmodell berechnet wurde
- erweitert durch zusätzliche Worte aus Marketplace-Trainingssendungen

Training der akustischen Modelle

- insgesamt 32 Mengen von akustischen Modellen
- Telephonmodelle trainiert mit bandbegrenzten WSJ Daten
- Musikmodelle mit künstlich durch Musik verunreinigten WSJ Daten trainiert
- für 9 Sprecher, für die ausreichend viele Daten in der Trainingsmenge vorhanden waren, Adaption der akustischen Modelle eines Basissystems über der Trainingsmenge
- Modelle für klare Sprache unbekannter Sprecher durch Adaption mit klarer Sprache aller Sprecher über der Trainingsmenge trainiert
- Telephonmodelle und Modelle für Sprache mit Hintergrundmusik durch überwachte Adaption über den Trainingssendungen an die akustischen Bedingungen in Marketplace-Sendungen angepaßt
- unüberwachte Adaption während des Tests, falls in den Testsendungen Segmente mit ausreichender Länge vorlagen

Vorverarbeitung

- Ausfilterung von Musik
- eigene Vorverarbeitung für Telephonsprache, Hintergrundmusik und klare Sprache

Dragon: Segmentierung

- in drei Stufen
- erste Stufe: Auf Stille basierend
- zweite Stufe: Erkennen und Entfernen von Musiksegmenten; hierbei Verwendung eines Maßes für Harmonie

- dritte Stufe: Segmentierung durch Klassifikatoren für die Klassen VB, RB, HM (siehe Abschnitt 7.1.1)

Sprachmodell

- erzeugt durch lineare Interpolation von Sprachmodellen, die über 2 Korpora berechnet wurden
- die zwei Korpora: 1. Zeitungstexte aus nordamerikanischen Zeitungen mit Schwerpunkt Wirtschaft (NAB Korpus) und 2. die 10 Marketplace-Trainingssendungen

Wortliste

- 60000 Worte aus dem NAB Korpus
- 297 zusätzliche Worte aus den 10 Marketplace-Sendungen

Training der akustischen Modelle

- sprecher- und geschlechtsunabhängige akustische Modelle mit *duration modeling*
- eigene Modelle für die Klassen VB, RB und HM
- Modelle für Klasse VB über WSJ Daten berechnet
- Modelle für Klasse RB mit bandbegrenzten WSJ Daten trainiert
- Modelle für Klasse HM ausgehend von den Modellen für Klasse VB durch Adaption über den Marketplace-Trainingssendungen bestimmt

Vorverarbeitung

- gleich für alle Modelle

BBN: Segmentierung

- durch ein Maß, das die Änderung der akustischen Bedingungen angibt
- anschließend Klassifizierung der Segmente bzgl. der Klassen *männlicher Sprecher*, *weiblicher Sprecher*, *Ansager* und *reine Musik*
- Entfernen der Segmente, die nur Musik enthalten
- Setzen von Segmentgrenzen an Stellen, wo in einem ersten Spracherkennungslauf ausreichend lang Stille erkannt wird. Für die verwendeten Erkenner wurden die akustischen Modelle über Daten männlicher Sprecher, weiblicher Sprecher und des Ansagers adaptiert.

Sprachmodell

- erzeugt durch lineare Interpolation von Sprachmodellen, die über 4 Korpora berechnet wurden

- die Korpora: 1. Der WSJ Korpus, 2. ein Korpus von nordamerikanischen Zeitungsnachrichten (NA Korpus), 3. ein Korpus von Transkriptionen amerikanischer Radiosendungen (BN Korpus) sowie 4. die 10 Marketplace-Nachrichtensendungen

Wortliste

- 45000 Worte aus den drei Korpora WSJ, NA und BN
- Worte aus den Marketplace-Trainingssendungen wurden nicht ergänzt

Training der akustischen Modelle

- drei Mengen von akustischen Modellen
- Adaption der akustischen Modelle eines Basissystems über den Marketplace-Trainingssendungen für die Klassen *männlicher Sprecher*, *weiblicher Sprecher* und *Ansager*

Vorverarbeitung

- keine Angaben

CMU (SPHINX): Segmentierung

- in drei Schritten
- im ersten Schritt Setzen von Segmentgrenzen, wenn mindestens eines von drei Klassifikatorpaaren eine Änderung der Klassenzugehörigkeit der Audiodaten anzeigt und an einer Stelle in einem 1s Fenster ein Kriterium für Stille erfüllt ist
- falls die so entstehenden Segmente eine zu große Länge haben, weiteres Zerteilen mit Hilfe eines Stillekriteriums
- in einem ersten Spracherkennungslauf Stellen, die als Stille erkannt werden, sammeln; aus diesen Stellen anhand verschiedener Kriterien neue Segmentgrenzen aussuchen
- es werden Klassifikatorpaare für *männliche/weibliche Sprecher*, *Telephonkanal ja/nein* und *Musik ja/nein* verwendet
- Beseitigen von irrtümlich erkannten Worten, falls nur Musik ohne Sprache vorliegt, durch Zurückweisen von Worten, bei denen die Wahrscheinlichkeit für die Dauer der Phoneme sehr gering ist (*long word rejection based on duration modeling*)

Sprachmodell

- erzeugt durch Interpolation von Sprachmodellen, die über drei Korpora berechnet wurden
- die Sprachmodelle: 1. Das offizielle ARPA HUB-3 Sprachmodell, 2. ein Sprachmodell berechnet über Transkriptionen von Nachrichtensendungen amerikanischer Sender aus dem Zeitraum der Evaluation und 3. ein Modell berechnet über den 10 Marketplace-Nachrichtensendungen

Wortliste

- 60000 Worte für die Erkennung von Sprache bei voller Bandbreite
- 30000 Worte für die Erkennung von Sprache bei reduzierter Bandbreite
- Auswahl der Worte: Die häufigsten 30000, bzw. 60000 Worte aus dem HUB-3 Korpus ergänzt durch in den Marketplace-Sendungen vorkommende Worte

Training der akustischen Modelle

- Modelle für reduzierte Bandbreite und volle Bandbreite
- Keine Angaben, ob die Modelle an die Marketplace Bedingungen angepaßt wurden

Vorverarbeitung

- Verwendung eines Verfahrens zur Kompensierung der akustischen Unterschiede zwischen Trainings- und Testumgebung (CDCN)

8.2 Einordnung des gewählten Ansatzes

Im Rahmen dieser Diplomarbeit standen für die Entwickelung des Spracherkennungssystems nur die Daten zur Verfügung, die auch von den Teilnehmer der Evaluation verwendet wurden. Die Spracherkennungsversuche VI und VII (siehe Abschnitt 7.1) wurden unter Evaluationsbedingungen durchgeführt, die Ergbenisse sind daher vergleichbar mit denen der anderen Systeme.

Die Wortfehlerraten des entwickelten Systems und der Systeme der Teilnehmer der Evaluation mit Ausnahme von IBM unterscheiden sich um weniger als 1%.

Da in dieser Diplomarbeit der Schwerpunkt auf die Segmentierung gelegt wurde, konnte auf die Entwicklung des Spracherkennungssystems nicht sehr viel Zeit verwendet werden. Hinzu kommt, daß für die Anpassung eines Spracherkennungssystems an eine neue Domäne viele Aufgaben zu bewältigen sind:

- Transkriptionen, Zeitmarken und Markierungen müssen in eine Form gebracht werden, die in einem Spracherkennungssystem verwendbar ist.
- Die Parameter des Systems müssen auf die neue Domäne abgestimmt werden. Das ist ein sehr zeitaufwendiger Prozeß, da viele Spracherkennungsläufe notwendig sind.
- Für manche in der neuen Domäne vorkommenden Worte müssen Phonemtranskriptionen erzeugt oder beschafft werden.

- Die Berechnung eines Sprachmodells kostet viel Zeit, da große Datenmengen zu bewältigen sind.
- Die Erstellung einer Wortliste ist kein trivialer Vorgang und erfordert einige Untersuchungen der Trainings- und Entwicklungstesttexte.

Viele Maßnahmen zur Verbesserung des Spracherkennungssystems konnten daher, und auch weil die Spracherkennungsexperimente selbst sehr lang dauern, nicht durchgeführt werden. Vor diesem Hintergrund ist die Wortfehlerrate von 41.9% im Vergleich mit den anderen Systemen sehr positiv zu bewerten.

8.2.1 Segmentierung

In [44, 15, 13] wird von zum Teil erheblichen Zunahmen der WF bei automatischer Segmentierung gegenüber dem Fall vorgegebener Segmentgrenzen berichtet. Im vorliegenden Fall konnte keine negativen Auswirkungen auf die Wortfehlerrate nachgewiesen werden (siehe Kapitel 7).

Vor- und Nachteile anderer Segmentierungsansätze

In [44, 15, 13, 8] werden keine Auswertungsergebnisse für die Segmentierungskomponenten der Systeme gegeben, die einen quantitativen Vergleich zulassen. Daher werden im folgenden in aller Kürze die Vor- und Nachteile der einzelnen Ansätze aufgelistet. + bedeutet hierbei einen Vorteil, - einen Nachteil und o eine fehlende Information.

BBN:

+ Kein Training notwendig, nur ein Einstellen von Schwellwerten für die Empfindlichkeit.
+ Kann Übergänge aller Art feststellen (Sprecherwechsel, Änderung der Kanaleigenschaften, Änderung der Hintergrundgeräusche).
+ Keine Segmentmarkierungen notwendig.
- Bei niedriger Schwelle viele falsche Segmentgrenzen, bei hoher Schwelle viele Auslassungen.
- Spracherkennungslauf notwendig; dadurch vermutlich langsam.
- Es wird von einer sehr großen Erhöhung der WF bei automatischer Segmentierung berichtet.

CMU (SPHINX):

+ Bei Auswahl der Klassifikatorpaare für den Segmentierer wurde die Verwechselbarkeit von Klassen berücksichtigt. Das Klassenpaar *klare/verunreinigte Sprache* wurde deshalb nicht verwendet.

- Korrekte Segmentmarkierungen zum Trainieren der Klassifikatoren notwendig.
- Keine Sprecherwechsel erkennbar.
- Keine Übergänge von durch Lärm verunreinigter zu klarer Sprache erkennbar.
- Durch dreistufigen Segmentierungsvorgang, der einen Spracherkennungslauf einschließt, sehr komplex und vermutlich zeitaufwendig.
- Kein explizites Entfernen von Musiksegmenten.
- Deutliche Verschlechterung der WF durch automatische Segmentierung.

Dragon:

- Der dreistufige Segmentierungsvorgang, der einen Spracherkennungslauf notwendig macht, ist komplex und vermutlich zeitaufwendig.
- Entfernen von Musik durch ein Maß für Harmonie scheint viel Sprache wegzuschneiden.
- Nur drei verschiedene Arten von akustischen Bedingungen werden durch Klassifikatoren repräsentiert. Hierdurch sind viele Änderungen der Akustik nicht wahrnehmbar.
- Klassenwahl für Klassifikatoren intuitiv.
- Segmentmarkierungen zum Trainieren der Klassifizierer notwendig.
- Es wird von einer starken Zunahme von Worterkennungsfehlern bei automatischer Segmentierung berichtet.

IBM:

+ Feine Unterscheidung der verschiedenen akustischen Bedingungen.
+ Berücksichtigung einer Großzahl von Sprecherwechseln.
+ Kein Spracherkennungslauf notwendig, daher vermutlich schnell.
+ Speziell auf die einzelnen akustischen Klassen abgestimmte Vorverarbeitung möglich.
- Aufgrund der großen Verwechselbarkeit einiger Klassen werden offenbar viele Segmentierungsfehler gemacht.
- Segmentmarkierungen zum Trainieren der Klassifizierer notwendig.
- Klassenwahl intuitiv. Es wird nicht berücksichtig, ob Klassen überhaupt akzeptabel trennbar sind.

o Es wird von keiner Verbesserung oder Verschlechterung durch automatische Segmentierung berichtet.

Vor- und Nachteile des gewählten Segmentierunsansatzes

Der Vorteil des im Rahmen dieser Diplomarbeit gewählten Segmentierungsansatzes ist, daß durch die automatische Klassenbildung die Verwechselbarkeit von akustischen Bedingugen berücksichtigt wird. Die Klassenbildung beruht nicht auf möglicherweise falschen Annahmen, sondern basiert auf den statistischen Eigenschaften der Audiodaten. Es konnte gezeigt werden, daß bezüglich mehrerer Maße, die die Güte einer Segmentierung angeben, die automatisch gefundene Klasseneinteilung besser für eine auf Klassifikation basierende Segmentierung geeignet ist als eine intuitive Wahl von Klassen.

Ein Nachteil des Ansatzes ist, daß genaue Segmentmarkierungen für eine ausreichend große Anzahl von Klassen notwendig sind. Jedoch kann der Segmentierer leicht durch zusätzliche Markierungsfelder erweitert werden. Es könnten z.B. Markierungen bezüglich Sprecherakzent, Echo oder andere akustische Bedingungen beim Klassenbildungsprozeß berücksichtigt werden.

Durch den einfachen Aufbau und die Tatsache, daß kein erster Spracherkennugslauf nötig ist, ist der Segmentierungsvorgang schnell. Der automatische Klassenbildungsprozeß macht es erforderlich, daß bei der Segmentierung für alle Klassen dieselbe Vorverarbeitung verwendet werden muß.

Ein negativer Einfluß der automatischen Segmentierung auf die Spracherkennung konnte nicht nachgewiesen werden. Die Wortfehlerrate wurde sogar minimal verringert, da sich bei stark durch Musik verunreinigter Sprache die längeren Segmente bei automatischer Segmentierung günstig ausgewirkt haben.

8.2.2 Auswahl von Klassen für die Spezialerkenner

Die automatische Klassenbildung zur Bestimmung von Klassen für die Entwicklung spezialisierter Spracherkenner hat nur eine sehr geringe Verbesserung erbracht. Das liegt sicher zum Teil daran, daß nur ein geringer Aufwand für die Anpassung von speziellen Erkennern betrieben werden konnte. Ein anderer Grund ist die sehr geringe Menge an Trainingsdaten für die einzelnen Spezialerkenner.

Die Möglichkeit, Klassen automatisch zu bilden, bietet den Vorteil, daß keine Klassen von Hand erzeugt werden müssen.

8.2.3 Die Spezialerkenner

Durch Adaption konnte die WF auf mit Musik verunreinigter Sprache um 14.5% und bei Sprache über Telephonverbindungen um 12.2% absolut verringert werden. Bei klarer Sprache wurde nur eine Verbesserung um ca. 4% erreicht. Die geringe Verbesserung im Fall klarer Sprache ist sicher darauf zurückzuführen, daß die akustischen Modelle des Basissystems schon verhältnismäßig gut zu der klaren Sprache in den Marketplace-Sendungen paßten.

In die Spezialerkenner konnte nur ein sehr geringer Entwicklungsaufwand gesteckt werden. Weiterhin mußten aus Zeitgründen eine kleine Wortliste

und ein stark verkleinerter Suchraum verwendet werden. Zudem enthielten die Daten, über denen das Sprachmodell errechnet wurde, einige Fehler. Es konnten nur 5 der 10 Trainingssendungen für das Training verwendet werden, da für die übrigen Sendungen keine Segmentgrenzen und Markierungen vorgegeben waren. Auch enthielten die Transkriptionen der 5 Sendungen einige Fehler, die Zeitmarken waren ungenau und die Markierungen inkonsistent. Aus Mangel an Zeit und Resourcen konnten weiterhin keine akustischen Modelle mit bandbegrenzen Daten trainiert werden. Auch mußte aus Zeitgründen auf Adaption während des Tests verzichtet werden, da die Dauer für einen Erkennungslauf hierdurch verdoppelt würde. Der naheliegende Schritt, für das Sprachmodell die Marketplace-Sendungen mit zu berücksichtigen, konnte auch aus Mangel an Zeit und vor allem Resourcen nicht durchgeführt werden.

Die verwendeten Spracherkenner können aus all diesen Gründen nur als grober erster Versuch betrachtet werden, die verschiedenen akustischen Bedingungen zu berücksichtigen. Es gibt noch viele weitere Maßnahmen, durch die die Erkennungsleistung sicher deutlich gesteigert werden könnte, hierzu zählen geschlechtsabhängige akustische Modelle oder Normalisierung bezüglich der Vokaltraktlänge sowie spezielle Vorverarbeitungen für die verschiedenen akustischen Bedingungen.

Unter Berücksichtigung dieser Einschränkung ist es sehr ermutigend, daß ein Ergebnis erzielt werden konnte, daß nicht hinter den Ergebnissen der Teilnehmer der Evaluation mit Ausnahme des Systems von IBM zurückbleibt.

8.3 Ausblick

Viele mögliche Verbesserungen der Spezialerkenner wurden schon im letzten Abschnitt aufgezählt. Diese und weitere sinnvolle Anpassungen sowie Möglichkeiten, den Segmentierungsansatz zu verfeinern, werden im folgenden genannt.

- Bei der Segmentierung wird ein Links-Rechts HMM verwendet, bei dem alle Zustände gleich modelliert werden. Eine Möglichkeit, den Segmentierungsansatz zu verfeinern, wäre, verschiedene HMM Stukturen mit unterschiedlich modellierten Zuständen zu untersuchen, um Zusammenhänge innerhalb der Segmente zu erfassen (wie z.B. Ein- und Ausblenden von Musik).

- Für den Segmentierungsansatz werden konsistente Markierungen für eine ausreichende Anzahl von akustischen Bedingungen benötigt, um die Segmentierungsklassen zu bestimmen. Das hat sich als Problem erwiesen. Ein Lösungsansatz wäre, unüberwacht Klassen zu bilden.

- Viele Verbesserungen des Sprachmodells sind möglich. Der naheliegendste nächste Schritt ist, das Sprachmodell BN92-95 mit einem Sprach-

modell über den 10 Marketplace-Nachrichtensendungen zu interpolieren.

- Im Sprachmodell werden weiterhin Satzbegrenzungen berücksichtigt. Aufgrund der möglicherweise fehlerhaften Segmentierung ist es unter Umständen sinnvoll, Satzbegrenzungen unberücksichtigt zu lassen oder andere Strategien zu entwickeln, wie in diesem Zusammenhang mit fehlerhafter Segmentierung umgegangen werden kann.

- Die im Kapitel 7 beschriebenen Versuche zeigen, daß *long word rejection* und *duration modeling*, wie es im System der CMU (SPHINX) verwendet wird (siehe Seite 80), insbesondere für mit Hintergrundmusik verunreinigte Sprache angebracht sind.

- Die Größe der Wortliste und der Suchraum sollten vergrößert werden.

- Die Vorverarbeitung des als Ausgangspunkt gewählten WSJ Systems ist für klare Sprache bei gleichbleibenden akustischen Bedingungen gut geeignet. Da sich in Radiosendungen jedoch die Bedingungen häufig ändern, ist eine robustere Vorverarbeitung unerläßlich.

- Neben einer generell robusteren Vorverarbeitung sollte für Sprache über Telephonverbindungen eine spezielle Vorverarbeitung mit Berücksichtigung der Besonderheiten des Telephonkanals verwendet werden.

- Geschlechtsabhängige akustische Modelle oder Vokaltraktlängennormalisierung könnten verwendet werden.

- Unüberwachte Adaption während des Tests hat sich in ersten Versuchen als sehr erfolgreich erwiesen und sollte daher eingesetzt werden. Das wird auch durch die guten Ergebnisse des Systems von IBM bestätigt.

- Die Vergrößerung der sehr kleinen Trainingsmenge würde sicher Verbesserungen bringen.

- Als Basissystem für den Telephonerkenner sollte ein mit bandbegrenzten Daten trainiertes System verwendet werden.

- Weiterhin haben die Versuche in [13] gezeigt, daß Kompensationsalgorithmen wie CDCN eine deutliche Verbesserung bringen. Es wäre zu überprüfen, ob durch eine Kombination von Adaption und z.B. CDCN die Wortfehlerrate weiter reduziert werden kann.

Anhang A

Basisklassen

In den folgenden zwei Tabellen sind die in der gesamten Trainingsmenge sowie in der Marketplace-Sendung vom 15. März 1994 vorkommenden Basisklassen aufgelistet. Die Sendung vom 15. März ist diejenige, die zum Testen des Segmentierers verwendet wurde. Die gesamte Trainingsmenge umfaßt die 5 Marketplace-Sendungen, für die Markierungen und Zeitmarken zur Verfügung standen.

Für jede Basisklasse ist die Gesamtdauer aller Segmente dieser Klasse sowie die Anzahl der vorkommenden Segmente aufgeführt. Die Notation zur Beschreibung der Basisklassen weicht von der Darstellung, die in Abschnitt 4.2.1 beschrieben ist, ab, weil in den im Rahmen der Diplomarbeit durchgeführten Untersuchungen nicht zwischen Hintergrundsprache und Lärm unterschieden wurde. Der Grund dafür ist das relativ seltene Vorkommen von Lärm und die schlechte Unterscheidbarkeit von Lärm und Hintergrundsprache, da mit Lärm auch Audiodaten markiert wurden, die unverständliche Sprache vieler Sprecher enthalten.

Analog zu der in Abschnitt 4.2.1 beschriebenen Notation sind drei Felder dargestellte. Das erste Feld gibt das Geschlecht des Sprechers an, das zweite Feld die Kanaleigenschaften und das dritte Feld die Hintergrundgeräusche. Die Hintergrundgeräusche können eine beliebige Kombination aus *Musik*, *Sprache* und *Lärm* sein oder $-$, falls keine Hintergrundgeräusche vorliegen. Der Übersichtlichkeit wegen wurden für die Einträge in den Feldern keine Abkürzungen gewählt.

Die 34 Basisklassen		
Menge von Basisklassen	Gesamtdauer	# Segmente
Mann- klar -–	2841.07	474
Mann- klar -{Sprache}	6.409	2
Mann- klar -{Sprache, Lärm}	80.791	16
Mann- klar -{Lärm}	22.91	6
Mann- klar -{Musik}	802.248	157
Mann- klar -{Musik, Sprache}	5.749	2
Frau- klar -–	829.418	148
Frau- klar -{Lärm }	297.187	50
Frau- klar -{Sprache}	7.365	2
Frau- klar -{Sprache, Lärm}	41.891	5
Frau- klar -{Musik}	530.645	97
Frau- klar -{Musik, Lärm}	14.433	4
–- klar -–	11.699	10
–- klar -{Sprache, Lärm}	13.12	4
–- klar -{Lärm}	5.43	2
–- klar -{Musik}	476.152	75
–- klar -{Musik, Lärm}	11.868	2
–- klar -{Musik, Sprache, Lärm}	7.727	1
Mann- verzerrt -–	815.805	128
Mann- verzerrt -{Sprache}	5.885	3
Mann- verzerrt -{Sprache, Lärm}	89.437	10
Mann- verzerrt -{Lärm}	121.226	17
Mann- verzerrt -{Musik}	61.192	11
Mann- verzerrt -{Musik, Lärm}	10.2	1
Frau- verzerrt -–	427.401	60
Frau- verzerrt -{Lärm}	102.825	15
–- verzerrt -–	1.219	1
–- verzerrt -{Sprache, Lärm}	15.266	2
–- verzerrt -{Musik}	1.279	1
Mann- Telephon- –	851.544	111
Mann- Telephon- {Lärm}	53.505	8
Mann- Telephon- {Sprache, Lärm}	22.664	2
Frau- Telephon- –	33.498	7
–- Telephon- –	3.072	3

Die Basisklassen in der Testsendung		
Menge von Basisklassen	Gesamtdauer	# Segmente
Mann- klar --	169.429	41
Mann- klar -{Sprache}	6.409	2
Mann- klar -{Musik}	47.916	10
Mann- klar -{Musik, Sprache}	5.749	2
Frau- klar --	374.582	66
Frau- klar -{Musik}	237.268	50
—- klar --	9.084	7
—- klar -{Musik}	118.919	17
Mann- verzerrt --	369.007	68
Mann- verzerrt -{Sprache}	2.504	1
Mann- verzerrt -{Musik}	11.486	4
Frau- verzerrt --	185.706	31
Mann-Telephon--	135.617	18
Mann-Telephon-{Lärm}	29.26	5
Frau-Telephon--	21.048	5

In folgender Tabelle ist eine Zuordnung der in der Testmenge vorkommenden Basisklassen zu den Klassen A1-A9 und H1-H9 angegeben. Hieran wird deutlich, daß sich die automatisch gefundene und die von Hand gewählte Klasseneinteilung sehr unterscheiden.

Zuordnung: Basisklassen der Testsendung zu H1-H9 und A1-A9		
Menge von Basisklassen	Klassen H1-H9	Klassen A1-A9
Mann- klar --	H1	A2
Mann- klar -{Sprache}	H2	A2
Mann- klar -{Musik}	H3	A1
Mann- klar -{Musik, Sprache}	H3	A2
Frau- klar --	H1	A7
Frau- klar -{Musik}	H3	A4
—- klar --	H1	A9
—- klar -{Musik}	H4	A5
Mann- verzerrt --	H5	A9
Mann- verzerrt -{Sprache}	H6	A9
Mann- verzerrt -{Musik}	H7	A9
Frau- verzerrt --	H5	A6
Mann-Telephon--	H8	A3
Mann-Telephon-{Lärm}	H9	A3
Frau-Telephon--	H8	A8

Menge von Fragen 1 (Fortsetzung)				
Einschränkung bzgl. Kanal und Hintergrund				
Geschlecht	-	Kanal	-	Hintergrund
*	-	*K*	-	*M*
*	-	*K*	-	*M(L)*
*	-	*K*	-	–
*	-	*K*	-	*L*
*	-	*T*	-	–
*	-	*V*	-	*M*
*	-	*V*	-	–
*	-	*V*	-	*M(L)*
*	-	*V*	-	*L*
Einschränkung bzgl. Sprecher, Kanal und Hintergrund				
Geschlecht	-	Kanal	-	Hintergrund
–	-	*K*	-	–
–	-	*V*	-	–
–	-	*T*	-	–
–	-	*K*	-	*M*
–	-	*V*	-	*M*
–	-	*K*	-	*M*
M	-	*K*	-	–
F	-	*K*	-	–
M	-	*K*	-	*M*
F	-	*K*	-	*M*

Die Formulierung der Fragen in der folgenden Tabelle entspricht der Konvention, die in Abschnitt 4.2.1 beschrieben ist. Hier wird nicht zwischen männlichen und weiblichen Sprechern unterschieden. Eine Frage, die für das Geschlecht des Sprechers ein *MF* enthält, wird für Basisklassen mit einem *M* oder einem *F* im entsprechenden Feld positiv beantwortet.

Menge von Fragen 2				
Einschränkung bzgl. eines Felds				
Geschlecht	-	Kanal	-	Hintergrund
MF	-	*	-	*
–	-	*	-	*
*	-	*	-	–
*	-	*	-	*M*
*	-	*	-	*M(L)*
*	-	*	-	*L*
*	-	*K*	-	*
*	-	*T*	-	*
*	-	*V*	-	*
Einschränkung bzgl. Sprecher und Kanal				
Geschlecht	-	Kanal	-	Hintergrund
MF	-	*K*	-	*
MF	-	*T*	-	*
MF	-	*V*	-	*
–	-	*K*	-	*
–	-	*T*	-	*
–	-	*V*	-	*
Einschränkung bzgl. Sprecher und Hintergrund				
Geschlecht	-	Kanal	-	Hintergrund
MF	-	*	-	*M*
MF	-	*	-	*M(L)*
MF	-	*	-	*L*
MF	-	*	-	–
–	-	*	-	*M*
–	-	*	-	*M(L)*
–	-	*	-	*L*
–	-	*	-	–

Menge von Fragen 2 (Fortsetzung)				
Einschränkung bzgl. Kanal und Hintergrund				
Geschlecht	-	Kanal	-	Hintergrund
*	-	*K*	-	*M*
*	-	*K*	-	*M(L)*
*	-	*K*	-	–
*	-	*K*	-	*L*
*	-	*T*	-	–
*	-	*V*	-	*M*
*	-	*V*	-	–
*	-	*V*	-	*M(L)*
*	-	*V*	-	*L*
Einschränkung bzgl. Sprecher, Kanal und Hintergrund				
Geschlecht	-	Kanal	-	Hintergrund
–	-	*K*	-	–
–	-	*V*	-	–
–	-	*T*	-	–
–	-	*K*	-	*M*
–	-	*V*	-	*M*
–	-	*K*	-	*M*
MF	-	*K*	-	–
MF	-	*K*	-	*M*

Literaturverzeichnis

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, NY, 1984.

[2] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, NY, 1995.

[3] G.D. Cook, J. D. Christie, P. R. Clarkson, M. M. Hochberg, B. T. Logan, and A. J. Robinson. Real-time recognition of broadcast radio speech. In *Proceedings of the IEEE*, pages 141–144, 1996.

[4] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, NY, 1973.

[5] G. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer Verlag, New York, NY, 1972.

[6] J. Fritsch. Modular neural networks for speech recognition. Diplomarbeit, Universität Karlsruhe (Germany) und Carnegie Mellon University (USA), 1996.

[7] H. Gish, M.-H. Siu, and R. Rohlicek. Segregation of speakers for speech and speaker identification. In *Proceedings of the IEEE*, pages 873–876, 1991.

[8] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L.Polymenakos, H. Printz, and M. Franz. Transcription of radio broadcast news with the IBM large vocabulary speech recognition system. In *Proceedings of the ARPA Speech Recognition Workshop*, pages 72–76, 1996.

[9] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, and L. Polymenakos. Acoustic models used in the IBM system for the ARPA HUB 4 task. In *Proceedings of the ARPA Speech Recognition Workshop*, pages 77–80, 1996.

[10] P. S. Gopalakrishnan, D. Nahamoo, M. Padmanabhan, and L. Polymenakos. Suppressing background music from music-corrupted data of the ARPA HUB 4 task. In *Proceedings of the ARPA Speech Recognition Workshop*, pages 81–84, 1996.

[11] D. E. Hall. *Musical Acoustics: An Introduction.* Wadsworth Publishing Company, Belmont, CA, 1980.

[12] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Company, Redwood City, CA, 1991.

[13] U. Jain, M. Siegler, S.-J. Doh, E. Gouvea, J. Huerta, P. J. Moreno, B. Raj, and R. M. Stern. Recognition of continuous broadcast news with multiple unknown speakers and environments. In *Proceedings of the ARPA Speech Recognition Workshop*, pages 61–66, 1996.

[14] R. Kneser and H. Ney. Improved backing-off from m-gram language modeling. In *Proceedings of the ICASSP*, 1995.

[15] F. Kubala, T. Anastasakos, H. Jin, J. Makhoul, L. Nguyen, R. Schwartz, and N. Yuan. Toward automatic recognition of broadcast news. In *Proceedings of the ARPA Speech Recognition Workshop*, pages 55–60, 1996.

[16] F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. Schwartz. Transcribing radio news. In *Proceedings of the ICSLP*, volume 2, 1996.

[17] C.J. Leggetter and P.C. Woodland. Speaker adaptation of continuous density HMMs using linear regression. In *Proceedings of the ICSLP*, volume 2, pages 451–454, 1994.

[18] C.J. Leggetter and P.C. Woodland. Speaker adaptation of HMMs using linear regression. Technical report CUED/F-INFENG/TR.181, Cambridge University Engineering Department, 1994.

[19] C.J. Leggetter and P.C. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, pages 110–115, 1995.

[20] Q. Lin and C. Che. Normalizing the vocal tract length for speaker independent speech recognition. In *Proceedings of the IEEE*, pages 201–203, 1995.

[21] F.-H. Liu, R.M. Stern, A. Acero, and P.J. Moreno. Efficient joint compensation of speech for the effects of additive noise and linear filtering. In *Proceedings of the IEEE*, volume 1, pages 257–260, 1992.

[22] F.-H. Liu, R.M. Stern, A. Acero, and P.J. Moreno. Environment normalization for robust speech recognition using direct cepstral comparison. In *Proceedings of the IEEE*, volume 2, pages 61–64, 1994.

[23] M. Maier. Dimensionalitätsreduktion von Sprachsignalen mit statistischen und neuronalen Methoden. Diplomarbeit, Universität Karlsruhe (Germany), 1994.

[24] J. D. Miller. Auditory-perceptual interpretation of the vowel. *Journal of the Acoustic Society of Amerika*, 5(85):2114–2134, 1989.

[25] C. Mokbel, D. Jouvet, and J. Monné. Blind equalization using adaptive filtering for improving speech recognition over telephone. In *Proceedings of the EUROSPEECH*, pages 1987–1990, 1995.

[26] H. Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. In *Proceedings of the IEEE*, 1994.

[27] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependencies in stochastic language modeling. *Computer, Speech and Language*, 8:1–35, 1994.

[28] L. Neymeyer, A. Sankar, and V. Digalakis. A comparative study of speaker adaptation techniques. In *Proceedings of the Eurospeech95*, pages 1127–1130, 1995.

[29] J.J. Odell. *The Use of Context in Large Vocabulary Speech Recognition.* Phd thesis, University of Cambridge, 1995.

[30] A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[31] D. O'Shaughnessy. *Speech Communication.* Addison-Wesley Publishing Company, Reading, MA, 1987.

[32] J. K. Ousterhout. *Tcl and the Tk Toolkit.* Addison-Welsley Publishing Company, Reading, MA, 1994.

[33] PRI Homepage `http://www.pri.org`. World Wide Web. Beschreibung von PRI und der von PRI produzierten Sendungen.

[34] L.R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proceedings of the IEEE*, 1989.

[35] L.R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[36] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, NJ, 1978.

[37] R. Reddy. Speech recognition by machine: A review. In *Proceedings of the IEEE*, pages 520–531, 1976.

[38] I. Rogina and A. Waibel. The Janus speech recognizer. In *Proceedings of the ARPA SLT Workshop*, pages 166–169, 1995.

[39] D.E. Rumelhart, G.E.Hinton, and R.J.Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[40] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proceedings of the IEEE*, pages 993–996, 1996.

[41] M.S. Spina and V.W. Zue. Automatic transcription of general audio data: Preliminary analyses. In *Proceedings of the ICSLP*, volume 2, 1996.

[42] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phonem recognition using time-delay neural networks. In *Proceedings of the IEEE Acoustic Speech and Signal Processing*, volume 37, pages 328–339, 1989.

[43] A. Waibel and K.-F. Lee, editors. *Readings in Speech Recognition.* Morgen Kaufmann Publishers, San Mateo, CA, 1990.

[44] S. Wegmann, L. Gillick, J. Orloff, B. Peskin, R. Roth, P. van Mulbregt, and D. Wald. Marketplace recognition using Dragon's continuous speech recognition system. In *Proceedings of the ARPA Speech Recognition Workshop*, pages 67–71, 1996.

[45] B. Welch. *Practical Programming in Tcl and Tk.* Prentice Hall, 1995.

[46] M. Wittmann, O. Schmidtbauer, and A. Aktas. Online channel compensation for robust speech recognition. In *Proceedings of the EUROSPEECH*, pages 1251–1254, 1993.

[47] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J.Young. Large vocabulary continuous speech recognition using HTK. In *Proceedings of the ICASSP*, 1994.

[48] M. Woszczyna and M. Finke. Minimizing search errors due to delayed bigrams in real-time speech recognition systems. In *Proceedings of the ICASSP*, 1996.

[49] S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 307–312, 1994.