Technische Grundlagen für eine empirisch fundierte Phonetik des Deutschen

Florian Schiel

27.11.2000

Inhaltsverzeichnis

Kapitel 1

Einleitung

Der etwas lang geratene Titel dieser Schrift, *Technische Grundlagen für eine empirisch fundierte Phonetik des Deutschen*, bedarf einiger Erläuterung vorab. Zunächst soll der genannte Hauptgegenstand der folgenden Darlegungen, *Technische Grundlagen*, darauf hinweisen, daß es sich hier nicht um eine vollständige Monographie über die Phonetik des Deutschen handelt, sondern um die Beschreibung einer wissenschaftlichen Methode, die es (in zukünftiger Arbeit) hoffentlich ermöglichen soll, zu einer vollständigen *Theorie der deutschen Phonetik*[1] zu gelangen. Gleichzeitig will das Wörtchen *empirisch* den Leser darauf aufmerksam machen, daß diese Arbeit auf einen strikt exploratorischen Ansatz abzielt, wie er für die klassischen Geisteswissenschaften in der hier angestrebten Form ungewöhnlich ist. Näheres dazu werde ich im Abschnitt 2.3 ausführen. Der vorliegende Text versteht sich daher weniger als Lehrbuch, sondern vielmehr als eine Anleitung zur Erzeugung von statistisch signifikanten Daten aus gemessenen physikalischen Signalen, welche dann wiederum als Basis für weitere Untersuchungen bzw. Theoriebildungen dienen mögen.

Der Text, so wie er in dieser Schrift vorgelegt wird, faßt Ergebnisse und Methoden zusammen, welche von einer interdisziplinär gemischten Gruppe von engagierten jungen Wissenschaftlern in den Jahren 1995 bis 2000 unter meiner wissenschaftlichen Leitung erarbeitet wurden. Zu dieser Gruppe gehörten und gehören neben Phonetikern auch Ingenieure, Computerlinguisten, Psycholinguisten und Physiker, welche sich zu diesem ungewöhnlichen Thema an der Ludwig-Maximilians-Universität München zusammengefunden hatten. Das Resultat ihrer Bemühungen ist in der Fachwelt in zahlreichen Publikationen unter dem Namen 'MAUS' ('Munich AUtomatic Seg-

[1] engl. *Complete Phonetic Theory* (CPT).

mentation') bekannt geworden ([41, 3, 40, 22, 21, 55]); folglich wird sich ein großer Teil dieser Schrift mit diesem Analyse-Werkzeug als Manifestation der hier dargelegten Ideen beschäftigen.

Zunächst aber möchte ich in diesem einleitenden Kapitel kurz den weit gespannten Kontext skizzieren, bevor im zweiten Abschnitt auf die konkreten Motive für diese Arbeit eingegangen wird. Die nachfolgenden drei Hauptkapitel schildern das Grundprinzip, nach welchem die MAUS-Methode funktioniert, erläutern, welche Ergebnisse damit zu erzielen sind, und in welchen Forschungszweigen diese Ergebnisse bereits erfolgreich eingesetzt wurden bzw. ein Einsatz denkbar und sinnvoll wäre.

Ich habe mich bemüht, diese Hauptkapitel bei aller Liebe zum technischen Detail verständlich und übersichtlich zu gestalten. Es kam mir mehr auf das Verständnis des Grundprinzips an, weniger auf die doch relativ komplizierten Detaildarstellungen. Aus diesem Grunde verzichten wir dort auch weitgehend auf mathematische Formulierungen oder technische Anleitungen. Der an diesen Details interessierte Leser findet jedoch im Anhang eine umfangreiche Darstellung der Methoden in einer formalen Form, welche es im Prinzip erlauben sollte, alle hier dargestellten Experimente nachzuprogrammieren. Ein großer Teil dieser rein technischen Darstellungen ist der Arbeit von Dr.-Ing. Andreas Kipp entnommen, der bis 1998 Mitglied meiner Gruppe war. Auch die den normalen Leser eher weniger interessierenden Einzelheiten der verwendeten empirischen Sprachkorpora habe ich aus dem Haupttext in den Anhang verbannt.

Die wissenschaftliche Arbeit ist keineswegs abgeschlossen; das MAUS-System entwickelt sich in ganz neue Formen, auf die hier nicht im Detail eingegangen werden kann, weil sie noch Gegenstand von mehreren Dissertationen sind, die sich in Arbeit befinden. Der historische Stand der wissenschaftlichen Arbeit an der MAUS-Methode, wie er hier dargelegt wird, ist daher ungefähr der Beginn des Jahres 1999. Im Literaturverzeichnis sind jedoch auch sämtliche Veröffentlichungen zu späteren Arbeiten aufgeführt.

Schließlich bleibt noch anzumerken, daß die MAUS-Methode mittlerweile von anderen Kollegen, insbesondere an der University of California, Berkeley, adaptiert und weiterentwickelt wurde. Diese Tatsache zeigt uns einerseits, daß MAUS offensichtlich eine überzeugende und auf andere Sprachen übertragbare Methode darstellt, und andererseits, daß unser Ansatz, in der ersten Stufe auf dem Weg zu einer empirisch fundierten Theorie der Phonetik zunächst exploratorisch vorzugehen, auch an anderen Forschungseinrichtungen verfolgt wird.

1.1 Phonetik und Phonologie

Über das Verhältnis von Phonologie und Phonetik ist im vergangenen Jahrhundert so viel geschrieben worden, daß eine auch nur ansatzweise abdeckende Übersicht den Rahmen dieses einführenden Kapitels bei weitem sprengen würde. Da diese Arbeit sich aber genau mit dem Kernproblem, nämlich der Relation von Symbolen und Signalen beschäftigt, besteht die Gefahr, in diesen Strudel der wissenschaftlichen Auseinandersetzungen hineingezogen zu werden. Deshalb muß hier ganz kurz auf die Situation eingegangen werden.

Die Unterscheidung zwischen dem *Sprechen* als einem immer einmaligen Vorgang in der gegebenen Wirklichkeit und der *Sprache*, derer sich der Sprecher bedienen muß und die er deshalb als ein mehr oder weniger feststehendes Regelsystem etc. erlernt, d.h. erworben haben muß, wird bei Trubetzkoy ausführlich erläutert ([49], S. 5). Bezugnehmend auf Ferdinand Sausure unterscheidet er zwischen dem *Sprechakt* und dem *Sprachgebilde*:

> "Jedesmal, wenn ein Mensch einem anderen etwas sagt, liegt ein *Sprechakt* vor. Der Sprechakt ist immer konkret, findet an einem bestimmten Ort und zu einer bestimmten Zeit statt. Er setzt voraus: einem bestimmten Sprecher (einen "Sender"), einen bestimmten Angesprochenen (einen "Empfänger") und einen bestimmten Sachverhalt, worauf er sich bezieht.[2]
>
> ...
>
> Im Gegensatz zum immer einmaligen Sprechakt ist die Sprache oder das *Sprachgebilde* etwas Allgemeines und Konstantes. Das Sprachgebilde besteht im Bewußtsein aller Mitglieder der gegebenen Sprachgenossenschaft und liegt unzähligen konkreten Sprechakten zugrunde. Andererseits hat aber das Sprachgebilde keine andere Existenzberechtigung als die Ermöglichung der Sprechakte und besteht nur insoweit sich die konkreten Sprechakte darauf beziehen, d.h. insoweit es sich in den konkreten Sprechakten aktualisiert. Ohne konkrete Sprechakte würde auch das Sprachgebilde nicht bestehen. Somit setzen Sprechakt und Sprachgebilde beide einander voraus. Sie sind untrennbar miteinander verbunden und dürfen als zwei aufeinander bezogene Seiten desselben Phänomens "Sprache" betrachtet werden. Ihrem Wesen nach sind sie aber ganz verschieden und müssen daher auch gesondert untersucht werden."

[2] vgl. dazu K. Bühler [6].

Darauf aufbauend verwendet Trubetzkoy die beiden Begriffe *Phonetik* für die Lautlehre der konkreten Sprechakte und *Phonologie* für die Lehre über das Sprachgebilde ([49], S. 14):

> "Somit kann die Phonetik als die Wissenschaft von der materiellen Seite der (Laute der) menschlichen Rede definiert werden."

und

> "Der Phonologe hat am Laut nur dasjenige ins Auge zu fassen, was eine bestimmte Funktion im Sprachgebilde erfüllt."

Die einschlägige Literatur zeigt, daß es im Verhältnis dieser beiden wissenschaftlichen Disziplinen immer wieder zu unterschiedlichen Auffassungen kam. Insbesondere ist die strikte Abtrennung der Phonologie von den konkreten Sprechakten sehr schwierig, weil dadurch die Gefahr besteht, ein völlig von der Wirklichkeit abgekoppeltes Theoriegebäude zu entwickeln.

Besonders deutlich hat Vennemann diese Problematik in Darstellung gebracht ([51], S.32-33):

> "Für manchen Linguisten ist Phonologie auch in jüngerer Zeit noch nichts anderes als ein Zweig der Phonetik, derjenige, der außer den physikalischen Aspekten auch die Rolle der phonetischen Elemente im Sprachsystem berücksichtigt, nämlich im Hinblick darauf, ob sie kontrastbildend wirken. Ganz in diesem Sinne hielt Martinet 1946 Vorträge über *Phonologie as functional phonetics* (1974)[vgl. [50]]. Man muß hier aber doch sehr genau unterscheiden: Die Wortphonologie hat Sprachsysteme zum Untersuchungsgegenstand, das sind abstrakte Objekte, die sich jedem unmittelbar beobachtenden Zugang verschließen; die Phonetik hingegen hat die akustische Form lautsprachlicher Signalisierungen sowie die physiologischen Vorgänge an den sie hervorbringenden Organismen zum Untersuchungsgegenstand; das sind konkrete Ereignisse, die unmittelbarer Beobachtung, nämlich sogar instrumenteller Analyse zugänglich sind. Die beiden Disziplinen haben also gänzlich disjunkte Untersuchungsbereiche; es kann darum keine Rede davon sein, daß die eine ein Teil der anderen sei."

Und weiter:

> "Das Verhältnis zwischen phonologischen und phonetischen Theorien kann also ebenfalls nicht von der Art eines Inklusionsverhältnisses sein. Von welcher Art aber ist es dann?
> Eine extreme Position ist die, daß es überhaupt keine interessante Beziehung zwischen phonologischen und phonetischen Theorien gebe. Dieser Position kommt Foley 1977 nahe."

Natürlich ergibt sich aus solch "extremer Position" sofort das Problem, daß phonologische Theorien sich aber explizit mit Elementen beschäftigen, die eine wie auch immer geartete Beziehung zur Realität haben müssen. Schließlich ist die Phonologie kein absolut abstraktes Theoriegebilde wie zum Beispiel die Riemannsche Geometrie. Auch Vennemann sieht dieses Problem, indem er fortfährt:

> "Das Problem, welches Foley übersieht, ist, daß auf jeden Fall irgendwelche Elemente einer phonologischen Theorie zu irgendwelchen Elementen einer phonetischen Theorie in Beziehung gesetzt werden müssen, gleichgültig, ob es sich um "abstrakte" Elemente oder Relationen handelt oder nicht; denn sonst wäre es gar nicht möglich, phonologische Elemente, wie z.B. die Verschlußlaute g, d, b in verschiedenen Sprachen zu identifizieren, wie dies Foley tut (25). Tatsächlich fixiert er diese Elemente mit den traditionellen phonologischen Kategorien wie okklusiv, velar, dental, labial (25, 28 et passim), die in der Tat eine phonetische Interpretation haben. Insofern stellt Foley hier selbst eine Beziehung zwischen seiner phonologischen Theorie und einer nicht formulierten, aber vorausgesetzten phonetischen Theorie her.
> ...
> Eine umfassende Sprachtheorie muß sowohl eine phonologische als auch eine phonetische Theorie als Teile enthalten, und sie muß ferner verdeutlichen, welcher Zusammenhang zwischen den beiden besteht."

Ausgehend von dieser grundlegenden Einsicht möchte ich im nächsten Abschnitt auf die Veränderungen eingehen, die sich aus den technischen Entwicklungen des ausgehenden Jahrhunderts für die geschilderte Situation ergeben.

1.2 Phonetik/Phonologie im Spannungsfeld Mensch-Maschine-Kommunikation

Mit dem Eintritt ins Computerzeitalter in der zweiten Hälfte des zwanzigsten Jahrhunderts hat sich in beiden Disziplinen, der Phonologie und der Phonetik, vieles verändert. Im Bereich der Phonologie, vor allem im angelsächsischen Raum, fand eine explosionsartige Entwicklung statt, bei der versucht wurde, die regel-basierten Modelle auf den verschiedensten Ebenen der Sprachbeschreibung in generative Computerprogramme umzuwandeln. Dieser Teilbereich der Phonologie wurde so stark betont, daß sich kurz darauf der Begriff *Computational Phonologie* (in Anlehnung an *Computational Lingustics*) für diese Forschungsrichtung einbürgerte. Im Bereich der Phonetik dagegen wurden der Computer und die immer preiswerter werdenden Digital-to-Analog-Converter dazu eingesetzt, gemessene biometrische Signale (einschließlich des Schalldrucksignals) in digitaler Form zu erfassen und vor allem weiter zu verarbeiten. Die Computertechnik diente dem Phonetiker also im weitesten Sinne als ein neues hochkompliziertes und hochflexibeles Meßinstrument. Dabei war interessanterweise zu beobachten, daß nicht nur in der klassischen Phonetik, sondern auch bei der in der Mitte des 20. Jahrhunderts aufblühenden *Kybernetik* der Computer im wesentlichen von Ingenieuren zur Erfassung und Manipulation von Sprache eingesetzt wurde – wenn auch mit unterschiedlicher Zielsetzung: Während die Phonetiker weiterhin an der grundsätzlichen Frage 'Wie funktioniert die sprachliche Kommunikation?' interessiert waren, arbeiteten die Kybernetiker ganz pragmatisch daraufhin, *benutzbare* sprechende und hörende Computerprogramme zu schaffen. Nach und nach entstanden aus diesen Aktivitäten weiter spezialisierte Ingenieurszweige wie *Automatische Spracherkennung, Sprachsynthese, Sprechererkennung* und *Sprecherverifikation.*

Anhand der Entwicklungen in Bereich *Automatischer Spracherkennung* lassen sich zwei interessante Beobachtungen machen, die im Prinzip auch für die anderen erwähnten Disziplinen gelten:

- Zunächst einmal zeigte sich, daß die erste naive Hypothese der Kybernetik 'Die Erkennung gesprochener Sprache ist ein reines Mustererkennungsproblem!' nicht zum Erfolg führt. Allenfalls einzelne Wörter aus einen kleinen Vokabular, von einem Sprecher in Isolierung und überdeutlich gesprochen, lassen sich mit Methoden der *linearen Klassifikation, Dynamic Time Warping* oder später mit *Hidden Markov Modellen* erfolgreich erkennen. Erst die Kombination mit

syntaktisch/pragmatischem Wissen in Form von *Syntaxmodellen, endlichen Automaten* oder statistischen *Language Modellen*, welche Wahrscheinlichkeiten für das Auftreten bestimmter *Wortfolgen* in der zu erkennenden Sprache berücksichtigen, führte zu einigermaßen befriedigenden Resultaten.[3] Ähnliche Tendenzen sind auch in den anderen Teildisziplinen, wie zum Beispiel der Sprachsynthese, zu beobachten[4].

Interessant ist hierbei, daß die Ingenieurstechnik die Phonologie quasi übersprungen hat, indem sie die akustische Mustererkennung (Phonetik) mit der Modellierung von Wortfolgen bzw. Syntax-Modellen (Computerlinguistik) kombinierte; Techniken, die phonologisches Wissen in Form von Regeln in den Spracherkennungsprozess mit einbeziehen, sind weitgehend unbekannt. Ein Grund hierfür mag sein, daß dieses Wissen fast ausschließlich regelbasiert vorliegt, und die Ingenieurswissenschaften sowohl in Bereich der akustischen Modellierung als auch im Bereich der syntaktisch/pragmatischen Modellierung feststellen mußten, daß selbst relativ primitive statistische Verfahren den regelbasierten im Sinne von meßbarer Erkennungsleistung bei weitem überlegen waren.

- Ein weiterer wesentlicher Punkt: Theorien sowohl im akustischen als auch im linguistischen Bereich können mit Hilfe der oben genannten Techniken einem pragmatischen Test unterworfen werden. Ist nämlich eine Theorie richtig in dem Sinne, daß sie die Wirklichkeit korrekt wiedergibt bzw. vorhersagt, dann läßt sich dies mit Hilfe eines entsprechenden Algorithmus und psychophysikalischen Experimenten verifizieren. Darüber hinaus ist sogar eine quantitative Evaluation einer Theorie möglich, indem man den daraus resultierenden Algorithmus einem genau definierten Test, einer sog. *Bench Mark* unterzieht, also den 'Erfolg' verschiedener Theorien im Experiment vergleicht.

Im Laufe der Zeit wurden die oben genannten Teildisziplinen der Kybernetik – ein Begriff, der heutzutage ein wenig aus der Mode gekommen ist – unter dem Begriff

[3]Ich kann hier in dieser Einleitung nicht weiter auf die soeben genannten Techniken eingehen; der interessierte Leser findet in der Literatur (z.B. [39]) oder in meinem Skriptum zur Vorlesung 'Einführung in die automatische Spracherkennung' ([42]) umfangreiches Material.

[4]Z.B. genügt es für eine natürlich klingende Sprachsynthese nicht, akustische Einheiten bzw. Generationsmodelle für akustischen Einheiten hintereinander zu schalten. Erst eine umfassende Analyse des zu synthetisierenden Satzes bis hinauf zur Pragmatik ermöglicht es, die prosodischen Parameter zu berechnen, welche den intendierten Satz wirklich richtig klingen läßt. Vgl. dazu z.B. das kommerzielle Sprachsynthese-System der Fa. Infovox (*www.zygo-usa.com/infovox.html*).

Mensch-Maschine-Kommunikation subsumiert, teilweise betrachtet man sie inzwischen auch nur noch als einen Teil des allgegenwärtigen Modebegriffs *Multi-Media*. An der grundsätzlichen Situation jedoch hat sich wenig geändert: Nach wie vor werden phonologische Theorien erarbeitet, die in der Anwendung keinen Niederschlag finden. Anders in der Phonetik: Dort schlägt sich zumindest Grundlagenwissen, wie die Physiologie des Ohres, in Form von allgemein akzeptierten Standard-Algorithmen wieder[5]. Aber auch hier ist der 'Wissenstransfer' eher als mager zu bezeichnen.

Warum dieser Ausflug in die Ingenieurswissenschaft? Er erlaubt es uns jetzt, wieder auf das ambivalente Verhältnis zwischen Phonologie und Phonetik zurückkommen. Es wäre vielleicht denkbar, daß die 'Verwertbarkeit' von Forschungsergebnissen beider Teildisziplinen der Sprachwissenschaft nur deshalb so gering erscheint, weil der von Vennemann angemahnte 'Zusammenhang' zwischen Phonologie und Phonetik völlig fehlt. Sicher, beide Disziplinen beziehen sich jeweils auf Teile des anderen, aber meist nur aus praktischen Gründen, und vor allem nicht empirisch abgesichert.[6]

> Zur Illustration mag eine kleine Anekdote dienen, die ich selber erlebt habe: Vor einigen Jahren wurde ich aufgefordert, auf einem Phonologen-Kongreß einen Gastvortrag über das MAUS-System zu halten. Der Veranstalter warnte mich vorneweg, möglichst auf jeglichen 'mathematischen Kram' zu verzichten und mich auf das wesentliche Prinzip und die damit erzielten Ergebnisse zu konzentrieren. Als Beispiel für den Output des Systems legte ich eine Folie auf mit statistischen Ergebnissen zur Aussprache bestimmter Wörter in spontaner Sprache, die MAUS ermittelt hatte. Der Vortrag wurde freundlich aufgenommen und in der anschließenden Diskussion gab es kaum ungewöhnliche Fragen. Nach dem Vortrag allerdings trat ein Wissenschaftler an mich heran und erkundigte sich noch einmal nach den statistischen Ergebnissen zur Aussprache. Unter anderen war dabei das Wörtchen 'ich' und laut der Auswertung des MAUS-Outputs wurde die Aussprachevariante [ɪç] fünfmal häufiger in Material gefunden als die kanonische Aussprache [ʔɪç], mit anderen Worten,

[5] Z.B. basiert die gängigste Art der Merkmalsextraktion (vgl. Anhang B), die *Mel-Frequeny-Cepstral-Coefficiants*, auf dem Ohrmodell von Zwicker ([57]).

[6] Auf den Realitätsbezug der Phonologie wurde bereits eingegangen; in der Phonetik ist zum Beispiel die Benutzung von phonologischen Einheiten durchaus gängige Praxis. Ein Phonetiker spricht nicht von 'einem Kontinuum, das zu Beginn sehr wenig Energie enthält, dann einen steilen Transienten aufweist mit flachem, zu tieferen Frequenzen abfallendem Kurzzeitspektrum und anschließend für kurze Zeit einen frikativen Bereich zeigt', er spricht schlichtweg von einem /t/.

> der Glottal-Verschluß vor dem [ɪ] wird in spontaner Sprache meistens zugunsten einer Glottalisierung oder sogar komplett aufgegeben. Der Wissenschaftler starrte eine Sekunde lang auf das Blatt und sagte dann spontan: "Aber im Deutschen steht vor einem wortinitialen Vokal *immer* ein Glottal-Stop."

In dieser Arbeit geht es hauptsächlich um eine neue Methode, Extensionen von physikalischen, d.h. gemessenen Signalen aus einem Kontinuum auf perzeptive Kategorien, die durch Symbole repräsentiert werden, abzubilden, wobei dies mit Hilfe von Methoden der Mustererkennung, also weitgehend deterministisch und unabhängig von subjektiven Einflüssen geschehen soll. In Verbindung mit sehr großen Sprachkorpora, wie sie für die wissenschaftliche Arbeit inzwischen in zunehmenden Maße zur Verfügung stehen, bedeutet dies, daß zum ersten Mal Basismaterial zur Verfügung steht, auf welchem der notwendige 'Zusammenhang' zwischen phonologischer und phonetischer Theorie in Form eines Vorhersagemodells verifiziert werden kann.

Kapitel 2

Motivation

In diesem Kapitel soll, bevor wir uns in späteren Abschnitten mit den technischen Details beschäftigen, eine Einordnung unserer Arbeit in den bisher diskutierten Rahmen versucht werden. Dabei wird auch die dahinter verborgene, zum Teil sehr pragmatische Motivation für unsere Vorgehensweise zur Darstellung kommen.

Im Rahmen dieser Arbeit ist unter dem Begriff *Segmentierung und Etikettierung* das Ergebnis einer Analyse eines konkreten Sprachsignals im Hinblick auf ein vorher festgelegtes Lautkategoriensystem gemeint. Dabei ist der Weg, der zu diesem Ergebnis führt, vorerst nicht spezifiziert; er kann traditionell über die manuelle Bearbeitung durch einen Phonetiker, oder über einen halbautomatischen oder vollautomatischen Algorithmus erfolgen. Das Analyseergebnis, das sich für jeden konkreten untersuchten Fall in Zahlen und Symbolen festhalten läßt, bildet somit eine real existierende (weil belegbare) Verbindung von kategorialem Wahrnehmungssymbol zu physikalischem Ereignis. Das allein würde nicht unser Interesse erklären, da solche Daten bereits seit mehreren Jahrzehnten zur Verfügung stehen. Eine neue Dimension erhält die Sachlage jedoch durch den Umstand, daß durch den Einsatz automatischer Verfahren erstmals Daten in so großer Menge erzeugt werden können, daß bereits existierende Theorien auf dieser Grundlage verifiziert und neue, empirisch motivierte Theorien aufgestellt und mit Zahlen belegt werden können.

Nach einer grundsätzlichen Klärung der Begriffe *Segmentierung* und *Etikettierung* im Kontext dieser Arbeit, werde ich kurz auf die dahinter verborgene problematische Beziehung zwischen lautsprachlicher Kategorie und physikalischem Signal eingehen. Desweiteren muß genauer begründet werden, warum vor diesem Hintergrund die Arbeit auf eine empirische Herangehensweise zielt und schließlich müssen ein paar Dinge zur Verwendung von statistischen Methoden gesagt werden.

2.1 Segmentierung und Etikettierung

Die *Segmentierung* eines Sprachsignals ist die Aufteilung des kontinuierlichen Schalldruckverlaufs bzw. einer anderen Repräsentation, z.B. einer Folge von Merkmalsvektoren, in zusammenhängend aufeinanderfolgende Abschnitte, die *Segmente*. Was die Abgrenzung betrifft, so können aufeinanderfolgende Segmente bündig oder nichtbündig, d.h. mit Überlappungen und/oder Lücken angeordnet sein. Von einem Segment spricht man sinnvollerweise immer nur dann, wenn damit eine wie auch immer geartete Bedeutung verknüpft ist, z.B. "Das Segment A enthält das Wort 'heute'". Der Ausdruck, mit dem die Bedeutung eines Segments in einem alphanumerischen Ausdruck (z.B. 'heute', '#99', 'hOYt@') festgehalten wird, soll *Etikett*[1] genannt werden.

Bildet die Menge aller möglichen Etiketten eine endliche geschlossene Menge, so spricht man von einem *Inventar* oder *Alphabet*. Basiert letzteres wiederum auf einer phonologischen oder phonetischen Theorie, so nennt man eine Kette von Etiketten im allgemeinen eine *phonetische/phonologische Transkription*.

Oder wie Kohler ([24], S. 15) es ausdrückt:

> "(Eine phonetische Transkription ist) die Beschreibung einer Äußerung nach Klassifikationskategorien, die in einer wissenschaftlichen Theorie über lautliche Phänomene verankert sind und das wissenschaftliche Objekt 'Laut' zugrundelegen."

Vom formalen Standpunkt ist man in der Wahl der "wissenschaftlichen Theorie" zunächst völlig frei. In der Praxis hat sich jedoch eingebürgert, von 'enger' bzw. 'breiter' Transkription zu sprechen, wenn man eine detailgetreue, signalnahe von einer phonologisch orientierten Transkription unterscheiden möchte:

> "Die lautliche Umschrift eines Wortes kann eng oder weit sein (engl. *narrow* vs. *broad transcription*). In einer engen Umschrift bemüht man sich um die Repräsentation von lautlichen Details; in einer weiten Umschrift sieht man von der Repräsentation lautlicher Details ab, wenn man meint, daß diese nur Aspekte der Manifestation identischer phonologischer Einheiten sind." ([51], S. 35)

[1] Wir werden in dieser Arbeit statt *Etikett* in bestimmten Zusammenhängen auch den Begriff *Symbol* verwenden.

In Abschnitt 3.1 werde ich auf die spezifischen Probleme der Auswahl des richtigen Alphabets für die automatische Segmentierung und Etikettierung eingehen.

Traditionell wird beim Begriff Transkription davon ausgegangen, daß diese von einem durch analytisches Hörtraining geschulten Ohrenphonetiker erstellt wird. Der Ohrenphonetiker bedient sich dabei bestenfalls nur eines geeigneten Wiedergabegerätes, um den zu analysierenden Sprachschall wiederholt abhören zu können. Tritt zur Transkription bzw. Etikettierung auch noch die Segmentierung, benötigt der Fachmann Hilfsmittel wie spezielle Signaleditoren zur Markierung der Grenzen im Signal und eine Darstellung des Oszillogramms und Sonagramms in verschiedenen Auflösungen. Heutzutage bilden geeignete Computerprogramme integrierte Werkzeuge, die alle diese Hilfsmittel zur Verfügung stellen.

Eine *automatische Etikettierung und Segmentierung* wird ohne jegliche Intervention seitens des Menschen mittels eines deterministischen Algorithmus aus den Eingabedaten berechnet. Die Vor- und Nachteile der automatische Etikettierung und Segmentierung gegenüber der manuellen Analyse werden im folgenden aufgeführt und mit einem positiven oder negativen Vorzeichen bewertet:

- Geschwindigkeit (+)
  Dank der immer preiswerteren Rechenkapazität auf modernen 'Personal Computern' ist es mittlerweile möglich, selbst sehr komplexe Berechnungen in relativ kurzer Zeit durchzuführen. Eine Analyse der manuellen Transkription im *Verbmobil I* Projekt ([13]), die am Institut für Phonetik in München durchgeführt wurde, ergab einen *Echtzeitfaktor* von ca. 1:800. Das bedeutet, daß ein geschulter Phonetiker zu Segmentierung und Etikettierung von 10 Sekunden aufgezeichneter Sprache im Schnitt ca. 8000 Sekunden benötigte (spontane Dialogsprache). Der entsprechende Wert für gelesene Sprache liegt bei ca. 1:100-200. Ein automatischer Algorithmus wie MAUS benötigt auf einer konventionellen Linux-Workstation (PII 450 MHz) dagegen nur einen Echtzeitfaktor von 1:2.

- Konsistenz (+)
  Üblicherweise werden bei der manuellen Segmentation und Etikettierung von Sprachkorpora aus Zeitgründen mehrere Spezialisten eingesetzt, die jeweils disjunkte oder auch teilweise überlappende Teile des Sprachkorpus bearbeiten. Dabei sind zwei Effekte zu beobachten ([10]): Erstens kommt es auch bei nur einem einzelnen Bearbeiter zu unsystematischen Abweichungen von einer wie

auch immer zu definierenden *Referenz-Segmentierung und -Etikettierung*[2]. Das heißt, die Ergebnisse weisen Idiosynkrasien der jeweiligen Bearbeiter auf, wobei noch erschwerend hinzukommt, daß sich diese im Laufe der Zeit auch noch verändern. Wir sprechen in diesem Fall vom *Intra-Labeler-Agreement*, welches idealerweise bei 100% liegen sollte. Bearbeiten mehrere Spezialisten disjunkte Bereiche des Sprachkorpus, kommt dazu noch das sog. *Inter-Labeler-Agreement* hinzu, welches, falls es von 100% abweicht, ausdrückt, um wieviel sich die jeweiligen Idiosynkrasien der Bearbeiter unterscheiden. Die Folge ist, daß sich unsystematische Fehler in das Ergebnis einschleichen, über deren statistische Verteilung meistens nichts ausgesagt werden kann, und welche sich daher bei der weiteren Nutzung der Daten als *Rauschen*[3] störend bemerkbar machen.

Es gibt mehrere Strategien, dieser Fehlerquelle gegenzusteuern: Regelmäßiges gemeinsames Training der Bearbeiter, Diskussion von Problemfällen, überlappende Bearbeitung zur Identifizierung von Problemstellen (welche dann evtl. nachbearbeitet werden müssen), Kontrolle aller Ergebnisse durch eine Person (um wenigstens das Inter-Labeler-Agreement zu erhöhen), etc.
Bei der automatischen Methode entfallen natürlich Inter-Labeler- und Intra-Labeler-Agreement, da der Algorithmus deterministisch und daher reproduzierbar vorgeht. Die Folge ist, daß für bestimmte Fehlerquellen per statistischer Auswertung möglicherweise ein systematischer Fehler bestimmt und korrigiert werden kann[4]

- Genauigkeit (-)
  Automatische Systeme sind derzeit nicht in der Lage, phonetische Feinheiten wie die Diacritica des IPA-Alphabets ([33]) zu erkennen. Dafür gibt es mehrere Gründe, auf die ich teilweise im Abschnitt 2.5 eingehen werde. Handelt es sich also um die Erstellung einer engen Transkription, wird keines der derzeit bekannten automatischen oder halbautomatischen Systeme akzeptable Ergebnisse liefern. Ähnliches gilt für die Qualität der Segmentierung: Besonders in problematischen Fällen – wie z.B. Übergängen von vokalischen zu nasalen Segmenten, Unterscheidung echter Pausen von Aspirationen, Abgrenzung von

[2]Auf das prinzipielle Problem der Referenz wird in Abschnitt 4.3 genauer eingegangen.

[3]Rauschen im Sinne der Verminderung von Information.

[4]Das Wort 'möglicherweise' deutet an, daß es nicht so einfach ist: Auch ein deterministischer Algorithmus liefert unsystematische Fehler, wenn er nach statistischen Methoden vorgeht, welche vom Eingangssignal bedingte Wahrscheinlichkeiten benutzen. Daher ist es nicht möglich, alle Fehler des Systems durch eine einfache statistische Analyse der Ergebnisse auszugleichen.

stimmhaften Frikativen zu umgebenden vokalischen Bereichen – wird ein automatisches Verfahren Schwierigkeiten haben, dem menschlichen Spezialisten das Wasser zu reichen.

- Inventargröße (-)
  Die Größe des verwendeten phonetischen oder phonologischen Alphabets hat direkte Auswirkungen auf die Wirksamkeit der statistischen Erkennungsverfahren, wie sie in automatischen Systemen Verwendung finden: Je größer das Inventar der Symbole, desto schlechter wird die Erkennungsleistung. Auch aus diesem Grunde eignen sich automatische Verfahren beim derzeitigen Entwicklungsstand nur bedingt für die Erstellung enger Transkriptionen.

Unter einer *halbautomatischen Etikettierung und Segmentierung* verstehen wir in diesem Zusammenhang eine hybride Form der beiden bereits geschilderten Varianten. In den meisten Fällen leistet dabei der menschliche Spezialist eine Hilfestellung, entweder vor der eigentlichen Analyse durch Bereitstellung von apriori-Wissen in Form einer Transkription, oder während und nach der Analyse durch Korrektur von Problemstellen.

2.2 Kategorie vs. Kontinuum

Wahrnehmungskategorien sind *per definitionem* diskrete Ereignisse, d.h. entweder nimmt eine Versuchsperson einen bestimmten Laut wahr als einer antrainierten Kategorie zugehörig oder nicht. Wichtig ist in diesem Zusammenhang das Wort 'antrainiert', weil diese Fähigkeit nicht automatisch bei jedem der Sprache mächtigen Menschen im vollen Umfang gegeben ist. Nur wenige Mitglieder der deutschen Sprachgemeinschaft sind sich z.B. darüber bewußt, daß die Laute [ç] in 'ich' und [x] in 'ach' phonetisch gesehen unterschiedliche Lautkategorien sind. Wenn man sie aber darauf hinweist – z.B. in einem Einführungskurs zur phonetischen Transkription –, wird die Wahrnehmung dieser Lautkategorien problemlos akzeptiert (es gibt natürlich Fälle, in denen die Akzeptanz nicht so hoch ist; meistens handelt es sich dann aber um Laute, die im Sprachsystem des Lernenden nicht vorkommen). Setzen wir aber zunächst der Einfachheit halber eine trainierte Versuchsperson voraus, dann handelt es sich bei jeder Zuordnung eines Sprachsignals zu einer Kategorie um ein diskretes Ereignis aus einer endlichen Menge (nämlich der Menge der antrainierten Lautkategorien).

Das konkrete Sprachsignal andererseits kann nicht Element einer endlichen Menge diskreter Ereignisse sein (obwohl gewisse technische Ansätze in der automatischen Spracherkennung in der jüngeren Vergangenheit genau in diese Richtung abzielten[5]). Das gemessene Sprachsignal entsteht durch einen physikalisch-akustischen Prozeß im Artikulationsapparat des Sprechers, wenn man von weiteren Einflüssen auf den Übertragungskanal, wie Störungen, Raumakustik etc. zunächst einmal absieht. Da die Muskelstellungen in Kehlkopf und Artikulationsapparat sowie die Lungenfunktion keineswegs nur endlich viele, diskrete Stellungen einnehmen können, sondern jedes Organ ein Kontinuum von möglichen Konfigurationen abdeckt, muß folglich auch das gemessene Schalldrucksignal Teil eines komplexen Kontinuums sein.

In dieser Arbeit geht es letztendlich um die empirische Analyse von Relationen zwischen Symbolen und gemessenen, einmaligen Ereignissen, genauer gesagt: zwischen Kategorien, die mehr oder weniger den wahrgenommenen Lautkategorien des Menschen entsprechen, und Stücken von gemessenen Schalldrucksignalen. Dabei stellt sich natürlich schon im Vorfeld die Frage, welcher Art diese Beziehung überhaupt sein kann.

Eine Möglichkeit wäre es, die gemessenen Zeitfunktionen, die in digitaler Form als endliche Kette von Zahlenwerten repräsentiert werden, als komplexe phonetische Merkmale zu definieren. Das ist nicht so weit hergeholt, wie es im ersten Moment klingt, wenn man z.B. Untersuchungen wie [12, 45] betrachtet, bei denen eine komplexe Verarbeitung des Sprachsignals Zeitfunktionen liefert, welche unter anderem den Grad der Stimmhaftigkeit oder den Grad der Nasalität zu bestimmten Zeitpunkten des gemessenen Ereignisses repräsentieren.

In [51], S. 35, schreibt Vennemann dazu sinngemäß (über das Vorgehen der amerikanischen Strukturalisten, zwei Ebenen der 'breiten' und 'engen' Transkription einzuführen):

> "Diesem Vorgehen (der Verwendung einer weiten und engen Umschrift), das anfangs nur transkriptionspraktische Gründe hatte, wurde im amerikanischen Strukturalismus theoretische Bedeutung beigemessen: Die Endpunkte der breit (weit) transkribierten Vereinfachung wurden zu Entitäten höheren Typs, den eigentlichen phonologischen Entitäten, erhoben, zu Phonemen, während die eng transkribierten Entitäten als phone-

[5] Vgl. z.B. die zahlreichen Versuche in den 70iger und 80iger Jahren des vergangenen Jahrhunderts, Erkennung von Sprache mittels hochdimensionaler Vektorquantisierung zu simulieren (z.B. [16]).

> tische Manifestationen der Phoneme verstanden wurden, ihre Allophone.
> ...
> Die wortphonologische Repräsentation erfolgt also auf zwei Ebenen, der phonemischen und der phonetischen...“

Weiter unten kritisiert Vennemann diesen Ansatz wie folgt:

> “Nach dem, was oben über das Verhältnis von Phonologie und Phonetik gesagt wurde (vgl. 1.1), dürfte deutlich sein, daß der angedeuteten strukturalistischen Konzeption, verstanden nicht als Methode, sondern als Ansatz zu einer phonologischen Theorie, ein Irrtum zugrunde liegt. Die “phonetischen“ Repräsentationen, wie detailiert sie auch immer ausgeführt sein mögen, erfolgen stets mit endlich vielen, diskreten, aneinandergereihten Elementen; sie sind also gar keine phonetischen Beschreibungen, sondern phonologische Beschreibungen. Beide “Ebenen“ sind also phonologische Ebenen. Die Beziehung zur Phonetik besteht nur implizit, indem die Elemente der “phonetischen“ Repräsentationen phonetisch gedeutet werden können. Insofern auf der “phonemischen“ Ebene nur ausgedrückt wird, welche Distributionsregularitäten für das beschriebene Sprachsystem angenommen werden, leistet sie in theoretischer Hinsicht überhaupt nichts, sondern erfüllt nur den praktischen Zweck einer bequemeren Umschrift. Insofern den Elementen dieser Ebene eine unabhängige Existenz zugeschrieben wird, entstehen die Aporien der sog. Phonemtheorie, über die es ein umfangreiches Schriftum gibt, das aber die Entwicklung der phonologischen Theorie um keinen Schritt weiter gebracht hat.“

Vor dem Hintergrund der beginnenden 80iger Jahre hat Vennemann natürlich vollkommen recht, weil er bei den “endlich vielen, diskreten, aneinandergereihten Elementen“ an Allophon-Kategorien oder allenfalls an phonetische Merkmals-Kategorien nach K.N. Stevens denkt (z.B. in [47]). In Lichte der digitalen Signalverarbeitung muß Vennemanns Aussage insofern eingeschränkt werden, als auch kontinuierliche physikalische Signale sich vollständig durch eine “Aneinanderreihung von endlich vielen, diskreten Elementen“ beschreiben lassen, nur haben diese Elemente eine völlig andere Qualität: Sie beschreiben die Stützwerte einer nach Nyquist ordnungsgemäß abgetasteten Schalldruckkurve, oder – nach einer weiteren eineindeutigen Abbildung – die Stützwerte einer zeitlichen Trajektorie in einem sehr

hoch-dimensionalen Raum (z.B. dem 'Mel-Frequenz-Cepstral-Koeffizienten'-Raum, vgl. Anhang B). Vennemanns Kritik bleibt also nach wie vor zutreffend, wenn man hinzufügt, daß sich die "phonetische" Repräsentation hier auf diskrete phonetische Klassen (z.B. IPA-Symbole) oder phonetische Merkmalsklassen beschränkt.

Im Kontext dieser Arbeit ist hierzu anzumerken, daß genau diese von Vennemann aufgedeckte Diskrepanz sich vollständig vermeiden läßt, wenn man tatsächlich eine empirisch abgedeckte Relation von phonologischer Klasse (wie immer man diese definieren möchte) und physikalischem Signal herzustellen vermag. Auch wenn diese Relation – wie Vennemann hervorhebt – aus sich heraus zunächst überhaupt nicht zur Theoriebildung beiträgt, so kann sie doch die *notwendige Grundlage* bilden, auf der sich Hypothesen erstmals erfolgreich falsifizieren lassen.

Also bleibt die Frage nach der grundsätzlichen Art der Relation zwischen dem Symbol und dem gemessenen Signal. Letztendlich besteht sie nur darin, daß der Beobachter in bestimmten Fällen eine Korrelation von Symbolen zu bestimmten Signalen beobachtet und in anderen Fällen eben nicht. Vergleicht man jedoch die gemessenen Schalldrucksignale der Segmente, bei denen sich eine repräsentative Gruppe von Beobachtern darüber einig ist, daß es sich um Realisierungen einer bestimmten lautlichen Kategorie handelt, so wird man feststellen, daß diese sich scheinbar in keinster Weise ähneln. Erst die Weiterverarbeitung, z.B. in Form eines Sonagramms, enthüllt dem fachmännischen Auge gewisse Merkmale, aus denen sich Rückschlüsse auf die Beschaffenheit des Signals ziehen lassen. Wie wir weiter unten noch sehen werden, herrscht jedoch auch unter Fachleuten in vielen Fällen durchaus keine Einigkeit darüber, was für eine Kategorie vorliegt, besonders wenn es sich um sogenannte Grenzfälle handelt. Die einfache Kausalitätsbeziehung 'Wenn Signal A, dann Kategorie B', wie sie in den klassischen Naturwissenschaften als Basis für nicht weiter ableitbare Naturgesetze herhalten muß, ist wegen der reichen Variabilität der gemessenen Signale nicht einfach aufzustellen, weil sich dann sofort die nächste Frage stellt: 'Aber wie muß A beschaffen sein?' Als Ausweg aus diesem Dilemma schlage ich vor, der Beziehung zwischen einer Wahrnehmungskategorie und der praktisch unendlichen Menge möglicher Signale, die dieser Kategorie im allgemeinen zugeordnet werden, vorerst den pragmatischen Begriff *empirische Relation* zuzuordnen[6].

Zur Frage, wie diese erfahrbare, aber nicht direkt beobachtbare empirische Relation zustande kommt, faßt Vennemann ([51], S. 48 ff) Vorschläge aus Tillmann und

[6]Tillmann ([48], siehe weiter unten) schlägt den Begriff *retinale Diskretisierung* vor, wobei dieser sich jedoch schon aus einer neurophysiologischen Hypothese über das Zustandekommen dieser Relation ergibt, die allerdings erst in der Dissertation von P. Hoole [14] verifiziert werden konnte.

Mansell ([48]) in einer sehr komprimierten Form zusammen:

> "Bei diesem Stand der Dinge hat neuerdings Tillmann (1980) eine hochinteressante Hypothese aufgestellt, die diese Diskrepanz zwischen physikalisch-phonetischen 'Flux' und wahrnehmungsmäßiger Segmentiertheit erklären soll. Das Problem besteht genausolange, wie man annimmt, daß die physikalischen Vorgänge und die wahrnehmungsmäßigen Vorgänge metrisch parallel strukturiert sind. Für diese Annahme gibt es aber keinen Grund. Vielmehr wissen wir aus anderen Bereichen, daß die Wahrnehmung nicht metrisch getreu ist, daß zum Beispiel objektiv gleiche Reizauslöser auf Hand und Mund stärker wahrgenommen werden als auf anderen Körperteilen (vgl. den oft abgebildeten Wahrnehmungshomunculus, z.B. in [48]: 197) und daß Bewegungen und Änderungen in Bewegungen in der visuellen Wahrnehmung hervorstechen gegenüber stationären Zuständen bzw. gleichförmigen Bewegungen. Tillmann nimmt nun an, daß das Nervensystem die durch den selben Organismus verursachten artikulatorischen Vorgänge wahrnehmend nicht metrisch getreu abbildet, sondern daß bei dieser Abbildung jede Änderung stark vergrößert wahrgenommen wird, und daß hierdurch die Diskretisierung der objektiv kontinuierlichen Vorgänge zustande kommt."

Vennemann weist dann auf Tillmanns Bezugnahme auf den Homunculus hin und zitiert die Passagen aus [48], die auch für unseren Zusammenhang nichts von ihrer Bedeutung verloren haben und deshalb genau in der von Vennemann gestrafften Gestalt hier wiedergegeben sind:

> "Die artikulatorische Retina ist außerordentlich reich mit taktilen Rezeptoren der verschiedensten Formen bestückt... Wir haben bereits gesehen, daß die gesamte Körperoberfläche zu einen zusammenhängenden Bild auf den sensorischen Kortex projeziert wird. Wir dürfen nun hinzufügen, daß die Projektion der artikulatorischen Retina etwa ein ganzes Drittel dieser Projektionsfläche für sich beansprucht. Der Sprachapparat ist also rein sensorisch ausgesprochen stark repräsentiert. Wir müssen festhalten, daß auf der Hirnrinde ein zusammenhängendes Bild der artikulatorischen Retina existiert." (210)
>
> Es ist heute allgemein bekannt, daß das afferente Nervensystem, durch das ja auch die Reafferenzen übertragen werden, die an der Peripherie

> vorliegenden retinalen Bilder nicht nur einfach überträgt, sondern auf dem Weg zur Großhirnrinde schon so vorverarbeitet, daß Konturen und Bewegungen und andere Verlaufseigenschaften wie das Einsetzen eines Druckes usw. verstärkt und hervorgehoben werden... Wir dürfen also annehmen, daß die bei jedem Sprechakt sich auf der artikulatorischen Retina sich abspielenden Kontakte und Kontaktverläufe auf der sensorischen Großhirnrinde sich in einer in den Konturen wesentlich verschärften Abbildung wiederfinden... Die Abbildung der im Bereich IM [der intramuskulären Aktivität] efferent erzeugten Artikulationsbewegungen auf die artikulatorische Retina ist extrem nicht-linear. Während das efferent erzeugte Artikulationsverhalten in jedem betrachteten Punkt kontinuierlich verläuft, ist das retinale Bild dieser Bewegungen jedoch kein Kontinuum, sondern es zerfällt in Segmente. Das kontinuierliche Oberflächenverhalten produziert nämlich auf der artikulatorischen Retina eine Sequenz diskreter Reizvorgänge. Die Reizorte wechseln und zwar in abrupten Sprüngen.“ (216)
>
> “[Wir] wollen... für die Überführung der äußeren (analogen) Stimuli in diskrete Reizmuster eine eigene Bezeichnung einführen und... von einer retinalen Diskretisierung von analogen Oberflächenreaktionen bzw. kurz von retinaler Diskretisierung sprechen.“ (216)
>
> “[Wir] können... nun sagen, daß sich aus der retinal diskretisierten Form dieses Verhaltens die alphabetische Beschreibung des Gesprochenen ergibt. Und zwar ergeben sich die konsonantischen Prädikate mehr oder weniger direkt aus der taktilen Reizform, während die vokalischen Prädikate aus den Varianten der taktilen Reizmuster abgeleitet werden können.“ (219)

Vennemann zieht aus diesen Zitaten die Schlußfolgerung, der wir uns auch für diese Arbeit anschließen dürfen:

> “Damit ist eine neurophysiologische Erklärung dafür angebahnt, daß wir diskret wahrnehmen, was wir kontinuierlich produzieren. Das Ziel einer Integration einer grammatikalischen (“geisteswissenschaftlichen“) mit einer naturwissenschaftlichen Theorie ist ein Stück nähergerückt.“

2.3 Warum empirisch?

Inzwischen ist das Wort 'empirisch' hier in unterschiedlichen Zusammenhängen so häufig gefallen, daß der Leser inzwischen ein Bild davon haben dürfte, was damit im Kontext dieser Arbeit gemeint ist. Trotzdem sollte noch kurz der Frage nachgegangen werden, warum hier mit der Empirie begonnen wird, anstatt zunächst eine vernünftige Hypothese aufzubauen, die anschließend verifiziert bzw. falsifiziert werden kann.

Ich denke, es gibt mehrere Gründe, bei der Frage, welcher Art die Beziehung von phonologischem Symbol und phonetischen Signal denn wirklich ist, zunächst explorativ vorzugehen. Zum einen könnte es ja sein, daß der Zusammenhang sich so komplex darstellt, daß wir erfolgreiche Hypothesen erst 'sehen' werden, wenn wir genügend Material vor Augen haben. Dies möchte ich die *optimistische* Annahme nennen, weil sie davon ausgeht, daß sich der gesuchte Zusammenhang tatsächlich in Form von explizit formulierbaren Regeln oder Naturgesetzen[7] erfolgreich darstellen läßt. Natürlich gibt es auch eine *pessimistische* Annahme, zu der ich persönlich neige, nämlich die Annahme, daß infolge der unglaublichen Variabilität der Sprache ein solches explizit formulierbares Gesetz nicht existiert, sondern vielmehr eine komplexe statistische Bindung zwischen den beiden Welten, Symbol und Kontinuum, besteht, die es aufzudecken und mit Hilfe geeigneter statistischer Werkzeuge zu modellieren gilt, und deren mathematische Parameter nur mit Hilfe von sehr viel empirisch erhobenem und aufbereitetem Material abgeschätzt werden können.

Nun läßt sich darüber streiten, ob denn eine komplexe statistische Bindung nicht letztendlich auch nur eine Gesetzmäßigkeit darstellt, die sich aufgrund ihrer Komplexität der sprachlichen oder mathematischen Formulierung entzieht. Wie dem auch sei, ich denke, daß es durchaus sinnvoll ist, sich zunächst 'von unten herauf', also über physikalisch belegbares Material an das Problem heranzutasten, und dann zu entscheiden, ob der Optimist oder der Pessimist recht behält.

2.4 Regeln vs. Statistik

Die Diskussion der Dichotomie von regelbasierten vs. statistischen Sprachverarbeitungsverfahren ist auch für unser Vorhaben von zentraler Bedeutung. Während die

[7] der Form 'Wenn A, dann B'.

Anhänger der sogenannten *Natural Language Processing (NLP)*, die laut Prof. Adrian Fourcin besser *Non-spoken Language Processing* heißen könnte, sich vor allem in der Tradition von Chomsky und Halle mit regelbasierten Systemen beschäftigen, bevorzugen Vertreter der *Spoken Language Processing (SLP)* fast ausschließlich statistische Verfahren.

Auch die bisherigen Ausführungen könnten vielleicht den Eindruck hervorrufen, hier würde letztendlich nur für den sogenannten *statistischen Ansatz*, welcher in den ingenieurwissenschaftlichen Disziplinen favorisiert wird, eine Lanze gebrochen. Dieser Eindruck ist falsch, auch wenn es zutrifft, daß wir im Bereich der akustischen Modellierung bei der Entwicklung des MAUS-Systems (bisher) auf regelbasierte Methoden weitgehend verzichtet haben[8]. Auf der Ebene der Aussprache-Modellierung wurde in MAUS von Anfang an mit regelbasiertem Wissen gearbeitet, welches später lediglich durch zusätzliches statistisches Wissen angereichert wurde. MAUS ist also in seinem hier darzustellenden Zustand[9] eines der wenigen technischen sprachverarbeitenden Systeme, in denen tatsächlich statistisches und regelbasiertes Wissen erfolgreich zur Kombination gebracht wird.

2.5 Das Gesetz der großen Zahlen

Trotzdem darf an dieser Stelle der Hinweis nicht fehlen, daß die Entwicklung des MAUS-Systems in erster Linie die statistische Auswertung der *Ergebnisse* eben dieses Systems im Auge hat. Dies läßt sich wie folgt begründen:
Betrachtet man die Erzeugung von gesprochener Sprache als eine physikalische Quelle, welche eine eindimensionale Zeitfunktion hervorbringt, so wird man feststellen, daß diese Quelle eine erstaunlich breit gestreute Variabilität aufweist. Am besten läßt sich dies illustrieren, wenn man nicht das analoge oder das digitalisierte Schalldrucksignal selber, sondern die in der Sprachverarbeitung übliche spektrale Repräsentation in Form von Kurzzeitspektren betrachtet, welche in relativen kurzen Abständen von 10-20 Millisekunden aus dem Schalldrucksignal berechnet werden. Betrachten wir zum Beispiel eine grundfrequenzbereinigte Darstellung in 12 *Mel Frequency Cepstral*

[8] Wobei nicht auszuschließen ist, daß sich dies in naher Zukunft noch ändern wird, nachdem Prof. Tillmann den Beschluß gefaßt hat, sich in naher Zukunft persönlich mit dem Problem der regelbasierten Grenzverbesserung von Segmenten in MAUS (durch die Entwicklung entsprechender MatLab-Programme) beschäftigen zu wollen.

[9] Entwicklungsstand 1999

Coefficiants[10] zusammen mit der jeweils gegebenen lokalen Energie, so ergibt sich mathematisch ein (nicht mehr anschaulich vorstellbarer) Vektorraum der Dimension 13. Kommt dazu noch die erste und zweite zeitliche Ableitung der Vektortrajektorie, die sich aus den zeitlich benachbarten Stützpunkten einfach berechnen läßt, so erhält man sogar einen 39-dimensionalen Vektorraum.[11] Bei einem derart weit aufgespannten Vektorraum würde man erwarten, daß die Trajektorien der Sprache in relativ kleinen, gut separierbaren Bereichen zu finden sind. Das Gegenteil ist der Fall: Sogar nach der bereits informationsreduzierend wirkenden Vorverarbeitung[12] streuen die 39-dimensionalen Vektorverläufe immer noch über einen weit ausgedehnten Bereich. Grund hierfür ist zum einen die erstaunliche Bandbreite innerhalb der einzelnen individuellen menschlichen Stimme, zum zweiten die noch weitaus größere Variation bei der Berücksichtigung verschiedener Sprecher und Situationen: Alter, Geschlecht, Erziehung, Physiologie, Wohnort, Geburtsort, Gesundheitszustand, ja sogar externe Einflüsse, wie Hintergrundgeräusche etc. bilden alle zusammen eine Quelle der Variation, bei der es sich technisch gesprochen um eine Rauschquelle handelt, die sich der eigentlichen Sprachinformation als Störung überlagert. Aus diesem Sachverhalt folgen zwei Dinge:

1. Die Beschreibung des Gesamtprozesses ist ohne statistische Methoden kaum möglich. Warum? Weil die überlagernden Teilprozesse nicht orthogonal sind und daher nicht einfach zum Beispiel durch eine Hauptachsentransformation in Richtung der Eigenvektoren getrennt werden können.

2. Da die das Sprachsignal beschreibenden Vektorräume sehr groß werden, muß, um das der Statistik zugrunde liegende *Gesetz der großen Zahlen* wenigstens einigermaßen zu erfüllen, die untersuchte Datenmenge hinreichend groß sein, und das heißt im Vergleich mit traditionellen Untersuchungen der Phonetik extrem groß.

Das oben Gesagte gilt zunächst nur für die akustische Beschreibungsebene, setzt sich aber analog in den Bereich der symbolisch-phonologischen Repräsentation fort, ja wird unter Umständen sogar noch verschärft. Nehmen wir zum Beispiel an, man

[10] Details zur Vorverarbeitung in *Mel Frequency Cepstral Coefficiants* finden sich in Anhang B.

[11] Solche Dimensionen sind in der Technik durchaus üblich und eher an der unteren Grenze anzusiedeln; viele praktische Applikationen arbeiten mit bis 100 Komponenten.

[12] Die Information über die gesamte A-Prosodie nach Tillmann wird bei der Verarbeitung in MFCC ausgefiltert.

interessiere sich für die tatsächliche Realisierung von deutschen Präfixen. Um zu wirklich fundierten und nicht nur exemplarischen wissenschaftlichen Aussagen zu kommen, benötigt man für jedes einzelne der ca. 800 möglichen Präfixe des Deutschen gemessene Beispiele der hier angesprochenen großen Zahl, so daß nicht nur die Variation innerhalb der einzelnen Sprecher (= genügend Beispiele eines einzelnen Sprechers), sondern auch innerhalb der Sprachgemeinschaft (= genügend große Zahl von Sprechern) abgedeckt wird. Da die Faktoren aber wie gesagt nicht orthogonal sind und sich daher multiplikativ verstärken, kann sich jeder leicht vorstellen, wie groß das Datenmaterial allein für diese einfache Fragestellung sein müßte.

Vor diesem Hintergrund wird eine Hauptmotivation für die Entwicklung der MAUS-Methode nunmehr klar verständlich: Für einen empirisch fundierten Ansatz in der Untersuchung der Beziehungen von phonologischen Symbolen und Signalen ist für eine wissenschaftlich zuverlässige Klärung sehr viel Material erforderlich. Dieses Material kann nur durch weitgehend automatisierte Verfahren wie das MAUS-System gewonnen werden.

Kapitel 3

MAUS

In diesem Kapitel soll die grundlegende Methode, nach der das MAUS-System arbeitet, beschrieben werden. Wie bereits vorab erwähnt, habe ich mich bemüht, die Darstellung knapp und übersichtlich zu gestalten, damit dem Leser der Blick auf die wesentliche Grundidee nicht durch zu viele Details verbaut wird. An den entsprechenden Stellen wird daher auf die Anhänge A.1 und A.2 verwiesen, wo man die ausführlichen Details über den mathematischen Formalismus und das Design des MAUS-Systems findet. Für das bessere Verständnis des von uns entwickelten Systems erscheint es mir außerdem sinnvoll, historisch vorzugehen, d.h. zunächst einige bereits bekannte automatische bzw. halbautomatische Verfahren der automatischen Segmentierung und Etikettierung zu skizzieren, dann die Grundidee des MAUS-Systems zu beschreiben und schließlich die Entwicklung des MAUS-Systems in seinen beiden prinzipiellen Varianten nachzuzeichnen. Zunächst aber muß ich noch auf ein grundlegendes Problem näher eingehen, nämlich die Wahl des geeigneten Symbol-Alphabets.

3.1 Das Problem der Kategorien

Durch die Auswahl des Symbol-Alphabets wird die Art und Weise, implizit auch die Qualität der automatischen Segmentierung und Etikettierung festgelegt. Aus dem bisher Gesagten ist klar geworden, daß es sich um phonologische Kategorien handeln muß, die sinnvollerweise Segmenten aus einem Schalldrucksignal zugeordnet werden können (auch wenn in der Literatur manche der folgenden Einheiten als 'phonetische' bezeichnet werden). Als phonologische Einheiten kommen in Betracht: Wörter, Morpheme, Silben, Laute oder Teile von Lauten.

Wörter sind hier nur der Vollständigkeit halber genannt. Obwohl eine Segmentierung und Etikettierung in Wörter (= Spracherkennung) natürlich sehr wünschenswert wäre, scheitert dies natürlich an der nach oben offenen Menge der möglichen Wörter einer Sprache[1]. Ein so großes Inventar läßt sich unmöglich mit individuellen akustischen Modellen bearbeiten; daher scheidet die Einheit Wort von vorne herein aus.

Morpheme, Silben und Laute sind alles mögliche Kandidaten für eine automatische Segmentierung und Etikettierung. Morpheme und Silben hätten den Vorteil, daß sie aufgrund ihrer komplexeren intrinsischen Struktur leichter (als Gesamteinheit) erkannt werden können. Allerdings wird dieser Vorteil fast wieder aufgehoben durch die immer noch beträchtliche Anzahl[2] der zu modellierenden Einheiten. Ein weiterer Nachteil ist natürlich, daß das Segmentierungsergebnis für viele Untersuchungen und Anwendungen zu grob ist. Teile von Lauten andererseits sind meistens zu kurz, um mit Methoden der digitalen Sprachsignalverarbeitung noch vernünftig erfaßt werden zu können. Beispielsweise verhindert das *Reziprozitätsgesetz der Systemtheorie* (z.B. [28]), daß sehr kurze Signalstücke im spektralen Bereich mit genügender Auflösung dargestellt werden können.

Damit bleibt als vernünftigste alphabetische Einheit der Sprachlaut: Er ist in der Regel von genügender Länge (> 20 ms), um noch verarbeitet werden zu können, und eignet sich nicht nur für Untersuchungen im Lautbereich selbst, sondern auch – bei Zusammenfassung mehrerer Laute zu Clustern – für die Untersuchung von Silben, Morphemen oder Wörtern.

Die nächste Frage, die sich sofort anschließt, ist, welches Lautinventar verwendet werden soll. Vom phonologischen Standpunkt wäre natürlich ein streng gültiges Phoneminventar der entsprechenden Sprache die nächstliegende Wahl. Dabei ergeben sich jedoch auf der Seite der akustischen Modellierung einige Probleme. Zunächst kann eine akustische Modellierung keine *gefüllten Lücken* zulassen; der Schalldruckverlauf muß als Ganzes ohne Ausgrenzungen betrachtet werden. Gefüllte Lücken[3] – in Ge-

[1] Im PHONOLEX-Projekt, das in meiner Arbeitsgruppe durchgeführt wird, geht es um die Erstellung eines möglichst umfassenden Aussprache-Lexikons des derzeit gesprochenen Deutschen. Die dabei erstellte Liste unterschiedlicher Wortformen hat inzwischen (Stand 2000) die Marke von 1600000 überschritten und es ist keine Sättigung absehbar.

[2] Über die tatsächliche Anzahl von möglichen Morphemen und Silben im Deutschen gehen die Meinungen ein wenig auseinander; die meisten Literaturstellen sprechen jedoch von mehr als 10000 Einheiten.

[3] Nicht zu verwechseln mit dem englischen Ausdruck 'filled pauses' für stimmhafte Hesitationen.

gensatz zu nicht gefüllten Lücken, also Pausen – entstünden immer dann, wenn ein Laut auftritt, der nach einem strengen Phonemsystem in keinem denkbaren Wortpaar zu einer phonologischen Distinktion führt. Typisches Beispiel im Deutschen wäre der Schwa-Laut[4]. Obwohl dieser streng genommen kein Phonem darstellt, braucht die akustische Modellierung dennoch dafür ein Modell, um die Lücke zu schließen. Das zweite Problem ergibt sich aus der zum Teil erheblichen phonetischen Variation innerhalb einer Phonem-Kategorie. Nehmen wir zum Beispiel die Frikative [ç] in 'ich' und [x] in 'ach', so sind diese genaugenommen Allophone der Phonem-Klasse /ch/, also Mitglieder der gleichen Kategorie. Nun unterscheiden sich aber die Signale von [ç] und [x] erheblich: Das Spektrum von [ç] ist flacher und in [x] sind infolge der größeren artikulatorischen Tiefe meistens deutliche Formantstrukturen zu erkennen. Der Versuch, solche unterschiedlichen physikalischen Signale mit ein und demselben akustischen Modell nachzubilden, führt unweigerlich zu Konflikten, die nur durch suboptimale Kompromißlösungen umgangen werden können[5]. Sinnvoller – von der akustischen Seite her gesehen – ist es, solche Signale als separate Kategorien zu behandeln.

Folgt man diesem pragmatischen Ansatz, so erhält man letztendlich ein Inventar, das genaugenommen weder eine 'breite' noch eine 'enge' Transkription ermöglicht, sondern eine Mischform: bestimmte Laute werden tatsächlich als phonologische Einheit behandelt, andere – wo die Akustik es erfordert – als Allophone. Die Entscheidung darüber, welche Allophone gemeinsam und welche getrennt behandelt werden, wird anhand der qualitativen phonetischen Ähnlichkeit getroffen. Von phonologischen Standpunkt aus ist dies zunächst mit keinerlei Nachteilen verbunden, weil durch eine einfache Nachbearbeitung die entsprechenden Allophongruppen eindeutig zu Phonemen zusammengefaßt werden können.

Nun ist dieser Aspekt der MAUS-Entwicklung nichts Neues. Bereits Anfang der 90iger Jahre wurde mit dem europäischen Forschungsprojekt 'Speech Assessment Methods' (SAM, vgl. [29]) für jede europäische Sprache[6] ein solches phonologisch-phonetisches Sub-Inventar des IPA-Alphabets ([33]) festgelegt, das inzwischen im Bereich der SLP zum Standard geworden ist. Ein großer praktischer Vorteil des *SAM Phonetic Alphabets* (SAM-PA) besteht darin, daß es mit dem standardisierten ASCII-Zeichensatz im Computer dargestellt werden kann.

[4]Nicht ganz unumstritten; manche Autoren erklären auch den Schwa-Laut zum Phonem

[5]z.B. durch Verwendung von parallelen Verarbeitungssträngen.

[6]Mittlerweile auch für die wichtigsten außereuropäischen Sprachen.

Tabelle 3.1 zeigt das in unseren Untersuchungen verwendete Lautinventar mit Beispielen und dem Bezug zum IPA-Alphabet. Abweichungen vom SAM Standard sind mit einem Sternchen gekennzeichnet; diese erfolgten jedoch nur aus Gründen der einfacheren Handhabung und können jederzeit durch einfachen Austausch der Symbole rückgängig gemacht werden.

IPA Nummer	BEISPIEL	IPA NAME	SAM-PA
Vokale			
304, 503	Kahn	Lower-case A, Length Mark	a:
304	kann	Lower-case A	a
302, 503	Beet	Lower-case E, Length Mark	e:
302	Meteor	Lower-case E	e
303	Bett	Epsilon	E
301, 503	riet	Lower-case I, Lenght Mark	i:
301	Politik	Lower-case I	i
319	ritt	Small Capital I	I
307, 503	bog	Lower-case O, Length Mark	o:
307	Politik	Lower-case O	o
306	Bock	Open O	O
308,503	Mus	Lower-case U, Length Mark	u:
308	Kulisse	Lower-case U	u
321	muss	Upsilon	U
309, 503	H"ute	Lower-case Y, Length Mark	y:
309	kyrillisch	Lower-case Y	y
320	H"utte	Small Capital Y	Y
303, 503	K"ase	Epsilon, Length Mark	E:
310, 503	H"ohle	Slashed O, Length Mark	2:
310	"Okonom	Slashed O	2
311	H"olle	O-E Digraph	9
Nasalierte Vokale			
304, 424	Restaurant	Lower-case A, Superposed Tilde	a~
303, 424 *	Teint	Epsilon, Superposed Tilde	E~
306, 424 *	Saison	Open O, Superposed Tilde	O~
311, 424	Parf"um	O-E Digraph, Superposed Tilde	9~

Diphthonge

304, 319	zwei	Lower-case A, Small Capital I	aI
304, 321	Bauch	Lower-case A, Upsilon	aU
306, 320	neun	Open O, Small Capital Y	OY

Unbetonte Vokale

322	lesen	Schwa	@
324	Leser	Turned A	6

Konsonanten

133	lesen	Lower-case Z	z
134	Tasche	Esh	S
135	Loge	Yogh	Z
138	dich	C Cedilla	C
140	Dach	Lower-case X	x
119	Junge	Eng	N
113 *	ich	Upper-case Q	Q
102	bei	Lower-case B	b
104	du	Lowr-case D	d
128	verfahren	Lower-case F	f
110	Gast	Lower-case G	g
146	Hast	Lower-case H	h
153	ja	Lower-case J	j
109	Kahn	Lower-case K	k
155	Licht	Lower-case L	l
114	Mann	Lower-case M	m
116	neun	Lower-case N	n
101	Platz	Lower-case P	p
122	Rauch	Lower-case R	r
132	las, Ma"s	Lower-case S	s
103	Torte	Lower-case T	t
129	Vase, wann	Lower-case V	v

Diacritica

503	Lengthening	Length Mark	(Vowel):
501 *	Primary Stress	Vertical Stroke(Superior)	'(Vowel)
502 *	Second. Stress	Vertical Stroke(Inferior)	"(Vowel)
406 *	Glottalization	Subscript Tilde	q(Item)
424	Nasalization	Superposed Tilde	(Item)~

Tabelle 3.1: SAM Phonetic Alphabet für die deutsche Sprache ([29]). Erläuterung: Auf der Kiel-Convention der *International Phonetic Association* wurden in der von Barry und Tillmann geleiteten Arbeitsgruppe 'Computer representation of individual language' folgende Erweiterungen zu den traditionellen IPA-Lautsymbolen vereinbart: *IPA-Nummern*, welche alle IPA-Zeichen und IPA-Diacritica als eindeutige Nummer identifizieren, und *IPA-Namen*, welche die Form des IPA-Zeichens drucktechnisch eindeutig beschreiben. Auf diese Weise wurde es erstmals möglich, in herkömmlichen Computer-Systemen mit IPA-Zeichen zu rechnen.

3.2 Automatische Verfahren – Überblick

Forschungsanstrengungen zur Erreichung einer halb- oder vollautomatischen Etikettierung und vor allem einer Segmentierung von Sprache haben eine lange Tradition. Sie wurden einerseits von Grundlagenforschern, andererseits auch von anwendungsorientierten Ingenieurswissenschaftlern unternommen, und dies schlägt sich in der Art und Weise des jeweiligen Vorgehens deutlich nieder. Während Linguisten, Phonologen und Phonetiker dazu tendieren, mit regelbasierten Methoden auf der Ebene von phonetischen Merkmalen zu operieren, findet man auf der technischen Seite fast ausschließlich die sogenannten statistischen Maximierungsansätze, bei denen es darum geht, einen bestimmten statistischen Ausdruck in geeigneter Weise gezielt zu maximieren.

In der jüngeren Vergangenheit hat sich gezeigt, daß analog zu den Anstrengungen im Bereich der automatischen Spracherkennung auch bei der automatischen Etikettierung und Segmentierung die regelbasierten Verfahren den statistischen deutlich unterlegen sind. Insbesondere zeigt sich bei regelbasierten Ansätzen fast immer das Problem, daß das System nur für die Äußerungen eines einzelnen Sprechers optimiert werden kann. Darüber hinaus bedeutet die Optimierung von Regeln – besonders

wenn es sich um Ansätze der in Japan favorisierten Fuzzy-Logik handelt – in den meisten Fällen mühsame Handarbeit bei der Festsetzung der Regelschwellwerte[7], während statistische Verfahren meistens selbstlernend sind. Konsequenterweise haben sich auch die wissenschaftlichen Ansätze zur automatischen Etikettierung und Segmentierung in den letzten beiden Dekaden auf statistische Verfahren konzentriert. Daher werde ich mich hier auf diese Ansätze beschränken.

3.2.1 Statistischer Formalismus

Folgt man den statistischen Ansatz zur Lösung eines Segmentierungsproblems, wird man in den meisten Fällen den folgenden Formalismus verwenden:

Ein Sprachsignal sei durch eine Folge O kleinster, nicht mehr teilbarerer Atome $O = o_1, o_2, ...o_N$ beschrieben. Solche Atome können z.B. die Abtastwerte eines nach Nyquist korrekt abgetasteten Schalldrucksignals sein; in diesem Falle entspricht die Länge eines Atoms, oder besser des kleinsten sinnvoll darstellbaren Segments, der Inversen der doppelten höchsten Frequenz im Spektrum des Sprachsignals (*Nyquist-Bedingung* oder auch *Abtasttheorem*). Meistens jedoch handelt es sich bereits um sehr viel längere Einheiten, die aus einer Vorverarbeitung des Sprachsignals, einer *Merkmalsextraktion*, stammen.

> Beispiel:
>
> In einen typischen Sprache verarbeitenden System wird das Sprachsignal mit mit einem Spektrum der Bandbreite 8kHz nach Nyquist mit 16kHz abgetastet. Daraus ergibt sich ein Abstand der Abtastwerte von 0.0000625 Sekunden. Im ersten Verarbeitungsschritt werden jedoch zur Berechnung des Kurzzeitspektrums jeweils überlappende Bereiche von ca. 20 Millisekunden Breite und in einem Abstand von jeweils 10 Millisekunden zusammengefaßt und zu einer Folge von Merkmalsvektoren umgeformt. Ein 'Atom' ist folglich nach der Merkmalsextraktion 10 Millisekunden breit[8].

Da die Atome nicht die Länge Null haben können, folgt daraus, daß O aus einer endlichen Anzahl $N(O)$ Atomen besteht. Ein Segment innerhalb unseres Sprachsi-

[7] Eine Ausnahme bildet in Bereich der Spracherkennung das in [2] beschriebene System, welches ein selbstlernendes Regel-System verwendet.

[8] Vgl. dazu auch Anhang B.

gnals besteht ebenfalls aus einer endlichen Anzahl von Atomen, die größer als Null, aber auf jeden Fall kleiner oder gleich $N(O)$ sein muß. Geht man davon aus, daß das Signal in eine bündige Folge von Segmenten aufgeteilt werden soll, wird sofort klar, daß es dazu prinzipiell zwar sehr viele, aber endlich viele Kombinationsmöglichkeiten (Segmentierungen) R gibt, die in einer Menge aller möglichen Segmentierungen Ψ zusammengefaßt werden können. Der eine Extremfall einer Segmentierung ist ein einziges Segment mit $N(O)$ Atomen; der andere Extremfall sind $N(O)$ Segmente der Länge von jeweils einem Atom.

Ein statistischer Ansatz setzt nun ein Modell (eine Wahrscheinlichkeitsdichtefunktion) als gegeben voraus, welches jeweils die Wahrscheinlichkeit wiedergibt, daß bei gegebenem Sprachsignal O eine Segmentierung R korrekt ist: $P(R|O)$. Woher dieses Modell stammt, interessiert vorerst nicht. Die Lösung des Segmentierungsproblems läßt sich nunmehr in ein Optimierungsproblem umformulieren, derart daß die Segmentierung $\hat{R}$ gesucht wird, für die gilt:

$$\hat{R} = \underset{R \in \Psi}{\operatorname{argmax}} P(R|O) \tag{3.1}$$

Bevor wir weiter gehen, zunächst zwei Anmerkungen:

1. Diese Herleitung betrachtet zur Vereinfachung der Darstellung tatsächlich nur die Optimierung über die Menge aller möglichen *Segmentierungen* eines Signals. Die Etikettierung der Segmente, d.h. die Entscheidung über die Klassenzugehörigkeit, wurde bisher gar nicht formalisiert und muß tatsächlich als eine zusätzliche statistische Größe mit berücksichtigt werden. Dies ändert jedoch nichts an der prinzipiellen Vorgehensweise beim statistischen Ansatz. Daher seien in der weiteren Herleitung die Begriffe 'Segmentierung' und 'Lautfolge' als synonym anzusehen.

2. Das Modell $P(R|O)$ kann man sich als Dichtefunktion über einer 'Hyperebene' vorstellen, welche – als mathematische Menge gesehen – alle möglichen Segmentierungen repräsentiert. $P(R|O)$ wird dort, wo die 'wahre' Segmentierung[9] liegt, ein Maximum haben.

$P(R|O)$ läßt sich direkt weder beobachten noch messen. Der Grund hierfür ist einsichtig: eine empirische Messung dieser Funktion wäre nur möglich, wenn man alle

[9]im Sinne der Optimierung.

denkbaren O, also alle denkbaren Sprachsignale untersucht hätte. Dies ist selbstverständlich unmöglich. Insofern scheint uns der statistische Ansatz auf ein unlösbares Problem geführt zu haben.

Nach dem *Bayes Theorem* läßt sich dieses empirisch unlösbare Problem jedoch in eine andere Form bringen, die uns weiterhilft. Es gilt nämlich die Identität:

$$P(O|R)P(R) = P(R|O)P(O) \tag{3.2}$$

und somit wird der Ausdruck in 3.1 zu:

$$\hat{R} = \underset{R \in \Psi}{\operatorname{argmax}} \frac{P(O|R)P(R)}{P(O)} \tag{3.3}$$

Die in diesem Ausdruck enthaltenen Terme lassen sich wie folgt interpretieren:

- $P(R)$ gibt die *apriori Wahrscheinlichkeit* für das Auftreten einer bestimmten Segmentierung R wieder. Dies mag überraschen, da man zunächst annehmen möchte, alle Segmentierungen seien gleichermaßen zugelassen und damit auch gleich wahrscheinlich.
  Die oben erwähnten Extrembeispiele für mögliche Segmentierungen machen sofort klar, daß dem nicht so ist: eine Segmentierung, bei der das gesamte Sprachsignal aus Segmenten von Atomlänge besteht, ist extrem unwahrscheinlich. Folglich müßte $P(R)$ für diesem Fall nahe bei Null anzusiedeln sein.
  Ein anderes Gedankenexperiment: Stellen wir uns vor, eine Segmentierung R beschreibe die Phonemefolge /ktrslt/. Eine Lautkombination dieser Art ist im Deutschen, selbst unter Einbeziehung von Dialekten und Spontansprache, nicht zu erwarten. Auch hier müßte $P(R)$ einen Wert nahe Null annehmen.
  $P(R)$ ist somit ein Ausdruck, welcher die *vom Sprachgebilde* definierten, allgemeinen Wahrscheinlichkeiten für konkrete Phonem-Folgen beliebiger Art wiedergibt. Man könnte auch sagen, $P(R)$ repräsentiert das *phonologische Wissen* in der Suche nach der optimalen Segmentierung einer Äußerung. Interessant in diesem Zusammenhang ist, daß $P(R)$ zwar formal eine Wahrscheinlichkeitsfunktion darstellt, aber ohne Einschränkung der Allgemeinheit auch eine binäre 'Wahr/Falsch' Entscheidung für bestimmte Phonemfolgen zur Anwendung kommen könnte, z.B. als das Ergebnis einer phonologischen Grammatik. In letzterem Falle würden ganz einfach die Wahrscheinlichkeitswerte, die $P(R)$ annehmen kann, entweder 1 (für 'Wahr') oder 0 (für 'Falsch') sein.

- $P(O|R)$ beschreibt formal die Wahrscheinlichkeit für ein bestimmtes Signal O gegeben eine bestimmte konkrete Phonemfolge R. Insofern könnte man diesen

Term auch als *akustisches Quellenmodell* bezeichnen. Da die Segmentierungen $R \in \Psi$ diskrete, abzählbare Ereignisse darstellen, läßt sich dieser Term anschaulich als eine Sammlung vom $|\Psi|$ verschiedenen Wahrscheinlichkeitsdichtefunktionen über O vorstellen, wobei $|\Psi|$ gleich der Anzahl der Elemente von Ψ, also aller möglichen Segmentierungen ist. Die individuelle Dichtefunktion $P(O|\bar{R})$ für eine konkrete Segmentierung $\bar{R}$ beschreibt dann die Wahrscheinlichkeiten für jedes akustische Signal O, daß es von einer Äußerung mit der Segmentierung $\bar{R}$ stammt.
$P(O|\bar{R})$ läßt sich empirisch abschätzen, indem man eine genügend große Menge von Äußerungen der gleichen Lautfolge $\bar{R}$ beobachtet. Es mag zunächst unrealistisch erscheinen, daß sich auf diese Weise auch $P(O|R)$ abschätzen läßt, weil dazu sehr viele Realisierungen von jeweils sehr vielen möglichen Äußerungen beobachtet werden müssen. In der Praxis läßt sich der Aufwand jedoch auf ein realisierbares Maß reduzieren, unter der Annahme, daß die akustischen Realisierungen einzelner Phoneme statistisch voneinander unabhängig sind.
Der Term $P(O|R)$ repräsentiert also in der Suche nach der optimalen Segmentierung die *akustische Modellierung*, unabhängig von der Auftretenswahrscheinlichkeit der Äußerung (Segmentierung) an sich. Man könnte daher auch sagen, $P(O|R)$ ist der Beitrag des *phonetischen Wissens* bei der Suche nach der optimalen Segmentierung.

- Der Term $P(O)$ im Nenner des zu maximierenden Ausdrucks steht analog zu $P(R)$ für die *apriori Wahrscheinlichkeit* für das Auftreten eines Signals O an sich und ist völlig unabhängig von der Art der Segmentierung R. Schon rein formal wird klar, daß dieser Term – da nicht von R abhängig – für alle R konstant bleiben muß. Da wir aber über alle $R \in \Psi$ maximieren, kann dieser Term entfallen und es bleibt:

$$\hat{R} = \underset{R \in \Psi}{\operatorname{argmax}} P(O|R)P(R) \tag{3.4}$$

  als Optimierungsbedingung.

Betrachtet man diesen Formalismus im Lichte der Dinge, die in Abschnitt 1.1 über das problematische Verhältnis von Phonologie und Phonetik gesagt wurden, ist es bemerkenswert, wie problemlos hier phonetisches und phonologisches Wissen Hand in Hand gehen.

3.2.2 Die Extremfälle

Vergleicht man die existierenden Verfahren zur automatischen Etikettierung und Segmentierung, so stellt man fest, daß mit wenigen Ausnahmen die *akustische Modellierung*, also der Term $P(O|R)$, in prinzipiell ähnlicher Weise, nämlich durch Verkettung von *Hidden Markov Modellen (HMM)* realisiert wird. Auf die Theorie der HMM möchte ich aus Platzgründen in dieser Arbeit nicht eingehen. Eine sehr gute Einführung findet sich z.B. in [39]. Für das weitere Verständnis genügt es zu wissen, daß sich mit Hilfe von segmentalen HMM, in unserem Falle Phonem-HMM, jede beliebige Äußerung einer Sprache durch Verkettung dieser Modelle akustisch modellieren läßt. D.h. konkret, daß bei vorgegebener Phonemfolge $\bar{R}$ für jedes unbekannte Signal O eine Wahrscheinlichkeit $P(O|\bar{R})$ berechnet werden kann.

Der zweite Term $P(R)$ in unserer Optimierungsformel wird dagegen sehr unterschiedlich behandelt. Das Spektrum reicht dabei von der *freien Phonemerkennung* bis hin zum sogenannten *Forced Alignment*[10].

Freie Phonemerkennung

Bei der *freien Phonemerkennung* handelt es sich genau genommen um eine automatische Spracherkennung auf Phonembasis[11]. Nur das akustische Signal ist in diesem Falle entscheidend für das Ergebnis; der Term $P(R)$ wird daher einfach zu 1 gesetzt.

Vom phonetischen Standpunkt aus wäre ein *freier Phonemerkenner* natürlich der ideale Algorithmus für die automatische Etikettierung und Segmentierung – wenn er funktionieren würde. Wenn man davon ausgeht, daß der Mensch nach mehreren Millionen Jahren evolutionärer Entwicklung eine optimale Fertigkeit beim Verstehen von Sprache entwickelt hat – und nichts scheint im Moment auf das Gegenteil hinzudeuten –, dann ist es bemerkenswert, daß er Logatome, also Wörter ohne Semantik, nicht fehlerfrei zu transkribieren vermag. Vermutlich ist die vollständige Information über die phonologisch-phonetische Lautfolge im Schalldrucksignal eben nur teilweise enthalten. Es überrascht daher nicht, daß nach dem heutigen Stand der Technik in der der *freien Phonemerkennung* bestenfalls *Akuratheiten*[12] von 65-70% zu erzielen

[10] Am ehesten mit 'erzwungene Abbildung' wiederzugeben.

[11] Gewöhnlich wird automatische Spracherkennung auf Wortbasis durchgeführt. Ersetzt man jedoch das Lexikon bekannter Wörter durch die Liste bekannter Phoneme einer Sprache, erhält man analog einen Spracherkenner, der Phoneme erkennt, also transkribiert.

[12] Zur Definition von *Akuratheit* siehe Abschnitt 4.1.1

sind.

Aus diesem Grunde entspricht dieser Extremfall nur einer Idealvorstellung und wird für die automatische Etikettierung und Segmentierung nicht verwendet. Trotzdem gibt es Einzelfälle, in denen *freie Phonemerkennung* – auch wenn sie fehlerhaft ist – eingesetzt wird, z.B. zum automatischen Auffinden von neuen Aussprachevarianten ([37]) oder zur Transkription oder zum Lernen unbekannter Namen (z.B. [46]).

Berücksichtigung der Phonotaktik

Die nächste Stufe in Richtung des anderen Extremfalles, des *Forced Aligments*, ist ein Phonemerkenner, der Wissen über die *Phonotaktik*, also die erlaubten Abfolgen von Lauten, in der Zielsprache verwendet.

In seiner einfachsten Form kann dies durch ein *statistisches Bigramm-Modell* geschehen, welches für jedes Phonem X eine diskrete Auftretenswahrscheinlichkeit $P(X|Y)$ berechnet, unter der Bedingung, daß vor X das Phonem Y aufgetreten ist. Ein solches Bigramm-Modell läßt sich aus einer großen Menge von Transkripten automatisch generieren.

Eine andere Möglichkeit ist die Verwendung einer phonologischen Grammatik, welche binäre Aussagen über das mögliche Auftreten des Phonems X im Kontext VXN liefert, wobei Prä- und Postkontexte V und N vor und nach dem Element X in diesem Fall nicht auf ein Phonem beschränkt sein müssen.

In beiden Fällen wird der Term $P(R)$ in unserer Optimierungsformel durch die entsprechende apriori Wahrscheinlichkeit der Phonemfolge R ersetzt. Untersuchungen der Verbmobil-Forschungsgruppe an der Universität Hamburg haben gezeigt, daß sich auf diese Weise die Qualität der freien Phonemerkennung signifikant steigern läßt[13].

Aussprachemodellierung

Die Phonotaktik berücksichtigt kein Wissen über die Wortstruktur der Zielsprache. Durch eine explizite Modellierung der möglichen Aussprachеformen von Wörtern läßt sich der Suchraum der Optimierung weiter einschränken. Ist zum Beispiel bekannt, daß die Endung 'er' im Deutschen fast ausschließlich als tiefer Schwa [ɒ] realisiert wird, so kann man dieses Wissen nutzen, um den Term $P(R)$ für alle

[13]Leider liegen dazu keine Veröffentlichungen vor.

anderen Fälle entsprechend einzuschränken[14]. Wie man sich leicht vorstellen kann, gibt es sehr viele verschiedene Möglichkeiten der Aussprachemodellierung. Das Spektrum reicht von rein statistischen Verfahren, welche für jedes mögliche Wort mehrere mögliche Aussprachen zusammen mit deren Auftretenswahrscheinlichkeiten modellieren (z.B. in [46, 1]), über regelbasierte Systeme, die entweder phonologisch oder statistisch basiert sein können (an dieser Stelle läßt sich auch das MAUS-System einordnen), bis hin zur expliziten Vorgabe der jeweiligen Aussprache in Form einer vorgefertigten Transkription (z.B. [35]).

In allen Fällen wird der Term $P(R)$ für eine konkrete Realisierung R einer Äußerung derart berechnet, daß er die apriori Wahrscheinlichkeit für die gewählte Aussprache wiedergibt.

Forced Alignment

Als Kontrapunkt bleibt, wie bereits erwähnt, die Möglichkeit, ein unbekanntes Sprachsignal auf eine bereits vorgegebene Folge von Lauten zu segmentieren. Der Suchraum (und damit der technische Aufwand) reduziert sich für diesen Fall enorm, weil sich nur noch die einzelnen Segmentgrenzen innerhalb der Segmentierung verschieben können. In den meisten Fällen handelt es sich um ein sog. *kanonisches Alignment*, womit eine Segmentierung auf die Standardaussprache, meistens die Aussprache isoliert gesprochener Wörter, gemeint ist. *Forced Alignment* wird in Trainingsverfahren der automatischen Spracherkennung häufig eingesetzt, um das Problem der fehlenden Segmentierung in den Trainingsdaten zu umgehen. Dabei wird in einem iterativen Prozeß das Sprachmaterial immer wieder neu segmentiert und anschließend die akustischen Modelle anhand der dabei gefundenen Segmentierung solange nachgeschätzt, bis das System auf einer Teststichprobe konvergiert (siehe z.B. [42]).

Für die automatische Etikettierung und Segmentierung ist *Forced Aligment* nur von geringer Bedeutung, weil sich das vorgegebene Transkript nicht ändern kann. Hat z.B. ein Sprecher das Wort 'haben' mit [haːm] anstatt kanonisch [haːbən] ausgesprochen, so wird dies beim Segmentieren mit *Forced Aligment* nicht detektiert. Im Gegenteil muß zwangsläufig eine Fehlsegmentierung und -etikettierung stattfinden, weil das tatsächlich realisierte finale [m] auf die vorgeschriebenen Etiketten [bən] verteilt wird. Trotzdem findet diese Methode auch in der phonetischen Forschung

[14] In der englischsprachigen Literatur spricht man in diesem Falle auch ganz richtig von 'constraints'.

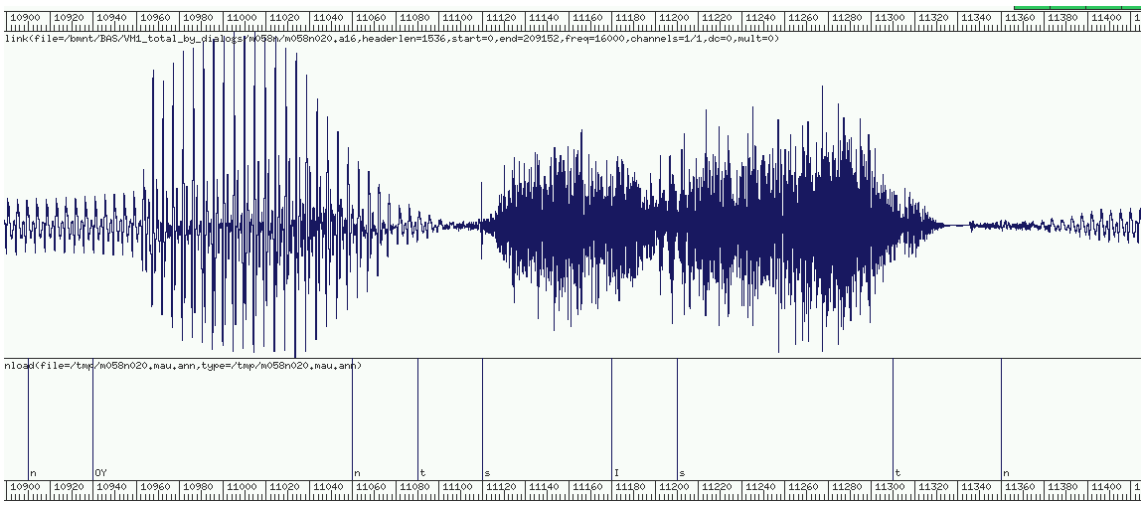


Abbildung 3.1: Schalldrucksignal des Wortes 'neunzigsten'; Ausschnitt aus einer spontanen Äußerung im Verbmobil-Korpus (Darstellung mit SfS)

vor allem zur zeitsparenden Segmentierung von akustischen Einzelereignissen (z.B. Silben, Wörtern) ihre Anwendung.

3.3 Eine einfache Idee: Grundprinzip von MAUS

Nachdem der Kontext für die MAUS-Methode klar geworden ist, möchte ich in diesem Abschnitt das Grundprinzip an einem einfachen Beispiel erläutern, bevor ich in den weiteren Abschnitten auf die verschiedenen Varianten von MAUS eingehen werde. Abbildung 3.1 zeigt das Schalldrucksignal des Wortes 'neunzigsten', welches aus einem spontan geäußerten Satz herausgeschnitten wurde. Dieses Signal zusammen mit der Information, daß irgendwo in diesem Bereich das Wort 'neunzigsten' gesprochen wurde, bildet den Ausgangspunkt für unser Beispiel.

Die weitere Verarbeitung erfolgt zunächst in zwei parallelen Zweigen[15]:

Signalverarbeitung des Schalldrucksignals

Das Schalldrucksignal wird einer Reihe von digitalen Verarbeitungschritten unterworfen[16]:

[15] Man könnte sogar sagen, in einem phonetischen und einem phonologischen Zweig.

[16] Mathematische Details in Anhang B.

1. Das Signal wird mit einem Hochpaß erster Ordnung gefiltert, um den typischen Abfall des Sprachspektrums auszugleichen. Dieser Abfall zu höheren Frequenzen wird einerseits durch den Abfall des Anregungsspektrums der menschlichen Glottis und andererseits durch die akustische Abstrahlungscharakteristik einer Kugelquelle (von der Mundöffnung in den Raum) verursacht. Beide Effekte überlagern sich nach der Systemtheorie ([28]) im Spektrum multiplikativ, d.h. das Anregungsspektrum der Glottis wird mit der Übertragungsfunktion der Abstrahlung multipliziert, was zu einem Abfall des Spektrum von etwa 6 dB/Oktave führt. Für die nummerische Verarbeitung der spektralen Werte ist es natürlich günstiger, wenn diese über den gesamten betrachteten Spektralbereich ungefähr in den gleichen Zahlenbereich zu liegen kommen; die Hochpaßfilterung erzielt durch Anhebung der höheren Frequenzbereiche genau den gewünschten Effekt.

2. Das Signal wird um einen etwaigen Gleichwertanteil (arithmetischer Mittelwert) ungleich Null bereinigt. Solche Gleichwertanteile können aus der analogen Verstärker- bzw. Filterschaltung vor der Analog/Digital-Wandlung stammen. Sie bewirken eine Verfälschung in der Berechnung der lokalen Energie des Schalldrucksignals (s. nächster Punkt). Daher wird ein etwaiger Gleichwertanteil vorab berechnet und vom Schalldrucksignal subtrahiert. Das Resultat dieser Operation ist, daß das Schalldrucksignal vollkommen symmetrisch um die Nulllage schwingt.

3. Das Signal wird 'gefenstert', d.h. mit einer glockenförmigen Fensterfunktion (*Hamming*-Fenster der Breite 20 Millisekunden) multipliziert, und innerhalb dieses Fensters werden aus dem verbleibenden Signalstück spektrale Parameter (*Merkmale*) berechnet und zu einem *Merkmalsvektor* zusammengefaßt (Kurzzeitanalyse, vgl. dazu auch Abb. B.1 auf Seite 165 sowie [39]). Im Falle des MAUS-Systems sind die extrahierten Merkmale

   - die lokale Energie (durch Quadration und Integration des Signals im Fensterbereich)
   - die erste und zweite zeitliche Ableitung der lokalen Energie
   - 12 gehörgerecht berechnete Cepstral-Koeffizienten (*Mel-Frequency-Cepstral-Coefficiants (MFCC)*, siehe Anhang B). Cepstrale Parameter stellen eine Sonderform von spektraler Darstellung dar, wie sie in der digitalen Sprachverarbeitung sehr häufig zur Anwendung kommt. Für das weitere Verständnis genügt es zu wissen, daß cepstrale Parameter in erster

Näherung das Spektrum eines Sprachsignals, bereinigt um seine Periodizität, beschreiben, also eine Abschätzung der Übertragungsfunktion des Vokaltraktes darstellen.

- die erste und zweite zeitliche Ableitung der 12 cepstralen Parameter

Die Dimensionalität der Merkmalsvektoren in MAUS ist somit 39[17]

Die oben skizzierte *Merkmalsextraktion* transformiert also das ein-dimensionale, relativ dicht abgetastete[18] Schalldrucksignal in eine relative grob abgetastete[19] Trajektorie im 39-dimensionalen Raum. Soweit der *phonetische* Zweig der Verarbeitung.

Wie können wir uns diese Trajektorie für unser Beispiel 'neunzigsten' vorstellen? Eigentlich übersteigt dies bereits unser räumliches Vorstellungsvermögen bei weiten; dennoch möchte ich versuchen, einige anschauliche Hinweise zu geben, indem wir nur zwei der Parameter, nämlich die lokale Energie e und den ersten Cepstral-Koeffizient c_1 betrachten.

Im initialen Nasal [n] unseres Beispielwortes bleiben die Lippen geschlossen, es wird folglich wenig Energie abgestrahlt, und der erste Parameter unseres Merkmalsvektors, die Energie, wird relativ kleine Werte annehmen. Im anschließenden Diphthong [ɔʏ] öffnen sich die Lippen, die abgestrahlte akustische Energie steigt an, erreicht im Silbenkern ihr Maximum und fällt zum zweiten Nasal wieder ab. Der Parameter e wird also eine Auf-Ab-Bewegung ausführen.
Der erste Cepstral-Koeffizient c_1 beschreibt die 'Welligkeit' des Mel-Spektrums für Schwingungen, deren Wellenlänge die Doppelte der Gesamtlänge des Spektrum (8kHz) beträgt. Anschaulich heißt dies, daß der Parameter c_1 hohe Werte annimmt, wenn z.B. das Mel-Spektrum einen breiten Hügel oder ein breites Tal über die ganze Länge formt, und niedrige Werte, wenn das Mel-Spektrum eher flach verläuft oder nur 'schnellere' Welligkeiten aufweist. Das Spektrum eines Nasals hat ein überwiegend niederfrequentes Spektrum, einerseits bedingt durch die Hauptresonanz der großvolumigen nasalen Passage und die starke Verengung an den Nasenöffnungen, andererseits durch die Dämpfung der Anti-Formanten oberhalb von 500 Hz ([27]). Deshalb ist zu erwarten, daß der Parameter c_1 relativ groß sein wird. Im Diphthong [ɔʏ] dagegen ändert sich die spektrale Form auch innerhalb des Silbenkerns durch

[17] $E + E' + E'' + 12C + 12C' + 12C'' = 39$

[18] Im Falle des MAUS-Systems: 16000 Abtastwerte pro Sekunde.

[19] 100 Abtastwerte pro Sekunde.

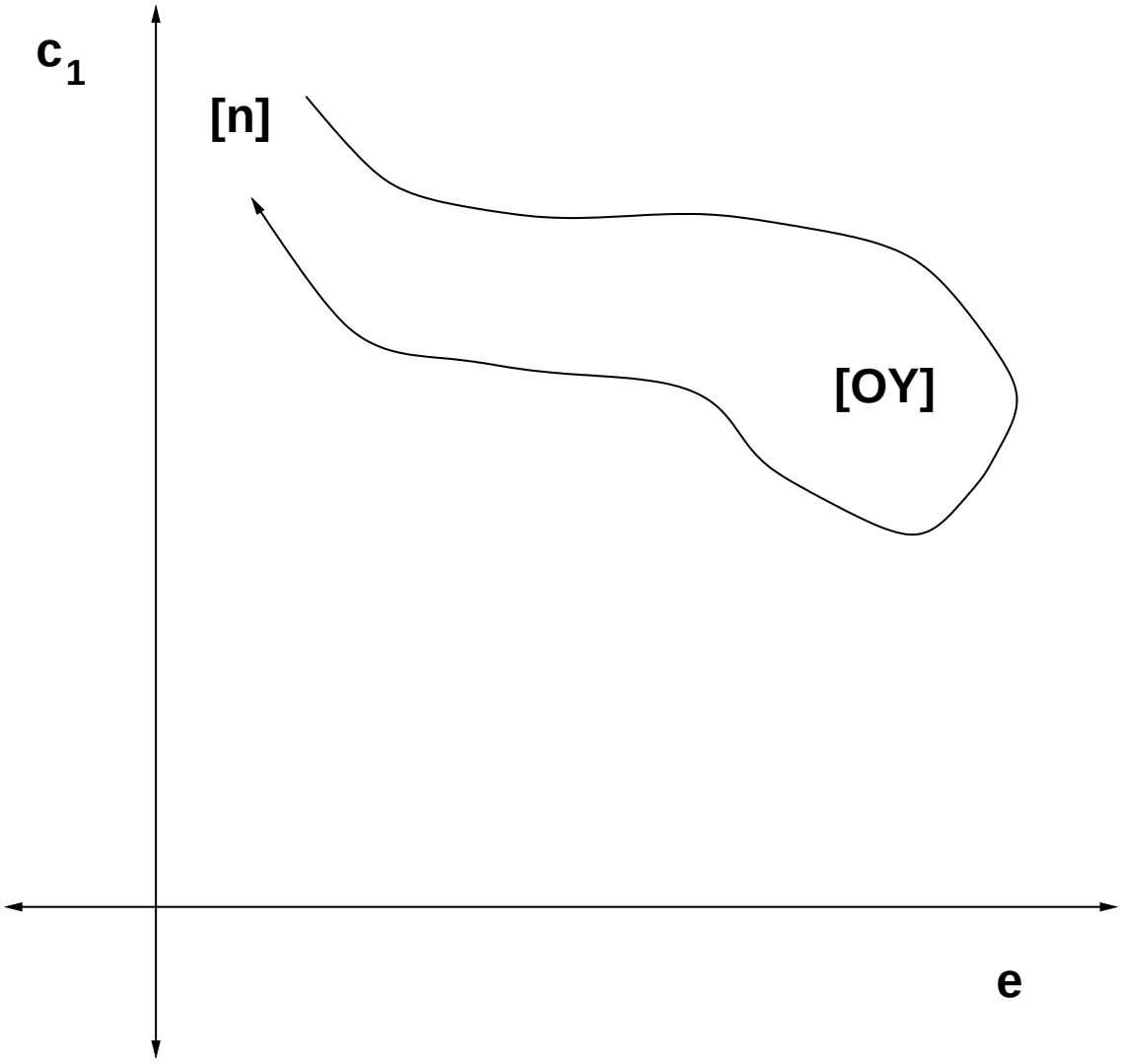


Abbildung 3.2: Trajektorie im zwei-dimensionalen Vektorraum der Parameter e und c_1 für die Phonemfolge [nɔʏn]

den Übergang vom [ɔ] zum [ʏ]; eine Vorhersage, wie c_1 sich verhalten wird, ist nur schwer möglich. Vermutlich wird er jedoch nicht nahe Null sein, da beide vokalischen Spektren ein großräumiges Gipfelverhalten aufweisen, aber geringer als der Wert in den Nasalen.
Betrachtet man nun das zeitliche Zusammenspiel der beiden Parameter e und c_1, findet man eine Trajektorie im zwei-dimensionalen Raum, die ungefähr so aussehen könnte, wie in Abbildung 3.2 wiedergegeben.

Modellierung der Aussprache-Hypothesen

Der zweite Verarbeitungszweig betrifft die Modellierung der zu erwartenden Aussprache der Äußerung. Wie bereits in Abschnitt 3.2.2 angedeutet, stellt der MAUS-Algorithmus Hypothesen über die Aussprache der zu segmentierenden Äußerung auf, welche dann als Einschränkungen im eigentlichen Optimierungsprozess eingesetzt werden. Ausgangspunkt ist hier die orthographische Repräsentation der Äußerung.

1. Die orthographische Repräsentation[20] muß zunächst in eine Kette von rein lexikalischen Einheiten gefiltert werden. Beispielsweise werden sämtliche Interpunktionen entfernt, Großschreibung zu Beginn einer Äußerung wird beseitigt (es sei denn, es handelt sich um ein Nomen bzw. einen Namen), und sämtliche zusätzlichen Markierungen werden entfernt.

   Beispiel[21]:

   `'... -/das #knock <"ah>/- das neun-und-neunzigste Kapitel!'`

   wird zu:

   `'... das "ah das neunundneunzigste Kapitel'`

2. Die verbleibenden lexikalischen Einheiten werden durch ihre kanonische Standard-Aussprache ersetzt. Dies kann entweder passiv durch ein möglichst abdeckendes Aussprache-Lexikon[22] oder aktiv durch ein automatisches 'Text-to-Phoneme'-System[23] erfolgen. Das Resultat ist eine Kette von SAM-PA Symbolen zusammen mit der Markierung der Wortgrenzen.

   Beispiel:

   `'... das "ah das neunundneunzigste Kapitel'`

   wird zu:

   `'... da:s#E:#das#nOYnUntnOYntsICst@#ka:pIt@l'`[24]

---

[20] Die orthographische Repräsentation kann z.B. auch in Form einer Transliteration gemäß Verbmobil-Konventionen vorliegen. In diesem Falle enthält sie außer den eigentlichen Wörtern noch zahlreiche Marker für Worttypen, syntaktische Ereignisse, Geräusche, etc.

[21] Aus *Verbmobil I* Transliteration.

[22] Am Institut für Phonetik wurde unter meiner Leitung in den letzten Jahren ein sehr großes, computerlesbares Aussprache-Lexikon *PHONOLEX* entwickelt.

[23] Z.B. das von M. Libossek in ihrer Magisterarbeit entwickelte silbenbasierte P-TRA System ([26]).

[24] Symbole in SAM-PA; '#' markiert eine Wortfuge.

3. Diese Kette von SAM-PA Symbolen läßt sich als ein linearer, gerichteter Graph oder Markov-Modell darstellen, wobei die SAM-PA-Symbole in den Knoten stehen, während an den gerichteten Kanten Übergangswahrscheinlichkeiten eingetragen werden können. Die Summe aller von einem Knoten abgehenden Übergangswahrscheinlichkeiten muß *per definitionem* 1 sein. Da der Graph zunächst nur von jedem Knoten genau einen Übergang zum nächsten Knoten hat, sind alle Übergangswahrscheinlichkeiten gleich 1. Der Graph repräsentiert in diesem Zustand genau eine Aussprache.
   Ich werde einen Graphen dieser Art in Zukunft *Aussprache-* oder *Variantengraph* nennen.

4. Im nächsten Schritt wird die kanonische Aussprache – die ja im Grunde nur eine mögliche, recht unwahrscheinliche Aussprache darstellt – durch andere Hypothesen erweitert. Dies geschieht formal dadurch, daß an beliebiger Stelle Knoten und Kanten sowie deren Übergangswahrscheinlichkeiten in den Graphen eingefügt werden. Nach wie vor muß dabei die Summe aller von einem Knoten abgehenden Übergangswahrscheinlichkeiten auf 1 normiert sein.

   Beispiele:

   Anstatt [nɔʏntsɪçstə] könnte auch infolge des überwiegend vokalischen Kontextes [nɔʏndzɪçstə] mit stimmhaften [dz] realisiert werden.
   Oder der erste Plosiv könnte elidiert worden sein: [nɔʏnsɪçstə]
   Das gleiche mit stimmhaften [z]: [nɔʏnzɪçstə]
   Der Frikativ [ç] könnte bei rascher Artikulation elidieren: [nɔʏntsɪstə]
   usw.

   Abbildung 3.3 zeigt ein Beispiel für einen Variantengraphen des Wortes 'neunzigste', der all diese und andere Variationen zuläßt.

   Der Term $P(R)$ in Gleichung 3.4 wird durch die Multiplikation aller Übergangswahrscheinlichkeiten längs eines möglichen Pfades durch den Variantengraphen bestimmt.[25]

Die Modellierung der Aussprache-Hypothesen erzeugt also aus der orthographischen Repräsentation einen statistisch gewichteten Variantengraphen, welcher für

[25] Zur formalen Beschreibung der Aussprache-Modellierung siehe Anhang A.1.

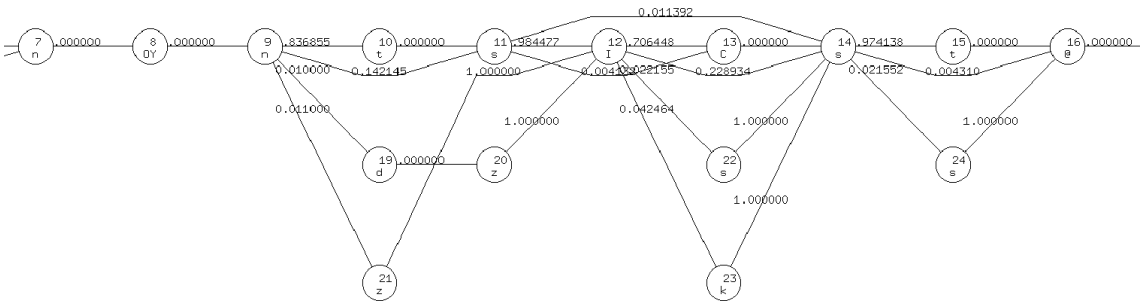


Abbildung 3.3: Variantengraph des Wortes 'neunzigste' in spontaner Sprache (Symbole in SAM-PA)

die betreffende Äußerung eine Vielzahl möglicher Aussprachen zusammen mit deren apriori Wahrscheinlichkeiten modelliert. Soweit der phonologische Zweig der Verarbeitung.

Integrierte Suche

Wie verbinden sich nun phonetische und phonologische Verarbeitung im weiteren Verlauf der MAUS-Analyse? Im Prinzip sehr einfach: Die Folge von Merkmalsvektoren aus dem phonetischen Verarbeitungstrang wird in einen *Viterbi*-Algorithmus gespeist, welcher wiederum als Einschränkung des Suchraums den Variantengraphen des phonologischen Verarbeitungsstranges verwendet.[26]

Dieser Satz bedarf einiger Erläuterungen:

- Der *Viterbi*-Algorithmus ist eine Methode, wie sie ganz allgemein zur Lösung von Dekodierungsproblemen verwendet wird[27]. Er basiert auf dem Prinzip der *Dynamischen Programmierung*, welches besagt, daß sehr komplexe Probleme mit aufgrund der Kombinatorik großer Suchräume exponentiell steigendem Aufwand sich dennoch mit linearem Aufwand lösen lassen, wenn die Berechnung im Suchraum schrittweise und rekursiv, d.h. mit Rückgriff auf bisher erzielte Teilergebnisse, erfolgen kann. Eine erschöpfende Behandlung dieser

[26] Diese Methode ist unter dem Begriff 'Integrierte Suche' bekannt.

[27] Im Englischen daher auch z.T. *Viterbi Decoding* genannt.

Methoden würde den Rahmen dieser Arbeit bei weitem sprengen. Eine sehr gute Einführung in die *Dynamischen Programmierung* ist zum Beispiel in [39] zu finden. Für das weitere Verständnis genügt es zu wissen, was der Viterbi-Algorithmus in Kontext der automatischen Spracherkennung zu leisten vermag: Er berechnet die Aufteilung der Merkmalsvektoren auf die akustischen Phonem-HMM derart, daß das Produkt aus akustischer Modellierung (der Term $P(O|R)$ in Gleichung 3.4) und der apriori Wahrscheinlichkeit einer bestimmten Phonemfolge R (der Term $P(R)$) ein Maximum erreicht. D.h. der Viterbi-Algorithmus löst genau die gewünschte Maximierung aus Formel 3.4.

- Der *Viterbi*-Algorithmus ist für sich alleine genommen zunächst genau das, was in Abschnitt 3.2.2 als 'Freie Phonemerkennung' bezeichnet wurde. D.h. daß theoretisch alle denkbaren Phonemketten dekodiert werden können. Durch die Kombination mit dem Variantengraphen wird der Suchraums des *Viterbi*-Algorithmus jedoch auf die Phonemketten eingeschränkt, die im Variantengraphen enthalten sind. Gleichzeitig werden die dabei berechneten apriori Wahrscheinlichkeiten berücksichtigt.

Zur Veranschaulichung ist es hilfreich, sich anhand des Beispiel-Wortes 'neunzigste' die einzelnen Arbeitsschritte der integrierten Suche zu vergegenwärtigen:

1. Zunächst werden die Knoten im Variantengraph durch entsprechende Phonem-HMM expandiert. Dies ist formal ohne weiteres möglich, da ein HMM ebenfalls aus Knoten und gerichteten Kanten besteht, an welchen sich Übergangswahrscheinlichkeiten befinden. Abbildung 3.4 zeigt den resultierenden expandierten Graphen für die ersten sechs Knoten des Variantengraphen im Beispiel 'neunzigste'. Jedes HMM enthält ein statistisches Modell des korrespondierenden Symbols, welches anhand von manuell segmentiertem Sprachmaterial trainiert wurde. Zusammen genommen bildet der expandierte Graph ein einziges großes HMM, welches ein Modell für die gesamte Äußerung darstellt.

2. Der Viterbi-Algorithmus weist jedem Merkmalsvektor der Äußerung genau einen Zustand in diesem Modell zu und bewertet die gesamte Zuweisung, indem er die Modellwahrscheinlichkeiten der HMM-Zustände ($P(O|R)$) mit allen Übergangswahrscheinlichkeiten des dabei durchlaufenen Pfades durch das Modell ($P(R)$) multipliziert. Dabei kommt es auch vor und ist sogar der Regelfall, daß ein Zustand nacheinander mehrere Merkmalsvektoren *konsumiert*; in diesem Fall wird die Übergangswahrscheinlichkeit für den Selbstübergang

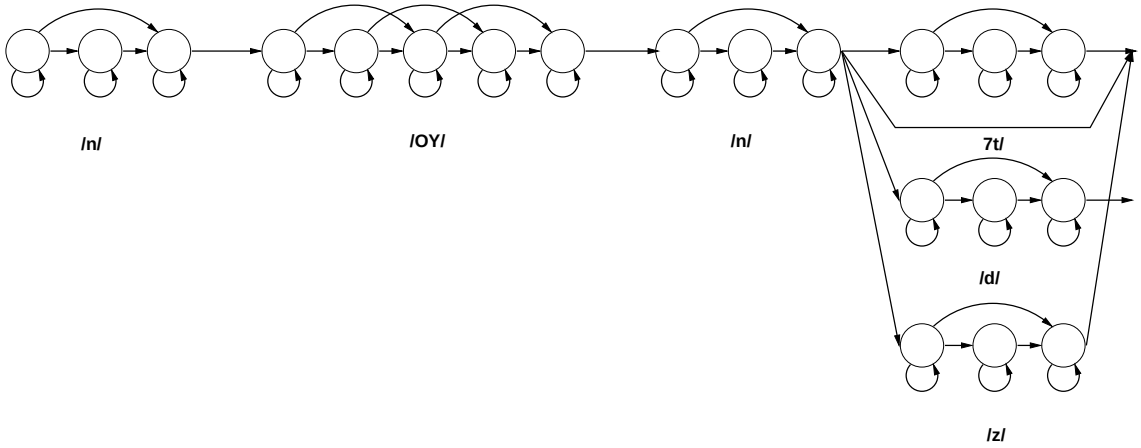


Abbildung 3.4: Auf HMM expandierter Graph für die ersten 6 Knoten des Beispiels 'neunzigsten'

aufmultipliziert. Eine solche Zuordnung von allen Merkmalsvektoren der Äußerung auf HMM-Zustände nennt man auch *Mapping* oder *Alignment*.

3. Alle möglichen Mappings werden bewertet und das mit der höchsten Gesamtwahrscheinlichkeit bestimmt. Im MAUS-System kommt dazu eine Variante des Viterbi-Algorithmus zum Einsatz, der an der University of Cambridge für das sog. *Hidden Markov Toolkit* (HTK, [56]) entwickelt wurde. Das Grundprinzip ist das eines *Token Passing Algorithm*: Zum Beginn der Berechnung wird genau ein Token in den ersten Knoten des HMM gespeist. Das Token bewertet den ersten Merkmalsvektor der Äußerung anhand des statistischen Modells des Zustandes, in dem es sich befindet, und speichert die erzielte Wahrscheinlichkeit intern ab. Dann stellt es fest, wieviele Übergänge aus diesem Zustand entspringen und klont sich selbst in ebensoviele Kopien. Jeder Klon folgt jeweils einen möglichen Übergang, multipliziert die dabei beobachtete Übergangswahrscheinlichkeit auf die bereits vorhandene Gesamtwahrscheinlichkeit auf und betritt den nächsten Zustand. Dann wiederholt sich der eben geschilderte Zyklus für alle Token und den zweiten Merkmalsvektor. Jedes Token speichert außer der aufmultiplizierten Gesamtwahrscheinlichkeit auch die bisher zurückgelegte Wegroute und repräsentiert daher genau ein mögliches Mapping bis zu dem Zustand, in dem es sich gerade befindet.
   Dieser Zyklus wird solange wiederholt, bis der letzte Merkmalsvektor bewertet wurde. Dann befinden sich selbstverständlich nur im letzten Zustand des HMM Token, die ein vollständiges Mapping von Merkmalsvektoren auf Zustände enthalten. Von diesen enthält das Token mit der höchsten aufmultiplizierten

Gesamtwahrscheinlichkeit das gesuchte Mapping.
Natürlich ist der Algorithmus in dieser Form nicht praktikabel, weil schon nach wenigen Millisekunden die Anzahl der Token explodiert. Deshalb werden in jedem Verarbeitungszyklus in jedem Zustand des HMM alle Token bis auf das mit der höchsten bisher aufmultiplizierten Wahrscheinlichkeit gelöscht. Dies ist ohne Einschränkung der Allgemeinheit möglich, weil die Token in einem Zustand und zu einem bestimmten Zeitpunkt alle möglichen Teil-Mapping bis zu diesem Zustand repräsentieren, die überhaupt möglich sind. Falls also das gesuchte Mapping tatsächlich zu diesem Zeitpunkt durch diesem Zustand verlaufen sollte (was zu diesem Zeitpunkt aber noch nicht bekannt ist!), so kann nur das Token mit der höchsten Wahrscheinlichkeit das korrekte Teil-Mapping enthalten. Alle anderen können also ohne Verlust entfernt werden[28].

4. Als letzter Schritt bleibt die Auswertung der Routen-Information des gefundenen Tokens, auch *Backtracking* genannt. Aus dem *Backtracking* wird ersichtlich, welche Zustände (und damit auch welche Phonem-HMM) zu welchen Zeitpunkten betreten und wieder verlassen wurden. Dies entspricht zusammen mit der Information, um welche Modelle es sich gehandelt hat, genau der gesuchten Etikettierung und Segmentierung.

Bis jetzt wurde nichts weiter über die Art der akustischen Modellierung innerhalb der Phonem-HMM ausgesagt. Das MAUS-System verwendet hierzu eine bewährte Standard-Technik, nämlich sogenannte *Links-Rechts-Modelle*[29] mit multi-modalen, multi-dimensionalen Gaußverteilungen, deren Parameter ungebunden sind. Konkret heißt dies, daß jeder HMM-Zustand bis zu 16 verschiedene 39-dimensionale Gaußglocken enthalten kann, welche über Gewichtungsfaktoren, die sich zu 1 addieren, linear zu einer Wahrscheinlichkeitsdichtefunktion überlagert werden. Obwohl die einzelnen Gaußglocken nur diagonalisierte Kovarianzmatrizen besitzen und somit immer achsenparallel zum Koordinatensystem zu liegen kommen, läßt sich durch die lineare Superposition mehrerer Gaußglocken eine große Mannigfaltigkeit von Dichteverteilungen modellieren.
Eine ausführliche Behandlung der Hidden Markov Modelle und ihrer vielen Varianten muß hier aus Platzgründen entfallen. Der interessierte Leser findet z.B. in [56] oder [39] weiterführende Literatur zu diesem Thema.

[28] Sog. Viterbi-Bedingung.

[29] Es sind keine 'Rücksprünge' erlaubt.

Zusammenfassend läßt sich sagen, daß in der integrierten Suche mit dem Viterbi-Algorithmus die beiden Wissensquellen der Phonetik (akustische Modellierung) und der Phonologie (Vorhersage über die Aussprache) miteinander kombiniert und am konkreten empirischen Schalldrucksignal *verifiziert*[30] werden. Es sollte hierbei betont werden, daß es sich bei beiden Wissensquellen nicht allein um traditionelles Wissen in Form von Regeln und Strukturen, sondern zusätzlich auch um statistisches Wissen in Form von Wahrscheinlichkeitsverteilungen handelt. Die beiden nächsten Abschnitte behandeln jeweils unterschiedliche Ansätze, wie dieses Wissen in Bereich der Aussprache-Modellierung gewonnen werden kann.

Eine ausführliche Beschreibung des System-Designs findet sich im Anhang unter Punkt A.2.

3.4 Experten-basiertes MAUS

Es ist naheliegend, zunächst zu versuchen, den phonologischen Zweig der Verarbeitung von MAUS mit klassischem phonologischen Wissen auszustatten. Zu diesem Zweck hat Barbara Wesenick ([53]) in ihrer Arbeit von 1994 eine umfangreiche Studie zur deutschen Aussprache erstellt. Basis für ihre Untersuchung waren zum einen die einschlägige Fachliteratur zum Thema Aussprache-Variation, zum anderen eigene Auswertungen und Beobachtungen am manuell segmentierten und etikettierten Sprachkorpus *PhonDat 2* (vgl. auch Anhang C). Neben der phänomenologischen Beschreibung und Belegung der Beobachtungen wurde außerdem versucht, diese in ein formales Regelwerk überzuführen, welches sich maschinell weiterverarbeiten läßt[31].

Die formale Struktur einer Regel zur Aussprache-Variation ist $LMR > E$, wobei L, M, R und E Folgen von SAM-PA-Symbolen beliebiger Länge (auch der Länge Null) darstellen. L und R bilden den linken und rechten *Kontext* der Regel, M ist die Lautfolge, die durch E ersetzt werden soll. L und R können auch leer, d.h. beliebig, sein; symbolisiert wird dies durch das Zeichen $\emptyset$. M und E können ebenfalls eine leere Lautfolge enthalten; in diesem Falle ist dies jedoch wörtlich zu nehmen.

Beipiel:

[30] Daher wird im Zusammenhang mit der MAUS-Technik z.T. auch von *Sprachverifikation* gesprochen.

[31] Ein ähnlicher Ansatz wurde früher schon von Ute Jekosch in [17] verfolgt. Eine eigene, mehr technisch orientierte Vorarbeit zu diesem Thema ist [44].

> Die Regel ∅ n b > m beschreibt eine regressive Assimilation des Artikulationsortes: der Nasal [n] wird zu [m], wenn der rechte Kontext den bilabialen Plosiv [b] enthält. Der linke Kontext ist hierbei beliebig (∅):
>
> 'anbinden' : [?anbɪndən] wird zu [?ambɪndən]

Die kombinierte Lautfolge LMR in jeder Regel bezeichnet immer Lautfolgenmuster in der *kanonischen Aussprache* der fraglichen Äußerung. Das bedeutet, die Regeln sind so formuliert, daß sie immer nur auf den kanonischen Aussprachestring angewendet werden dürfen. Mit anderen Worten, eine rekursive Anwendung der Regeln ist hier – im Gegensatz zu den meisten bekannten phonologischen Theorien – nicht intendiert.

> Tatsächlich verlief die Entwicklung des Wesenickschen Regelsets zunächst über die klassische Anwendung rekursiver Regeln. Es ist einsichtig, daß ein rekursiv angewandtes Regelset mächtiger ist in dem Sinne, daß es mit weniger Regeln mehr Phänomene zu erklären vermag. Aus diesem Grunde – vgl. *Ockham's razor* – bevorzugen vor allem Anhänger der *computational phonology* die Verwendung rekursiver Strukturen.
> Was dabei oft übersehen wird – und tatsächlich findet sich bemerkenswerterweise kaum ein Hinweis dazu in der einschlägigen Literatur – ist die Tatsache, daß ein rekursiv angewandtes Regelsystem mit der gleichen Anzahl von Regeln zwar *mehr beobachtete Phänomene erklären* kann, aber in den meisten Fällen *noch sehr viel mehr nicht beobachtete Phänomene* beschreibt. Wenn man daher das Regelset als ein *Erklärungsmodell* erstellt und es dann durch simple Umkehrung des Analyse-Prozesses in eine *Vorhersagemodell* umdreht, findet man plötzlich eine Vielzahl von Vorhersagen, die so nie beobachtet und in den meisten Fällen auch leider vollkommen unwahrscheinlich sind. Schlimmer noch erweist es sich als fast unmöglich, bei rekursiver Regelanwendung vorherzusehen, was das System alles vorhersagen wird, d.h. auch die Regelformulierung durch den Experten wird erheblich erschwert, sobald es sich um mehr als eine handvoll Regeln handelt.
> Der erste Variantengenerator, den ich im Zusammenhang mit Wesenicks Untersuchung entwickelte, produzierte zwar alle intendierten Aussprachevarianten, aber auf jede sinnvolle Variante wurden im Schnitt mehr als das Zehnfache an völlig unsinnigen Lautfolgen generiert. Aus diesem Grunde wurde das gesamte Regelsystem so umformuliert, daß die linke

Seite der Regeln nur noch auf die kanonische Aussprache angewandt werden durfte. Als logische Schlußfolgerung ergab sich daraus, daß teilweise mehrere phonologische Prozesse, die der Theorie nach sequentiell zu verlaufen hatten, in eine einzige Regel zusammengefaßt wurden, welche – quasi nur auf der *Oberflächenform* operierend – lediglich den Ausgangspunkt und das realisierte Endergebnis enthält.
Dies erklärt auch zum Teil die für eine phonologische Theorie abnorm hohe Anzahl von Regeln (1495) in der Arbeit von Wesenick.
Ein kleiner positiver Nebeneffekt dieses Paradigmenwechsels war, daß der Algorithmus des Generators sehr viel einfacher formuliert werden konnte: Während vorher sämtliche erzeugten Aussprache-Varianten solange wieder in das Regelwerk eingespeist werden mußten, bis keine neue Variante mehr erzeugt wurde[32], genügt nunmehr eine einzige Anwendung sämtlicher Regeln auf die kanonische Lautfolge der fraglichen Äußerung.

In [20], S. 20ff, findet sich eine Übersicht der wichtigsten Regelgruppen:

Regressive Assimilation des Artikulationsortes: Der Artikulationsort eines Nasals oder Plosivs kann an den eines nachfolgenden labialen oder dorsalen Lautes angeglichen werden, wie z.B. in

anbinden : ʔanbɪndən ⟶ ʔa**mb**ɪndən
mitkriegen : mɪtkriːgən ⟶ mɪ**kk**riːgən

Progressive Assimilation des Artikulationsortes: Umgekehrt wie im vorhergehenden Fall kann sich ein Nasal oder Plosiv auch an den vorhergehenden Labial oder Dorsal angleichen. Dies geschieht hauptsächlich dann, wenn ein dazwischenliegender /ə/-Laut elidiert wurde (siehe /ə/-Elision weiter unten).

eben : ʔeːbən ⟶ ʔeː**bm**
sagen : zaːgən ⟶ zaː**gŋ**

Regressive Assimilation der Artikulationsart: Die regressive Assimilation der Artikulationsart betrifft hauptsächlich Nasale.

Signal : zɪgnaːl ⟶ zɪ**ŋn**aːl
Abend : ʔaːbənt ⟶ ʔaː**mm**t

Im zweiten Beispiel muß zuvor noch der /ə/ entfallen und der Artikulationsort assimiliert werden.

[32] Um unabhängig von der Reihenfolge der Regelanwendung zu sein.

Progressive Assimilation der Artikulationsart: Auch bei dieser Art der Assimilation ist meist ein Nasal beteiligt

Bundes : bʊndəs ⟶ bʊ**nn**əs

Assimilation der Stimmhaftigkeit/Stimmlosigkeit: Hierbei gleicht ein Laut seine Phonationsart an angrenzende Laute an, d.h. ein eigentlich stimmhafter Laut wird stimmlos oder umgekehrt.

absonderlich : ʔapzɔndɐlɪç ⟶ ʔa**ps**ɔndɐlɪç
absonderlich : ʔapzɔndɐlɪç ⟶ ʔa**bz**ɔndɐlɪç

/ə/-Elision: Vom Ausfall des Reduktionsvokals /ə/ sind vor allem Endsilben wie -en, -em, und -el betroffen. Die /ə/-Elision ist oft Voraussetzung für das Auftreten von Assimilationen, wie aus den obigen Beispielen hervorgeht.

haben : haːbən ⟶ haː**bn**
Igel : ɪgəl ⟶ ɪ**gl**

/t/-Elision: Die Auslassung des apikalen Plosives ist vor allem zwischen Konsonanten zu beobachten.

Glanz : glants ⟶ gla**ns**
schriftlich : ʃrɪftlɪç ⟶ ʃrɪ**fl**ɪç

Elision von /ç/,/x/,/h/: In dem Datenmaterial, das [53] zugrunde lag, wurde die Elision der hinteren Frikative /ç/,/x/ und /ç/,/x/,/h/ beobachtet. Der Reduktionsprozeß ist kontinuierlich, und die totale Elision ist dabei die Endstufe der Abschwächung.

Nachmittag : naxmɪtaːk ⟶ namɪtaːk
nächste : nɛçstə ⟶ nɛstə
drumherum : drʊmhərʊm ⟶ drʊmərʊm

Reduktion von Geminaten: Geminaten (zwei direkt aufeinanderfolgende gleiche Laute) können zu einem Laut reduziert werden. Dieser Fall kann auch dann eintreten, wenn die Geminate erst durch Assimilation oder Elision entsteht, wie das erste Beispiel zeigt.

Bundes : bʊndəs ⟶ bʊnnəs ⟶ bʊ**n**əs
mitteilen : mɪttailən ⟶ mɪ**t**ailən

Unter der Verwendung von 8 zusammenfassenden Lautklassen[33], welche anstelle einzelner SAM-PA-Symbole verwendet werden dürfen, fand Wesenick 1495 Aussprache-Regeln, welche sich nach Ersetzung der Lautklassen durch ihre Mitglieder zu insgesamt 5545 einfachen Ersetzungsregeln expandieren. Das Regelsystem ist unter-

[33]Konsonanten, Vokale, Diphthonge, Affrikate, Liquide, Nasale, Frikative und Plosive.

spezifizierend in dem Sinne, daß auch Varianten erzeugt werden, die nie im Untersuchungsmaterial beobachtet wurden. Dies kann ein Vorteil sein, wenn diese – als Generalisierung eines Spezialfalls – tatsächlich in der Realität vorkommen. Andererseits besteht die Gefahr, daß das Regelwerk sich übergenerativ verhält und zu viele unsinnige oder sehr unwahrscheinliche Varianten verarbeitet werden müssen.

Mit Hilfe der von Wesenick erarbeiteten Regeln läßt sich für jede deutsche Äußerung eine Liste von hypothetischen Aussprachen R erzeugen; das System macht jedoch keinerlei Aussagen über die apriori Wahrscheinlichkeit einer einzelnen Aussprachevariante $P(R)$. Der Experte, der eine Regel formuliert, könnte allenfalls seine Meinung in qualitativen Bewertungen, wie 'sehr häufig' oder 'wenig wahrscheinlich' ausdrücken. Dazu kommt noch das Problem, daß die von Wesenick zusammengetragenen Regeln aus unterschiedlichen Quellen, also von unterschiedlichen Experten stammen, deren jeweiliges idiosynkratisches Wertesystem eine solche Quantifizierung beeinflussen würde. Aus diesem Grunde wurde zunächst jeder Variante die gleiche bedingte Wahrscheinlichkeit $P(R) = \frac{1}{|\Sigma|}$ zugewiesen, d.h. die Wahrscheinlichkeit aller Varianten ist gleich dem Kehrwert der Anzahl aller möglichen Varianten.

Die Umformung einer Liste gleichwahrscheinlicher Lautfolgen in einen statistischen Hypothese-Graphen, wie er in in Abschnitt 3.3 eingeführt wurde, ist nicht trivial. Es genügt nicht, lediglich alle m abgehenden Kanten von einem Knoten mit der gleichen Wahrscheinlichkeit $\frac{1}{m}$ zu belegen.
Betrachten wir noch einmal den einfachen Variantengraphen unseres Beispiels 'neunzigste' (Abb. 3.3). Folgt man dem kanonischen Pfad durch diesen Graphen und multipliziert die Übergangswahrscheinlichkeiten nach obiger einfacher Regel auf, so erhält man: $P(R_{kan}) = 1 * 1 * \frac{1}{4} * 1 * \frac{1}{3} * \frac{1}{4} * 1 * \frac{1}{3} * 1 * 1 = \frac{1}{138}$. Daraus würde man folgern, daß dieser Graph genau 138 mögliche Aussprachen des Wortes 'neunzigste' modelliert. Tatsächlich sind es weniger als 138. Warum? Weil gewisse Varianten sich in diesem Graphen disjunktiv überlagern. Betrachten wir z.B. den Verlauf der Variante /nOYndzICst@/ und der Variante /nOYntsCst@/, so wird klar, daß diese beiden Varianten sich niemals zu /nOYndzCst@/ kombinieren können, weil die Elision des /I/ nur für Varianten mit dem ersten Frikativ /s/ möglich sind. Daraus folgt, daß die volle Kombinatorik von 138 Varianten nicht erreicht wird, und somit eine Belegung der abgehenden Kanten mit dem Kehrwert der Anzahl zu Gesamtwahrscheinlichkeiten ungleich dem Kehrwert der möglichen Varianten führt. In Abschnitt A.1.2 des Anhangs A ist ein Formalismus und sein Beweis wiedergegeben, welcher die Wahrscheinlichkeiten für beliebige Variantengraphen korrekt berechnet.

Damit ist eine erste Methode der Generation von Aussprache-Hypothesen in Form

von regelbasierten, statistisch gleich gewichteten Ausprachevarianten in MAUS möglich. Eine vergleichende Diskussion der verschiedenen Methoden sowie die Darstellung der damit erzielbaren Ergebnisse findet sich im Kapitel 4 und 5. Zunächst aber soll mit der zweiten prinzipiellen Möglichkeit der empirischen Modellierung der theoretische Teil abgeschlossen werden.

3.5 Empirisch basierte Verbesserung von MAUS

Wie bereits mehrfach angedeutet, zielt unsere wissenschaftliche Arbeit daraufhin ab, die Leistung des MAUS-Prinzips dadurch zu erhöhen, daß die Vorhersagefähigkeit des Aussprache-Modells durch Hinzufügen empirischer Daten verbessert wird. Dabei sind mehrere Fälle zu unterscheiden, weil das Aussprache-Modell aus zwei prinzipiell unterschiedlichen Informationsquellen besteht. Zum einen sind dies die Aussprache-Hypothesen selber (d.h. die konkreten Lautfolgen R) und damit natürlich die zugrundeliegenden Regeln, zum anderen sind dies die assoziierten apriori Wahrscheinlichkeiten $P(R)$. Bei der im vorangegangenen Abschnitt beschriebenen Methode handelte es sich dabei um das in Regeln formulierte Wissen eines bzw. mehrerer Experten und – in Ermangelung von Daten – um eine einfache Gleichgewichtung aller damit generierten Hypothesen. Daraus folgt, daß es prinzipiell drei Möglichkeiten zur Verbesserung[34] des Aussprache-Modells aus Abschnitt 3.4 gibt:

1. Die Verbesserung der generierten Hypothesen R und damit der zugrundeliegenden Regeln unter Beibehaltung der gleichverteilten Regelstatistik,
2. die Verbesserung der apriori Wahrscheinlichkeiten $P(R)$ unter Beibehaltung des Regelsets,
3. die Verbesserung von beiden gleichzeitig, R und $P(R)$,

sowie die Kombinationen der obigen Punkte. Die Punkte 1 und 2 wurden in unserer Arbeit nicht systematisch untersucht, obwohl die Möglichkeit dazu prinzipiell besteht. Im wesentlichen liefen die weiteren Untersuchungen in MAUS darauf hinaus, ein optimal abgestimmtes Regelset sowie die dazu assoziierten apriori Wahrscheinlichkeiten automatisch aus einem finitem Korpus von etikettierten Sprachmaterial automatisch zu lernen (Punkt 3). Zur Zeit der Niederschrift dieser Arbeit erfolgen

[34] Verbesserung im pragmatischen Sinne, daß dadurch die beobachtbare Wirklichkeit erfolgreicher modelliert wird.

weitere Untersuchungen, in welchen versucht wird, das Experten-basierte Wissen aus Abschnitt 3.4 mit automatisch gelerntem, statistischem Wissen zu kombinieren (Vorveröffentlichungen z.B. in [3]). Da diese Arbeiten im Rahmen einer weiteren Dissertation gesondert erörtert werden sollen, wird im Rahmen dieser Arbeit nicht darüber berichtet.

3.5.1 Automatisches Lernen und Datenabhängigkeit

Für die weitere Untersuchung zielen wir darauf ab, das nötige Wissen für unser Aussprache-Modell – Regeln und Statistik – aus empirischen Daten zu gewinnen. Damit wird impliziert, daß es sich um ein automatisches Lernverfahren handelt und daß die Ergebnisse – also die Parameter des aus dem Lernprozeß resultierenden Modells – wiederum eine statistische Abhängigkeit von den verwendeten Daten aufweisen. Letzteres wird sofort klar, wenn man sich vor Augen hält, daß die Prämisse aller statistischen Modelle, nämlich daß die abgeschätzten Modellparameter auf einer hinreichend großen, gegen Unendlich strebenden Datenmenge basieren, natürlich niemals erfüllt sein kann. Als Gedankenexperiment könnte man sich vorstellen, daß die Analyse des Datensatzes nur eines einzelnen Sprechers aus dem süddeutschen Raum eventuell nur zur Ausbildung einiger weniger Regeln führt, welche eben die typische Aussprache dieser Region widerspiegeln. Die Verwendung dieses Modells mit einem Hamburger Sprecher wird nur wenig sinnvolle Ergebnisse liefern. Es besteht also im Gegensatz zum Experten-basierten Modell in Abschnitt 3.4 bei allen empirischen Methoden eine Abhängigkeit zwischen dem Modell und dem zugrundeliegenden Datenmaterial. Es mag zunächst überraschen, daß sich diese Abhängigkeit außer auf Sprecherspezifika – wie Alter, Geschlecht, Herkunft, etc – auch auf den *Inhalt der aufgezeichneten Sprache* erstreckt[35]. D.h. ein Aussprache-Modell, das Daten der gleichen Sprechergruppe, aber unterschiedliche Inhalte als das Zielkorpus[36] enthält, wird voraussagbar schlechtere Ergebnisse liefern[37].

[35] Sog. *Domänenabhängigkeit.*

[36] Hierbei handelt es sich um das Korpus, welches analysiert werden soll.

[37] Ein ganz ähnlicher Effekt ist im Bereich der Spracherkennung seit langen bekannt: Um wirklich gute Erkennungsleistung zu erzielen, benötigt man unbedingt umfangreiches Sprachmaterial aus der *Domäne* der Anwendung.

3.5.2 Lernen von Aussprache-Regeln

Ganz egal, ob es sich um einen Experten oder einen automatischen Lernalgorithmus handelt, das Lernen von Ausspracheregeln setzt voraus, daß etikettiertes Sprachmaterial in genügendem Umfang als Referenz zur Verfügung steht. Da dieses Material die Basis aller daraus gelernten Modelle bzw. Regeln darstellt, wurde in unseren Untersuchungen nur manuell transkribiertes Material aus den Sprachkorpora *Phondat2* (gelesen) und *Verbmobil 1* (spontan) verwendet. Das bedeutet, daß zu jeder Äußerung eine Lautkette – notiert in SAM-PA-Symbolen – für die tatsächlich geäußerten Sprachlaute S sowie eine Lautkette der kanonischen Einzelwörter K vorliegt. Vergleicht man S und K, die nicht von gleicher Länge sein müssen, so weichen diese natürlich voneinander ab, weil niemand eine zusammenhängende Äußerung so wie einzelne Zitierformen ausspricht. Der Einfachheit halber betrachten wir die kanonische Symbolkette K als die Referenz und die tatsächlich gesprochene Lautkette S als mit Abweichungen von dieser behaftet. Bildet man S auf K derart ab, daß ein Maximum an Korrespondenz einzelner Lautsymbol-Sequenzen auftritt[38], stellt man fest, daß genau drei systematische Abweichungen möglich sind:

1. *Ersetzungen*: In S findet sich ein anderes Symbol als in K, z.B. /#**h**a:b@n#/ vs. /#**C**a:b@n#/.
2. *Auslassungen*: Ein Symbol, das in K vorkommt, fehlt in S, z.B. /#ha:b**@**n#/ vs. /#ha:bn#/
3. *Einfügungen*: Ein Symbol in S taucht in K gar nicht auf, z.B. /#ha:b@n#/ vs. /#ha:b@**r**n#/

Formal läßt sich bei Auslassungen und Einfügungen in der jeweils anderen Lautsymbol-Kette ein Null-Symbol $\emptyset$ einfügen und so S und K auf gleiche Länge bringen. In den obigen Beispielen also:

1. *Ersetzung*: /#**h**a:b@n#/ vs. /#**C**a:b@n#/.
2. *Auslassung*: /#ha:b**@**n#/ vs. /#ha:b$\emptyset$n#/
3. *Einfügung*: /#ha:b@$\emptyset$n#/ vs. /#ha:b@**r**n#/

[38] *Longest Common Subsequence Alignment*; dies läßt sich z.B. durch symbolische Dynamische Programmierung erreichen; vgl. dazu die Bewertungsverfahren im Bereich der automatischen Spracherkennung.

Damit vereinfacht sich die Abbildung nur noch auf Bereiche, die korrespondieren, und Bereiche, die voneinander abweichen. Letztere entsprechen methodologisch direkt der Lautfolge M einer Aussprache-Regel[39], während sich in der Lautfolge K links und rechts davon Kontexte L und R beliebiger Länge (auch Null!) definieren lassen. Eine solche Aufteilung einer Sequenzabweichung nennen wir *Instanz einer Ausspracheregel* und kann als solche direkt aufgezeichnet werden. Um auf unsere obigen Beispiele zurückzukommen, bilden diese bei einem Links-Rechts-Kontext-Bereich von 1 Symbol jeweils eine Instanz der folgenden Ausspracheregeln[40]:

1. *Ersetzung*: # h a > C
2. *Auslassung*: b @ n > ∅
3. *Einfügung*: @ ∅ n > r

Entsprechend ergeben sich bei einem Links-Rechts-Kontext von Null Symbolen die folgenden Instanzen:

1. *Ersetzung*: h > C
2. *Auslassung*: @ > ∅
3. *Einfügung*: ∅ > r

An diesem Beispiel wird bereits deutlich, daß ein Verzichten von Kontext-Information zu recht merkwürdigen Regeln führt. Die 3. Regel in diesem Beispiel bedeutet nichts anderes, als: "Füge an jeder beliebigen Stelle den Laut /r/ ein!", was sicher nicht sinnvoll ist.

Die geschilderte Methode läßt sich gleichermaßen auf alle Bereiche der Abweichung in allen Äußerungen des Sprachmaterials anwenden und die dabei beobachteten Regelinstanzen festhalten. Schon an diesen einfachen Beispielen wird folgendes sofort klar:

- Die Weite des Links-Rechts-Kontext ist kein analytisch festzulegender Parameter, d.h. es ist uns keine Methode bekannt, die den optimalen Kontextbereich definiert.

[39] Zur Erinnerung: eine Ausspracheregel hat die Form $LMR > E$.

[40] Das Symbol # steht für eine Wortgrenze.

- Je größer der Links-Rechts-Kontext gewählt wird, desto vielfältiger die gefundenen Regelformen. Man sollte sich an dieser Stelle klar vor Augen halten, was die willkürliche Wahl eines Kontextbereiches bedeutet: Während der Experte in Abschnitt 3.4 den Kontext einer Regel aus seinem Theorieverständnis heraus begründen konnte – "Ich weiß, daß der Nasal /n/ *im Kontext eines nachfolgenden bilabialen Plosivs* dazu tendiert, die Qualität des Nasals /m/ anzunehmen." –, wählt der Lernalgorithmus völlig blindlings einen Kontext, ob dieser nun für die Regel relevant ist oder nicht.

- *Jede* Abweichung von realisierter Lautfolge zu kanonischer Lautfolge führt zu einer Regelinstanz, egal, wie groß der Links-Rechts-Kontext gewählt wurde. Das bedeutet, daß auch Versprecher, ja sogar Idiosynkrasien der Etikettierer sich in Regeln niederschlagen, auch wenn sie keinerlei Anspruch auf Allgemeinheit haben. Dies widerspricht natürlich vollkommen der allgemeinen Auffassung einer 'Regel'. Mit anderen Worten, 'Regelinstanzen', wie sie hier gefunden werden, bedeuten noch lange keine 'Regel'.

Aus alledem läßt sich folgern, daß es mit dem *Auffinden* der Regelinstanzen nicht getan ist; man benötigt zusätzlich ein *Auswahlverfahren*, das festlegt, welche der gefundenen Regelinstanzen wirklich sinnvoll sind.

Der naheliegenste Ansatz für ein Auswahlkriterium ist die *Häufigkeit* einer gefundenen Regel und damit ganz einfach die Gesamtzahl der Instanzen, die im Sprachmaterial zu einer bestimmten Regel gefunden wurde. Dies ist einsichtig, wenn man annimmt, daß es sich bei den 'falschen' Regeln um statistische Abweichungen ohne zugrundeliegende Tendenz handelt, z.B. um seltene Versprecher oder Fehler der Etikettierer. Die Anwendung solcher *Schwellen*, welche seltene Instanzen von vorneherein ausklammern, kommen oft in automatischen Lernverfahren zur Anwendung, ohne daß damit ein theoretischer Hintergrund verbunden wäre (z.B. in [43]). Im Kontext von MAUS können wir das Problem glücklicherweise zunächst hintanstellen, weil sich durch die statistische Gewichtung der gefundenen Regeln automatisch eine Bewertung im Sinne von 'sinnvoll ' vs. 'weniger sinnvoll' ergeben wird, wie der folgende Abschnitt zeigen wird.

Fassen wir zusammen: Aus manuell etikettiertem Material läßt sich relativ leicht algorithmisch eine Menge von Regelinstanzen ermitteln, wobei als entscheidende Parameter die fixe Weite des Links- und des Rechts-Kontext der Regeln anzusehen ist. Regelinstanzen sind in diesem Kontext zunächst nichts anderes als sortierte

Beobachtungen von Abweichungen der tatsächlich beobachteten Lautfolge zur kanonischen Lautfolge der verketteten Zitierformen.

3.5.3 Lernen von apriori Wahrscheinlichkeiten

Von Hypothesenwahrscheinlichkeiten zu Regelwahrscheinlichkeiten

Abstrahieren wir jetzt zunächst einmal davon, woher oder mit welchen Regeln eine Hypothese für eine Aussprachevariante R gebildet wurde, so geht es jetzt darum, den Term $P(R)$, nämlich die apriori Auftretenswahrscheinlichkeit unabhängig einer Realisierung der Variante R, zu bestimmen. Ein Generator produziere zu einer orthographisch oder kanonisch gegebenen Äußerung die Variante R, diese werde in einen weiteren Algorithmus eingespeist, und dieser produziere $P(R)$.

Dabei stoßen wir sofort auf folgendes Problem:

Bisher wurde gesagt, daß $P(R)$ sich direkt bestimmen lasse, indem man den Sprachprozeß beobachtet und empirisch die Häufigkeit der Variante R mißt. Prinzipiell ist das möglich, aber leider nicht praktikabel.
Praktisch würde dies nämlich bedeuten, daß wir jede einzelne von Millionen Aussprache-Hypothesen genügend oft, also sicher einige hundert Male beobachten müßten, um zu statistisch sicheren Aussagen kommen zu können. Bei der Vielfältigkeit der menschlichen Sprache ist dies, auch nur ansatzweise, vollkommen unmöglich. D.h. die oben geschilderte *unabhängige statistische Bewertung von Aussprache-Hypothesen* ist nicht realisierbar. Daher treffen wir folgende Vereinfachung und sagen: Angenommen, das Auftreten einer lokalen Variation in einer Lautfolge ist statistisch unabhängig vom Auftreten einer gleichen oder anderen Variation *an anderer Position in der selben Lautfolge*, dann können wir eine Dekomposition der apriori Wahrscheinlichkeit $P(R)$ derart durchführen, daß $P(R)$ gleich dem Produkt der apriori Auftretenswahrscheinlichkeiten von *Teilsequenzen* von R ist[41].

Praktisch bedeutet dies, daß eine hypothetische Lautfolge der Äußerung R in Teilsequenzen segmentiert wird, die zum Teil Lautfolgen analog zur kanonischen Aus-

[41] Wichtig ist dabei, daß diese Teilsequenzen R vollständig abdecken; das bedeutet, daß es sich nicht nur um die Wahrscheinlichkeiten von Abweichungen handelt, sondern in ganz genau gleicher Manier auch um die Wahrscheinlichkeiten von *Nicht-Abweichungen*. Ein Blick auf unsere einfachen Varianten des Wortes 'haben' macht das sofort klar: Wenn unser Modell die von der kanonischen Aussprache abweichenden Hypothesen /Ca:b@n/, /ha:bn/ und /ha:b@rn/ voraussagt, muß die kanonische Aussprache /ha:b@n/ eine Gesamtwahrscheinlichkeit von kleiner 1 aufweisen.

sprache enthalten, zum Teil aber auch abweichende Lautfolgen, wobei in beiden Fällen lokale apriori Wahrscheinlichkeiten für diese Teilsequenzen berechnet werden können. $P(R)$ ist dann ganz einfach das Produkt aller lokalen apriori Wahrscheinlichkeiten.

> Nehmen wir noch einmal unser Beispiel 'haben' und die drei Regeln mit Links-Rechts-Kontext der Länge 1:
>
> - *Kanonisch*: /#ha:b@n#/
> - *Ersetzung*: /#Ca:b@n#/ Regel: # h a > C
> - *Auslassung*: /#ha:bn#/ Regel: b @ n > ∅
> - *Einfügung*: /#ha:b@rn#/ Regel: @ ∅ n > r
>
> Mit diesem Regelsystem ist die folgende Aussprache-Hypothese R' generierbar:
>
> /#Ca:bn#/ (Regel: # h a > C und Regel b @ n > ∅)
>
> Stimmt die oben getroffene Annahme der Unabhängigkeit von Teilsequenzen, kann uns nichts daran hindern, folgende Segmentierung der Hypothese vorzunehmen:
>
> [#Ca:] [bn] [#]
>
> und die Gesamt-Wahrscheinlichkeit $P(R')$ wäre dann das Produkt aus den drei Teilsequenz-Wahrscheinlichkeiten[42]:
>
> $$P(R') = P(/\#Ca:/) * P(/bn/) * P(/\#/) \tag{3.5}$$
>
> Diese Segmentierung entspricht natürlich nicht nur zufällig genau dem Kontext-Bereich der beiden verursachenden Regeln.

Fassen wir zusammen: Anstelle der statistischen Modellierung von kompletten Äußerungen, welche aus praktischen Gründen nicht möglich ist, reduzieren wir unter der Annahme der statistischen Unabhängigkeit von variierenden Teiläußerungen die Modellierung auf Teilsequenzen, welche genau den Kontext-Bereichen des gelernten Regelsets entsprechen. Damit verschiebt sich die Problematik, praktisch unendlich viele Aussprache-Hypothesen beobachten zu müssen, auf das Problem, möglichst

[42] Wobei die Teilsequenz-Wahrscheinlichkeit P(/#/) für das Auftreten einer Wortgrenze hier nur aus formalen Gründen aufgeführt ist.

viele Teilsequenzen zu beobachten. Glücklicherweise hat sich dies bereits mit der Detektion der Regelinstanzen im vorherigen Abschnitt erledigt. Denn eine gefundene Regelinstanz ist nichts anderes als die Beobachtung von zwei Teilsequenzen: eine abweichende und die dazugehörige kanonische Teilsequenz der Aussprachehypothese. Die apriori Wahrscheinlichkeit für das Auftreten einer Regelinstanz werde ich im folgenden als *Regelwahrscheinlichkeit* bezeichnen.

Parameterabschätzung

Die Basis für die Abschätzung der Regelwahrscheinlichkeiten sind zunächst einfache Histogramme für gefundene Regelinstanzen. Regeln, die häufige Phänomene, wie z.B. eine Schwa-Elision in Endsilben, beschreiben, werden hohe Zähler aufweisen; Regeln, die sehr seltene Ereignisse oder gar *Fehler* im manuell etikettierten Material repräsentieren, werden sehr niedrige Zähler aufweisen, meistens sogar nur 1 sein.

Um aus empirischen Häufigkeiten zu einem aussagekräftigen statistischen Modell zu kommen, bietet sich zunächst als intuitiv einsichtigstes Verfahren die *Maximum-Likelihood*-Methode an. Diese Methode besagt nichts anderes, als daß die tatsächliche $P(u_i)$ eines diskreten Ereignisses u_i aus der Menge U durch die *relative Häufigkeit* des Auftretens des Ereignisses $N(u_i)$ in einer hinreichend großen und repräsentativen Stichprobe mit N Experimenten abgeschätzt wird:

$$\hat{P}(u_i) = \frac{N(u_i)}{N} \tag{3.6}$$

Die *Maximum-Likelihood*-Methode erfüllt zwei wichtige Kriterien für die statistische Parameterabschätzung:

1. Die Summe aller Abschätzungen $\hat{P}(u_i)$ über sämtliche diskrete Ereignisse der Menge U ist 1.

2. Es läßt sich zeigen (vgl. Anhang A.1.3), daß eine Abschätzung der apriori Wahrscheinlichkeiten durch die relative Häufigkeit die Entropie für die resultierende statistische Quelle minimiert. Da die Entropie den Grad der "Überraschtheit" eines Beobachters der Quelle ausdrückt, ist dies ein gutes Kriterium für die Voraussagefähigkeit des Modells, denn ein Beobachter sollte natürlich von dem Modell möglichst wenig "überrascht" werden.

In unserem speziellen Falle der Regelwahrscheinlichkeiten bedeutet relative Häufigkeit, daß für jede gelernte Regel $LMR > E$ die Zahl der gefundenen Regelinstanzen

ins Verhältnis zu allen Vorkommnissen der Lautfolge LMR in der kanonischen Lautfolge gesetzt werden muß.

> Beispiel:
>
> Im Datenmaterial wird 467 mal die Regelinstanz b @ n $>$ $\emptyset$ gefunden, während die Lautfolge /b@n/ im gesamten Material in der kanonischen Umschrift 8768 mal vorkommt. Daraus ergibt sich ein Schätzwert für die Regelwahrscheinlichkeit

$$\hat{P}(b@n > \emptyset) = \frac{467}{8768} = 0.0533 \tag{3.7}$$

Allerdings ergibt sich ein Problem, weil die meisten Regelinstanzen keineswegs häufig genug im Datenmaterial vorkommen, als daß man von "einer hinreichend großen und repräsentativen Stichprobe" sprechen könnte. D.h. wir haben es hier mit einem klassischen Fall von *sparse data*[43] zu tun, wie er vor allem auch in der automatischen Spracherkennungstechnik häufig gegeben ist. In solchen Fällen wird der Schätzwert den tatsächlichen Wahrscheinlichkeitswert nur sehr schlecht wiedergeben. Die Statistik bietet uns eine ganze Reihe von Verfahren zur Abschätzung der *Sicherheit*[44] einer statistischen Abschätzung. Alle diese Verfahren haben die Eigenschaft, daß die Konfidenzmaße stark invers proportional zur Anzahl der beobachteten Ereignisse bzw. der durchgeführten Experimente sind. Die Frage ist also, wie in solchen Fällen, in denen die Konfidenzmaße in Größenordnungen der geschätzten Wahrscheinlichkeiten anwachsen, also zu wenig beobachtete Ereignisse vorliegen, zu verfahren ist.

Geht man noch einen Schritt weiter, so muß man sich auch mit der Frage beschäftigen, wie es sich mit der Abschätzung der Auftretenswahrscheinlichkeit von Ereignissen verhält, die überhaupt nicht beobachtet wurden, weil sie in der Datenstichprobe zufällig nicht aufgetreten sind. Diese Frage ist keineswegs nur von theoretischer Bedeutung, weil man sicher damit rechnen muß, daß das Modell über Äußerungen Vorhersagen liefern muß, die im Datenmaterial niemals vorkamen.

Um trotzdem zu vernünftigen Schätzwerten für unsere Regelwahrscheinlichkeiten zu kommen, schlägt Kipp ([20]) für das MAUS-System die folgenden beiden kurz skizzierten Methoden[45] vor:

---

[43] etwa "spärliche Daten".

[44] engl. *confidence*.

[45] Formale Darstellung in den Abschnitten A.1.3 und A.1.3.

Unabhängigkeit von Kontexten

Zum einen läßt sich durch die Annahme der *statistischen Unabhängigkeit von Links- und Rechts-Kontext einer Regel* erreichen, daß die Abschätzungen der Regelwahrscheinlichkeiten auf größeren Ereignismengen beruhen (und damit automatisch sicherere Schätzwerte darstellen) und sich gelernte Regeln mit Links-Rechts-Kontext L und R auf weitere nicht beobachtete Fälle generalisieren, sofern die Kontexte wechselweise übereinstimmen.

> Beispiel:
>
> Im Datenmaterial wurden zwar die Regelinstanzen b @ n $>$ $\emptyset$ und d @ m $>$ $\emptyset$ gefunden, aber niemals b @ m $>$ $\emptyset$ beobachtet. Nach der klassischen *Maximum-Likelihood*-Methode würde somit $\hat{P}(b@m > \emptyset)$ zu Null geschätzt. Unter obiger Annahme ergibt sich durch die Kombination der Kontexte auch für die nicht beobachtete Regel b @ m $>$ $\emptyset$ ein Schätzwert, der nicht Null ist.

Effektiv bedeutet diese Methode ein "Verschmieren" der spärlichen empirischen Information auf einen größeren Beobachtungsbereich. Die Schätzwerte für die apriori Wahrscheinlichkeiten werden dadurch zwar statistisch gesehen "besser", büßen aber gleichzeitig wiederum ihren scharfen Fokus ein, indem sie ohne Rücksicht auf phonologische Theorie verallgemeinern. Trotzdem sind ähnliche Methoden in der technischen Sprachverarbeitung sehr erfolgreich, da sie einen pragmatischen Ausweg aus dem Dilemma der *data sparsity* bilden. Ob sich diese Methode auch im Falle der automatischen Etikettierung und Segmentierung positiv auswirkt, muß ein Experiment erst nachweisen.

Absolute Discounting auf Hintergrundmodell

Eine weitere pragmatische Verfahrensweise im Umgang mit spärlichen Daten ist unter dem englischen Begriff *Discounting* bekannt. Die zugrundeliegende Idee dabei ist die einer *Umverteilung der Wahrscheinlichkeitsmasse* auf nicht beobachtete Ereignisse. Praktisch bedeutet dies, daß Schätzwerte von Ereignissen, bei denen Häufigkeiten vorliegen, um eine bestimmte Funktion, die *Discounting Function*, dekrementiert werden, und diese Differenzbeträge dann nach einem bestimmten Schema auf Ereignisse verteilt werden, deren Häufigkeit aus den Daten zunächst einmal Null

war (sogenanntes *Hintergrundmodell*). Beim *Absolute Discounting* wird die *Discounting Function* durch einen konstanten, pragmatisch gewählten Wert[46] ersetzt. Dies führt dazu, daß hohe Schätzwerte von häufig beobachteten Ereignissen *relativ gesehen weniger dekrementiert* werden als seltenere Ereignisse. Anschaulich ist dies auch sinnvoll, weil häufig beobachtete Regelinstanzen zu sicheren Schätzwerten führen, während selten beobachtete Regelinstanzen eher unsichere Schätzwerte zur Folge haben. Es wäre folglich schlecht, die guten Schätzwerte durch den Abzug von Wahrscheinlichkeitsmasse zu verfälschen, während die Verfälschung bei den schlechteren Schätzwerten ohnehin in den Bereich des hohen Konfidenzintervalls fällt.

Es bleibt die Frage, auf welche nicht beobachtete Ereignisse die Wahrscheinlichkeitsmasse abgegeben wird. Theoretisch würde *Discounting* in unserem Falle der Aussprachemodellierung bedeuten, daß sich die Wahrscheinlichkeitsmasse auf sämtliche denkbaren Aussprachevarianten verteilt. Infolge der Möglichkeit, an beliebiger Stelle beliebig viele Laute einzufügen, führte dies aber dazu, daß eine Verteilung auf unendlich viele Ereignisse vorgenommen werden müßte, was natürlich nicht sinnvoll ist. Als einfachste praktikable Variante bietet sich daher an, den Variantengraph so zu erweitern, daß an jeder beliebigen Stelle, genau *ein* Lautsymbol eingefügt, gelöscht oder ausgetauscht werden kann. Durch die Verwendung von Sub-Graphen, die in sogenannten 'Joker'-Knoten repräsentiert werden, ist dies technisch relativ einfach durchzuführen. Eine weitere Möglichkeit wäre es, die Menge der Aussprachen, auf die Wahrscheinlichkeitsmasse verteilt wird, auf diejenigen des Expertenbasierten Modells aus Abschnitt 3.4 zu beschränken. Unsere Experimente zeigten jedoch keinen signifikanten Unterschied in der Performanz im Vergleich zum einfachen Hintergrundmodell. Daher – und wegen des erheblichen Aufwands – wurde diese Variante nicht weiter verfolgt.

In diesem Kapitel wurden abschließend mehrere Methoden und Wissensquellen für die Modellierung von Aussprache in MAUS erörtert. Fassen wir kurz die wichtigsten zusammen:

1. *Auf kanonischem Lexikon basierende Modellierung*[47]:
   - *Regelwissen*: wird hier nicht angewandt

[46] sog. *pragmatischer Parameter*.

[47] Eine einzige Aussprache-Hypothese, nämlich die Verkettung der Zitierformen der beteiligten Wörter.

- *Regelwahrscheinlichkeiten*: die einzige Variante übernimmt die gesamte Wahrscheinlichkeitsmasse (minus etwaiges Discounting)
- *Hintergrundmodell durch Absolute Discounting*: ist optional möglich

2. *Auf Expertenwissen basierende Modellierung*:

   - *Regelwissen*: stammt von Experten (bzw. aus der Literatur)
   - *Regelwahrscheinlichkeiten*: sind gleichverteilt auf alle Hypothesen
   - *Hintergrundmodell durch Absolute Discounting*: ist optional möglich

3. *Empirisch basierte Modellierung*:

   - *Regelwissen*: gelernt aus manuell etikettiertem Korpus
   - *Regelwahrscheinlichkeiten*: gelernt aus manuell etikettiertem Korpus
     (a) *Maximum Likelihood* Methode
     (b) *Maximum Likelihood* Methode mit Unabhängigkeit von Links- und Rechts-Kontext
   - *Hintergrundmodell durch Absolute Discounting*: ist optional möglich

Das folgende Kapitel beschäftigt sich mit der experimentellen Evaluierung dieser Methoden.

Kapitel 4

Evaluierung

In diesem Kapitel werden Methoden und Experimente dargestellt, welche eine datenbasierte Beurteilung der *Performanz* des MAUS-Systems mit seinen verschiedenen Varianten der Aussprache-Modellierung erlauben. Darin enthalten sind auch quantitative Aussagen darüber, wie gut sich die Ergebnisse des MAUS-Systems – d.h. die Etikettierung und Segmentierung des Schalldrucksignals in diskrete phonologische Ereignisse – bei Anwendung auf gelesene und spontane Sprache im Vergleich zu menschlichen Spezialisten darstellen. Beide Qualitätsbeurteilungen sind in dieser Arbeit unter dem Begriff der *Evaluierung* zusammengefaßt.

Nicht enthalten sind in dieser Arbeit unsere Arbeiten zur theoretischen Beurteilung der Aussprache-Modellierung anhand von Perplexität und Entropie der verschiedenen Ansätze. Dazu sei auf die ausgezeichnete Darstellung in [20], Kapitel 8, verwiesen.

4.1 Grundlagen

Bei der Evaluierung der MAUS-Ergebnisse muß prinzipiell zwischen der Qualität der Etikettierung, d.h. der Genauigkeit der kategorialen Klassifikation (*phonologisch*), und der Qualität der Segmentierung, d.h. der meßbaren Abweichung der Segmentgrenzen zu einer – wie auch immer definierten – Referenzsegmentierung (*phonetisch*), unterschieden werden. Beide sind von ihrem Wesen her vollkommen unterschiedliche Qualitäten und können daher nicht sinnvoll in ein gemeinsames Qualitätsmaß vereint werden. Aus diesem Grunde werden beide unabhängig voneinander beurteilt.

4.1.1 Qualität von Etikettierungen – Akuratheit

Bei der Beurteilung von Etikettierungen handelt es sich prinzipiell darum, die Abweichung einer Kette von Lautsymbolen, dem Transkript, von einer anderen Kette von Lautsymbolen, der Referenz, quantitativ zu messen. Dabei sind mehrere Punkte zu beachten:

- Referenz und Transkript können verschiedener Länge sein.
- Auch wenn dies nicht der Fall ist, existieren unter Umständen mehrere Möglichkeiten der Zuordnung (Abbildung) von Lautsymbolen des Transkripts auf Lautsymbole der Referenz[1].
- Jeder Versuch einer Abbildung des Transkripts auf die Referenz führt automatisch zu *Auslassungen, Einfügungen und Ersetzungen* (vgl. Abschnitt 3.5.2), die als qualitätsmindernd gewertet werden müssen.
- Das Qualitätsmaß sollte *symmetrisch* sein, d.h. den selben Wert ergeben, unabhängig davon, ob man das Transkript mit der Referenz oder die Referenz mit dem Transkript vergleicht. Dies ist nicht trivial, weil – wie sich weiter unten zeigen wird – die Frage, was eigentlich die Referenz darstellt, nicht immer eindeutig geklärt werden kann.

Ein ganz ähnliches Problem stellt sich bei der Beurteilung von Ergebnissen der automatischen Spracherkennung. Auch hier kann das Resultat eines Spracherkennungsalgorithmus zusätzliche, fehlende oder vertauschte Wörter gegenüber der tatsächlich getätigten Äußerung aufweisen. Das allgemein gebräuchliche Bewertungsmaß, die *word accuracy*, wurde von A. Kipp in [20] für die MAUS-Evaluation wie folgt angepaßt:

> Ein Maß für die Abweichung bzw. die Übereinstimmung zweier Strings voneinander kann daraus berechnet werden, wieviele Symbole in einem String verändert, getilgt oder eingefügt werden müssen, um den jeweils anderen zu erhalten. Ist
>
> - N_{ref} die Anzahl der Symbole im Referenzstring $\boldsymbol{c}$
> - N_{vgl} die Anzahl der Symbole im zu vergleichenden String $\boldsymbol{r}$

[1] Wenn z.B. mehrere Ersetzungen hintereinander erfolgen.

- N_{ins} die Anzahl der Symbole, die in $\boldsymbol{r}$ eingefügt werden müssen,
- N_{del} die Anzahl der Symbole, die aus $\boldsymbol{r}$ getilgt werden müssen,
- N_{rep} die Anzahl der Symbole, die in $\boldsymbol{r}$ ersetzt werden müssen,

so ist die sogenannte *Akkuratheit*[2] ein Maß für die Übereinstimmung von Folgen. Sie ist definiert als Abbildung $\boldsymbol{\Sigma} \times \boldsymbol{\Sigma} \rightarrow [0, 1]$:

$$A(\boldsymbol{r}, \boldsymbol{c}) = \frac{N_{\text{ref}} - N_{\text{del}} - N_{\text{ins}} - N_{\text{rep}}}{N_{\text{ref}}} \tag{4.1}$$

Die Abbildung ist nicht symmetrisch, d.h. $A(\boldsymbol{r}, \boldsymbol{c}) \neq A(\boldsymbol{c}, \boldsymbol{r})$. Da aber beispielsweise beim Vergleich zweier manueller Transkriptionen eine Referenz nur willkürlich festgelegt werden könnte, ist ein symmetrisches Übereinstimmungsmaß ebenfalls sinnvoll und nötig. Dies kann leicht durch Mittelwertbildung erreicht werden. Die symmetrische Akkuratheit ist demnach definiert als:

$$\overline{A}(\boldsymbol{r}, \boldsymbol{c}) = \frac{A(\boldsymbol{r}, \boldsymbol{c}) + A(\boldsymbol{c}, \boldsymbol{r})}{2} \tag{4.2}$$

Das Testset besteht aus vielen Paaren $(\boldsymbol{r}, \boldsymbol{c})$. Die Akkuratheit über Paare aus einer Menge $\mathbf{X} \subset \boldsymbol{\Sigma} \times \boldsymbol{\Sigma}$ ist definiert als

$$A(\mathbf{X}) = \frac{\sum\limits_{(\boldsymbol{r}, \boldsymbol{c}) \in \mathbf{X}} |\boldsymbol{c}| A(\boldsymbol{r}, \boldsymbol{c})}{\sum\limits_{(\boldsymbol{r}, \boldsymbol{c}) \in \mathbf{X}} |\boldsymbol{c}|} \tag{4.3}$$

und die symmetrische Akkuratheit als

$$\overline{A}(\mathbf{X}) = \frac{1}{2} \left(\frac{\sum\limits_{(\boldsymbol{r}, \boldsymbol{c}) \in \mathbf{X}} |\boldsymbol{c}| A(\boldsymbol{r}, \boldsymbol{c})}{\sum\limits_{(\boldsymbol{r}, \boldsymbol{c}) \in \mathbf{X}} |\boldsymbol{c}|} + \frac{\sum\limits_{(\boldsymbol{r}, \boldsymbol{c}) \in \mathbf{X}} |\boldsymbol{r}| A(\boldsymbol{c}, \boldsymbol{r})}{\sum\limits_{(\boldsymbol{r}, \boldsymbol{c}) \in \mathbf{X}} |\boldsymbol{r}|} \right) \tag{4.4}$$

Damit geht jede Äußerung gewichtet mit der Anzahl der Transkriptionssymbole im jeweiligen Referenzstring ein.

[2]Der englische Begriff *Accuracy* wird auch in der deutschen Literatur häufig verwendet.

Zur Berechnung dieser Maße muß die Anzahl der nötigen Einfügungen, Tilgungen und Ersetzungen pro Äußerung festgestellt werden. Hierzu bedient man sich der sogenannten *dynamischen Programmierung*. Dabei werden die Symbole der beiden zu vergleichenden Strings $\boldsymbol{c} = \gamma_0\gamma_1\ldots\gamma_{I-1}$ und $\boldsymbol{r} = \rho_0\rho 1\ldots\rho_{J-1}$ einander optimal zugeordnet und zwar durch Anwendung von Operationen aus der Menge Ω = {"Tilgung","Einfügung","Zuordnung"}. Jede dieser Operationen verursache Kosten:

- Die Zuordnung zweier Symbole wird mit den Kosten $m_{\text{mat}}(\alpha,\beta)$ beaufschlagt
- Die Tilgung eines Symbols verursacht die Kosten m_{del}
- Die Einfügung eines Symbols verursacht die Kosten m_{ins}

Das Optimalitätskriterium sind minimale Kosten. Eine Zuordnung mit insgesamt minimalen Kosten kann durch einen rekursiven Algorithmus, nämlich die dynamische Programmierung, gefunden werden. Man führt dabei eine Hilfsgröße $d(i,j)$ ein, die die minimalen Kosten für eine Zuordnung der Teilstrings $\gamma_0\ldots\gamma_i$ und $\rho_0\ldots\rho_j$ ausdrückt. Diese Hilfsgröße kann rekursiv berechnet werden, indem eine bislang optimale Zuordnung um eine Operation aus Ω erweitert wird und die Kosten zu den bisherigen addiert werden:

$$d(i,j) = \max_{\omega\in\Omega}\begin{cases} d(i-1,j) + m_{\text{del}} & \text{falls } \omega = \text{"Einfügung"} \\ d(i,j-1) + m_{\text{ins}} & \text{falls } \omega = \text{"Tilgung"} \\ d(i-1,j-1) + m_{\text{mat}}(\gamma_i,\rho_j) & \text{falls } \omega = \text{"Zuordnung"} \end{cases} \tag{4.5}$$

Als konkrete Kosten werden in den Experimenten

$$m_{\text{mat}}(\alpha,\beta) = \begin{cases} 1 & \text{für } \alpha \neq \beta \\ 0 & \text{sonst} \end{cases} \tag{4.6}$$

$$m_{\text{del}} = m_{\text{ins}} = 1 \tag{4.7}$$

verwendet, d.h. alle Zuordnungsfehler werden gleich bewertet. Die Zuordnung von gleichen Symbolen verursacht keine Kosten. Es sei darauf

hingewiesen, daß die Kosten selbst nicht in die Akkuratheit eingehen, sondern nur der Steuerung der Zuordnung dienen.

Es wird gefordert, daß jeweils beim ersten und letzten Symbol in $\boldsymbol{c}$ und $\boldsymbol{r}$ gilt: $m(\gamma_0, \rho_0) = 0$ bzw. $m(\gamma_{I-1}, \rho_{J-1}) = 0$. Dies kann ohne Einschränkung der Allgemeinheit dadurch erreicht werden, daß den Strings jeweils ein identisches Startsymbol vorangestellt wird und ein identisches Endsymbol angehängt wird.

In diesem Falle sind die Kosten der optimalen Zuordnung dann $d(I-1, J-1)$. Die eigentlich interessierenden Größen $N_{\text{del}}, N_{\text{ins}}$ und N_{rep} können durch Zurückverfolgen der jeweiligen Maximierungsentscheidungen gewonnen werden. Diese Entscheidungen können in einer Hilfsgröße $b(i, j) \in \Omega$ gespeichert werden.

Das Resultat dieser Berechnung über eine Menge von Test-Äußerungen, die mittlere symmetrische Akuratheit, ist ein Wert, der bei vollkommener Übereinstimmung von Transkript zu Referenz den Wert 1 und bei Vertauschung sämtlicher Lautsymbole den Wert 0 annimmt, und wird daher im weiteren mit 100 multipliziert und in Prozent ausgedrückt. Es sollte beachtet werden, daß die symmetrische Akuratheit ein pragmatisches Qualitätsmerkmal darstellt und durchaus auch negative Werte annehmen kann, wenn z.B. sehr viele Einfügungen vorliegen.

4.1.2 Segmentierungen – Histogramm der Abweichungen

Für die quantitative Beurteilung der Segmentierung ergibt sich sofort das alte Problem, daß man Äpfel nicht mit Birnen vergleichen sollte: Die Messung der zeitlichen Abweichung einer Segmentgrenze gegenüber einer Referenz-Segmentierung macht nur Sinn, wenn es sich in beiden Fällen um ein Segment der gleichen Kategorie handelt. Aus diesem Grunde stellt die Evaluierung der Segmentgrenzen quasi eine logische Untermenge der Beurteilung der Etikettierung dar. Man könnte es als Anweisung so formulieren:
Wenn ein Segment gemäß der Referenz etikettiert worden ist, dann bestimme die zeitliche Abweichung der Vorder- und Hintergrenze dieses Segments im Vergleich zur Lage dieser Grenzen in der Referenz-Segmentierung und trage die Werte in ein Histogramm ein! Segmente, die von der Referenz abweichende Etikette tragen, werden ignoriert!
Vorteil dieser Methode ist, daß sich automatisch ein symmetrisches Qualitätsmaß

(Histogramm) ergibt. Nachteil ist, daß das Histogramm zur Beurteilung des Gesamtergebnisses nur indirekt beiträgt; es ist lediglich eine weiter spezifierende Beurteilung.

4.2 Sprachmaterial

4.2.1 Test- und Entwicklungsmaterial

Um die oben geschilderten experimentellen Werte zu ermitteln, benötigt man Sprachmaterial, das manuell von Experten etikettiert und segmentiert wurde. Desweiteren darf dieses Material weder zur Ermittlung der Aussprachemodelle noch zur Festlegung pragmatischer Parameter[3] verwendet werden. Um diese Trennung im weiteren Text deutlich zu machen, werde ich vom *Trainingsmaterial* sprechen, wenn das Sprachkorpus zur Ermittlung der phonetisch-akustischen Modelle und der phonologisch-empirischen Aussprache-Modelle gemeint ist, vom *Entwicklungsmaterial*, wenn von Sprachmaterial die Rede ist, das zur Festlegung pragmatischer Parameter benötigt wird, und vom *Testmaterial*, wenn ich das Material anspreche, welches ausschließlich zur Evaluierung des MAUS-Systems verwendet wurde.

Für die nachfolgend beschriebenen Experimente stammen alle drei disjunkten Teil-Korpora entweder aus dem Korpus *PhonDat 2*, einem anwendungsorientierten Sprachkorpus in der Domäne *Zugauskunft*, welches im Zusammenhang mit dem BMB+F Projekt *ASL* in den Jahren 1989 - 1992 erhoben wurde, oder aus dem deutschen *Verbmobil I* Korpus, einer großangelegten Sammlung von Dialogaufnahmen im Rahmen des vom BMB+F in den Jahren 1993 - 1996 geförderten Projektes *Verbmobil* zur Entwicklung einer vollautomatischen Übersetzungshilfe.

Für das weitere Verständnis ist es wichtig zu bemerken, daß Teile des Testmaterials mehrfach etikettiert und segmentiert wurden.

- *Verbmobil*
  Im spontansprachlichen Teil waren mehrere für die Aufgabe ausgebildete Spezialisten an der Bearbeitung der Daten beteiligt. Vier von diesen haben unter kontrollierten Bedingungen jeweils komplette Dialoge bearbeitet; diese werden im weiteren, damit die Anonymität gewahrt bleibt, mit den Abkürzungen *man1-4* bezeichnet. Es wurden genau zwei komplette Dialoge – *m116d* und

[3] Z.B. der Gewichtung des *Absolute Discounting*.

m231d – von jeweils drei Spezialisten unabhängig voneinander bearbeitet. Die übrigen 9 Dialoge des Test- und Entwicklungsmaterials wurden jeweils von einer gemischten Gruppe bearbeitet, wobei jeweils nur eine Etikettierung und Segmentierung pro Dialog vorliegt. Dieser Teil des Korpus wird im weiteren mit *manx* bezeichnet.
Der mehrfach bearbeitete Teil enthält 73 Dialogbeiträge mit 9587 phonologischen Lautsymbolen in der kanonischen Referenztranskription, der gesamte Test- und Entwicklungskorpus enthält 372 Dialogbeiträge mit 44465 Symbolen.

- *PhonDat 2*
  Das Testmaterial des gelesenen Korpus wurde von drei Spezialisten etikettiert und segmentiert; diese werden im folgenden mit den Abkürzungen *man5-7* bezeichnet. Das mehrfach bearbeitete Material umfaßt in diesem Fall 200 Äußerungen mit insgesamt 10495 phonologischen Lautsymbolen.

Beide Korpora werden im Anhang C kurz beschrieben.

4.2.2 Trainingsmaterial

Für die Bestimmung der empirischen Aussprachemodelle wird neben dem Test- und Entwicklungsmaterial noch entsprechend bearbeitetes Trainingsmaterial benötigt, aus welchem die Parameter der statistischen Modelle – Regeln und Regelwahrscheinlichkeiten – nach Abschnitt 3.5 gelernt werden.

In unseren Untersuchungen wurden hierfür Segmentierungen aus dem sog. 'Kiel Corpus of Spontaneous Speech' ([25]), einen Teil des Verbmobil I Korpus, entnommen. Der Umfang dieser Daten beträgt 72 Dialoge mit 1245 Dialogbeiträgen. Das Material wurde in SAM-PA und nach gleichen Vorgaben wie im Falle des Test- und Entwicklungskorpus manuell bearbeitet.

4.3 Das Problem der absoluten Wahrheit

Jede experimentelle Evaluierung eines Verfahrens benötigt eine objektive Referenz, an welcher die Leistung des Verfahrens, die *Performanz*, objektiv gemessen werden kann. Bei der empirischen Beurteilung von Etikettierung und Segmentierung

von Sprache in phonologische Kategorien stellt sich sofort das Problem, daß diese absolute Referenz für eine gegebene Sprachäußerung nicht existiert. Selbst bei Beschränkung auf die Etikettierung, d.h. die tatsächlich produzierte Kette von phonologischen Phonem-Symbolen, die einer sprachlichen Äußerung zugrunde liegen soll, wird man fast immer unterschiedliche Meinungen bei unterschiedlichen beurteilenden Personen finden. Dies gilt in ganz besonderem Maße für spontane Sprache, weil sich dort die von der Theorie geforderten Phoneme oft infolge von Reduktionsprozessen im Signal gar nicht mehr physikalisch nachweisen lassen, obwohl der Beobachter die Lautkategorien im Kontext immer noch zu "hören" meint. Das im vorangegangenen Abschnitt skizzierte Evaluierungsverfahren nützt uns also wenig, solange keine absolute Klarheit über die Referenz geschaffen werden kann.

Aus diesem Dilemma gibt es mehrere mögliche Auswege. Am einfachsten wäre die willkürliche Festlegung eines Spezialisten[4] zur Referenz. Der Nachteil hierbei wäre, daß dieser dann die gesamte Aufgabe allein, in möglichst konsistenter Verfassung und unter höchster Selbstdisziplin durchführen müßte. Untersuchungen (z.B. [10]) haben gezeigt, daß sich allerdings auch das sog. *Intra-Labeler-Agreement*, also die Übereinstimmung mit den eigenen Urteilen, im Laufe der Zeit unter 100 % absinkt. Ein anderer Ausweg wäre die Einberufung einer möglichst große Gruppe von Spezialisten, welche sich in Klausur begeben und – eventuell auch mittels demokratischer Mehrheitsentscheidung – die Problemfälle bereinigen. Dies ist aus Kostengründen meistens nicht durchführbar, ganz zu schweigen von den Problemen, die man sich einhandelte, wenn die Auswahl der Spezialisten nicht vollkommen ausgewogen erfolgen würde.

Abgesehen von solchen technischen Problemen würde ein solches Vorgehen auch bedeuten, das Problem unter den wissenschaftlichen Teppich zu kehren. Letztendlich ist es für den Wissenschaftler ja auch interessant, warum dieses Problem überhaupt besteht.

In [54] haben meine beiden Mitarbeiter M.-B. Wesenick und A. Kipp untersucht, inwieweit die Etikettierungen und Segmentierungen verschiedener Bearbeiter in Bezug auf das Testmaterial des *PhonDat 2* Korpus voneinander abweichen. Untersucht wurden in dieser Arbeit Plosive, Frikative und Nasale. Das Problem der nicht entscheidbaren Referenz wurde in diesem Ansatz dadurch gelöst, daß in einem sogenannten *leave-one-out* Verfahren jeweils die Ergebnisse eines jeden Bearbeiters mit allen Ergebnissen der übrigen Bearbeiter verglichen wurde. Da es sich bei unserem

[4]Möglichst des anerkannt besten Spezialisten.

Qualitätsmaß um ein symmetrisches Maß handelt (vgl. Formel 4.4), ergeben sich z.B. bei 3 Bearbeitungen genau 3 experimentelle Werte für die Inter-Labeler-Akuratheit, aus welchen der arithmetische Mittelwert gezogen wird.

Die Untersuchung der o.g. Konsonantengruppen im Testmaterials aus *PhonDat 2* ergab eine mittlere Inter-Labeler-Akuratheit von 94,8% ([54]). D.h. im Mittel stimmen bei allen drei Bearbeitern die Phonem-Kategorien zu 94,8% überein.

In [21] wurde das gleiche Verfahren auf alle Phoneme des Testmaterials erweitert; die mittlere Inter-Labeler-Akuratheit reduzierte sich auf 93.7%. Dies ist konsistent mit Aussagen in [10], nämlich daß die Beurteilung der Vokalkategorie im allgemeinen schwerer fällt als z.B. die Unterscheidung eines Plosivs von einem Frikativ.

Das gleiche Prinzip kann auch auf die Beurteilung der Segmentierung übertragen werden. Wertet man die übereinstimmend etikettierten Konsonanten des *PhonDat 2* Korpus Testmaterials aus, so ergibt sich für die Abweichung der Segmentgrenzen verschiedener Bearbeiter das folgende Bild ([54]):

63% der Grenzen stimmen exakt[5] überein
73% weichen um weniger als 5 msec ab
87% weichen um weniger als 10 msec ab
93% weichen um weniger als 15 msec ab
96% weichen um weniger als 20 msec ab
99% weichen um weniger als 32 msec ab

Eine detailiertere Auswertung nach den verschiedenen Phonem- und Grenzklassen[6] zeigt darüber hinaus, welche Fälle bei der manuellen Segmentierung am wenigsten konsistent gehandhabt werden.

Für die im folgenden dargestellte experimentelle Evaluierung des MAUS-Systems werden wir ganz analog als Referenz die Daten von möglichst vielen verschiedenen menschlichen Bearbeitern heranziehen und über diese das arithmetische Mittel ziehen. Darüberhinaus wird zu jedem Experiment die entsprechende Leistung zwischen mehreren menschlichen Bearbeitern untereinander bestimmt (Inter-Labeler-Akuratheit) und als *pragmatisches Maximum* neben dem Ergebnis von MAUS dargestellt.

[5]Die Bearbeiter waren angewiesen, jeweils im positiven Nulldurchgang des Schalldrucksignals zu schneiden; dies erklärt den relativ hohen Anteil "exakter" Schnitte.

[6]Je nachdem, welche Segmente links und rechts der Grenze etikettiert wurden.

4.4 Experimentelle Ergebnisse

Die folgenden quantitativen Ergebnisse wurden im Laufe der Jahre 1996 - 2000 in verschiedenen Projekten erarbeitet ([22, 21, 20, 54, 55]) und werden hier zum ersten Mal in einer gebündelten Form beschrieben.

4.4.1 Gelesene Sprache

Die Ergebnisse für gelesene Sprache basieren auf dem bereits beschriebenen Korpus *PhonDat 2* aus der Domäne *Zugauskunft*. Da dieses Korpus nur 16 Sprecher enthält, kann keine empirisch gestützte Aussprache-Modellierung durchgeführt werden. Der Umfang des Materials ist gerade ausreichend für das *Entwicklungs-* und *Testmaterial*. Die empirische Evaluation beschränkt sich daher im Falle der gelesenen Sprache auf die Modellierung der Aussprache durch Expertenwissen, welche kein größeres empirisches Material benötigt.

Modellierung durch Expertenwissen

Dieser Evaluation liegt das in Abschnitt 3.4 beschriebene Regelwerk zugrunde. Wir behandeln zunächst die Etikettierung und dann die Segmentierung.

Etikettierung: Die folgende Tabelle zeigt die experimentellen Bedingungen und die erzielte Akuratheit der Etikettierung.

Basis:	200 Äußerungen aus *Phondat 2*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere symmetrische Inter-Labeler-Akuratheit man5-7:	93,7 %
Anzahl der Regeln:	5545
Mittlere symmetrische MAUS-Akuratheit man5-7:	**87,93 %**

Bezogen auf das *pragmatische Maximum* von 93,7 % bedeutet dies für das MAUS-System eine *relative Performanz* von **93 %**.

In diesen Untersuchungen wurde kein *Hintergrundmodell* verwendet.

Segmentierung: Die Evaluierung erfolgte in Form eines Histogramms über die absoluten Abweichungen der Segmentgrenzen übereinstimmend etikettierter Segmente. Bild 4.1 zeigt die Verteilung der Abweichungen für das oben bezeichnete Testmaterial. Die Verteilung zeigt einen zentralen Gipfel um die Abweichung 0 msec

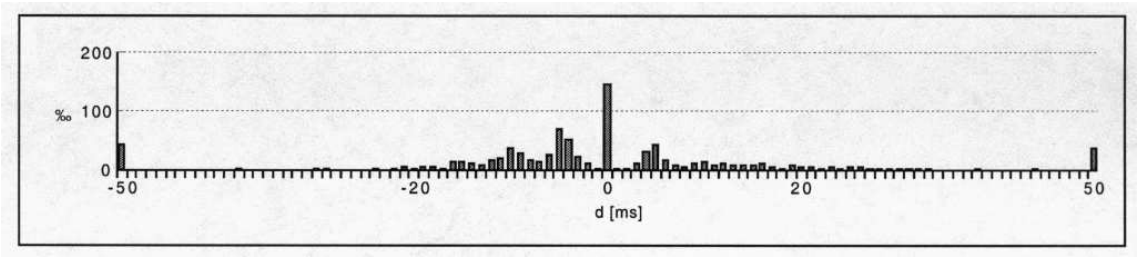


Abbildung 4.1: Histogramm über die Abweichungen der vom MAUS-System ermittelten Segmentgrenzen für gelesene Sprache und bei Modellierung durch Expertenwissen ([21])

sowie weitere Nebengipfel bei jeweils +/- 5 msec und +/- 10 msec. Die Histogrammwerte an den beiden Rändern des dargestellten Zeitbereichs fassen sämtliche weiter außen liegenden Werte zusammen. Es ist deutlich zu sehen, daß der Großteil der Abweichungen zwischen dem MAUS-System und den menschlichen Bearbeitern in einem Bereich von +/- 20 msec zu liegen kommt.

Die Nebengipfel entstehen durch Verschiebungen um jeweils einen Glottisschlag im Sprachsignal. Diese Ergebnisse wurden mit einer experimentellen Variante des MAUS-Systems erzielt, welches die Anleitung für die menschlichen Bearbeiter, nämlich immer im nächstliegenden positiven Nulldurchgang des Sprachsignals zu schneiden, mit einem einfachen Suchalgorithmus simulierte. Dadurch entsteht eine Art 'Rasterung' des Signals in Punkte, die auf Vielfachen der Inversen der Grundfrequenz des Sprechers – in diesem Fall einer Sprecherin – zu liegen kommen. Da das MAUS-System und die Referenz-Bearbeiter nur an diesen Rasterpunkten schneiden durften, ergibt sich automatisch eine mehrgipfelige Verteilung der Abweichungen. In den folgenden Versuchen verschwindet dieser Effekt, weil dort die Evaluierung über mehrere Sprecher, mit jeweils unterschiedlichen Grundfrequenzen, durchgeführt wurde.

4.4.2 Spontane Sprache

Die Ergebnisse für spontane Sprache basieren auf dem bereits beschriebenen Korpus *Verbmobil I* aus der Domäne *Terminabsprache*.

Inter-Labeler-Akuratheit

Zunächst wurde für das mehrfach bearbeitete Testmaterial und die entsprechenden menschlichen Bearbeiter die Inter-Labeler-Akuratheit der Etikettierung sowie die zugehörige Verteilung der Abweichungen in den Segmentgrenzen bestimmt.

Etikettierung:

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere sym. Inter-Labeler-Akuratheit man1-4:	**85.6 %**

Das heißt, die menschlichen Bearbeiter des Testmaterials sind sich im Mittel in 85,6 % der Fälle in ihrem Urteil einig. Dieser Wert liegt deutlich unter dem Wert von 93,7 % im vorangegangenen Experiment. Damit wird bestätigt, was zu erwarten war, nämlich daß die Beurteilung von Lautkategorien in spontaner Sprache auch dem menschlichen Spezialisten signifikant schwerer fällt als in gelesener Sprache.

Segmentierung: Abbildung 4.2 zeigt ein Histogramm über die Abweichungen der Grenzen übereinstimmend etikettierter Segmente aus den Daten der menschlichen Bearbeiter *man1* und *man2*. ([20]). Es fällt auf, daß auch hier wieder die bereits erwähnten Nebengipfel zu beobachten sind.

Modellierung durch kanonisches Lexikon

Dieses Experiment wurde durchgeführt, um die Wirkung der Aussprache-Modellierung besser beurteilen zu können. Bei der Modellierung durch ein kanonisches Lexikon wird auf eine Aussprache-Modellierung vollständig verzichtet und lediglich ein *Forced Alignment*[7] auf die verkettete Lautsymbolfolge der Zitierformen der beteiligten Wörter durchgeführt.

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere sym. Inter-Labeler-Akuratheit man1-4:	85.6 %
Mittlere sym. kanonische Akuratheit man1-4:	**73.7 %**

[7] Vgl. Abschnitt 3.2.2.

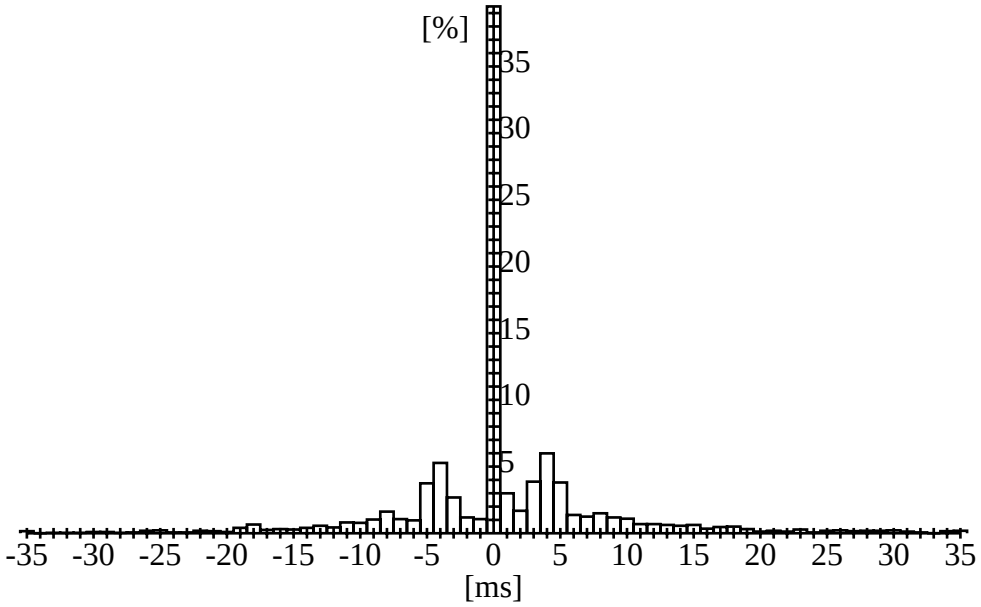


Abbildung 4.2: Histogramm über die Abweichungen der von menschlichen Bearbeitern ermittelten Segmentgrenzen für spontane Sprache

Die Inter-Labeler-Akuratheit liegt um absolut 11.9 % über der Akuratheit der kanonischen Aussprache. Mit anderen Worten, in dem verwendeten Testmaterial weicht die tatsächliche Aussprache nach dem Urteil der Spezialisten in mehr als einem Zehntel der phonologischen Kategorien von der Zitierform der Wörter ab.

Modellierung durch Expertenwissen

Dieser Evaluation liegt das in Abschnitt 3.4 beschriebene Regelwerk zugrunde. Nachfolgend sind die Versuchsbedingungen für dieses Experiment aufgelistet.

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere sym. Inter-Labeler-Akuratheit man1-4:	85.6 %
Anzahl der Regeln:	5545
Mittlere sym. Akuratheit man1-4:	**74.9 %**

Im Vergleich mit der symmetrischen *kanonischen* Akuratheit aus dem vorangegangenen Abschnitt von 73.7 % stellt dieser Wert nur eine schwach signifikante Verbesserung (0.05) dar. Vermutlich ist dies ein Hinweis darauf, daß das Expertenmodell

Spontansprache weniger gut modelliert als erwartet. Untersuchungen der Entropie und Perplexität des Expertenmodells ([20], Kapitel 8) bestätigen dies. Eine mögliche Erklärung, warum das Modell in Bezug auf Spontansprache schwächere Hypothesen generiert, ist die Tatsache, daß dieses Modell eben einerseits auf der Untersuchung eines gelesenen Sprachkorpus beruht und andererseits auch die Ergänzungen aus der einschlägigen Literatur zum Thema Aussprachevariation, die zusätzlich in das Modell eingeflossen sind, meistens auf Beobachtungen sorgfältig kontrollierter Sprache basieren. Insofern ist es nicht erstaunlich, daß das Modell bei Anwendung auf Spontansprache eine schlechte symmetrische Akuratheit zeigt.

Setzt man den Wert mit der Inter-Labeler-Akuratheit ins Verhältnis, so erhält man eine *relative Performanz* von **87.5 %**.

Modellierung durch Maximum-Likelihood-Methode

Dieser Evaluation liegt das in Abschnitt 3.5.3 beschriebene, automatisch gelernte Regelwerk zugrunde, wobei die Regelwahrscheinlichkeiten durch die klassische *Maximum-Likelihood*-Methode abgeschätzt werden.

Eine Analyse des Trainingsmaterials ergab insgesamt 13623 beobachtete Regelinstanzen, die sich durch 1807 Regeln beschreiben lassen.

Etikettierung: Die folgende Tabelle zeigt die experimentellen Bedingungen und die erzielte Akuratheit der Etikettierung.[8]

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere sym. Inter-Labeler-Akuratheit man1-4:	85.6 %
Anzahl der Regeln:	1807
Mittlere sym. Akuratheit man1-4 (manx):	**83.1 %** (81.3 %)

Das heißt, bei alleiniger Betrachtung der mittleren Akuratheit erreicht das MAUS-System **97 %** der Performanz der menschlichen Bearbeiter.

[8] In Klammern sind jeweils auch die Werte für das gesamte Test- und Entwicklungsmaterial angegeben. Letztere sind aber nicht vergleichbar mit der Inter-Labeler-Akuratheit und können allenfalls als Absolutwerte betrachtet werden.

Segmentierung: Die Evaluierung erfolgte auch hier in Form eines Histogramms über die absoluten Abweichungen der Segmentgrenzen übereinstimmend etikettierter Segmente. Bild 4.3 zeigt exemplarisch die Verteilung der Abweichungen der MAUS-Segmentierung von der eines menschlichen Bearbeiters für den Dialog *m116d* des oben bezeichneten Testmaterials ([20]). Im Gegensatz zur Evaluierung der Seg-

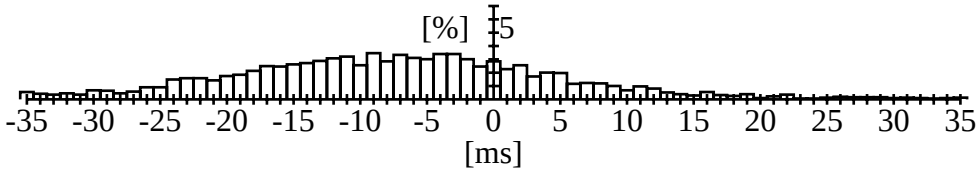


Abbildung 4.3: Histogramm über die Abweichungen der vom MAUS-System ermittelten Segmentgrenzen für spontane Sprache und bei Modellierung durch *Maximum-Likelihood*-Methode.

mentierung in gelesener Sprache ist hier die Verteilung deutlich breiter. Dies ist – wie bereits erwähnt – darauf zurückzuführen, daß diese Variante des MAUS-Systems nicht ausschließlich im positiven Nulldurchgang des Schalldrucksignals schneidet, und im Testmaterial mehr als ein Sprecher vorhanden ist.

Außerdem ist zu beobachten, daß der Schwerpunkt der Verteilung nicht bei Null liegt; d.h. das MAUS-System hat die Tendenz, Schnitte im Mittel früher zu setzen als der menschliche Bearbeiter. Für diesen Effekt gibt es bis dato keine befriedigende Erklärung. Ein Rechenfehler bei der Berechnung der Schnittgrenzen konnte ausgeschlossen werden, nachdem Kollegen an der Technischen Universität München mit einem völlig unabhängigen Viterbi-Dekoder und anderen Daten den Effekt reproduzieren konnten. Die Hypothese, die Kipp in seiner Arbeit äußert ([20], S. 139), nämlich daß dies eventuell auf die Tatsache zurückzuführen sei, daß in der von MAUS produzierten Etikettierung im Mittel mehr phonologische Lautsymbole enthalten sind als in den Referenzdaten der menschlichen Bearbeiter, halte ich für unbefriedigend, weil eine solche Unsymmetrie sich gleichmäßig auf die gesamte Äußerung verteilen müßte. Mehr Segmente bedeuten schließlich Abweichungen der vorderen Segmentgrenzen in die Vergangenheit und der hinteren Segmentgrenzen in die Zukunft, welches sich im Mittel wieder ausgleichen würde, es sei denn, die Einfügung geschieht vorzugsweise im vorderen Teil der Äußerung, wofür es jedoch keine Anhaltspunkte gibt[9]. Verschiedene Hypothesen, daß es mit der Berechnung

[9]Der Effekt von im Mittel mehr Segmenten in den MAUS-Daten gibt vielmehr ein weiteres Erklärungsmodell ab, warum die Histogrammverteilung so breit ist.

der Merkmalsparameter aus dem Schalldrucksignal zu tun haben könnte, wurden untersucht, ergaben aber kein schlüssiges Erklärungsmodell.

Zusammenfassend läßt sich sagen, daß die Segmentierung bei Anwendung auf Spontansprache deutlich schlechter ausfällt als bei gelesener Sprache. Dies geht konform mit der Beobachtung, daß auch die absolute Performanz in der Etikettierung von menschlichen Bearbeitern und MAUS infolge der starken Reduktionserscheinungen und Verschleifungen in spontaner Sprache deutlich unter der Performanz von gelesener Sprache liegt.

Verbesserung der Parameterabschätzung

In Abschnitt 3.5.3 wurde die Möglichkeit diskutiert, die methodischen Nachteile der klassischen *Maximum-Likelihood*-Methode, nämlich daß Ereignisse, die im Trainingsmaterial nicht beobachtet wurden, die Wahrscheinlichkeit Null bekommen (*data sparsity*), durch die Unabhängigkeit der Regelkontexte oder durch den Einsatz von *Absolute Discounting* auf ein Hintergrundmodell zu vermeiden.

Kipp ([20]) hat das hier verwendete Trainingsmaterial daraufhin untersucht, ob es sich tatsächlich in diesem Experiment um einen Fall von *data sparsity* (vgl. Abschnitt 3.5.2) handelt. Abbildung 4.4 zeigt das Anwachsen der Anzahl unterschiedlicher Regeln über der Anzahl der beobachteten Regelinstanzen im Trainingsmaterial. Zu Beginn steigt der Graph steil an, weil noch jede beobachtete Regelinstanz zu einer neuen Regel führt, später flacht die Kurve ab, weil jetzt größtenteils nur noch Wiederholungen der bereits bekannten Regeln beobachtet werden. Wäre das Trainingsmaterial tatsächlich repräsentativ im Sinne statistischer Modellierung, so sollte die Anzahl der Regeln innerhalb der verarbeiteten Datenmenge auf einem bestimmten Niveau konvergieren, d.h. keine weiteren Regeln mehr beobachtet werden. Es ist deutlich zu sehen, daß dies nicht der Fall ist: Die Anzahl der Regeln steigt bis zum Ende der Analyse stetig weiter an. Kipp folgert aus diesem Experiment ganz richtig, daß es sich also auch hier um einen Fall von *data sparsity* handelt, und deshalb pragmatische Methoden zur Verbesserung der statistischen Modellierung angebracht sind.

Unabhängigkeit der Kontexte: Das folgende Experiment folgt der in Abschnitt 3.5.3 beschriebenen Methode, den Links- und Rechts-Kontext einer Regel als statistisch voneinander unabhängig zu betrachten.

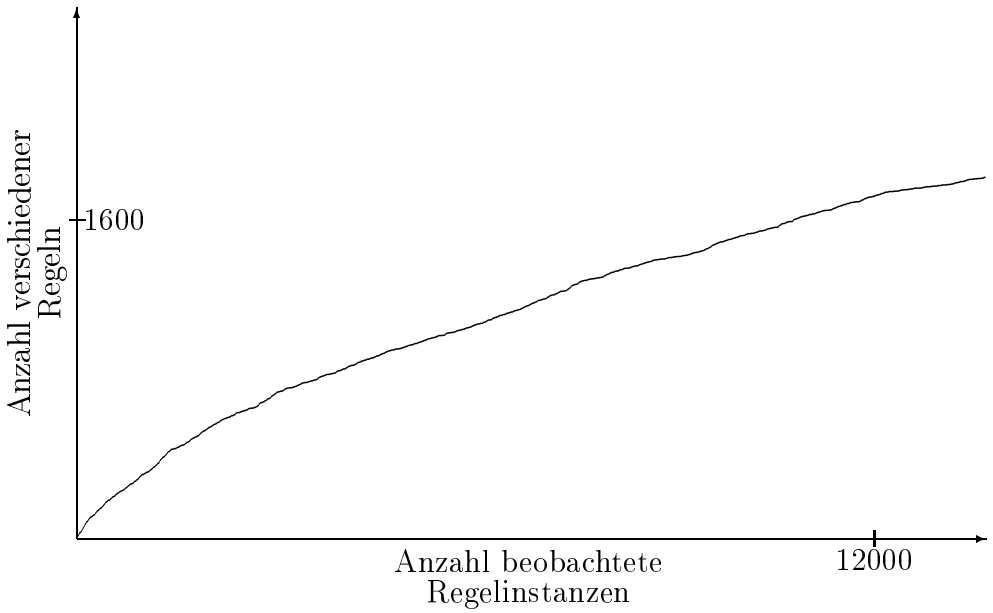


Abbildung 4.4: Anwachsen der Zahl unterschiedlicher Regeln über die im Trainingsmaterial beobachteten Regelinstanzen ([20]).

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere sym. Inter-Labeler-Akuratheit man1-4:	85.6 %
Anzahl der Regeln:	1807
Mittlere sym. Akuratheit man1-4 (manx):	**83.1 %** (81.3 %)

Das heißt, die Akuratheit hat sich gegenüber der klassischen Likelihood-Methode nicht verändert[10].

***Absolute Discounting* auf Hintergrundmodell:** Das folgende Experiment verwendet als Hintergrundmodell alle Varianten, die vom Expertenmodell erzeugt werden können, und verteilt 17% der Wahrscheinlichkeitsmasse[11] gleichförmig auf diese Varianten.

[10] Wenngleich innerhalb der Experimente die Werte signifikant verändert wurden, blieb der Mittelwert über alle Auswertungen gleich.

[11] Dieser pragmatische Parameter wurde heuristisch in einer Serie von Experimenten bestimmt.

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Mittlere symmetrische Inter-Labeler-Akuratheit man1-4:	85.6 %
Anzahl der Regeln:	1807
Mittlere symmetrische Akuratheit man1-4 (manx):	**83.0 %** (81.7 %)

Auch hier ist keine signifikante Veränderung gegenüber der klassischen *Maximum-Likelihood*-Methode zu beobachten.

4.4.3 Diskussion

Etikettierung

Betrachten wir zunächst die Etikettierung. Um zu einer angemessenen Bewertung der dargestellten Forschungsergebnisse zu gelangen, muß hier noch einmal die Problematik der fehlenden absoluten Referenz angesprochen werden.

Wir haben dieses Problem bisher dadurch umschifft, daß wir die Leistung des MAUS-Systems mit einer Gruppe von menschlichen Spezialisten vergleichen und zusätzlich die Inter-Labeler-Akuratheit der menschlichen Bearbeiter untereinander bestimmen. Setzt man diese Werte zueinander ins Verhältnis, erhält man die *relativen Performanz*-Werte von 93 % für gelesene Sprache im Expertenmodell, 87.5 % für spontane Sprache mit dem Expertenmodell und 97 % für spontane Sprache mit dem *Maximum-Likelihood*-Modell. Tatsächlich sagen aber diese Werte – auch wenn sie auf den ersten Blick beeindruckend wirken – nicht sehr viel aus, weil sie von der absoluten Lage der Akuratheiten abhängig sind. Ich möchte daher noch einmal auf das Experiment aus Abschnitt 4.4.2 zurückkommen, bei welchem das MAUS-System als einfaches Abbildungssystem auf die verketteten Zitierformen der Wörter agiert. In gewisser Weise läßt sich dieser Fall als das 'pragmatische Minimum' betrachten, ganz analog zum 'pragmatischen Maximum' in Abschnitt 4.3, weil dieser Fall trivial, d.h. ganz ohne jegliche Aussprache-Modellierung gelöst werden kann. Die Spannweite zwischen 'pragmatischem Minimum' und 'pragmatischem Maximum' ist daher als eine mögliche Bandbreite der Performanz der automatischen Etikettierung zu sehen (deren Grenzen allerdings überschritten werden können!). Normiert man die MAUS-Akuratheit auf diesen Bereich, an dessen einem Ende die menschliche Performanz steht und an dessen anderem Ende die rein maschinelle Mustererkennung, so erhält man für die MAUS-Etikettierung von Spontansprache die folgenden Werte:

Basis:	73 Dialogbeiträge aus *Verbmobil I*
Anzahl der unabhängigen Bearbeitungen:	3
Inter-Labeler-Akuratheit man1-4:	**100.0 %**
Forced Alignment (kanonisch):	**0.0 %**
Expertenmodell:	**10.1 %**
Maximum-Likelihood-Modell:	**79.0 %**
ML-Modell mit unabhängigen Kontexten:	**79.0 %**
ML-Modell mit Absolute Discounting:	**78.2 %**

Diese Darstellung erlaubt es jetzt auch, verschiedene Experimente, die auf unterschiedlichen Korpora beruhen, miteinander zu vergleichen.

Insgesamt zeigen die durchgeführten Evaluationen, daß das Expertenmodell mit 10.1 % der menschlichen Leistung bei der Bearbeitung von Spontansprache zwar für gelesene Sprache, sonst aber allenfalls als Hintergrundmodell einsetzbar ist. Die mangelnde Performanz dieses Ansatzes gerade bei Spontansprache ist vermutlich darauf zurückzuführen, daß die Regeln alle gleichwahrscheinlich behandelt werden.[12] Dies mag für die Bearbeitung von deutlich gesprochener Sprache ausreichend sein, weil in diesem Falle die akustische Abbildung mittels des Viterbi-Algorithmus stärker ins Gewicht fällt und die mangelnde statistische Modellierung der Aussprache kompensiert. Es ist bekannt, daß Spontansprache z.B. in der automatischen Spracherkennung um ein Vielfaches schwieriger zu erkennen ist als gelesene Sprache[13]. Wir sollten also erwarten, daß das MAUS-System – da es dieselben Methoden verwendet – bei der akustischen Modellierung von Spontansprache deutlich schlechtere Ergebnisse liefert. Das Expertenmodell kann diesen Mangel jedoch nicht ausgleichen, da es selber über kein statistisches Wissen verfügt.

Der Ansatz, sowohl Regeln als auch die zugehörigen Regelwahrscheinlichkeiten nach dem *Maximum-Likelihood*-Prinzip aus empirischen Daten zu lernen, zeigt dagegen mit 79 % der menschlichen Leistung ein befriedigendes Ergebnis. Kritiker mögen anmerken, daß eine Degradation von über 20 % für exakte wissenschaftliche Untersuchungen nicht tolerabel sei. Dem möchte ich zustimmen, aber dazu bemerken, daß

[12] Jüngere Untersuchungen von Beringer ([3]) deuten darauf hin, daß sich das Expertenmodell dennoch sehr gewinnbringend für die automatische, iterative Ausdünnung von Regeln anhand konkreter Daten verwenden läßt. Eine Dissertation zu diesem Thema ist derzeit in Vorbereitung; daher möchte ich hier nicht weiter auf dieses Thema eingehen.

[13] Vergleicht man z.B. die Erkennerleistung des Verbmobil-Spracherkennungsmoduls für gelesene Sprache (ca. 98 %) mit der für spontane Sprache (beste Werte bei 86 %), so ergibt sich eine Performanzunterschied von absolut 12 %.

dieser Mangel an Exaktheit in bestimmten, sehr interessanten Untersuchungsansätzen durch die um Größenordnungen höhere Anzahl der möglichen Einzelbeobachtungen ausgeglichen werden kann. Die Fehler, die das MAUS-System unzweifelhaft macht, könnten sich bei der Untersuchung von sehr vielen Etikettierungen als systematische Fehler herausstellen – da das MAUS-System ja ein deterministisches System ist – und somit durch entsprechende Korrekturen bei der Auswertung der Daten berücksichtigt werden. Das gleiche Verfahren ist nicht möglich bei der Auswertung selbst sehr vieler Etikettierungen, die von mehreren verschiedenen Einzelpersonen angefertigt wurden.

Die beiden Versuche, den nachgewiesenen Datenmangel bei der Parameterabschätzung des *Maximum-Likelihood*-Modells entweder durch Annahme der Unabhängigkeit von Links- und Rechts-Kontext oder durch *Absolute Discounting* zu verbessern, hat zu keiner signifikanten Verbesserung der Performanz geführt. Dies ist ein unerwartetes Ergebnis, vor allem vor dem Hintergrund der Untersuchungen von Kipp ([20]), welche zeigen, daß ein *Maximum-Likelihood*-Modell mit *Absolute Discounting* auf das Expertenmodell für steigende *Discounting*-Gewichtung durchwegs niedrigere Perplexität aufweist als das klassische *Maximum-Likelihood*-Modell[14]. Möglicherweise ist der Effekt zu gering, um sich in unseren Experimenten signifikant auszuwirken. Bemerkenswert in diesem Zusammenhang ist noch, daß ein Experiment mit kanonischer Aussprache, aber unter Benutzung der *Absolute Discounting*-Technik schwach signifikant bessere Ergebnisse aufweist als das Expertenmodell[15]. Dies ist ein weiterer Hinweis darauf, daß das Regelwerk des Expertenmodells für spontane Rede nicht adäquat ist.

Zusammenfassend läßt sich sagen, daß das MAUS-System unter Beibehaltung der Domäne von Trainings- und Testmaterial ausreichend gute Ergebnisse für statistische Untersuchungen an großen Korpora liefert, während das Expertenmodell vor allem für Untersuchungen an Korpora mit wenig oder gar keinem verfügbaren Trainingsmaterial von Vorteil sein könnte.

Segmentierung

Was die Segmentierungsleistung der MAUS-Ergebnisse betrifft, so ist die Genauigkeit der Grenzpositionierung in den meisten Fällen für phonetische Ansprüche voll-

[14]Niedrigere Perplexität bedeutet eine bessere Vorhersagefähigkeit des Modells.

[15]Unter Verwendung von Spontansprache.

kommen ungenügend. Der Grund hierfür ist in erster Linie in der Art der Merkmalsextraktion der akustischen Modellierung zu suchen: Die Fenstertechnik mit überlappenden Analysebereichen, welche um jeweils 10 msec verschoben werden, läßt keine feinere zeitliche Auflösung als ca. 15 msec zu. Erste Vorarbeiten haben jedoch gezeigt, daß die Fehler, die das MAUS-System bei der Grenzsetzung macht, weitgehend systematisch und vom Kontext der Etikettierungen links und rechts der betreffenden Grenze abhängig sind. Ein kontext-getriebenes Analyse-Verfahren, welches je nach den gefundenen angrenzenden phonologischen Klassen bestimmte Regeln zur Verbesserung der Grenzlage im Schalldrucksignal anwendet, sollte die Leistung des Systems erheblich verbessern. In einer der ersten Implementierungen der MAUS-Methode, welche später aus Gründen der Effizienz neu programmiert werden mußte, wurden solche Regeln, z.B. die Behandlung von äußerungsinitialen Aspiranten oder die Korrektur der Vordergrenzen von Plosiven, bereits experimentell erprobt, aber leider bis dato noch nicht systematisch untersucht. Hier liegt noch ein großes Potential, die Leistung der MAUS-Methode in Zukunft zu verbessern.

Trotzdem ist die Segmentierung des MAUS-Systems immerhin so gut, daß sich damit deutliche Verbesserungen beim Training von Systemen der automatischen Spracherkennung erzielen lassen (vgl. Abschnitt 5.3). Hier findet sich wahrscheinlich der bereits angesprochene Effekt wieder, daß nämlich der große Umfang der mit der MAUS-Methode möglich gewordenen Beobachtungen sich so positiv auswirkt, daß die relativ grobe Segmentierung im MAUS-Ergebnis nicht mehr ins Gewicht fällt.

Wie bereits erwähnt, ist eine kontextgetriebene Nachbearbeitung der Segmentierung mit Hilfe von einfachen Regeln möglich. Komplexere Verfahren werden es in Zukunft erlauben, die Leistung der MAUS-Methode an die der menschlichen Bearbeiter anzugleichen. Dies gewinnt zunehmend an Bedeutung vor allem für die automatische Sprachsynthese, weil es dort in bestimmten Verfahren sehr darauf ankommt, exakt und im positiven Nulldurchgang geschnittene Sprachsignalstücke zu erhalten (vgl. auch Abschnitt 5.5.1).

Kapitel 5

Anwendungen

Dieses Kapitel beschäftigt sich mit Arbeiten, die auf der Anwendung des MAUS-Systems basieren. Sie reichen von grundlagenorientierten Untersuchungen, wie die Phonem-Statistik des gesprochenen Deutschen, bis hin zu anwendungsorientierten Experimenten, wie die Verbesserung der automatischen Spracherkennung. Nicht alle der hier beschriebenen Arbeiten wurden am Institut für Phonetik und Sprachliche Kommunikation der Ludwig-Maximilians-Universität München durchgeführt. Teilweise handelte es sich um direkte Kooperationen, wie z.B. mit dem Forschungszentrum der Fa. Daimler-Crysler in Ulm (Dr. P. Regel-Brietzmann), teilweise um Projekt-Kooperationen wie im BMB+F-Projekt *Verbmobil*, oder um Arbeiten, die ich bei zwei Forschungsaufenthalten am *International Computer Science Institute* (ICSI) in Berkeley, Kalifornien, durchgeführt habe.

5.1 Phonem-Statistik

Eine sehr naheliegende Anwendung des MAUS-Systems ist die Erstellung einer Auftretensstatistik der deutschen Phoneme in gesprochener Sprache. Zu diesem Zweck wurden die *Verbmobil I und II* Korpora mit insgesamt 1295958 phonologischen Segmenten ausgewertet. Tabelle 5.1 zeigt das Ergebnis in der Form SAM-PA-Symbol S, absoluter Zähler N_S und relative Häufigkeit $P(S)$ bezogen auf das Gesamtkorpus. Da es sich bei der Etikettierung und Segmentierung des MAUS-Systems um eine bündige Segmentierung handelt, sind in dieser Tabelle mehrere Symbole enthalten, die keine phonologischen Einheiten repräsentieren:

SAM-PA S	N_S	$P(S)$	SAM-PA S	N_S	$P(S)$
n	126963	0.097968	k	23435	0.018083
t	86651	0.066863	r	21161	0.016328
s	79126	0.061056	g	19856	0.015321
6	62134	0.047944	U	19445	0.015004
m	54132	0.041770	O	17792	0.013729
d	51183	0.039494	x	15902	0.012270
I	50591	0.039038	o:	15163	0.011698
a	50041	0.038613	u:	12790	0.009869
<p:>	47724	0.036825	j	11198	0.008641
a:	42602	0.032873	N	11147	0.008601
<nib>	41182	0.031777	h	10903	0.008413
v	37187	0.028695	E:	10671	0.008234
@	37059	0.028596	S	10477	0.008084
i:	34133	0.026338	aU	10371	0.008003
f	32889	0.025378	p	9911	0.007648
e:	30113	0.023236	Y	7540	0.005818
aI	29732	0.022942	9	3701	0.002856
l	29309	0.022616	y:	2917	0.002251
Q	29163	0.022503	OY	2853	0.002201
E	29060	0.022424	2:	2157	0.001664
C	26768	0.020655	e	284	0.000219
b	24486	0.018894	Z	32	0.000025
z	24015	0.018531	i	9	0.000007

Tabelle 5.1: Phonem-Häufigkeiten im *Verbmobil I und II*-Korpus (siehe Text)

<p:> Pausen-Segment
<nib> Nicht-sprachlicher Bereich (z.B. Geräusche)

Außerdem ist zu beachten, daß das MAUS-System teilweise gelängte Vokale von ungelängten unterscheidet; um z.B. die Statistik des Phonems /a/ zu bestimmen, müssen die Werte von [aː] und [a] addiert werden.

Wie zu erwarten, ist der Nasal [n] mit 126963 Spitzenreiter in der Auftretenshäufigkeit dicht gefolgt vom Plosiv [t]. Überraschend ist Platz vier mit dem 'tiefen Schwa' [ɒ], dem Reduktionslaut, der typischerweise für die phonologische Lautkombination /er/ geäußert wird. Dieser ist fast doppelt so häufig etikettiert worden wie der Schwa-Laut [ə]. Vergleicht man diese Werte mit einer Auswertung der *kanonischen* Aussprache desselben Materials, so erhält man folgende Gegenüberstellung von *kanonischen* vs. *etikettierten* Häufigkeiten $P(S)$:

$P(S)$	*kanonisch*	*segmentiert*
ə	0.0542	0.0286
ɒ	0.0479	0.0432

Während also der 'tiefe Schwa' [ɒ] seine Häufigkeit zwischen kanonischer Aussprache und tatsächlich etikettierter Sprache kaum ändert, wird die Häufigkeit des Schwa-Lautes [ə] beim Übergang von Zitierform zu tatsächlicher Realisierung drastisch reduziert. Die hauptsächlich verantwortliche Ursache für diese beobachteten Werte ist vermutlich die Reduktion von wortfinalen Silben, die kanonisch einen Schwa-Laut enthalten.[1]

5.2 Wortübergreifende Assimilation

Assimilationen von Lauten, welche durch den Einfluß von angrenzenden Lauten in benachbarten Wörtern verursacht werden, sind in der deutschen Phonetik seit lan-

[1] Um Mißverständnisse zu vermeiden, muß an dieser Stelle darauf hingewiesen werden, daß die kanonische Aussprache, wie sie vom MAUS-System verwendet wird, nach den Konventionen der phonetischen Partner-Institute in den *PhonDat*-Projekten gebildet wird. Das vokalisierte /r/ ist nach dieser Konvention bereits kanonisch. Das Wort 'Eier' wird beispielsweise kanonisch als [aɪɒ] und nicht als [aɪər] beschrieben. Das bedeutet, daß MAUS keine Regel der Form ər>ɒ verwendet, weil diese 'Reduktion' bereits in der kanonischen Form enthalten ist. Vielmehr hat das MAUS-System i.a. Regeln, welche das vokalisierte /r/ wieder in seine explizite Aussprache [ər] zurückformen, da es ja vorkommen kann (wenn auch selten), daß ein Sprecher Bühnendeutsch artikuliert.

gem bekannt und beschrieben (z.B. [24]). Eine interessante Fragestellung ist, ob diese verschiedenen konkreten Assimilationen, die sich durch phonologische Regeln beschreiben lassen, in tatsächlich gesprochener spontaner Sprache vorkommen und wenn ja, wie häufig diese sind. In [41] habe ich für diese Fragestellung eine entsprechende Statistik für das *Verbmobil I*-Korpus untersucht.

Die folgenden Assimilationstypen wurden für die statistische Analyse ausgewählt:[2]

- *Regressive Assimilation des Artikulationsortes*
  Ein wortfinaler stimmloser Plosiv wird elidiert, wenn das darauffolgende Wort mit einem stimmlosen Plosiv beginnt:

  $\emptyset, p, \#k > \emptyset$ *"a***bk***aufen"*
  $\emptyset, t, \#p > \emptyset$ *"gu***t p***ariert"*

  Ein wortfinaler Konsonant wechselt seinen Artikulationsort je nachdem, welcher wortinitiale Konsonant auf ihn folgt:

  $\emptyset, t, \#m > p$ *"sieh***t m***an"*
  $\emptyset, n, \#p > m$ *"sei***n P***latz"*

- *Progressive Assimilation der Artikulationsart*
  Ein wortinitialer Konsonant wechselt seine Artikulationsart je nachdem, was für eine Artikulationsart der vorangegangene wortfinale Konsonant hatte, wobei beide Konsonanten den gleichen Artikulationsort haben:

  $\emptyset, n\#d, \emptyset > \#n$ *"we***nn d***en"*
  $\emptyset, m\#b, \emptyset > \#m$ *"de***m B***aum"*
  $\emptyset, s\#d, \emptyset > \#s$ *"da***s d***a"*

- *Assimilation der Stimmhaftigkeit*
  Die Stimmhaftigkeit bzw. Nicht-Stimmhaftigkeit wird zwischen wortfinalem und wortinitialen Konsonanten angeglichen:

  $\emptyset, t, \#v > d$ *"geh***t w***eiter"*
  $t\#, v, \emptyset > f$ *"geh***t w***eiter"*
  $\emptyset, t, \#d > \emptyset$ *"geh***t d***as"*
  $t\#, d, \emptyset > t$ *"geh***t d***as"*

- *Elision von stimmlosen Frikativen*
  Stimmlose Frikative werden an der Wortfuge getilgt.

[2]Die Formulierungen der Assimilationsregeln wurden an die in dieser Arbeit verwendete Konvention angepaßt.

$t\#, h, \emptyset > \emptyset$	*"seh***t h***er"*
$N\#, h, \emptyset > \emptyset$	*"la***ng h***er"*
$C\#, h, \emptyset > \emptyset$	*"si***ch h***aben"*
$\emptyset, x, \#t > \emptyset$	*"na***cht***arocken"*
$\emptyset, x, \#h > \emptyset$	*"na***chh***elfen"*

303446 gesprochene Wörter aus dem deutschen Teil des *Verbmobil I*-Korpus wurden mit der MAUS-Methode (*Maximum Likelihood*-Aussprachemodellierung) analysiert und die oben genannten Assimilationen gezielt gesucht, indem die kanonischen Wortformen mit der in Abschnitt 3.5.2 beschriebenen symbolischen Dynamischen Programmierung auf die vom MAUS-System gelieferten Etikettierungen abgebildet wurden. Tabelle 5.2 zeigt jeweils die Anzahl der in der kanonischen Aussprache gefundenen Kontexte N_{ges} und die Anzahl der Assimilationen N_{ass}, die an diesen Kontexten vom MAUS-System festgestellt wurden. Das Ergebnis zeigt, daß 5 von 16 Assimilationen überhaupt nicht von MAUS detektiert worden sind, drei sind so selten, daß es sich dabei möglicherweise auch um Fehlleistungen des Systems handeln kann, und nur 8 Assimilationen scheinen signifikant häufig aufgetreten sein.

Eine solches Ergebnis ist nur unter vielen Vorbehalten zu interpretieren:
Zunächst handelt es sich hier nicht um die statistische Analyse der gesprochenen Sprache an sich, sondern um die eines endlichen Korpus, mit endlich vielen Sprechern und einer eng begrenzten Domäne. Desweiteren muß berücksichtigt werden, daß das hier verwendete Aussprache-Modell auf einem Ausschnitt des MAUS-Systems trainiert wurde, und somit nur Aussprachevarianten finden kann, die in diesem Ausschnitt repräsentativ auftraten. Aus Abschnitt 3.5.3 wissen wir, daß das Trainingsmaterial aus *Verbmobil I* nicht vollständig repräsentativ sein kann, weil auch nach fast vollständiger Analyse des Korpus auf Regelinstanzen hin immer noch neue, noch nicht beobachtete Regeln auftraten. Drittens wissen wir aus den Ergebnissen der empirischen Evaluierung in Kapitel 4, daß das MAUS-System bei Anwendung auf spontane Sprache nur ca. 79 % der Leistung eines durchschnittlichen menschlichen Experten erreicht, also mit Fehlern zu rechnen ist.

Selbst unter Berücksichtigung aller dieser Vorbehalte ist es dennoch bemerkenswert, daß es mit Hilfe der MAUS-Methode zum ersten Mal möglich ist, mit relativ einfachen statistischen Methoden in der Literatur beschriebene Effekte der Aussprachevariation an sehr großen Datenmengen zu verifizieren und festzustellen, daß gewisse Assimilationen über Wortgrenzen hinweg zumindest so selten aufzutreten scheinen[3],

[3]oder wenn, dann unter anderen Rahmenbedingungen, z.B. in bestimmten Dialekten.

Assimilation	N_{ges}	N_{ass}
$\emptyset, p, \#k > \emptyset$	95	0
$\emptyset, t, \#p > \emptyset$	302	106
$\emptyset, t, \#m > p$	2290	0
$\emptyset, n, \#p > m$	560	16
$\emptyset, n\#d, \emptyset > \#n$	7804	360
$\emptyset, m\#b, \emptyset > \#m$	808	0
$\emptyset, s\#d, \emptyset > \#s$	3138	7
$\emptyset, t, \#v > d$	1834	1
$t\#, v, \emptyset > f$	1834	0
$\emptyset, t, \#d > \emptyset$	5457	3053
$t\#, d, \emptyset > t$	5457	0
$t\#, h, \emptyset > \emptyset$	893	852
$N\#, h, \emptyset > \emptyset$	49	45
$C\#, h, \emptyset > \emptyset$	998	947
$\emptyset, x, \#t > \emptyset$	325	1
$\emptyset, x, \#h > \emptyset$	177	3

Tabelle 5.2: Wortübergreifenden Assimilationen - absolute Häufigkeit im *Verbmobil I*-Korpus. Die Spalte N_{ges} enthält die absoluten Häufigkeiten der Regelkontexte in der kanonischen Aussprache; die Spalte N_{ass} enthält die absoluten Häufigkeiten der Regelinstanzen im Material, das vom MAUS-System bearbeitet wurde.

daß ihre Relevanz für die technische Verwertung fraglich ist. Solche wissenschaftlichen Ergebnisse können u.U. direkten Einfluß auf die akustische Modellierung in Methoden der automatischen Spracherkennung haben, in welchen die korrekte Behandlung von akustischem Kontext über Wortgrenzen hinweg ein großes Problem darstellt.

5.3 Automatische Spracherkennung

Für die technische Anwendung in Form von Systemen zur automatischen Erkennung gesprochener Sprache bestehen zahlreiche Möglichkeiten, die Analysedaten des MAUS-Systems einzusetzen. Als Beispiel möchte ich hier zunächst eine Untersuchung skizzieren, die ich 1997 als Gastwissenschaftler am *International Computer Science Institute* (ICSI) durchgeführt habe ([43]).

Es handelte sich um die – nach wie vor unter Sprachwissenschaftlern umstrittene – Frage. ob die Variation der Aussprache im Kontext eines technischen Systems für die Erkennungsleistung eine Rolle spielt oder nicht.

> Beispiel:
> Die akustischen Modelle[4] eines sprecherunabhängigen Spracherkennungssystem werden mit einer Trainingsstichprobe von 1000 Sprechern trainiert. Als Aussprache-Lexikon dient ein Liste von 500 kanonischen Formen des betreffenden Vokabulars.
> Die Frage ist nun: Läßt sich eine signifikante Verbesserung der Performanz des Systems erreichen, indem man statt der kanonischen Aussprache eine explizite Modellierung der Aussprache zum Einsatz bringt, und wie muß diese idealerweise beschaffen sein?

Tatsächlich scheint diese Fragestellung zunächst trivial, weil jeder, der sich mit fließend gesprochener Rede auseinander gesetzt hat, aus eigener Erfahrung weiß, wie stark ein Wort, eingebettet in eine Äußerung, sich von dem als Zitierform geäußerten gleichen Wort unterscheidet. Sehr früh schon gab es daher Bestrebungen, Wissen über Reduktionen, Assimilationen und andere Effekte der fließenden Sprache in technische Systeme zur automatischen Spracherkennung zu integrieren (z.B. eigene Arbeiten in [44]). Überraschend war, daß in den meisten Publikationen zu

[4] Die genaue Topologie ist für unser Beispiel zunächst nicht von Bedeutung.

diesem Thema über keine oder nur schwach signifikante Verbesserungen in der Performanz der Systeme berichtet wird. Begründet wurde dies häufig damit, daß der Suchraum des Dekodierungsproblems der automatischen Spracherkennung durch zusätzliche Variation der Aussprache vergrößert wird, und dadurch die Gefahr einer Fehlerkennung ansteigt.

> Zum Beispiel könnte in dem oben genannten System mit 500 Wörtern im Wortschatz eine Analyse der Wortrealisierungen ergeben, daß mit durchschnittlich 2,5 Aussprachevarianten pro lexikalem Eintrag zu rechnen ist[5]. Technisch bedeutet dies für den Mustererkennungsprozeß, daß im schlimmsten Falle statt mit 500 Worthypothesen an jeder Stelle des Schalldrucksignals nunmehr mit 1250 Worthypothesen gerechnet werden muß. Abgesehen von diesem deutlich erhöhten Aufwand kann die Hinzufügung von Aussprachevarianten auch die Ambiguität des Aussprachemodells erhöhen. D.h. die Anzahl der Homophone im Aussprachemodell steigt an, und diese können nur durch Einbeziehung z.B. einer Wortfolge-Statistik oder einer Syntax aufgelöst werden.

Von diesen zweifellos vorhandenen technischen Problemen abgesehen sind heutzutage zumindest zwei prinzipielle Gründe für das Versagen dieser frühen Ansätze bekannt. Der eine Grund ist, daß in den ersten Versuchen, explizites Wissen über die Variabilität gesprochener Sprache technisch zu verwerten, mit regelbasierten Systemen gearbeitet wurde, welche sich nur sehr schwer mit den statistischen Ansätzen der akustischen Modellierung vereinen ließen. Den zweiten Grund möchte ich vorerst noch etwas zurückstellen und zum besseren Verständnis zunächst kurz unser Experiment beschreiben.

Zunächst wurde ein Basissystem entworfen und getestet, welches sprecherunabhängig die jeweils wahrscheinlichste Wortkette zu Äußerungen aus dem deutschen *Verbmobil I*-Korpus bestimmte. Der Korpus wurde aufgeteilt in ca. 30 Stunden Trainingsmaterial[6] und ca. 4 Stunden Testmaterial[7]. Der Wortschatz des Erkenners hatte einen Umfang von 840 Wörtern. Die 46 akustischen Modelle des Erkenners[8] wurden durch kontinuierliche Hidden Markov Modelle (HMM) mit je 3 - 5 Zuständen

[5] Werte für Aussprachevarianten pro lexikalem Eintrag variieren in der Literatur zwischen 1,5 und 20.

[6] *Verbmobil* Volumes 1, 2, 3, 4, 5, 7, 12.

[7] *Verbmobil* Volume 14.

[8] Phonologische Symbole und spezielle Modelle für Pause, Anfangspause, Endpause, Geräusche.

realisiert. Die Parameter der HMM wurden aus 1 Stunde, 40 Minuten manuell etikettiertem und segmentiertem Material initialisiert und anschließend in einer iterativen Prozedur mit dem gesamten Trainingsmaterial solange neu abgeschätzt, bis die Performanz auf der Teststichprobe konvergierte[9]. Als *language model*, also der statistischen Modellierung der Wortfolgen, diente ein einfaches Bigramm-Modell, welches lediglich aus den Transkripten des Trainingsmaterials trainiert wurde.
Die Leistung des Basissystems wurde anhand der *Wortakuratheit*[10] beurteilt. Diese ist – ganz analog zur Akuratheit in Abschnitt 4 – die relative Häufigkeit von Ersetzungen, Auslassungen und Einfügungen von Wörtern in den Wortketten, welche der Erkenner auf dem Testmaterial produziert. Die Wortakuratheit konvergierte für das Basissystem nach 12 Trainings-Iterationen bei einem Wert von 63.44 %[11].

Das gesamte Trainingsmaterial wurde anschließend mit der MAUS-Methode etikettiert und segmentiert, und aus diesem Material wurde ein einfaches statistisches Modell zur Aussprache-Modellierung entwickelt. Dabei verzichteten wir vollständig auf die Verwendung von Regeln, sondern beschränkten uns auf die direkte Modellierung aller im Wortschatz vorkommenden Wörter. Im Trainingsmaterial wurden ca. 230000 Wortformen beobachtet, und deren jeweilige Varianten ausgezählt. Nachfolgend ein Ausschnitt aus diesem Material:

```
terminlich
adj
t E 6 m i: n l I C
t E 6 m i: n I C        3
t @ m i: l I C  3
t E 6 m i: n l I C      10
t E 6 m i: l I C        1
t @ m i: n l I C        7
&
```

[9]Diese Standard-Prozedur ist auch unter dem Namen *segmental-k-means*, oder *embedded training* bekannt (z.B. [18]).

[10]engl. *word acuracy*.

[11]Dies ist nicht der optimal erreichbare Wert in dieser Erkennungsaufgabe. Im Rahmen des Verbmobil-Projektes wurden mehrere verschiedene akustische Erkenner entwickelt und in regelmäßigen Abständen evaluiert. Die besten Ergebnisse aus diesen Evaluationen erreichten Wortakuratheiten von ca. 86 %. Dabei ist anzumerken, daß die Wortakuratheit auf der besten erkannten Wortkette wenig über die Performanz im integrierten Gesamtsystem aussagt, weil es dort weniger auf den *besten erkannten Satz*, als vielmehr auf das *beste Wortgitter (word lattice)* ankommt, welches an die weiteren Verarbeitungsstufen des Systems weitergereicht wird. Die Evaluierung von Wortgittern ist keine triviale Aufgabe, weshalb in dieser Untersuchung auf den Einsatz von Wortgittern verzichtet wurde.

```
...
Karfreitag
nou
k a: 6 f r aI t a: k
k a: 6 f r aI t a: k    15
k a: 6 f r aI t a x     3
&
...
weil
par
v aI l
v a l   11
v aI    108
v aI l  207
&
...
siebenundzwanzigsten
adj
z i: b @ n U n t t s v a n t s I C s t @ n
z i: b @ n U n s v a n t s I s t @ n    1
z i: b m U n s v a n t s I k s t n      2
z i: b m U n s v a n s I C s t n        1
z i: b m U n s v a n t s I C s t @ n    1
z i: m U n s v a n t s s t @ n  1
z i: m U n s v a n t s s n      1
... (weitere 48 Varianten nicht dargestellt)
&
...
Namen
nou
n a: m @ n
n a: m  30
n a: m @ n      15
&
...
Essen
nou
Q E s @ n
@ s n   2
E s n   16
E s @ n 6
s n     3
E s     1
Q E s @ n       7
```

```
Q E s   1
Q E s n 21
&
```

Die oben gezeigten Einträge pro lexikaler Einheit bestehen jeweils aus orthographischer Form, dem Wortklassenkürzel[12], der kanonischen Aussprache und einer Liste von gefundenen Aussprachevarianten, unter welchen sich natürlich auch die kanonische wiederfinden kann, zusammen mit der absoluten Häufigkeit dieser Varianten. Beobachtungen mit sehr geringer Häufigkeit, wie z.B. die 'Varianten' /Es/ und /QEs/ für das Wort 'Essen' sind häufig nur auf Fehler des MAUS-Systems zurückzuführen. Die Daten wurden daher nach folgender Methode ausgedünnt, um solche 'Ausreißer' zu vermeiden und nur die statistisch zuverlässigen Werte weiter zu verarbeiten:

- Varianten von lexikalen Einträgen, deren *Gesamtzahl* an Beobachtungen unter einer Schwelle von absolut 20 zu liegen kam, wurden nicht berücksichtigt. D.h. Wörter, die weniger als 20mal im Korpus zu beobachten sind, haben im Modell keine Variation.
- Aus den verbleibenden Beobachtungen wurde die bedingte Wahrscheinlichkeit $P(V|L)$, daß gegeben der lexikale Eintrag L die Variante V zu beobachten ist, berechnet. Anschließend wurden alle Varianten getilgt, deren bedingte Wahrscheinlichkeit kleiner als 10 % war.
- Die bedingten Wahrscheinlichkeiten $P(V|L)$ der verbleibenden Varianten pro lexikalem Eintrag L wurden auf eine Gesamtwahrscheinlichkeit von 1.0 normiert:

$$\sum_{V} P(V|L) \equiv 1 \tag{5.1}$$

Das Resultat dieser Prozedur für die oben gezeigten Daten besteht dann nur noch aus einer Liste von möglichen Aussprachevarianten V pro lexikalem Eintrag L zusammen mit ihrer bedingten Wahrscheinlichkeit $P(V|L)$:

```
terminlich          0.434783
t E 6 m i: n l I C
terminlich          0.130435
t E 6 m i: n I C
```

[12] *nou* = Nomen, *adj* = Adjektiv, *par* = Partikel etc.

```
terminlich           0.304348
t @ m i: n l I C
terminlich           0.130435
t @ m i: l I
Karfreitag           1.000000
k a: 6 f r aI t a: k
weil                 0.342857
v aI
weil                 0.657143
v aI l
siebenundzwanzigsten 0.509091
z i: b m U n s v a n t s I s t n
siebenundzwanzigsten 0.490909
z i: m U n s v a n t s I s t n
Namen                0.333333
n a: m @ n
Namen                0.666667
n a: m
Essen                0.320000
E s n
Essen                0.420000
Q E s n
Essen                0.120000
E s @ n
Essen                0.140000
Q E s @ n
```

In dem hier wiedergegebenen Auszug wurde z.B. dem lexikalen Eintrag 'Karfreitag' nur noch die kanonische Ausspracheform /ka:6fraIta:k/ zugewiesen, weil die Gesamtanzahl der Beobachtungen mit 18 als nicht zuverlässig eingestuft wurde. Konsequenterweise hat daher die einzige, kanonische Ausspracheform des Eintrags 'Karfreitag' die bedingte Wahrscheinlichkeit von 1,0.

Mit diesem einfachen statistischen Modell zur Variation der Aussprache und dem oben beschriebenen Basissystem wurden die folgenden Experimente durchgeführt:

1. Training der akustischen Modelle auf kanonische Wortformen[13] und Test mit dem statistischem Aussprachemodell.

[13] Analog zum Basissystem.

2. Training der akustischen Modelle auf MAUS-Etikettierungen und Test mit dem statistischem Aussprachemodell.

3. Training der akustischen Modelle auf MAUS-Etikettierungen und Test mit kanonischer Aussprache.

Mit Hilfe dieser drei Experimente – zusammen mit dem Ergebnis des Basissystems – sollte die folgende Frage geklärt werden:

> Kann eine Steigerung der Performanz erreicht werden, wenn *gleichzeitig* die akustische Modellierung *und* die Aussprachemodellierung auf der Basis von MAUS-Etikettierungen durchgeführt wird? *(Experiment 2 vs. Basissystem)*

Unsere Hypothese war, daß die meisten bisherigen Versuche, die Performanz von automatischen Spracherkennungssystem zu verbessern, gescheitert waren, weil im Trainingskonzept eine Inkonsistenz enthalten war: Entweder wurden dieselben akustischen Modelle verwendet und nur die Aussprachemodellierung ausgetauscht (wie in *Experiment 1*) oder es handelte sich um ein Aussprachemodell, welches auf völlig anderen Daten entwickelt wurde. Um außerdem auszuschließen, daß wir es hier nicht lediglich mit einem Seiteneffekt der besseren akustischen Modellierung verursacht durch die MAUS-Daten zu tun haben, wurde zusätzlich noch das *Experiment 3* durchgeführt.

Abbildung 5.1 zeigt den Verlauf der Wortakuratheit über den Trainingsiterationen für das Basissystem und *Experiment 2*, also der konsistenten Verwendung von akustischen und Aussprache-Modell. Es ist deutlich zu sehen, daß die Wortakuratheit in *Experiment 2* während des gesamten Trainingsablaufs über der des Basissystems liegt. Der Konvergenzwert von 66,35 % liegt signifikant (0.001) über dem Wert 63.44 % des Basissystems. Die Tatsache, daß die Performanz in *Experiment 2* mit weitaus besseren Werten startet als im Basissystem, ist auf die Initialisierung der akustischen Modelle auf die Segmentierungen des MAUS-Systems zurückzuführen.

Das *Kontroll-Experiment 1* wurde mit den Parametersätzen der Konvergenzpunkte durchgeführt. D.h. die akustischen Modelle am Konvergenzpunkt des Basissystems wurden mit dem Aussprachemodell aus *Experiment 2* kombiniert. Wie wir erwartet hatten, ergab sich für diese inkonsistente Kombination keine Verbesserung der Wortakuratheit, im Gegenteil lag diese sogar schwach signifikant unterhalb der des Basissystems.

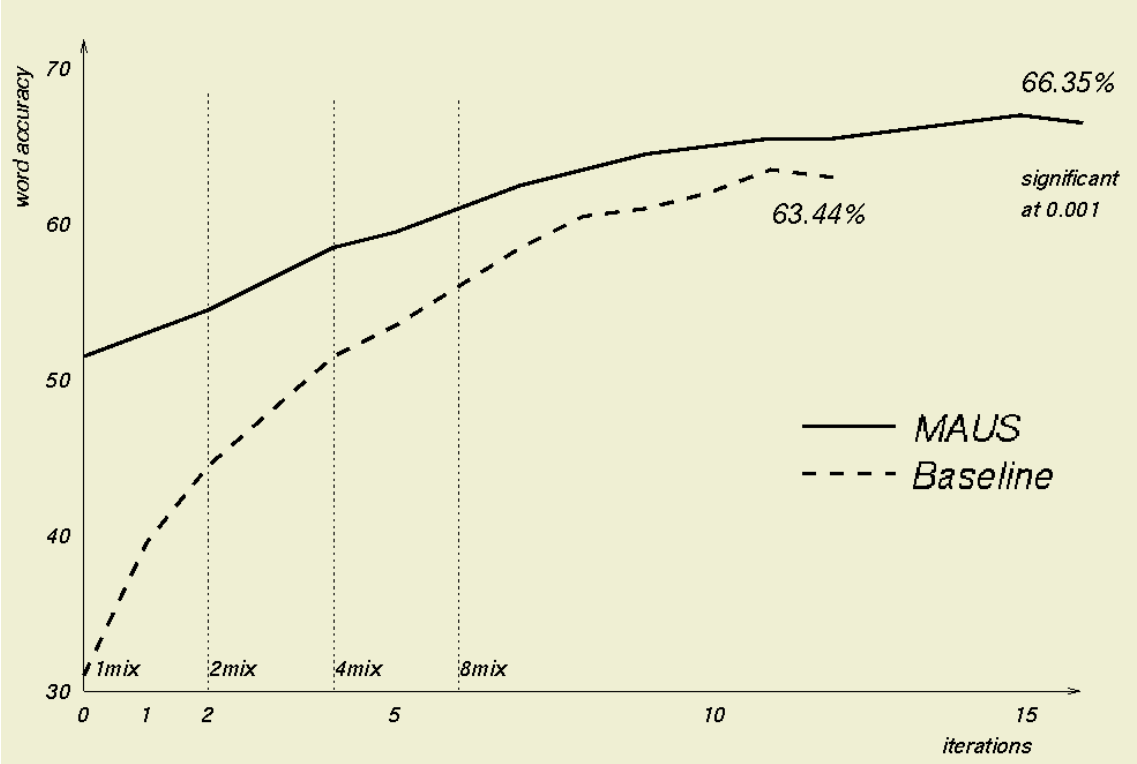


Abbildung 5.1: Verlauf der Testset-Performanz (Wortakuratheit) über Trainingsiterationen für das *Basissystem* (Baseline) und *Experiment 2* (MAUS)

Ganz analog wurde *Experiment 3* durchgeführt: Die akustischen Modelle am Konvergenzpunkt von *Experiment 2* wurden mit der kanonischen Aussprachemodellierung aus dem Basissystem kombiniert. Auch hier ergab sich keine signifikante Verbesserung gegenüber der Performanz des Basissystems. Das bedeutet, daß der Gewinn, den wir in *Experiment 2* beobachten, nicht allein auf eine bessere akustische Modellierung zurückgeführt werden kann.

Aus diesen Experimenten lassen sich folgende Schlußfolgerungen ziehen:

- Eine Verbesserung der Performanz in Systemen zur automatischen Spracherkennung läßt sich durch Modellierung der Aussprache erzielen.
- Bedingung dabei ist, daß akustisches und Aussprache-Modell konsistent und auf Basis derselben Daten erstellt werden.

Damit wird jetzt auch der erwähnte zweite Grund für das häufig berichtete Versagen der Aussprachemodellierung deutlich: In den meisten bisherigen Untersuchungen war die obige Bedingung nämlich in der einen oder anderen Form nicht erfüllt.

- Die erzielte Verbesserung ist relativ gering, wenn man den deutlich erhöhten Aufwand bei Training und Test des Systems in Betracht zieht.

Letzteres führt unweigerlich zur Frage, welche Informationsquellen dann wohl dem menschlichen Gehirn zur Verfügung stehen, wenn es in der Lage ist, diese Aufgabe so mühelos zu bewältigen. Eine der vielen möglichen Hypothesen ist, daß es sich um eine Adaption an den jeweiligen regionalen Dialekt handeln könnte. Der nächste Abschnitt befaßt sich mit einer Untersuchung, die genau dieser Fragestellung nachgeht.

5.4 Regionale Aussprachevariation

Die folgende Untersuchung wurde von N. Beringer in Zusammenarbeit mit Dr. P. Regel-Brietzmann vom Daimler-Crysler-Forschungsinstitut in Ulm durchgeführt ([5]). Es sollte geklärt werden, ob sich die Performanz des akustischen Spracherkenners im *Verbmobil I*-Projekt durch Berücksichtigung der regionalen Herkunft des jeweiligen Sprechers verbessern ließe.

Aus Platzgründen kann ich die Ergebnisse der Experimente hier nur stark gekürzt zusammenfassen:

Grundlage der Experimente war ein sog. *cheating experiment*, bei welchem angenommen wurde, das System sei dazu in der Lage, die sprachliche Herkunft eines unbekannten Sprechers mit 100%iger Genauigkeit zu bestimmen. Mit Hilfe dieser Information kann das System für jeweils eine von 12 Sprachregionen[14] ein spezielles Aussprachemodell auswählen, welches aus Trainingsdaten von Sprechern genau dieser Region trainiert worden war (siehe auch Abb. 5.2). Das Spracherkennungssystem 'erkennt' sozusagen zuerst, woher der unbekannte Sprecher stammt und 'adaptiert' sich daraufhin durch Anpassung seines Aussprachemodells. Die akustische Modellierung wurde in allen Experimenten beibehalten, d.h. es handelt sich hier auf jeden Fall um eine Inkonsistenz im Sinne des vorangegangenen Abschnitts. Der Grund hierfür war vorwiegend statistischer Natur: Eine Umschaltung auch der akustischen Modelle hätte zu einer starken Degradierung der akustischen Modellierung geführt, weil dann nur im Schnitt ein Zwölftel des Trainingsmaterials für jeweils einen Satz

[14] *Hessen, Berlin, Oberschwaben, Franken, Westphalen, Württemberg, Niedersachsen, Holstein, Bayern, Rheinland*, sowie zwei weitere Klassen *Nord* und *Süd*, in welchen nicht eindeutig zu identifizierende Sprecher eingeordnet wurden.

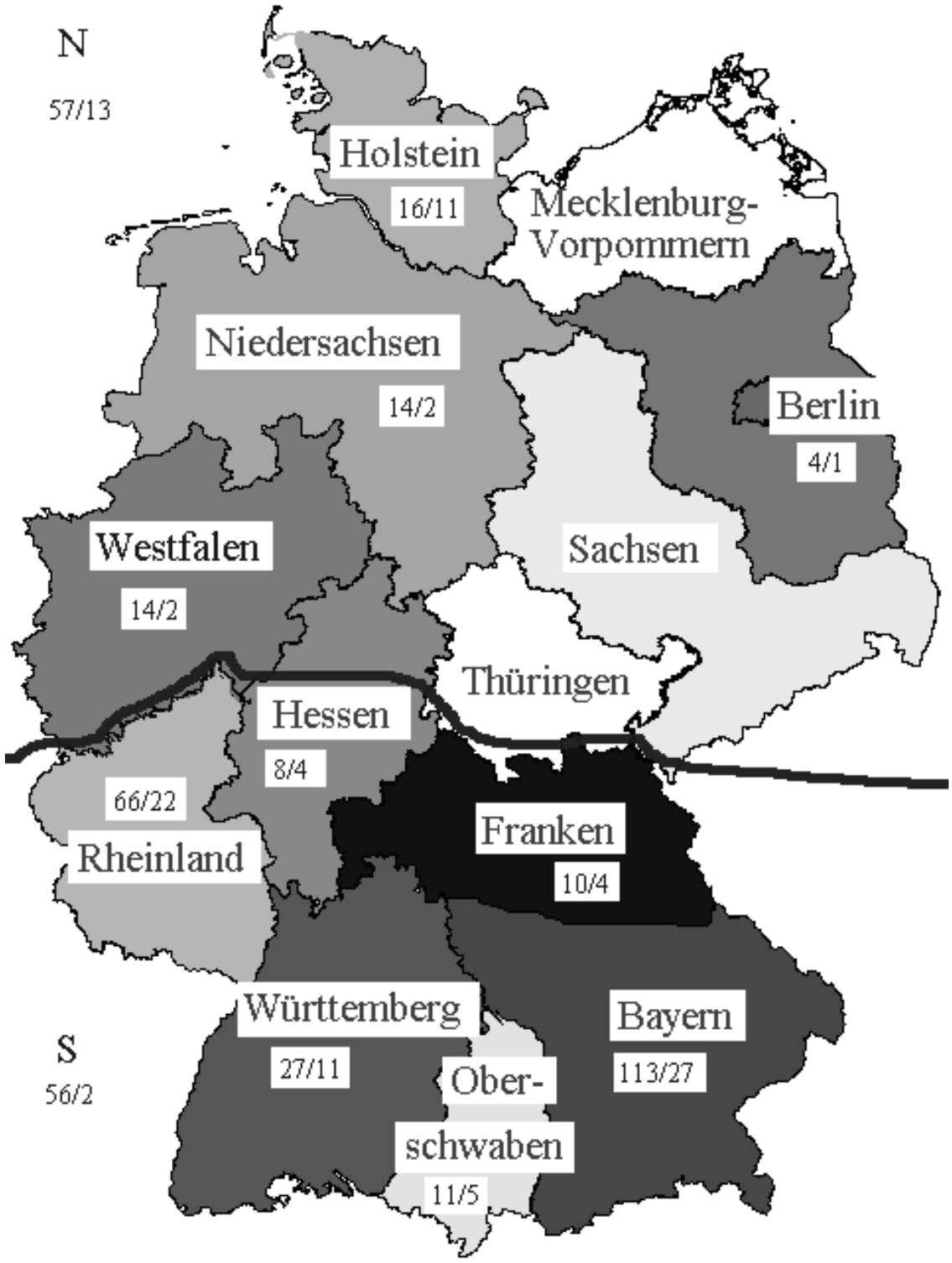


Abbildung 5.2: Verteilung der Trainings- und Testsprecher über die 12 verwendeten Sprachregionen. Im ostdeutschen Sprachraum standen keine ausreichenden Daten zur Verfügung.

dialektal spezifischer Modelle zur Verfügung gestanden hätte.
Das Basissystem, welches mit einem einheitlichen kanonischen Aussprachemodell für alle Testsprecher arbeitet, erreichte eine Wortakuratheit von 69.09 %.

In einem ersten Experiment wurden die regional-spezifischen Aussprachemodelle um sämtliche im entsprechenden Trainings-Subkorpus gefundenen Aussprachevarianten (im Schnitt 1.8 pro lexikalem Eintrag) erweitert. Ein Test über alle regionalen Testdaten ergab eine signifikante Degradation der Wortakuratheit auf 66.52 %.

In den weiteren Experimenten wurden die Aussprachevarianten der verschiedenen Regionen bestimmten Beschränkungen unterworfen, ähnlich dem Ausdünnungsverfahren im Abschnitt 5.3, um statistisch unzuverlässige Varianten auszuklammern. Eine Auswahl dieser Experimente und ihr jeweiliges Ergebnis ist hier kurz wiedergegeben.

- *Minimale Variantenzahl*
  Es wurden nur Varianten in das Aussprachemodell aufgenommen, deren gesamte Beobachtungszahl höher als 50 war. Erzielte Wortakuratheit: 69.60 %

- *Minimale bedingte Variantenwahrscheinlichkeit*
  Es wurden nur Varianten in das Aussprachemodell aufgenommen, deren bedingte Auftretenswahrscheinlichkeit $P(V|W)$ höher als 15 % war. Erzielte Wortakuratheit: 69.38 %

- *Beschränkung auf Funktionswörter*
  Es wurden nur Varianten sogenannter Funktionswörter[15] in die regional-spezifischen Aussprachemodelle aufgenommen. Hintergrund dieses Experiments war der Gedanke, daß diese die am häufigsten vorkommenden Wörter in gesprochener Sprache sind, und daher einerseits für die statistische Analyse in genügender Anzahl beobachtet werden können, andererseits möglicherweise durch ihre Häufigkeit besonders anfällige Kandidaten für dialektale Verschleifungen darstellen. Das Ergebnis des Experiments konnte diesen Gedanken allerdings nicht bestätigen: die erzielte Wortakuratheit reduzierte sich für dieses Experiment auf 68.41 %

Zusammenfassend läßt sich sagen, daß keine signifikanten Verbesserungen bei Berücksichtigung der regionalen Herkunft der Versuchspersonen erzielt werden konnten. Der Grund hierfür kann in der Natur der Sache an sich liegen, nämlich daß die

[15]Die Klassen Artikel, Pronomen, Konjugationen, Hilfsverben und Präpositionen.

– vermutlich in einer formalen Aufgabenstellung wie *Verbmobil* nur sehr schwache – regionale Variation der Aussprache bereits von den akustischen Modellen in ausreichendem Umfang berücksichtigt wird. Es könnte aber auch sein, daß allein durch die Tatsache, daß das Trainingsmaterial für die Erstellung der regional-spezifischen Aussprachemodelle in zwölf Teile geteilt werden mußte, die individuell zur Verfügung stehende Datenmenge so klein geworden war[16], daß keine zuverlässige statistische Modellierung mehr möglich war, und deshalb die Performanz-Werte degradierten. Es ist hier, wie immer in statistisch-empirischen Untersuchungen, schwer zu trennen, ob ein beobachteter Effekt direkt kausal oder indirekt aus der Datensituation entstanden ist.

5.5 Weitere Anwendungen für die MAUS-Technik

Die folgenden kurzen Skizzen beschreiben Arbeiten, welche entweder nicht an der Ludwig-Maximilians-Universität München durchgeführt wurden oder von Kollegen am Institut für Phonetik und Sprachliche Kommunikation durchgeführt wurden, welche nicht direkt an der Entwicklung des MAUS-Systems beteiligt sind, sondern dieses vielmehr als Hilfswerkzeug für ihre wissenschaftliche Arbeit benutzen.

5.5.1 Konkatenative Sprachsynthese

Die verschiedenen Methoden der automatischen Erzeugung fließender Sprache, kurz Sprachsynthese genannt, beruhen im wesentlichen auf zwei fundamental unterschiedlichen Ansätzen: der sogenannten Vollsynthese und der konkatenativen Synthese (sehr gute Übersicht beispielsweise in [12]).

Bei der – historisch älteren – Vollsynthese wird versucht, den gesamten akustischen Spracherzeugungsprozeß mit Hilfe eines Modell nachzubilden. Prototypisches Beispiel ist der *Klatt-Synthetisator* ([23]), welcher analog zum Quelle-Filter-Modell nach Fant ([11]) die Schallerzeugung in der Glottis durch eine Impulsquelle und das artikulatorische Ansatzrohr durch eine Serie von Resonatoren (Filtern) zu realisieren suchte.

Die konkatenative Synthese versucht statt dessen, die Sprache weitgehend unabhängig von einem Vokaltraktmodell nur durch Verknüpfung tatsächlich produzierter

[16] Minimum waren 19 Dialogbeiträge.

Sprachstücke zu erzeugen. Damit eine solche 'Verkettung'[17] eine natürlich klingenden A-Prosodie[18] erhält, sind verschiedene Maßnahmen wie Manipulation der Dauer und des Grundfrequenzverlaufs notwendig.

Konkatenative Verfahren liefern derzeit die am natürlichsten klingende Sprachsynthese und werden deshalb intensiv erforscht und ständig verbessert (z.B. das Synthese-System CHATR [8]). Das Spektrum der dabei verwendeten Spracheinheiten reicht von ganzen Wörtern über Silben, Halbsilben, Diphonen und Phonemen bis hin zu Teilen von Phonemen.

Alle konkatenativen Verfahren basieren jedoch auf einer gründlichen Aufbereitung des empirischen Sprachmaterials, aus welchem die zu synthetisierende Stimme zusammengesetzt wird. Alle diese Verfahren benötigen daher eine Etikettierung und – mehr oder weniger sorgfältige – Segmentierung des Sprachmaterials. Dies ist natürlich mit sehr viel manueller Arbeit verbunden, insbesondere wenn man bedenkt, daß die konkatenativen Verfahren im allgemeinen desto bessere Qualität produzieren, je größer das Inventar der verwendeten Signalstücke ist, und für jede Synthesestimme die Bearbeitung wiederholt werden muß.

Das MAUS-Verfahren kann für solche Systeme das 'Rohmaterial' liefern, insbesondere die interessierenden Teile einer Sprachaufzeichnung aus Trägersätzen herausschneiden und/oder bestimmte Typen von Verschleifungen automatisch etikettieren. Kollegen an der Universität Bonn haben dies für das dort entwickelte Synthesesystem HADIFIX ([34]) bereits vorexerziert. Um allerdings zu hochqualitativen Segmenten für die Synthese zu kommen, ist eine manuelle Nachbearbeitung immer noch notwendig.

5.5.2 Elektromagnetische Artikulographie

Im Rahmen der experimentalphonetischen Untersuchungen am Institut für Phonetik und Sprachliche Kommunikation mit Hilfe der *Elektro-Magnetischen Artikulographie* (EMMA[19], z.B. [15]) wird das MAUS-System routinemäßig zur Lokalisation von interessanten Lautkombinationen im Schalldrucksignal eingesetzt.

Die Elektromagnetische Artikulographie basiert auf der Messung von Spulenpositio-

[17] engl. *concatenation*.

[18] im Sinne von Tillmann und Mansell [48].

[19] engl. *electromagnetic midsagital articulography*

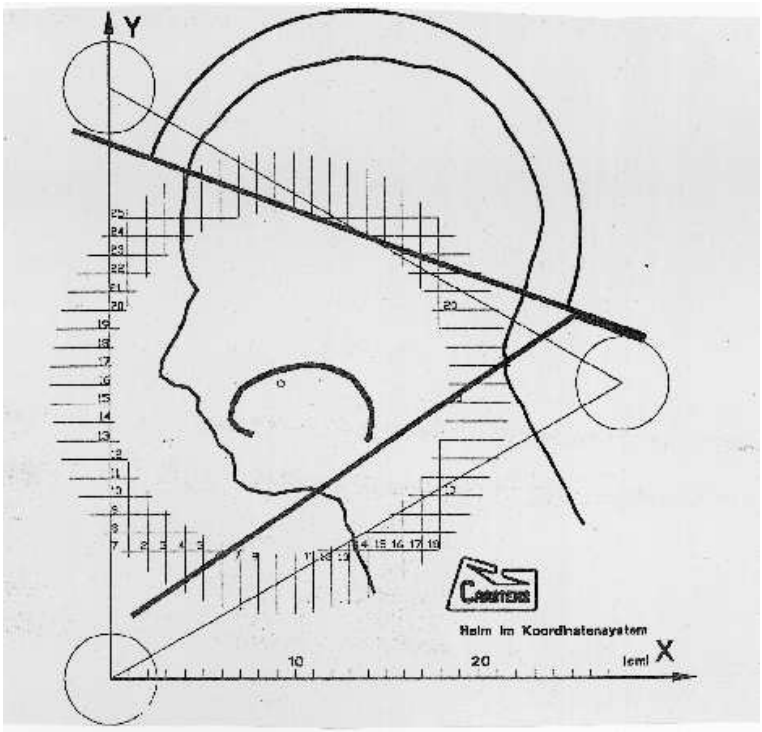


Abbildung 5.3: Schematischer Versuchsaufbau der Elektromagnetischen Artikulographie. Die drei Kreise an den Ecken des Dreiecks symbolisieren die Position der Erregerspulen. In der Mitte ist die Kontur der Zunge der Versuchsperson angedeutet. Meßspulen werden i.a. auf der Zunge, den Lippen und Zähnen plaziert (Abdruck mit freundlicher Genehmigung der Fa. Carstens Medizinelektronik)

nen in einer Ebene im Raum[20], welche auf die Artikulationsorgane der Versuchsperson aufgeklebt werden. Die Positionsbestimmung erfolgt durch Auswertung der magnetischen Induktion von drei niederfrequenten Feldern unterschiedlicher Frequenz in den Meßspulen. Während der Messung ausgeführte Bewegungen der Artikulationsorgane können auf diese Weise zumindest punktweise sehr genau bestimmt werden (s. Abb. 5.3). Synchron zu den Bewegungsparametern der Meßspulen wird das Schalldrucksignal der Sprache der Versuchsperson ausgezeichnet. Auf diese Weise ist sehr leicht möglich, mit Hilfe des MAUS-Systems bestimmte Lautkombinationen in den multidimensionalen Bewegungsdaten zu lokalisieren. Z.B. kann untersucht werden, wie sich die Artikulatoren bei der Produktion eines isolierten Lautes oder einer isolierten Silbe verhalten, und diese Daten können dann sofort mit Daten desselben Lautes oder derselben Silbe in verschiedenen Kontexten, bei verschiedener Sprechgeschwindigkeit, etc verglichen werden.

[20]Meistens der sagitale Schnitt durch den Kopf der Versuchsperson: derzeit wird am Institut für Phonetik und Sprachliche Kommunikation ein EMMA-Verfahren zur dreidimensionalen Artikulographie entwickelt.

Kapitel 6

Rückblick und Ausblick

In den beiden vorangegangenen Kapitel wurde sehr detailiert eine praktisch anwendbare Methode beschrieben, welche es ermöglicht, die Beziehung zwischen Sprachsymbol und Sprachsignal anhand von tatsächlichen gemessenen Ereignissen und ohne den subjektiven Einfluß einer Person herzustellen. Ich möchte nun zum Abschluß noch ein paar Gedanken zum größeren Rahmen skizzieren, in welchem die MAUS-Methode anzusiedeln ist, und schließlich noch einen kurzen Ausblick geben, wie sich die wissenschaftliche Arbeit mit der MAUS-Methode weiterentwickeln und auf welches Ziel sie, möglicherweise erst in ferner Zukunft, hinsteuern wird.

Von meiner Ausbildung her habe ich zu allererst ein Ingenieursstudium absolviert und bin erst später, über die praktischen Aspekte der Sprachwissenschaften, wie sie bei Spracherkennung und Sprachsynthese unverzichtbar sind, dazu gekommen, mich als Außenseiter mit den Inhalten von Fächern wie Linguistik, Phonologie und Phonetik zu beschäftigen. Dadurch habe ich möglicherweise eine etwas andere Herangehensweise an klassische Probleme der Sprachwissenschaft als jemand, der eine geisteswissenschaftliche Ausbildung genossen hat. Nach der langen wissenschaftichen Arbeit mit der MAUS-Methode komme ich zur Überzeugung, daß die strikte Trennung der Welt der phonologischen Symbole von der Welt der phonetischen Signale, wie sie Trubetzkoy 1938 in seinem Buch gefordert hat, sich heutzutage nicht mehr aufrechterhalten läßt. Eine phonologische Theorie ohne die Berücksichtigung der Realität kann es genauso wenig geben wie eine Phonetik ohne die Berücksichtigung der Sprachstrukturen. Alles, was ich in dieser vorliegenden Arbeit an technischen Methoden und Verfahren beschrieben habe, bestätigt dies.

Man nehme z.B. die in einer Fußnote in Abschnitt 3 erwähnte Tatsache, daß automatische Spracherkennung sich nur durch die Kombination von akustischem und

sprachlichem Wissen, z.B. in der Kombination von *Hidden Markov Modellen* und Wortfolgestatistik (*Language Models*), erfolgreich anwenden läßt. Hier haben wir genau die Situation, wie Vennemann sie (im übertragenen Sinne) anspricht, wenn er von *Realisations-Phonologie* spricht: die akustisch-phonetische Modellierung für sich alleine genommen ist nutzlos, dient nicht einmal erfolgreich zur phonetischen Analyse, wie wir in Abschnitt 3.2.2 gesehen haben. Genauso wenig kann uns das reine Wissen über die Syntax bzw. Anordnung der Wörter zu einer funktionierenden, Sprache erkennenden Maschine verhelfen. Erst die – zugegebenerweise nicht ganz triviale, aber elegante – Kombination beider Wissensquellen über eine statistische Optimierungsformel hat dazu geführt, daß man Diktiersysteme inzwischen kaufen kann. Allerdings würde ich noch einen Schritt weiter gehen und sagen: *tres faciunt collegium!*
Um wirklich zu Theorien zu kommen, die der Realität von sprachlicher Kommunikation genügen, muß noch ein dritter ins Boot kommen, nämlich die Signalverarbeitung.

Als ich 1995 von der Technischen Universität München an das Institut für Phonetik und Sprachliche Kommunikation der Ludwig-Maximilians-Universität München wechselte und dort eine neue wissenschaftliche Arbeit begann, war mein Anspruch der, das Wissen der sprachwissenschaftlichen Fächer für die technische Verbesserung von Verfahren wie Spracherkennung, Sprachsynthese und Sprecherverifikation nutzbar zu machen. Ich hoffe, daß mir dies zumindest teilweise gelungen ist. Noch wichtiger war für mich jedoch die Erkenntnis, daß der umgekehrte Anspruch die gleiche, wenn nicht sogar größere Bedeutung hat, nämlich die technischen Methoden der digitalen Signalverarbeitung für die Sprach- und Humanwissenschaften zu erschließen. Die MAUS-Methode ist letzendlich nichts anderes als eine Exemplifizierung dieses Prinzips: Mit technischen Methoden, die in den ingenieurwissenschaftlichen Disziplinen für praktische Anwendungen entwickelt wurden, betreiben wir im IPSK heute Grundlagenforschung, deren Ergebnisse wiederum auf der Anwendungsseite von den Ingenieurwissenschaften nutzbringend eingesetzt werden können.

Schon vor 41 Jahren hat Meyer-Eppler im Vorwort seines Hauptwerkes 'Grundlagen und Anwendungen der Informationstheorie' ([30]) geschrieben:

> "Zentrales Anliegen aller Betrachtungen ist die *menschliche Kommunikationskette* und der in ihr stattfindende *Zeichenverkehr*, der von *Signalen* getragen wird, die den Sinnesorganen zugänglich sind. Die meßbaren Eigenschaften dieser Signale bilden die Grundlage für alle weiteren Untersuchungen, wie etwa für die Frage nach den zur Signalübermittlung

> geeigneten Übertragungssystemen, die Statistik der hierbei verwendeten stereotypen Signalformen und den Einfluß von Störungen auf die Signalübermittlung sowie die mögliche Sicherung gegen Übertragungsfehler."

Nachdem wir selbst jetzt sechs Jahre an der Nahtstelle zwischen Geistes- und Naturwissenschaften gearbeitet haben, wird uns klar, daß wir die MAUS-Methode genau im Sinne Meyer-Epplers entwickelt haben, um anhand der physikalisch meßbaren Wirklichkeit *und* der Strukturen, die uns nur die sprachwissenschaftliche Analyse liefern kann, zu verstehen, wie sprachliche Kommunikation tatsächlich funktioniert.

Die weitere Entwicklung und Anwendung der MAUS-Methode, wird – ganz abgesehen von der praktischen Nützlichkeit für die wissenschaftliche Arbeit, wie in Abschnitt 5 beispielhaft skizziert – letzendlich auf eine empirisch abgesicherte Beschreibung der vollständigen sprachlichen Kommunikationskette führen. Diese Beschreibung wird phonologische und phonetische Teile haben, die sich ergänzen müssen – ergänzen in dem Sinne, daß damit die Realität so genau nachmodelliert wird, daß auch der dritte im Boot, der mit der technischen Verarbeitung von Sprache befaßte Ingenieur, davon profitieren kann. Von diesem hohen Ziel sind wir im Moment noch weit entfernt. Die nächsten Schritte, die unsere wissenschaftliche Arbeit mit dem MAUS-System vorsieht, nehmen sich vergleichsweise bescheiden aus.

Zur Zeit sind Arbeiten im Gange, das MAUS-System auf andere Sprachen, im wesentlichen Englisch und Japanisch, zu portieren ([4]). Von dieser Arbeit erhoffen wir uns die Beantwortung der Frage, ob die MAUS-Methode auch auf andere Sprach-Systeme anwendbar ist, oder ob unser Erfolg etwa auch darauf zurückzuführen ist, daß speziell das Deutsche sich für diese Art der Analyse besonders gut eignet.

In Verbindung mit diesen Arbeiten kommt eine neue Variante des MAUS-Systems zum Einsatz, welche zur Ermittlung des statistischen Aussprachemodells mit weniger manuell bearbeitetem Datenmaterial auskommt als bisher. Wenn diese Variante von MAUS sich als erfolgreich herausstellen sollte, wird es sehr viel leichter möglich sein, die MAUS-Methode auch für andere Sprachen zur Anwendung zu bringen.

Die notwendige Nachbesserung der Segmentierung wurde bereits in Abschnitt 4 angesprochen. Wir planen derzeit eine Lehrveranstaltung (Praktikum), in deren Rahmen solche Methoden mit dem mathematischen Programmpaket *MatLab* erarbeitet werden sollen. Die erfolgreichsten dieser Methoden werden in das MAUS-System einfließen. Auf diese Weise hoffen wir, bald sehr große Mengen an zuverlässig segmentiertem Material für die wissenschaftliche Arbeit zur Verfügung stellen zu können.

Als letzter Punkt sei die enge Kooperation mit dem *Bayerischen Archiv für Sprachsignale* (BAS) genannt. Ich selber habe in den letzten 5 Jahren in Personalunion diese Initiative an der Ludwig-Maximilians-Universität München geleitet und versucht, abgesehen von der reinen Bereitstellung und Produktion von Sprachressourcen auch die MAUS-Methode in die Aktivitäten des BAS einzubinden. Derzeit sind fast alle im BAS verfügbaren Sprachkorpora mit dem MAUS-System bearbeitet worden, und die Ergebnisse dieser Analysen werden der wissenschaftlichen Öffentlichkeit kostenlos über das Internet zur Verfügung gestellt. Im Gegenzug hat das BAS mit Einnahmen aus dem Vertrieb der Sprachressourcen, also der eigentlichen Signale, über die *European Language Resources Association* (ELRA) an z.T. auch industrielle Anwender die wissenschaftliche Arbeit am MAUS-System unterstützt. Soweit es die Umstände erlauben, hoffen wir, daß sich diese fruchtbare Symbiose auch in Zukunft fortsetzen lassen wird.

Literaturverzeichnis

[1] X. Aubert and C. Dugast. Improved acoustic-phonetic modelling in Philips' dictation system by handling liasons and multiple pronunciations. In *Proceedings of the EUROSPEECH*, pages 767–770, Madrid, Spain, 1995.

[2] M. Beham. *Merkmalsextraktion und Regelgewinnung für die automatische Spracherkennung*. PhD thesis, Technische Universität München, 1995.

[3] N. Beringer and F. Schiel. Independent automatic segmentation of speech by pronunciation modeling. In *Proceedings of the International Conference of Phonetic Sciences 1999*, pages 1653–1656, San Francisco, USA, 1999.

[4] N. Beringer and F. Schiel. The quality of multilingual automatic segmentation using German MAUS. In *Proceedings of ICSLP*, pages IV, 728–731, Bejing, China, 2000.

[5] N. Beringer, F. Schiel, and P. Regel-Brietzmann. German regional variants - a problem for automatic speech recognition? In *Proceedings of the International Conference on Spoken Language Processing*, pages 85 – 88, Sydney, Australia, 1998.

[6] K. Bühler. *Sprachtheorie*. Gustav Fischer Verlag, Stuttgart, 1965.

[7] S. Burger. Transliterations-lexikon. Verbmobil Techdok 36, Ludwig-Maximilians-Universität München, Institut für Phonetik, München, 1995.

[8] W.N. Campbell and A.W. Black. Chatr: a multilingual speech re-sequencing synthesis system. SP96-7 Tech Rept IEICE 7, ATR, Kyoto, Japan, 1996.

[9] F. Caroli, R. Nübel, B. Ripplinger, and J. Schütz. Transfer in VERBMOBIL. Verbmobil Report 11, 1994.

[10] B. Eisen and H.G. Tillmann. Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases. In *Proceedings of ICSLP*, pages 872–874, Banff, Canada, 1992.

[11] G. Fant. *Acoustic theory of speech production.* 's-Gravenhage, 1960.

[12] J.L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer Verlag, Berlin - Heidelberg - New York, 1972.

[13] W. Heß, K. J. Kohler, and H. G. Tillmann. The PHONDAT-VERBMOBIL speech corpus. In *Proceedings of the EUROSPEECH*, pages 863–866, Madrid, Spain, 1995.

[14] P. Hoole. Theoretische und methodische Grundlagen der Artikulationsanalyse in der experimentellen Phonetik. In *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM), Volume 34*, pages 3–173. Ludwig-Maximilians-Universität München, 1996.

[15] P. Hoole. Electromagnetic articulography. In W.J. Hardcastle and N. Hewlett, editors, *Coarticulation*, pages 260–269. Cambridge University Press, 1999.

[16] H. Iwamida, S. Katagiri, E. McDermott, and Y. Tohkura. A hybrid speech recognition system using HMMs with an LVQ-trained codebook. In *Proceedings of the ICASSP*, pages 489–492, 1990.

[17] U. Jekosch and T. Becker. Maschinelle Generierung von Aussprachevarianten: Perspektiven für Sprachsynthese- und Spracherkennungssysteme. *Informationstechnik it*, 6:400, 1999.

[18] B.H. Juang and L.R. Rabiner. The segmental k-means algorithm for estimating parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38:1639 – 1641, 1990.

[19] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on ASSP*, 35(3):400–401, 1987.

[20] A. Kipp. *Automatische Segmentierung und Ettikettierung von Spontansprache.* PhD thesis, Technische Universität München, 1998.

[21] A. Kipp, M.-B. Wesenick, and F. Schiel. Automatic detection and segmentation of pronunciation variants in german speech corpora. In *Proceedings of the ICSLP*, pages 106–109, Philadelphia,USA, 1996.

[22] A. Kipp, M.-B. Wesenick, and F. Schiel. Pronunciation modeling applied to automatic segmentation of spontaneous speech. In *Proceedings of the EUROSPEECH*, pages 1023–1026, Rhode, Greece, 1997.

[23] D.H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America (JASA)*, 37:971–995, 1980.

[24] K. J. Kohler. *Einführung in die Phonetik des Deutschen.* E. Schmidt, Berlin, 1995.

[25] K. J. Kohler. Labelled data bank of spoken standard german: The Kiel corpus of read/spontaneous speech. In *Proceedings of the ICSLP*, Philadelphia, USA, 1996.

[26] M. Libossek and F. Schiel. Syllable-based text-to-phoneme conversion for German. In *Proceedings of ICSLP*, pages 283–286, Bejing, China, 2000.

[27] K. Machelett. Das Lesen von Sonagrammen in der Phonetik. Magisterarbeit, Ludwig-Maximilians-Universität München, 1994.

[28] H. Marko. *Methoden der Systemtheorie. Die Spektraltransformationen und ihre Anwendungen.* Springer-Verlag, Berlin, 1997.

[29] EU-Project 'Speech Assessment Methods'. www.phon.ucl.ac.uk/home/sampa/home.htm.

[30] W. Meyer-Eppler. *Grundlagen und Anwendungen der Informationstheorie.* Springer Verlag, Berlin - Heidelberg - New York, 1969.

[31] A. Nadas. On turing's formula for word probabilities. *IEEE Transactions on ASSP*, 33(6):1414–1416, 1985.

[32] H. Ney and U. Essen. Estimating 'small' probabilities by leaving one out. In *Proceedings of the EUROSPEECH*, pages 2239–2242, Berlin, 1993.

[33] The Principles of the International Phonetic Association. International Phonetic Association, University College London, 1949 (reprint of 1984).

[34] T. Portele, B. Steffan, R. Preuss, W.F. Sendlmeier, and W. Hess. Hadifix – a speech synthesis system for German. In *Proceedings of the ICSLP*, pages 1227–1230, Banff, Canada, 1992.

[35] S. Rapp. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models.

[36] F. Reinhardt and H. Soeder. *dtv-Atlas zur Mathematik – Band I: Grundlagen, Algebra, Geometrie*. Deutscher Taschenbuch Verlag, München, 1994.

[37] M. Riley. A statistical model for generating pronunciation networks. In *Proceedings of the ICASSP*, pages 737–740, 1991.

[38] M. Riley, F. Pereira, and M. Mohri. Transducer composition for context-dependent network expansion. In *Proceedings of the EUROSPEECH*, pages 1427 – 1430, Rhode, Greece, 1997.

[39] G. Ruske. *Automatische Spracherkennung*. R. Oldenbourg Verlag München Wien, 1998.

[40] F. Schiel. Probabilistic analysis of pronunciation with MAUS. *The ELRA Newsletter*, 4:6–9, 1997.

[41] F. Schiel. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the International Conference of Phonetic Sciences 1999*, pages 607–610, San Francisco, USA, 1999.

[42] F. Schiel. *Einführung in die automatische Spracherkennung*. Skriptum zur Vorlesung, Ludwig-Maximilians-Universität München, 1999.

[43] F. Schiel, A. Kipp, and H.G. Tillmann. Statistical modeling of pronunciation: It's not the model, it's the data. In *Proceedings of the ESCA Tutorial and Research Workshop on 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, pages 31–36, Kerkrade/Netherlands, 1998.

[44] F. Schiel and F. Wolfertstetter. Regelbasierte Erzeugung von robusten Aussprachemodellen und deren Darstellung im Silbenraster. *Studientexte zur Sprachkommunikation*, 8:173–182, 1991.

[45] O. Schmidbauer. *Ein System zur Lauterkennung in fliessender Sprache auf der Basis artikulatorischer Merkmale*. PhD thesis, Technische Universität München, 1989.

[46] T. Sloboda and A. Waibel. Dictionary learning for spontaneous speech. In *Proceedings of the ICSLP*, Philadelphia, USA, 1996.

[47] K.N. Stevens, S.Y. Manuel, S. Shattuck-Hufnagel, and S. Liu. Implementation of a model for lexical access based on features. In *Proceedings of the ICSLP*, pages 499–502, Banff, Canada, 1992.

[48] H.G. Tillmann and Ph. Mansell. *Phonetik: Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß.* Klett-Cotta, Stuttgart, 1980.

[49] N. S. Trubetzkoy. *Grundzüge der Phonologie, 2. Auflage.* Vandenhoeck und Ruprecht, Göttingen, 1958.

[50] Th. Vennemann. Phonology as non-functional non-phonetics. In W. Dreßler, editor, *Phonologica 1980.* 1980.

[51] Th. Vennemann and J. Jacobs. *Sprache und Grammatik.* Wissenschaftliche Buchgesellschaft Darmstadt, 1982.

[52] S. Watanabe. *Knowing and Guessing.* John Wiley & Sons, New York, 1969.

[53] M.-B. Wesenick. Entwurf eines unterspezifizierenden Regelsystems der Aussprache des Deutschen als Basis für empirische Untersuchungen. Magisterarbeit, Ludwig-Maximilians-Universität München, 1994.

[54] M.-B. Wesenick and A. Kipp. Estimating the quality of phonetic transcriptions and segmentations of speech signals. In *Proceedings of the ICSLP*, pages 129–132, Philadelphia, USA, 1996.

[55] M.-B. Wesenick and F. Schiel. Applying speech verification to a large data base of German to obtain a statistical survey about rules of pronunciation. In *Proceedings of ICSLP*, pages 279–282, Yokohama,Japan, 1994.

[56] S. Young. *The HTK Book.* Cambridge University, 1995.

[57] E. Zwicker. *Psychoakustik.* Springer-Verlag, Berlin u.a., 1982.

Anhang A

MAUS: Technik

Die folgenden Abschnitte sind der Dissertation von Dr.-Ing. Andreas Kipp ([20], Abschnitte 6 und 7) entnommen, welcher in den Jahren 1995 bis 1998 unter meiner Betreuung maßgeblich an der Entwicklung der MAUS-Technik beteiligt war[1]. Die Dissertation wurde am 30.03.1998 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 14.09.1998 angenommen.

Der erste Abschnitt 'Modellierung der Aussprache' enthält die detailierte und mathematisch ausformulierte Beschreibung der Modellierung von Aussprache mit Hilfe eines gerichteten, azyklischen Graphen. Der zweite Abschnitt 'Aufbau des MAUS Systems' beschreibt die prinzipielle Realisierung des MAUS-Systems, ohne auf die zahlreichen programmiertechnischen Probleme bei der Implementierung in C++ näher einzugehen.

Der technisch interessierte Leser sei zusätzlich auf die bereits veröffentlichte Literatur zu diesem Thema verwiesen ([20, 21, 22, 43, 3, 55, 41]

A.1 Modellierung der Aussprache

Die Modellierung der Aussprache gewinnt immer größere Bedeutung in der automatischen Spracherkennung, besonders weil die Analyse von Spontansprache mittlerweile zum Standard geworden ist. Spontan geäußerte Sprache ist in größerem Ausmaße als z.B. gelesene Sprache Reduktionserscheinungen und Lautverschiebungen

[1] Textbezüge des Originals, soweit sie in dieser Arbeit nicht vorkommen, sind der besseren Lesbarkeit wegen nicht enthalten

unterworfen, die eine ökonomischere Artikulation erlauben. Begrenzt wird die auftretende Verschleifung und Koartikulation durch die Anforderung, daß die Äußerung weiterhin verständlich bleiben muß. Deshalb ist die Aussprachevariation gewissen Regeln unterworfen, die als phonetisch-phonologische Ausspracheregeln formuliert werden können. Ausgangspunkt ist dabei die abstrakte Referenzaussprache, die aus den kanonische Formen der Einzelwörter einer Äußerung gebildet wird.

Phonetisch-phonologische Ausspracheregeln, die auf eine Referenzaussprache angewendet werden, können die Grundlage einer regelbasierten Aussprachemodellierung sein. Dieses Vorgehen ist im Abschnitt A.1.2 beschrieben.

Manuell erstellte Transkriptionen drücken implizit ebenfalls Ausspracheregeln aus. Die Transkriptionen geben hierbei die Realisierung einer Äußerung wieder, deren orthographischer Gehalt bekannt ist. Aus dem orthographischen Gehalt kann die Referenzaussprache konstruiert werden. Setzt man Referenzaussprache und tatsächliche Realisierung vieler Äußerungen in Beziehung, kann man daraus mit statistischen Methoden Wahrscheinlichkeiten für Aussprachevarianten schätzen. Dieses Vorgehen ist im Abschnitt A.1.3 beschrieben.

Abschnitt A.1.1 stellt dar, wie die Information über Aussprachevariabilität – entweder aus Expertenwissen heraus postuliert oder durch die Beobachtung gewonnen – in Form von Symbolmanipulationen formuliert werden kann. Dazu müssen den Betrachtungen aber einige grundsätzliche Überlegungen zur symbolischen Repräsentation von Referenztranskriptionen und Aussprachevarianten vorausgeschickt werden.

Grundsätzlich soll es möglich sein, sowohl die Referenztranskription, die im folgenden mit $\boldsymbol{c}$ bezeichnet wird, als auch eine beliebige tatsächliche Realisierung, im folgenden mit $\boldsymbol{r}$ notiert, mit dem gleichen Lautinventar darzustellen und zwar in Form von Strings oder Symbolfolgen aus genau diesem Inventar. Im Rahmen dieser Arbeit wird deshalb ein phonemisches Lautinventar, nämlich das bereits erwähnte SAM Phonetic Alphabet für Deutsch, verwendet. Welches Alphabet aber konkret verwendet wird, spielt keine Rolle, solange die obige Bedingung erfüllt ist. Das Alphabet wird daher im folgenden auch nicht näher spezifiziert, sondern nur mit dem Buchstaben $\mathbf{S}$ bezeichnet und angenommen, daß es sich bei $\mathbf{S}$ um eine endliche Menge von Lautsymbolen handelt. Die Menge aller möglichen Strings über $\mathbf{S}$ wird mit dem entsprechenden griechischen Großbuchstaben $\mathbf{\Sigma}$ notiert. Für die Anwendung werden oft nur endliche Untermengen von $\mathbf{\Sigma}$ betrachtet. Diese werden mit $\mathbf{\Theta}$ und einem tiefgestellten Index bezeichnet.

Formal soll das Aussprachemodell die Wahrscheinlichkeit wiedergeben, daß eine Äußerung mit einer Referenzaussprache $\boldsymbol{c}$ in Wirklichkeit als $\boldsymbol{r}$ realisiert wird. Diese Wahrscheinlichkeit wird mit $P(\boldsymbol{r}|\boldsymbol{c})$ notiert. Dabei wird nur eine endliche Anzahl von Realisierungen $\boldsymbol{r}$ zu $\boldsymbol{c}$ geben. Für diese gilt dann $P(\boldsymbol{r}|\boldsymbol{c}) \neq 0$.

Die Wahrscheinlichkeiten $P(\boldsymbol{r}|\boldsymbol{c})$ werden entweder als gleichverteilt über alle möglichen Realisierungen angenommen oder aus den Wahrscheinlichkeiten für die Variation kleinerer Einheiten konstruiert, d.h aus den Wahrscheinlichkeiten für Mikro-Aussprachevarianten, die im nächsten Abschnitt eingeführt werden.

A.1.1 Mikro-Aussprachevarianten

Mit dem Konzept von Alphabeten und Strings kann Aussprachevariation in Form von Symbolmanipulationen formuliert werden, die eine Referenztranskription in eine Aussprachevariante transformieren. Eine Aussprachevariante entsteht dadurch, daß regelhaft – eventuell abhängig von den umgebenden Symbolen – Teilstrings der Referenztranskription ersetzt werden. Jede einzelne solche Ersetzung repräsentiert eine Verschleifungs- oder Reduktionserscheinung, und spezifiziert unter Umständen eine bestimmten Kontext.

Da jeweils nur Teilstrings einer Referenzaussprache betroffen sind, wird hier der Begriff der *Mikro-Aussprachevariante* eingeführt. Eine Mikro-Aussprachevariante $\mathbf{u}$ besteht aus einer Anwendbarkeitsbedingung $\mathbf{m}$ und einem Ersetzungsphonemstring $\boldsymbol{b}$. Anwendbarkeitsbedingung und Mikro-Aussprachevariante sind wie folgt definiert:

$$\begin{aligned} \mathbf{m} &= (\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}), \quad \mathbf{m} \in \mathbf{M} \\ \mathbf{M} &\subseteq \boldsymbol{\Sigma} \times \boldsymbol{\Sigma} \times \boldsymbol{\Sigma} \end{aligned} \tag{A.1}$$

$$\begin{aligned} \mathbf{u} &= (\boldsymbol{b}, \mathbf{m}), \quad \mathbf{u} \in \mathbf{U} \\ \mathbf{U} &\subseteq \boldsymbol{\Sigma} \times \boldsymbol{\Sigma} \times \boldsymbol{\Sigma} \times \boldsymbol{\Sigma} \end{aligned} \tag{A.2}$$

Dabei ist $\boldsymbol{x}$ der Prä- bzw. $\boldsymbol{y}$ Postkontext (der in $\boldsymbol{r}$ jeweils unverändert bleibt) und $\boldsymbol{a}$ der Phonemstring der durch $\boldsymbol{b}$ ersetzt wird, vorausgesetzt die Referenztranskription enthält eben den Phonemstring $\boldsymbol{xay}$.[2]

Beispiel: Referenzaussprache $\boldsymbol{c}$ = [Q a b @ n t], Mikro-Aussprachevariante $\mathbf{m}$ = ([m], [b], [@n], [t]). Erzeugte Variante $\boldsymbol{r}$ = [Q a b m t].

[2]Die Schreibweise $\boldsymbol{xay}$ bezeichnet die Konkatenation von Strings

Die Strings $\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}$ und $\boldsymbol{b}$ können auch leer sein. Dabei bedeutet

- $\boldsymbol{x} = \mathbf{0}$, daß kein linker Kontext spezifiziert, d.h. die Variante ist in beliebigem Linkskontext anwendbar,
- $\boldsymbol{y} = \mathbf{0}$, daß kein rechter Kontext spezifiziert, d.h. die Variante ist in beliebigem Rechtskontext anwendbar,
- $\boldsymbol{a} = \mathbf{0}$, daß eine Einfügung von $\boldsymbol{b}$ (wobei gelten muß $\boldsymbol{b} \neq \mathbf{0}$) stattfindet,
- $\boldsymbol{b} = \mathbf{0}$, daß eine Elision von $\boldsymbol{a}$ (wobei gelten muß $\boldsymbol{a} \neq \mathbf{0}$) stattfindet.

Für die Generierung der gesamten Menge $\mathbf{U}$ von Mikro-Aussprachevarianten sind in den Abschnitten A.1.2 und A.1.3 zwei Ansätze diskutiert. Aus der gesamten Menge der Mikro-Aussprachevarianten ist für die Referenztranskription einer bestimmten Äußerung dann eine Untermenge anwendbar.

Die Beziehung zwischen der Referenztranskription $\boldsymbol{r}$ und einer bestimmten tatsächlichen Realisierung oder auch *Makro-Aussprachevariante* $\boldsymbol{c}$ kann nun mit Hilfe der anwendbaren Mikro-Aussprachevarianten hergestellt werden: Die Makro-Variante entsteht durch Anwendung einer Untermenge der auf $\boldsymbol{c}$ anwendbaren Mikro-Variantenmenge. Umgekehrt können mit einer Menge von anwendbaren Mikro-Varianten mögliche Makro-Varianten zu einer Referenztranskription hypothetisiert werden. Genau das ist der Ansatzpunkt zur Verwendung von Mikro-Aussprachevarianten im Rahmen des hier diskutierten Systems zur automatischen Segmentierung und Etikettierung: Zur Referenztranskription einer sprachlichen Äußerung werden mögliche Aussprachevarianten hypothetisiert und durch eine statistische Lautmodellierung die am besten zur akustischen Repräsentation der Äußerung passende Hypothese ausgewählt.

Mikro-Aussprachevarianten zu einer Referenzaussprache

Zu einer gegebenen Referenzaussprache $\boldsymbol{c}$ existiert eine Menge von anwendbaren Mikro-Aussprachevarianten $\mathbf{u}$. Diese sind dadurch bestimmt, daß ihre Anwendbarkeitsbedingungen $\mathbf{m} = (\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ auf $\boldsymbol{c}$ zutreffen, d.h $\boldsymbol{c}$ kann als $\boldsymbol{c} = \boldsymbol{sxayt}$, mit $\boldsymbol{s} \in \boldsymbol{\Sigma}$ als ein Präfix von $\boldsymbol{c}$ und $\boldsymbol{t} \in \boldsymbol{\Sigma}$ als ein Suffix von $\boldsymbol{c}$ geschrieben werden. Da ein und dieselbe Anwendbarkeitsbedingung an mehreren Stellen zutreffen kann, ist eine konkrete Mikro-Aussprachevariantenanwendung $\mathbf{q}$ durch die Variante selbst und die Position i des Strings, an der $\mathbf{m}$ paßt, also durch ein Tupel $(i, \mathbf{u})$ gegeben. Schreibt

man $\boldsymbol{c}$ als Konkatenierung seiner Phonemsymbole, also $\boldsymbol{c} = \gamma_0\gamma_1 \ldots \gamma_{N-1}$, so kann eine Menge $\mathbf{Q}^{(\boldsymbol{c})}$ wie folgt definiert werden[3]:

$$\begin{aligned}\mathbf{Q}^{(\boldsymbol{c})} = \{\ (i, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})\ \big|\ (\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) \in \mathbf{U}\ \wedge\ \boldsymbol{x} = \gamma_{i-|\boldsymbol{x}|} \ldots \gamma_{i-1} \wedge \\ \boldsymbol{a} = \gamma_i \ldots \gamma_{i+|\boldsymbol{a}|-1}\ \wedge\ \boldsymbol{y} = \gamma_{i+|\boldsymbol{a}|+1} \ldots \gamma_{i+|\boldsymbol{a}|+|\boldsymbol{y}|-1}\ \}\end{aligned} \tag{A.3}$$

Diese Menge kann als Relation $i\mathbf{Q}^{(\boldsymbol{c})}\mathbf{u}$ verstanden werden, die aussagt, daß die Mikro-Aussprachevariante $\mathbf{u}$ im Referenzaussprachestring $\boldsymbol{c}$ an der Stelle i anwendbar ist. Die Menge ist endlich und ihre Elemente können indiziert werden $\mathbf{Q}^{(\boldsymbol{c})} = \{\mathbf{q}_k | 0 \leq k < K\}$.

Beispiel: Referenzaussprache $\boldsymbol{c} = [\overset{0}{\mathrm{Q}}\,\overset{1}{\mathrm{a}}\ \underbrace{\overset{2}{\mathrm{b}}}_{\boldsymbol{x}}\ \underbrace{\overset{3}{@}\,\overset{4}{\mathrm{n}}}_{\boldsymbol{a}}\ \underbrace{\overset{5}{\mathrm{t}}}_{\boldsymbol{y}}]$

Mikro-Aussprachevariante $\mathbf{m} = ([\mathrm{m}], [\mathrm{b}], [@\mathrm{n}], [\mathrm{t}])$

Konkrete Mikro-Aussprachevariantenanwendung $\mathbf{q} = (3, [\mathrm{m}], [\mathrm{b}], [@\mathrm{n}], [\mathrm{t}])$

Makro-Aussprachevarianten zu einer Referenzaussprache

Eine Makro-Aussprachevariante $\boldsymbol{r}$ zur Referenzaussprache $\boldsymbol{c}$ ist nun durch eine Untermenge $\mathbf{G}_j^{(\boldsymbol{c})} \subseteq \mathbf{Q}^{(\boldsymbol{c})}$ und der Ersetzung eines jeden der Variation unterworfenen Phonemstrings $\boldsymbol{a}$ aus $\mathbf{q} = (i, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$, $\mathbf{q} \in \mathbf{G}_j^{(\boldsymbol{c})}$ mit dem Phonemstring $\boldsymbol{b}$ gegeben.

Die Menge $\mathbf{G}_j^{(\boldsymbol{c})}$ kann dabei keine beliebige Untermenge von $\mathbf{Q}^{(\boldsymbol{c})}$ sein: Wenn durch die Anwendung einer Mikro-Aussprachevariante $\mathbf{u}'$ die Anwendbarkeitsbedingung einer anderen Mikro-Aussprachevariante $\mathbf{u}''$ verletzt würde, können nicht beide gleichzeitig in einer Menge $\mathbf{G}_j^{(\boldsymbol{c})}$ enthalten sein. Die Menge der sich paarweise ausschließenden Mikro-Aussprachevarianten läßt sich mit der Ausschlußrelation $\mathbf{R}_\mathrm{e}$ formulieren.

Definiert man die Hilfsgrößen $n_\mathrm{a}(\mathbf{q}), n_\mathrm{s}(\mathbf{q}), n_\mathrm{r}(\mathbf{q}), n_\mathrm{e}(\mathbf{q})$, die die Position von Beginn und Ende des Prä- und des Postkontexts von $\mathbf{q}$ im Referenzaussprachestring $\boldsymbol{c}$ angeben, so kann $\mathbf{R}_\mathrm{e}$ mit Gleichung A.3 wie folgt beschrieben werden:

$$n_\mathrm{s}(\mathbf{q}') < n_e(\mathbf{q}'') \quad \wedge \quad n_\mathrm{s}(\mathbf{q}') \geq n_\mathrm{a}(\mathbf{q}'') \quad \Rightarrow \quad \mathbf{q}'\mathbf{R}_\mathrm{e}\mathbf{q}'' \tag{A.4}$$

$$\mathbf{q}'\mathbf{R}_\mathrm{e}\mathbf{q}'' \quad \Rightarrow \quad \mathbf{q}''\mathbf{R}_\mathrm{e}\mathbf{q}' \tag{A.5}$$

[3]Die Notation $|\boldsymbol{s}|$ bezeichnet die Länge eines Strings.

mit

$$n_{\mathrm{s}}((i,\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})) = i - |\boldsymbol{x}| \tag{A.6}$$

$$n_{\mathrm{a}}((i,\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})) = i - 1 \tag{A.7}$$

$$n_{\mathrm{r}}((i,\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})) = i + |\boldsymbol{a}| \tag{A.8}$$

$$n_{\mathrm{e}}((i,\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})) = i + |\boldsymbol{a}| + |\boldsymbol{y}| - 1 \tag{A.9}$$

Beispiel: Referenzaussprache $\boldsymbol{c}$ = [Q a b @ n t], konkrete Mikro-Aussprachenanwendungen $\mathbf{q}' = (3, [\mathrm{b}], [@\mathrm{n}], [\mathrm{t}])$ und $\mathbf{q}'' = (3, [\mathrm{b}], \mathbf{0}, [\mathrm{n}])$ (Elision von [@]). $\mathbf{q}'$ und $\mathbf{q}''$ überlappen sich und können nicht gleichzeitig angewendet werden.

Für jede Menge $\mathbf{G}_j^{(\boldsymbol{c})}$, $j = 0 \ldots J-1$ muß stets gelten[4]:

$$\bigwedge_{\mathbf{q}' \in \mathbf{G}_j^{(\boldsymbol{c})}} \bigwedge_{\mathbf{q}'' \in \mathbf{G}_j^{(\boldsymbol{c})} \setminus \mathbf{q}'} (\mathbf{q}', \mathbf{q}'') \notin \mathbf{R}_{\mathrm{e}} \tag{A.10}$$

Jede Menge $\mathbf{G}_j^{(\boldsymbol{c})}$, für die Gl. A.10 erfüllt ist, erzeugt eine Makro-Aussprachevariante $\boldsymbol{r}_j$. Ist die Menge $\mathbf{G}_j^{(\boldsymbol{c})}$ die leere Menge, so wird die Referenzaussprache erzeugt. Im Prinzip können durch die Bestimmung aller möglichen $\mathbf{G}_j^{(\boldsymbol{c})}$ alle durch Mikro-Varianten zu einer Referenzaussprache generierbaren Makro-Aussprachevarianten angegeben werden. Darüber hinaus müssen sie – nach noch zu diskutierenden Verfahren – mit Wahrscheinlichkeiten $P(\boldsymbol{r}_j|\boldsymbol{c})$ gewichtet werden, so daß $\sum_{(j)} P(\boldsymbol{r}_j|\boldsymbol{c}) = 1$ gilt.

Mit der bereits früher beschriebenen Sprachlautmodellierung und den entsprechenden Suchverfahren kann die in Kapitel 3 formulierte Maximierungsaufgabe nach Gl 3.4 gelöst werden, indem für alle Segmentierungen $\{\boldsymbol{K}|\boldsymbol{s}(\boldsymbol{K}) = \boldsymbol{r}_j\ ,\ j \leq 0 < J\}$ der Term $P(\boldsymbol{r}_j|\boldsymbol{c})P(\boldsymbol{O}|\boldsymbol{K})$ berechnet wird. Die explizite Berechnung jedes dieser Terme beinhaltet jedoch etliche redundante Schritte, da viele $\boldsymbol{r}_i$ gleiche Substrings[5] enthalten.

Im folgenden Abschnitt ist daher eine wesentlich kompaktere Darstellung der zu einer Referenzaussprache möglichen Makro-Aussprachevarianten beschrieben, die die Lösung der Maximierungsaufgabe wesentlich vereinfacht.

---

[4] $\bigwedge_x$ bedeutet hierbei "für alle x gilt". Die Notationskonventionen folgen dem dtv-Atlas zur Mathematik [36].

[5] $\boldsymbol{a}$ ist ein Substring von $\boldsymbol{r}$, wenn gilt $\boldsymbol{r} = \boldsymbol{sat}$. $\boldsymbol{s}$ und $\boldsymbol{t}$ können dabei auch Strings der Länge 0 sein

Generierung eines Variantengraphen

Eine kompakte Darstellung aller zu einer Referenzaussprache $\boldsymbol{c}$ durch Mikro-Aussprachevarianten hypothetisierten Makro-Varianten $\boldsymbol{r}_j$ muß der Tatsache Rechnung tragen, daß sich viele der $\boldsymbol{r}_j$ sehr ähnlich sind, da sich die Mengen der angewendeten Mikro-Aussprachevarianten $\mathbf{G}_j^{(\boldsymbol{c})}$ überschneiden oder sich gar enthalten. Desweiteren soll die Struktur mit der auf HMMs erster Ordnung basierenden Sprachlautmodellierung kompatibel sein. Es bietet sich die Darstellung als endlicher Zustandsautomat erster Ordnung an.

Der Automat ist gekennzeichnet durch M diskrete Zustände. In jedem dieser Zustände wird entweder genau ein Symbol $\sigma = \rho(d_i)$ aus dem (phonetischen) Alphabet $\mathbf{S}$ emittiert oder kein Symbol (nichtemittierender Knoten).

Der Automat hat zwei ausgezeichnete Zustände, die kein Symbol emittieren: Einen Anfangszustand und einen Endzustand, die am Beginn bzw. am Ende jeder Zustandsfolge stehen müssen, sonst aber nicht auftreten dürfen. Der Anfangszustand hat den Index $i = 0$ und der Endzustand den Index $i_{\max}$.

Von jedem Zustand (mit Ausnahme des Endzustandes) kann ein Übergang zu bestimmten anderen Zuständen erfolgen. Diesen Zustandsübergängen sind Wahrscheinlichkeiten zugeordnet. Es werden nur Zustandsfolgen, die im Anfangszustand beginnen und im Endzustand enden, betrachtet.

Der endliche Automat kann durch einen gerichteten Graphen beschrieben werden (vgl. Abschnitt 3.3): Jeder Zustand entspricht einem Knoten des Graphen und jeder mögliche Zustandsübergang einer gerichteten Kante zwischen den entsprechenden Zustands-Knoten.

Wenn der den Automaten beschreibende Graph zyklenfrei ist, was im folgenden immer der Fall sein wird, kann nur eine endliche Menge von endlich langen Strings $\boldsymbol{r}_j$, $j = 0 \ldots J-1$ vom Automaten generiert werden. Im Gegensatz zu den HMMs, die ja in sehr ähnlicher Weise beschrieben sind, ist hier dann kein Selbstübergang möglich. Auch sind den Zustandsübergängen keine expliziten Zeitpunkte zugeordnet.

Es soll nun zur Aussprachemodellierung ein derartiger endlicher Automat erzeugt werden. Dieser sei beschrieben durch einen gerichteten, zyklenfreien Graphen $\mathcal{V}$, den sogenannten *Variantengraphen*. Der Graph ist durch eine Knotenmenge $\mathbf{D}$ und eine Kantenmenge $\mathbf{E}$ bestimmt, Die Kanten werden als Tupel von zwei Elementen der Knotenmenge notiert

Ein durch den Automaten generierter String $\boldsymbol{r}_j$ ist dann entweder die Referenzaus-

sprache $\boldsymbol{c}$ oder eine durch die Anwendung von Mikro-Aussprachevarianten erzeugte Makro-Variante.

Um $\boldsymbol{c} = \gamma_0\gamma_1 \ldots \gamma_{N-1}$ generieren zu können, enthält der Graph die Knotenmenge

$$\mathbf{D}_0 = \{h_i | 0 \leq i < N,\ \rho(h_i) = \gamma_i\} \tag{A.11}$$

Die Knoten der Menge $\mathbf{D}_0$ müssen so verbunden werden, daß die Referenzaussprache emittiert wird:

$$\mathbf{E}_0 = \left\{\ (h_{i-1}, h_i) \ \middle|\ 1 \leq i < N\ \right\} \tag{A.12}$$

Die weitere Struktur des Graphen ist intuitiv sofort klar: Für jede anwendbare Mikro-Aussprachevariante $\mathbf{q}_k \in \mathbf{Q}^{(\boldsymbol{c})}$ muß ein Alternativpfad angelegt werden, der an der entsprechenden Stelle vom Referenzaussprachepfad abzweigt und gegebenenfalls die Symbole der Variante emittiert. Dazu müssen Kanten und Knoten dem Graphen hinzugefügt werden. Im folgenden wird immer die Anwendung des k-ten Elements aus $\mathbf{Q}^{(\boldsymbol{c})}$ betrachtet, nämlich $\mathbf{q}_k = (i, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$. Die Knoten- und Kantenmenge nach der Anwendung der Mikro-Aussprachevarianten $\mathbf{q}_0, \mathbf{q}_1 \ldots \mathbf{q}_k$ sei $\mathbf{D}_{k+1}$ bzw. $\mathbf{E}_{k+1}$.

Wenn der Ersetzungsstring $\boldsymbol{b} = \beta_0 \ldots \beta_{L-1}$ eine Länge größer Null hat, werden jeweils Knoten $q_{k,l}$ mit der Eigenschaft $\rho(q_{k,l}) = \beta_l$ hinzugefügt. Diese Knoten sind doppelt indiziert (Index k von $\mathbf{q}_k$ und l der Position im String) und mit q (nicht fettgedruckt wie bei $\mathbf{q}_k \in \mathbf{Q}^{(\boldsymbol{c})}$) bezeichnet.

Es ist nun aber zu berücksichtigen, daß alle Elemente einer Mikro-Aussprachevariante $(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ mit Ausnahme von $\boldsymbol{a}$ auch der leere String $\mathbf{0}$ sein können. Daher müssen insgesamt acht Fälle unterschieden werden, die im folgenden jeweils mit einem Beispiel aufgeführt sind. Die Kreise in den Abbildungen repräsentieren die Knoten, also die Zustände. Die Symbole in den Kreisen werden im entsprechenden Zustand des Automaten emittiert. Nichtemittierende Zustände sind durch zwei konzentrische Kreise notiert. Die Verbindungen (Pfeile) stellen Kanten, also mögliche Zustandsübergänge dar.

Die Fälle, in denen die Kontexte $\boldsymbol{x}$ und $\boldsymbol{y}$ eine Länge größer Null haben, sind einfach zu behandeln. Wenn Mikro-Aussprachevarianten keinen linken oder rechten Kontext spezifizieren, werden nichtemittierende Knoten x_i zwischen zwei Knoten h_{i-1} und h_i eingefügt (Die x_i sind also indiziert nach dem Referenzaussprache-Knoten, auf den sie folgen).

Fall 1: Ersetzung in beidseitig spezifiziertem Kontext (Abb. A.1). $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \beta_0 \ldots \beta_{L-1} \wedge \boldsymbol{x} \neq \mathbf{0} \wedge \boldsymbol{y} \neq \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \cup \{q_{k,l} | 0 \leq l < L \wedge \rho(q_{k,l}) = \beta_l\} \tag{A.13}$$

$$\mathbf{E}_{k+1} = \mathbf{E}_k \cup \{(h_{i-1}, q_{k,0}), (q_{k,L-1}, h_{i+K})\} \cup \{(q_{k,j-1}, q_{k,j} | 1 < j < L\} \tag{A.14}$$

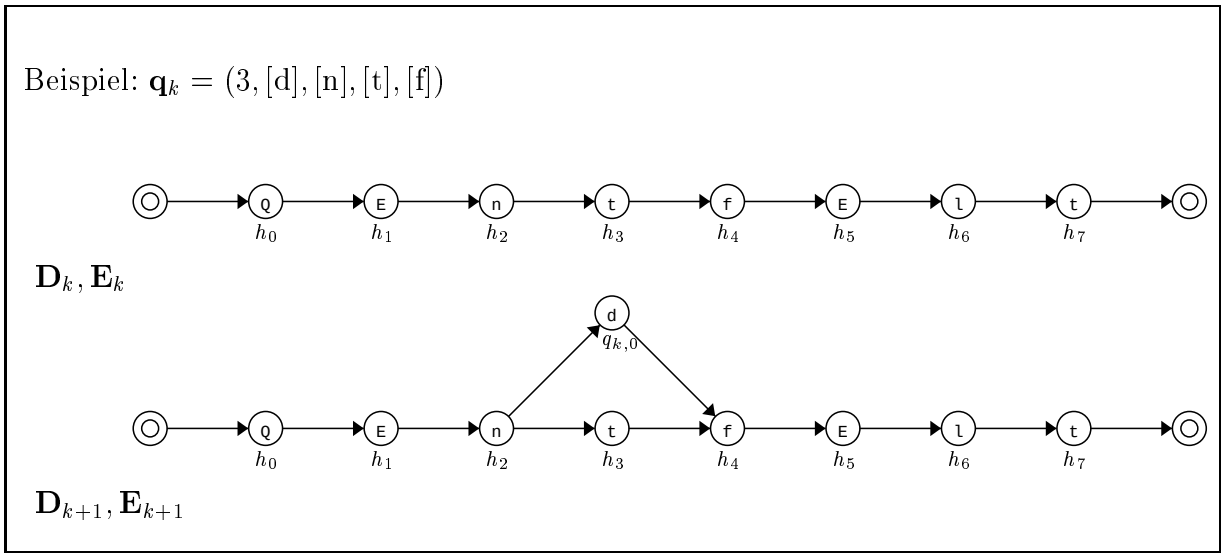


Abbildung A.1: Beispiel für Ersetzung in beidseitig spezifiziertem Kontext (Fall 1).

Fall 2: Elision in beidseitig spezifiziertem Kontext (Abb. A.2). $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \mathbf{0} \wedge \boldsymbol{x} \neq \mathbf{0} \wedge \boldsymbol{y} \neq \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \tag{A.15}$$

$$\mathbf{E}_{k+1} = \mathbf{E}_k \cup \{(h_{i-1}, h_{i+K}\} \tag{A.16}$$

Fall 3: Ersetzung in spezifiziertem Rechtskontext und beliebigem Linkskontext (Abb. A.3). $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \beta_0 \ldots \beta_{L-1} \wedge \boldsymbol{x} = \mathbf{0} \wedge \boldsymbol{y} \neq \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \cup \{x_{i-1}\} \cup \{q_{k,l} | 0 \leq l < L \wedge \rho(q_{k,l}) = \beta_l\} \tag{A.17}$$

$$\begin{aligned} \mathbf{E}_{k+1} = {} & \mathbf{E}_k \setminus \{(h_{i-1}, h_i)\} \cup \{(x_{i-1}, q_{k,0}), (q_{k,L-1}, h_{i+K})\} \\ & \cup \{(q_{k,j-1}, q_{k,j}) | 1 < j < L\} \\ & \cup \{(h_{i-1}, x_{i-1}), (x_{i-1}, h_i)\} \end{aligned} \tag{A.18}$$

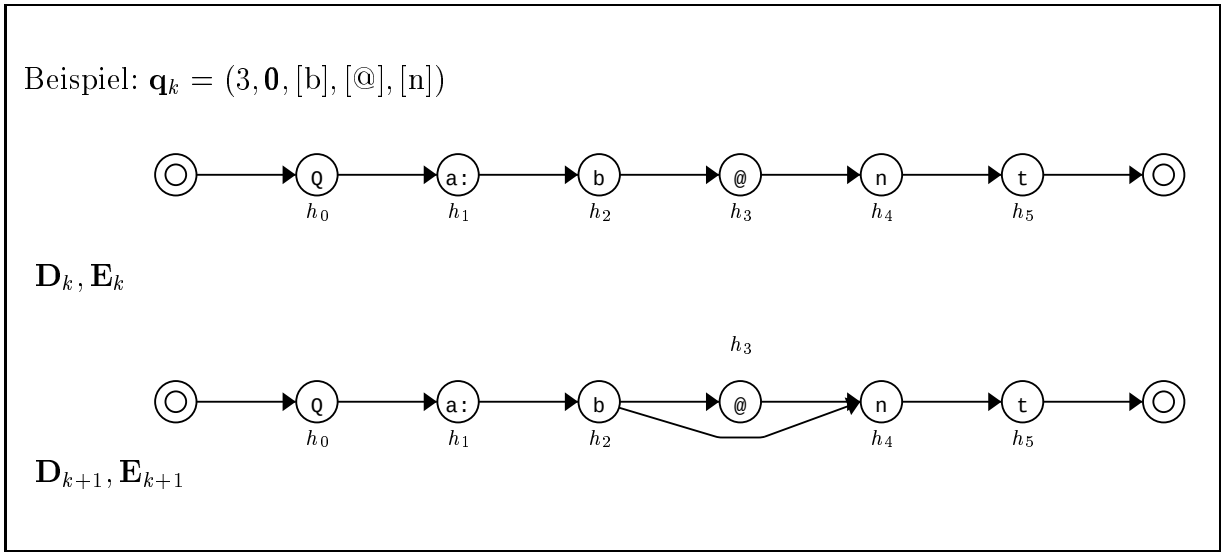


Abbildung A.2: Beispiel für Elision in beidseitig spezifiziertem Kontext (Fall 2). Die Elision wird durch die Einfügung einer Kante über den zu elidierenden Knoten hinweg berücksichtigt. Kanten emittieren kein Symbol, man kann also sagen, es wird der Leerstring $\mathbf{0}$ emittiert.

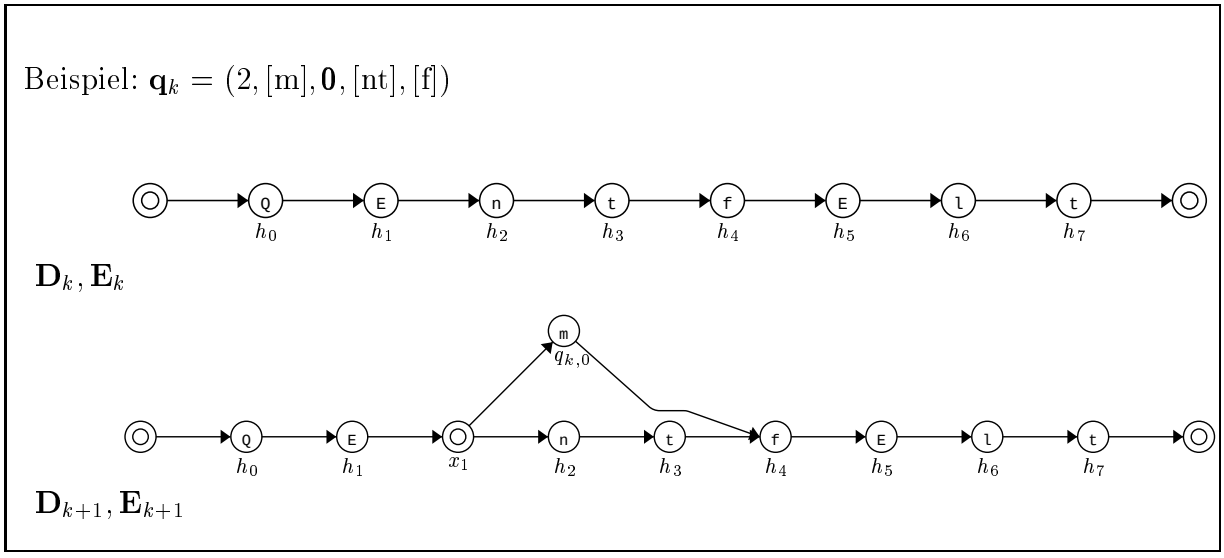


Abbildung A.3: Beispiel für Ersetzung in spezifiziertem Rechtskontext und beliebigem Linkskontext (Fall 3). Links wird ein nichtemittierender Knoten eingefügt.

Fall 4: Ersetzung in spezifiziertem Linkskontext und beliebigem Rechtskontext (Abb. A.4). $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \beta_0 \ldots \beta_{L-1} \wedge \boldsymbol{x} \neq \mathbf{0} \wedge \boldsymbol{y} = \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \cup \{x_{i+K-1}\} \cup \{q_{k,l} | 0 \leq l < L \wedge \rho(q_{k,l}) = \beta_l\} \tag{A.19}$$

$$\begin{aligned}\mathbf{E}_{k+1} = {} & \mathbf{E}_k \setminus \{(h_{i+K-1}, h_{i+K})\} \cup \{(h_{i-1}, q_{k,0}), (q_{k,L-1}, x_{i+K-1})\} \\ & \cup \{(q_{k,j-1}, q_{k,j} | 1 < j < L\} \\ & \cup \{(h_{i+K-1}, x_{i+K-1}), (x_{i+K-1}, h_{i+K})\}\end{aligned} \tag{A.20}$$

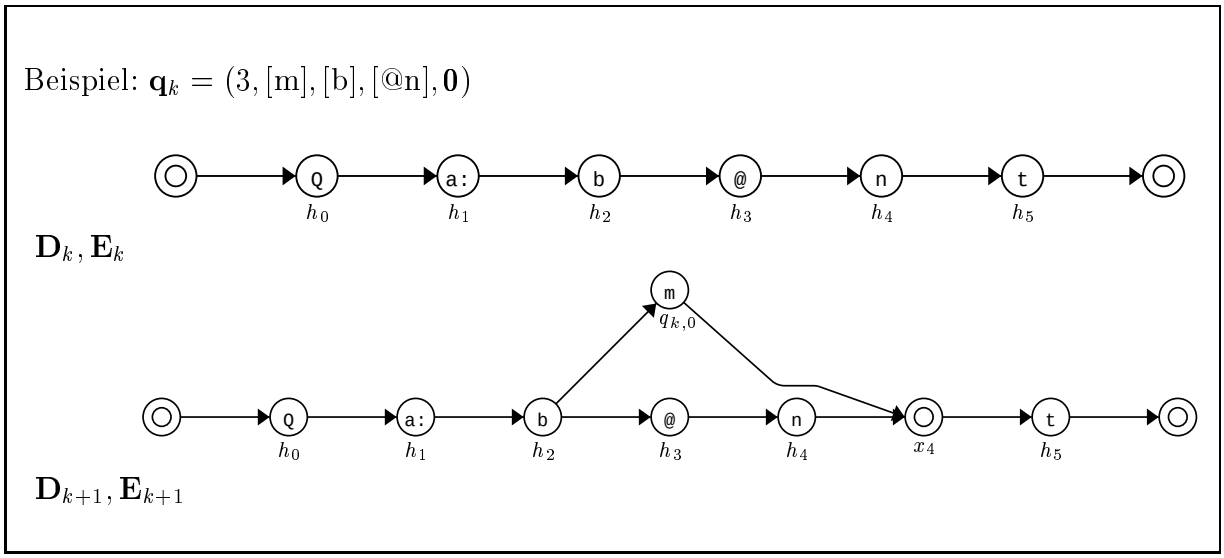


Abbildung A.4: Beispiel für Ersetzung in spezifiziertem Linkskontext und beliebigem Rechtskontext (Fall 4). Rechts wird ein nichtemittierender Knoten eingefügt.

Fall 5: Ersetzung in beliebigem Kontext (Abb. A.5). $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \beta_0 \ldots \beta_{L-1} \wedge \boldsymbol{x} = \mathbf{0} \wedge \boldsymbol{y} = \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \cup \{x_{i-1}, x_{i+K-1}\} \cup \{q_{k,l} | 0 \leq l < L \wedge \rho(q_{k,l}) = \beta_l\} \tag{A.21}$$

$$\begin{aligned}\mathbf{E}_{k+1} = {} & \mathbf{E}_k \setminus \{(h_{i-1}, h_i), (h_{i+K-1}, h_{i+K})\} \cup \{(x_{i-1}, q_{k,0}), (q_{k,L-1}, x_{i+K-1})\} \\ & \cup \{(q_{k,j-1}, q_{k,j}) | 0 < j < L\} \\ & \cup \{(h_{i-1}, x_{i-1}), (x_{i-1}, h_i), (h_{i+K-1}, x_{i+K-1}), (x_{i+K-1}, h_{i+K})\}\end{aligned} \tag{A.22}$$

Fall 6: Elision in spezifiziertem Linkskontext und beliebigem Rechtskontext (Abb. A.6. $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \mathbf{0} \wedge \boldsymbol{x} \neq \mathbf{0} \wedge \boldsymbol{y} = \mathbf{0}$

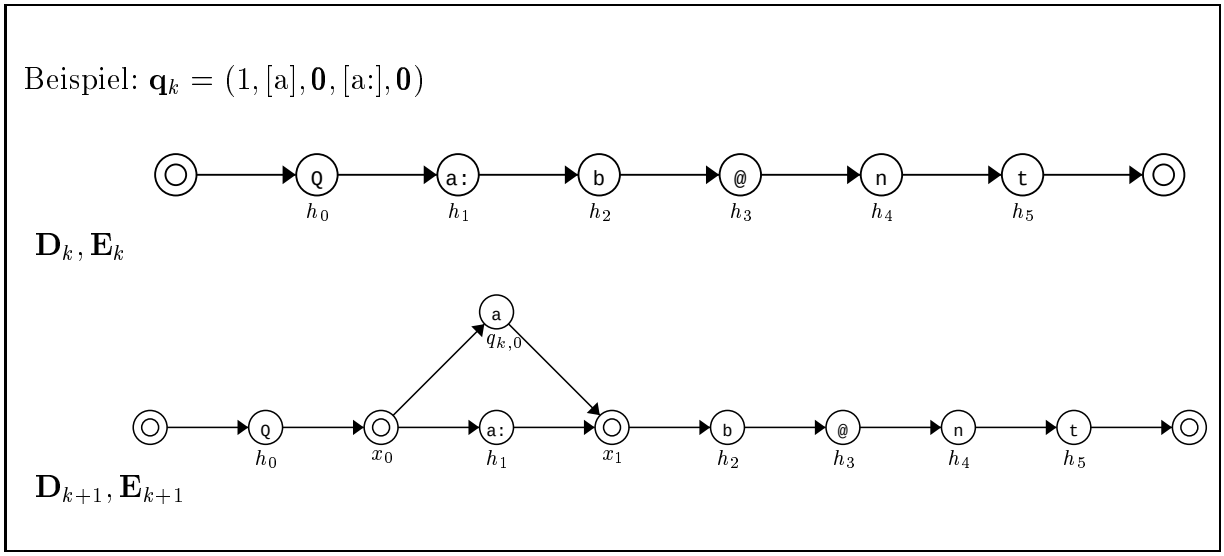


Abbildung A.5: Beispiel für Ersetzung in beliebigem Kontext (Fall 5). Auf beiden Seiten wird ein nichtemittierender Knoten eingefügt.

$$\mathbf{D}_{k+1} = \mathbf{D}_k \cup \{x_{i+K-1}\} \tag{A.23}$$

$$\mathbf{E}_{k+1} = \mathbf{E}_k \setminus \{(h_{i+K-1}, h_{i+K})\} \cup \{(h_{i-1}, x_{i+K-1})\} \cup \{(h_{i+K-1}, x_{i+K-1}), (x_{i+K-1}, h_{i+K})\} \tag{A.24}$$

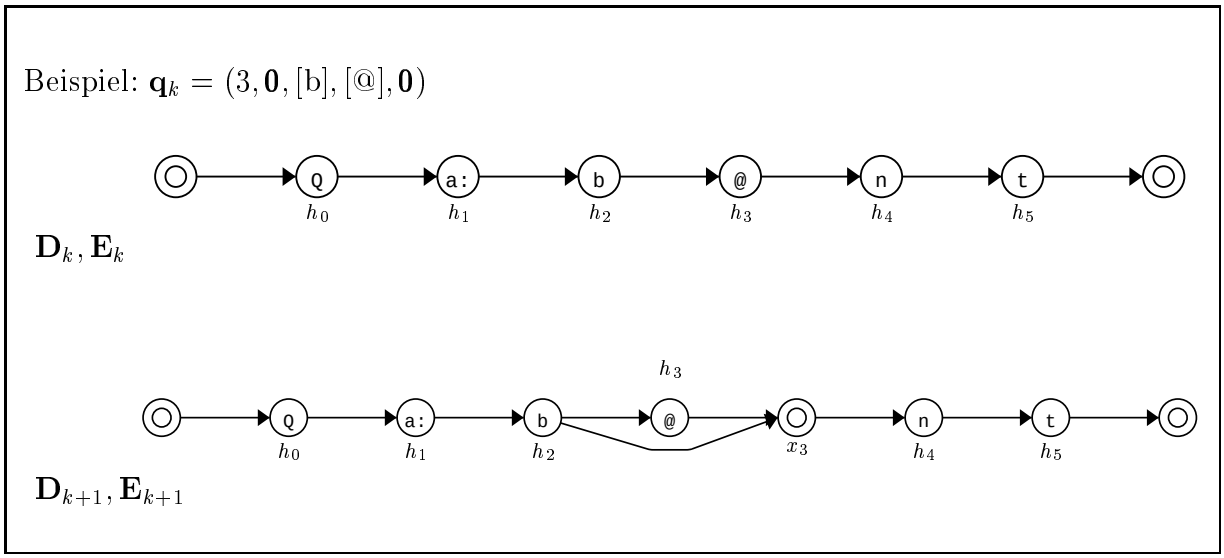


Abbildung A.6: Beispiel für Elision in spezifiziertem Linkskontext und beliebigem Rechtskontext (Fall 6). Rechts wird ein nichtemittierender Knoten eingefügt.

Fall 7: Elision in spezifiziertem Rechtskontext und beliebigem Linkskontext (Abb. A.7. $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \mathbf{0} \wedge \boldsymbol{x} = \mathbf{0} \wedge \boldsymbol{y} \neq \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \cup \{x_{i-1}\} \tag{A.25}$$

$$\begin{aligned}\mathbf{E}_{k+1} = \mathbf{E}_k \setminus \{(h_{i-1}, h_i) \cup \{(x_{i-1}, h_{i+K})\} \\ \cup \{(h_{i-1}, x_{i-1}), (x_{i-1}, h_i)\}\end{aligned} \tag{A.26}$$

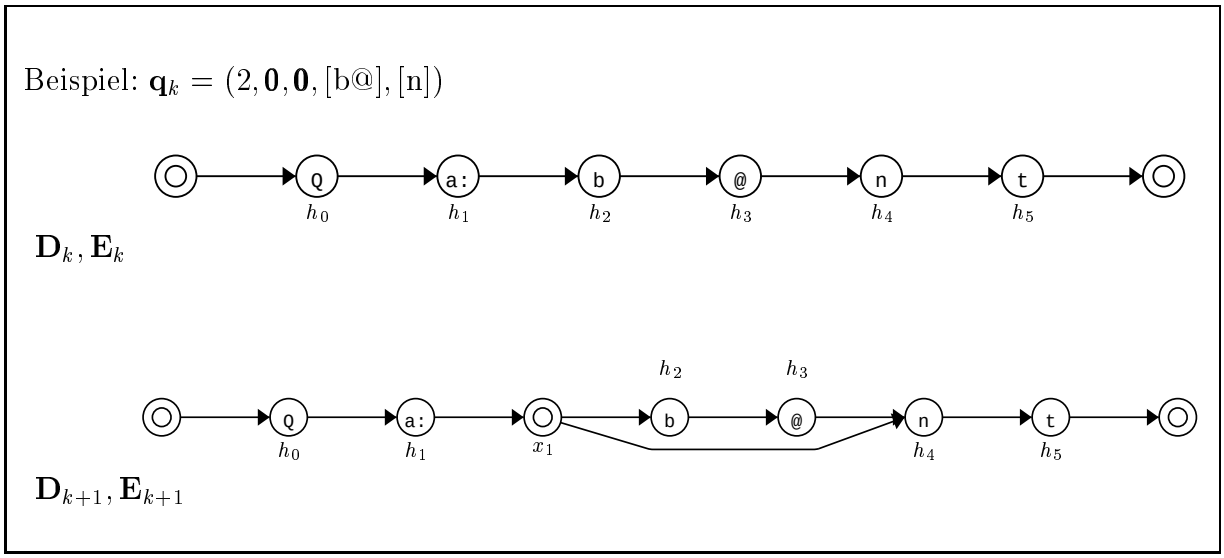


Abbildung A.7: Beispiel für Elision in spezifiziertem Rechtskontext und beliebigem Linkskontext (Fall 7). Links wird ein nichtemittierender Knoten eingefügt.

Fall 8: Elision in beliebigem Kontext (Abb. A.8. $\boldsymbol{a} = \alpha_0 \ldots \alpha_{K-1} \wedge \boldsymbol{b} = \mathbf{0} \wedge \boldsymbol{x} = \mathbf{0} \wedge \boldsymbol{y} = \mathbf{0}$

$$\mathbf{D}_{k+1} = \mathbf{D}_k \tag{A.27}$$

$$\begin{aligned}\mathbf{E}_{k+1} = \mathbf{E}_k \setminus \{(h_{i-1}, h_i), (h_{i+K-1}, h_{i+K})\} \cup \{(x_{i-1}, x_{i+K-1})\} \\ \cup \{(h_{i-1}, x_{i-1}), (x_{i-1}, h_i), (h_{i+K-1}, x_{i+K-1}), (x_{i+K-1}, h_{i+K})\}\end{aligned} \tag{A.28}$$

Die Einführung der nichtemittierenden Hilfsknoten ist nicht zwingend notwendig, erleichtert jedoch die Betrachtung von Fällen bei denen eine Mikro-Aussprachevariante ohne Linkskontext unmittelbar auf eine solche ohne Rechtskontext folgt. Ohne nichtemittierende Knoten könnte dann nur mit Kenntnis der übrigen $\mathbf{q}_k$ bestimmt werden, welche Kanten dem Graphen hinzugefügt werden müssen. Abb. A.9 verdeutlicht dies.

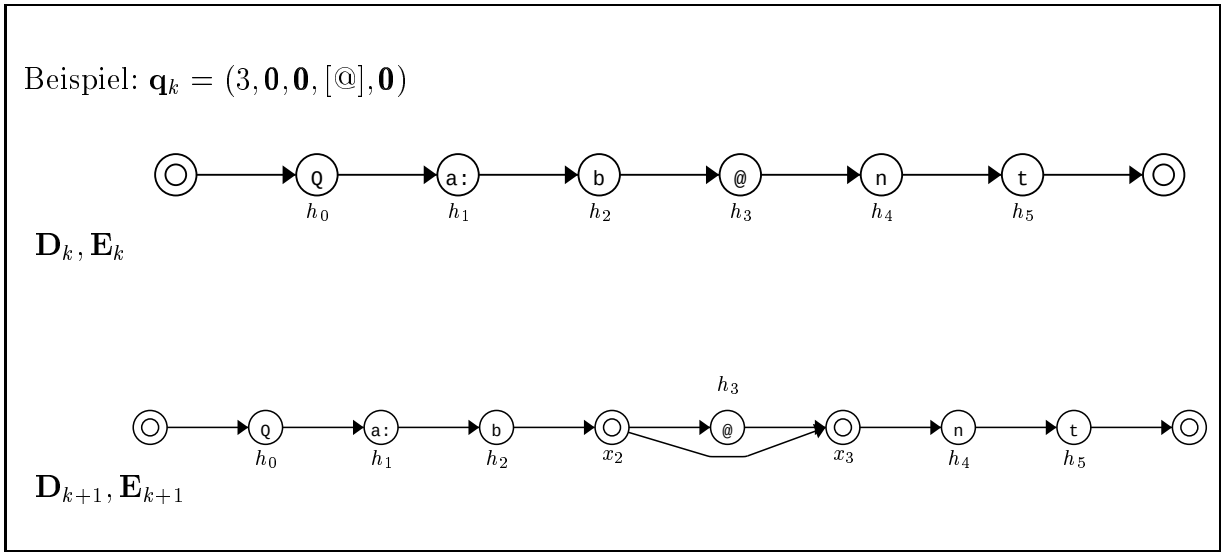


Abbildung A.8: Beispiel für Elision in beliebigem Kontext (Fall 8). Auf beiden Seiten wird ein nichtemittierender Knoten eingefügt.

Jede mögliche Mikro-Aussprachevariante ist durch die vorstehende Fallunterscheidung abgedeckt, und der Graph kann algorithmisch für jedes $\mathbf{q}_k \in \mathbf{Q}^{(c)}$ bearbeitet werden.

Der gesamte Graph ergibt sich dann als:

$$\mathbf{D} = \mathbf{D}_K \tag{A.29}$$

$$\mathbf{E} = \mathbf{E}_K \tag{A.30}$$

Damit ist die Struktur des endlichen Zustandsautomaten durch einen Variantengraphen beschrieben, der alle durch eine Menge von Mikro-Aussprachevarianten hypothetisierbaren Makro-Aussprachevarianten $\boldsymbol{r}_j$, $j = 0 \ldots I-1$ (einschließlich der Referenzaussprache, $\bigvee_{j=0}^{I-1} r_j = c$) beinhaltet[6]. In Abb. A.10 ist ein einfaches Beispiel (nur Fall 1 tritt auf) für einen Zustandsautomaten dargestellt, der insgesamt sechs Strings generieren kann.

Der Variantengraph kann als Lautfolge-Netzwerk interpretiert werden. In Abschnitt A.2.4 wird gezeigt werden, wie anhand eines solchen Netzwerkes eine komplexe HMM-Struktur aufgebaut werden kann, die dann ein statistisches Modell für die

[6] $\bigvee_x$ bedeutet hierbei "es gibt ein x für das gilt". Die Notationskonventionen folgen dem dtv-Atlas zur Mathematik [36].

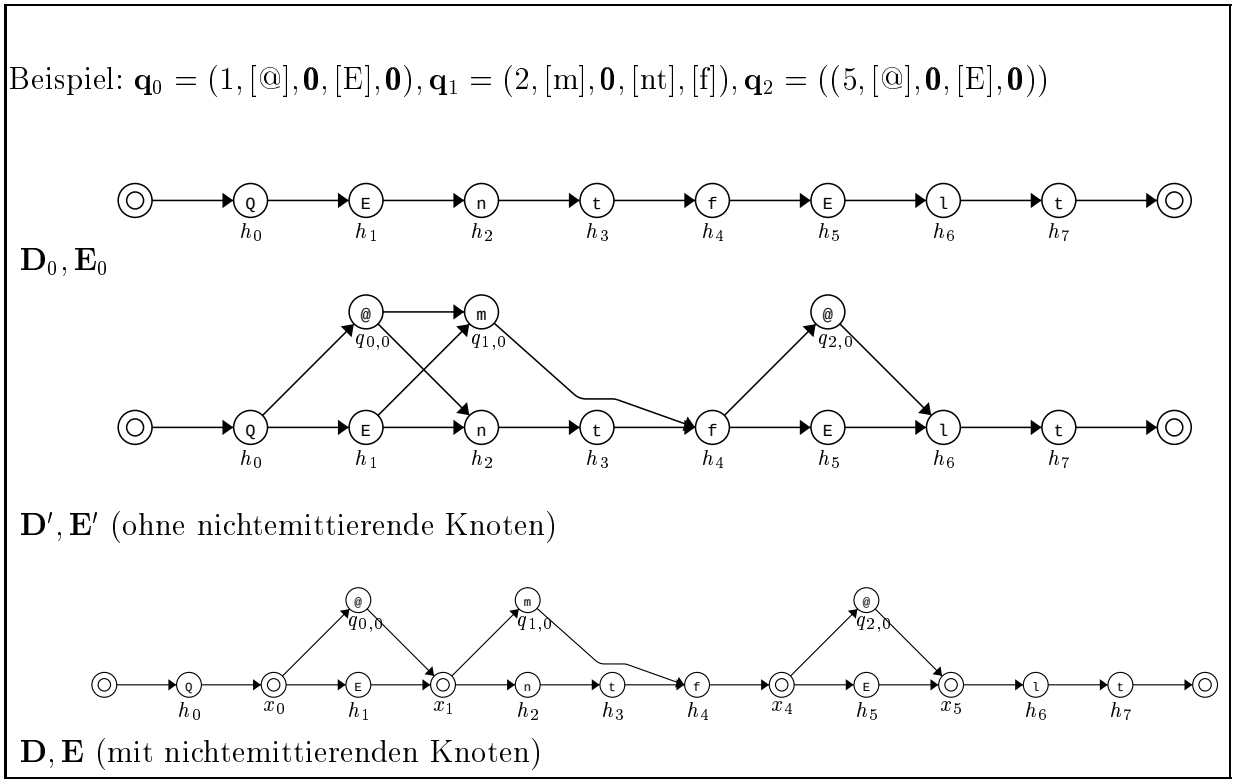
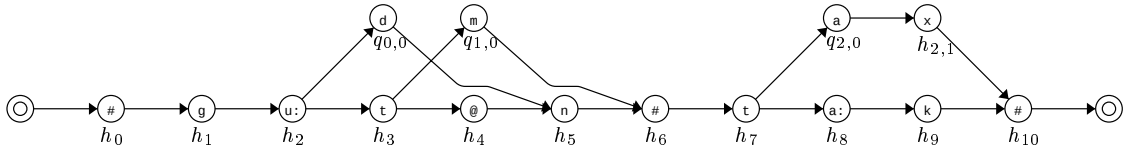


Abbildung A.9: Nichtemittierende Knoten erleichtern die Betrachtung von Fällen, in denen Mikro-Aussprachevarianten aufeinandertreffen, die keinen Kontext spezifizieren. Im gezeigten Beispiel sind dies $\mathbf{q}_0$ und $\mathbf{q}_1$ ($\mathbf{q}_2$ ist die gleiche Mikro-Aussprachevariante wie $\mathbf{q}_0$, nur mit anderer Anwendungsstelle). Die entstehenden Graphen ohne nichtemittierende Knoten (Mitte) und mit nichtemittierenden Knoten (Unten) sind äquivalent, aber im ersteren Fall kann erst nach der Anwendung aller $\mathbf{q}_2$ festgestellt werden, daß zur Erzeugung aller möglicher Varianten noch die Kante $(q_{0,0}, q_{1,0})$ notwendig ist. Die einzelnen Bearbeitungsschritte sind also nicht unabhängig voneinander. Mit nichtemittierenden Knoten können in jedem Bearbeitungsschritt alle jeweils notwendigen Knoten und Kanten hinzugefügt werden.

akustische Realisierung aller im Variantengraphen enthaltener Lautfolgen darstellt und unmittelbar zur Segmentierung und Etikettierung verwendet werden kann.

Bisher wurde lediglich die Struktur des Variantengraphen diskutiert. Um den Varianten Wahrscheinlichkeiten zuzuordnen, müssen die Zustandsübergänge, sprich die Kanten des beschreibenden Graphen, gewichtet werden. Die in den folgenden beiden Abschnitten A.1.2 und A.1.3 diskutierten Modelle geben dazu zwei verschiedene Ansätze an.



$\mathbf{G}_0^{(\boldsymbol{c})} = \emptyset$: $\boldsymbol{r}_0 = \boldsymbol{c} =$ [#gu:t@n#ta:k#]

$\mathbf{G}_1^{(\boldsymbol{c})} = \{(3, [d], [u:], [t@], [n])\}$: $\boldsymbol{r}_1 =$ [#gu:dn#ta:k#]

$\mathbf{G}_2^{(\boldsymbol{c})} = \{(4,[m],[t],[@n],[\#])\}$: $\boldsymbol{r}_2 =$ [#gu:tm#ta:k#]

$\mathbf{G}_3^{(\boldsymbol{c})} = \{(8,[ax],[t],[a:k],[\#])\}$: $\boldsymbol{r}_3 =$ [#gu:t@n#tax#]

$\mathbf{G}_4^{(\boldsymbol{c})} = \{(3,[d],[u:],[t@],[n]),$
$(8,[ax],[t],[a:k],[\#])\}$: $\boldsymbol{r}_4 =$ [#gu:dn#tax#]

$\mathbf{G}_5^{(\boldsymbol{c})} = \{(4,[m],[t],[@n],[\#]),$
$(8,[ax],[t],[a:k],[\#])\}$: $\boldsymbol{r}_5 =$ [#gu:tm#tax#]

Abbildung A.10: Darstellung eines Zustandsautomaten, der durch die Anwendung von drei Mikro-Aussprachevarianten entsteht. Darunter eine Liste aller möglichen Untermengen von $Q^{(\boldsymbol{c})}$ deren Elemente jeweils zusammen auf die Referenztranskription $\boldsymbol{c}$ angewendet werden können und die dabei entstehenden Varianten $\boldsymbol{r}$. Diese Varianten werden erzeugt, wenn alle Wege durch den Graphen vom Anfangs- zum Endknoten ausgewertet werden. Das Symbol # ist dabei das Worttrennungssymbol und wird bei der Indizierung mitgezählt.

A.1.2 Regelbasierte Aussprachemodellierung

Generierung von Aussprachеregeln

Die im Rahmen des regelbasierten Aussprachemodells verwendete Menge von Mikro-Aussprachevarianten $\mathbf{U}_{\text{man}}$ besteht aus phonetisch-phonologisch motivierten Aussprachеregeln. Diese Regeln beschreiben in abstrakter Weise lautsprachliche Phänomene und müssen zur konkreten Anwendung als Symbolmanipulationen formuliert werden.

Die Repräsentation phonetischen Expertenwissens als symbolische Aussprachеregeln in der Form, wie sie hier im regelbasierten Aussprachemodell verwendet werden, basiert auf der Arbeit von Wesenick in [53]. Dabei werden manuell segmentierte und etikettierte Sprachdaten hinsichtlich der auftretenden Aussprachevarianten systema-

tisch ausgewertet und Regeln generiert. Diese Regeln decken einerseits tatsächlich beobachtete Varianten ab. Andererseits werden aber auch im phonetischen Sinne plausibel erscheinenden Varianten berücksichtigt, für die zwar keine Evidenz, aber ähnliche Fälle in den Daten gefunden werden konnten.

Es werden Lautklassen verwendet, die sich durch das Zusammenfassen von Lauten anhand ihrer phonetischen Klassifizierung ergeben. Dadurch wird eine Verallgemeinerung der beobachteten Varianten erreicht. Weiterhin wurden Regeln für zahlreiche in [53] angeführten Phänomene hinzugefügt, falls sie nicht beobachtet worden waren.

Mit der Verwendung von 8 Lautklassen (Konsonanten, Vokale, Diphtonge, Affrikate, Liquide, Nasale, Frikative, Plosive) werden 1495 Regeln postuliert. Expandiert man die Regeln durch Ersetzung der Klassen mit allen Klassenvertretern, so ergeben sich 5545 Mikro-Aussprachevarianten.

Die Aufstellung von Regeln, die nicht auf tatsächlich beobachteten Varianten beruhen – viele davon entstehen bei der Expansion der Klassen – birgt natürlich die Gefahr in sich, daß auch ungewollte Varianten bei der Regelanwendung erzeugt werden. Das Regelsystem ist bewußt übergenerierend bzw. unterspezifizierend angelegt, in dem Sinne, daß die Erzeugung mehrerer falscher Varianten in Kauf genommen wird, wenn nur alle phonetisch sinnvollen Varianten miterzeugt werden. Die Intention ist dabei, mit einer guten akustischen Modellierung die falschen Varianten – die ja tatsächlich niemals realisiert werden – auszuschließen. Trotzdem kann die "Falschinformation", die durch die Übergeneration induziert wird, zu Problemen führen, wenn die akustische Modellierung nicht optimal ist (was in der Praxis zutrifft).

Anwendung von Ausspracheregeln

Mit der Menge $\mathbf{U}_{\text{man}}$ der generierten Mikro-Aussprachevarianten kann mit dem in Abschnitt A.10 dargestellten Verfahren zu jeder Referenzaussprache ein Variantengraph erzeugt werden, der einen endlichen Zustandsautomaten beschreibt. Jede Zustandsfolge des Automaten emittiert dabei genau einen String $\boldsymbol{r}_j$, $j = 0 \ldots J-1$.

Die Wahrscheinlichkeiten für Zustandsübergänge in diesem Automaten müssen auf irgendeine Art und Weise gewichtet werden, um die Beschreibung des Automaten zu vervollständigen. Liegt Information über die Wahrscheinlichkeit der einzelnen Mikro-Aussprachevarianten vor, so kann versucht werden, die Zustandsübergänge, die die Anwendung der jeweiligen Aussprachevariante ausdrücken, entsprechend zu gewichten.

Eine derartige Information ist bei der regelbasierten Aussprachemodellierung aber nicht vorhanden. Deshalb wird angenommen, daß jede Variante, die vom Zustandsautomaten generiert werden kann, gleichwahrscheinlich ist, d.h $P(\boldsymbol{r}_j|\boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}) = \frac{1}{J}$ für $j = 0 \ldots J-1$.

Diese Gleichverteilung der Wahrscheinlichkeiten über alle emmittierten Makro-Aussprachevarianten kann nicht dadurch erreicht werden, daß lokal alle Zustandsübergänge $P(d_j|d_i)$, $d_j \in \Gamma^+(d_i)$ gleich gewichtet werden, sondern es muß global die Struktur des Automaten, also der Variantengraph, berücksichtigt werden. In Abb. A.11 ist ein Ausschnitt aus einem möglichen Graphen dargestellt. Die verwendeten Nachfolger- bzw. Vorgängermengen und sonstigen Größen sind dort eingezeichnet.

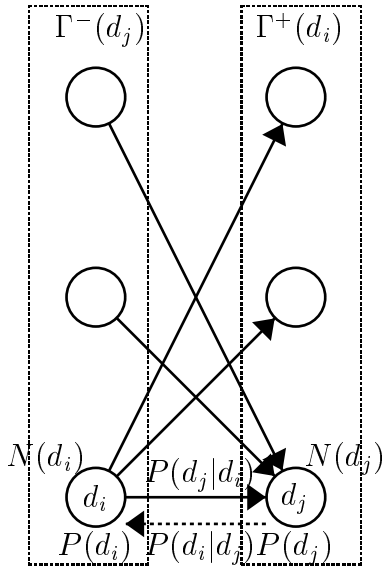


Abbildung A.11: Ausschnitt aus einem Graphen. Dargestellt sind die Knoten d_i und d_j und ihre Nachfolger- bzw. Vorgängermengen. Die jeweils zugehörigen Wahrscheinlichkeiten und die Anzahl der Wege sind mit eingezeichnet.

Es wird zunächst die Anzahl $N(d_i)$ der möglichen Zustandsfolgen, die in d_i enden, bestimmt. Dies ist mit einem trivialen rekursiven Verfahren möglich, indem die $N(d_i)$ in der Reihenfolge der Ränge $r(d_i)$ berechnet werden:

$$N(d_j) = \begin{cases} 1 & \text{für } \Gamma^-(d_j) = \emptyset \text{ (Anfangsknoten)} \\ \sum_{d_i \in \Gamma^-(d_j)} N(d_i) & \text{sonst} \end{cases} \tag{A.31}$$

Die Anzahl der Wege, die in einem Zustand enden, ist also die Summe der Wege, die in den möglichen Vorgängerzuständen enden. Setzt man voraus, daß diese Wege jeweils gleichwahrscheinlich sind, kann hieraus die Wahrscheinlichkeit, daß ein Zustand d_i der Vorgänger eines anderen d_j war, ermittelt werden:

$$P(d_i|d_j) = \frac{N(d_i)}{N(d_j)} \quad \text{mit } d_i \in \Gamma^-(d_j) \tag{A.32}$$

Sei $P(d_i)$ die Wahrscheinlichkeit, daß der Zustand d_i in einer beliebigen Zustandsfolge auftritt, so sind die gesuchten Zustandsübergangswahrscheinlichkeiten $P(d_j|d_i)$ nach Bayes gegeben durch:

$$P(d_j|d_i) = \frac{P(d_j)}{P(d_i)} P(d_i|d_j) = \frac{P(d_j)}{P(d_i)} \frac{N(d_i)}{N(d_j)} \quad \text{mit } d_i \in \Gamma^-(d_j) \tag{A.33}$$

Wenn diese Wahrscheinlichkeit für alle Zustände $d_j \in \Gamma^+(d_i)$ bekannt ist, läßt sich $P(d_i)$ mit der Rückwärts-Rekursion

$$P(d_i) = \sum_{d_j \in \Gamma^+(d_i)} P(d_j) P(d_i|d_j) = \sum_{d_j \in \Gamma^+(d_i)} P(d_j) \frac{N(d_i)}{N(d_j)} \tag{A.34}$$

berechnen. Die Wahrscheinlichkeit, daß der Anfangs- und der Endzustand in der Zustandfolge liegen, ist Eins, da ja gefordert wird, daß alle Zustandsfolgen im Anfangszustand beginnen und im Endzustand enden:

$$P(d_0) = 1 \tag{A.35}$$

$$P(d_{i_{\max}}) = 1 \tag{A.36}$$

Da alle Knoten mit einem bestimmten Rang nur Nachfolger mit höherem Rang haben können, ist es möglich, die $P(d_j)$ mit Gl. A.34 in absteigender Reihenfolge der zugehörigen Ränge $r(d_j)$ zu berechnen.

Damit können die Zustandsübergangswahrscheinlichkeiten gemäß Gl. A.33 bestimmt werden.

Die Wahrscheinlichkeit einer beliebigen Zustandsfolge $\boldsymbol{S} = s_0 s_1 ... s_{T-1}$, die im Anfangszustand beginnt ($s_0 = d_0$) und im Endzustand endet ($s_{T-1} = d_{i_{\max}}$), ist dann mit den Gln. A.31 – A.36:

$$P(\boldsymbol{S}) = \prod_{t=1}^{T-1} P(s_t|s_{t-1}) = \prod_{i=1}^{T-1} \frac{P(s_i)}{P(s_{i-1})} \frac{N(s_{i-1})}{N(s_i)} =$$
$$\frac{P(s_1)}{P(d_0)} \frac{N(d_0)}{N(s_1)} \frac{P(s_2)}{P(s_1)} \frac{N(s_1)}{N(s_2)} \cdots \frac{P(d_{i_{\max}})}{P(s_{T-2})} \frac{N(s_{T-2})}{N(d_{i_{\max}})} =$$
$$\frac{N(d_0)}{P(d_0)} \frac{P(d_{i_{\max}})}{N(d_{i_{\max}})} = \frac{1}{N(d_{i_{\max}})} = \frac{1}{J}$$

Da jede Zustandsfolge $\boldsymbol{S}$, die im Anfangszustand beginnt und im Endzustand endet, genau einen der Strings $\boldsymbol{r}_j$, $j = 0 \ldots J-1$ emittiert, ist $N(d_{i_{\max}})$ gleich der Anzahl J der erzeugbaren Strings. Die Wahrscheinlichkeit für die Emission von $\boldsymbol{r}_j$, ist dann – unter Berücksichtigung der Tatsache, daß eine Referenzaussprache $\boldsymbol{c}$ bei der Erzeugung des Automaten zugrunde lag – $P(\boldsymbol{r}_j|\boldsymbol{c}) = P(\boldsymbol{S}) = \frac{1}{J}$. Damit ist die angestrebte Gleichverteilung über alle $\boldsymbol{r}_j$ erreicht.

A.1.3 Statistische Aussprachemodellierung

Ein grundlegender Nachteil der bisher besprochenen Ansätze mit von Experten zusammengetragenen Ausspracheregeln besteht darin, daß entstehende Varianten zur Referenzaussprache nicht statistisch gewichtet werden können. Zwar ist eine Klassifizierung der Regeln nach Verschleifungsgrad der Sprache möglich, eine solide statistische Basis stellt dies indes nicht dar. Desweiteren werden derartige Regelwerke mit steigendem Umfang immer anfälliger für Inkonsistenzen und Fehler.

Bei der Modellierung von Sprachlauten haben sich statistische Verfahren wie HMMs und künstliche neuronale Netze als überlegen gegenüber regelbasierten Verfahren erwiesen. Bei statistischen Modellen für Sprachlaute werden die Modellparameter anhand von etikettierten Sprachdaten, sogenanntem Trainingsmaterial, geschätzt. Ein ähnliches Verfahren bietet sich hier ebenfalls an, nämlich die Wahrscheinlichkeiten für bestimmte Aussprachevarianten anhand von Sprachmaterial zu schätzen, das von Experten phonetisch transkribiert wurde. Auf diese Weise kann Expertenwissen in das Modell einfließen, indem Aussprachevarianten, die von den menschlichen Experten – im allgemeinen nach regelhaften Konventionen – annotiert wurden, ihrer Auftretenshäufigkeit entsprechend im Modell berücksichtigt werden.

Zur Repräsentation des Expertenwissens werden wiederum Mikro-Aussprachevarianten verwendet, die hier mit Wahrscheinlichkeiten gewichtet werden. Diese Wahrscheinlichkeiten können mit statistischen Methoden aus der Beobachtung der im handtranskribierten Trainingsmaterial von den menschlichen Experten verwendeten Mikro-Aussprachevarianten geschätzt werden.

Mit dieser statistisch basierten Gewichtung besteht der grundlegende Unterschied zu der in Abschnitt A.1.2 behandelten Modellierung darin, daß für jede beliebige Makro-Aussprachevariante einer Äußerung eine individuelle Wahrscheinlichkeit angegeben werden kann und keine Gleichverteilung angenommen wird. Dadurch wird die Entropie, die in der Gesamtheit aller möglichen Varianten enthalten ist, geringer. Mit anderen Worten ist die Information kompakter repräsentiert, was zu besseren Ergebnissen bei der automatischen Segmentierung (vgl. auch Kapitel 4) führt.

Gewichtete Mikro-Aussprachevarianten

Die in Abschnitt A.1.1 eingeführten Mikro-Aussprachevarianten $\mathbf{u} = (\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ beschreiben die Ersetzung des Phonemstrings $\boldsymbol{a}$ durch den Phonemstring $\boldsymbol{b}$, wenn in der Referenzaussprache der Phonemstring $\boldsymbol{s} = \boldsymbol{xay}$ aufgetreten ist. Um für dieses Ereignis eine Wahrscheinlichkeit angeben zu können, wird zunächst als Ergebnisraum die Menge $\mathbf{U}$ festgelegt: Jedes Element $\mathbf{u}$ aus $\mathbf{U}$ entspricht dem Elementarereignis, daß bei Vorliegen der Anwendbarkeitsbedingung $\mathbf{m} = (\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ *genau* der Phonemstring $\boldsymbol{a}$ durch $\boldsymbol{b}$ ersetzt wird. Die Wahrscheinlichkeit der Variation von $\boldsymbol{b}$ gegeben $\mathbf{m}$ kann damit notiert werden als

$$P(\boldsymbol{b}|\mathbf{m}) = P(\boldsymbol{b}|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) \tag{A.37}$$

Die Wahrscheinlichkeiten $P(\boldsymbol{b}|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ sollen nun aus hand-transkribiertem Trainingsmaterial geschätzt werden. Das Trainingsmaterial besteht dabei aus einer Menge von Äußerungen, für die sowohl die Referenzaussprache $\boldsymbol{c}$ als auch die tatsächliche Realisierung $\boldsymbol{r}$ (durch die manuelle Transkription) bekannt ist. Führt man für jede der Äußerungen eine Zuordnung von möglichst langen in $\boldsymbol{c}$ und $\boldsymbol{r}$ enthaltenen Phonemstrings (*longest common subsequece alignment*) durch, so können $\boldsymbol{r}$ und $\boldsymbol{c}$ dargestellt werden als:

$$\boldsymbol{c} = \boldsymbol{s}_0\boldsymbol{a}_0\boldsymbol{s}_1\boldsymbol{a}_1 \ldots \boldsymbol{a}_{L-1}\boldsymbol{s}_L \tag{A.38}$$

$$\boldsymbol{r} = \boldsymbol{s}_0\boldsymbol{b}_0\boldsymbol{s}_1\boldsymbol{b}_1 \ldots \boldsymbol{b}_{L-1}\boldsymbol{s}_L \tag{A.39}$$

Hierbei sind $\boldsymbol{s}_i$ die in der Referenzaussprache und der Realisierung gleichermaßen enthaltenen Phonemstrings, während die Phonemstrings $\boldsymbol{a}_i$ der Referenzaussprache durch die Phonemstrings $\boldsymbol{b}_i$ ersetzt werden, um die tatsächliche Realisierung zu erhalten. Jede dieser Ersetzungen kann als Mikro-Aussprachevariante interpretiert werden: Sei $\boldsymbol{x}_i$ ein beliebiges Suffix von $\boldsymbol{s}_i$ und $\boldsymbol{y}_{i+1}$ ein beliebiges Präfix von $\boldsymbol{s}_{i+1}$ so ist das Tupel $(\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{a}_i, \boldsymbol{y}_{i+1})$ eine Mikro-Aussprachevariante nach der Definition in Gl. A.2.

Durch diese Zuordnung können Häufigkeiten für das Auftreten von Mikro-Aussprachevarianten, also Häufigkeiten $n(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ angegeben werden. Das Vorgehen ist in Abb. A.12 verdeutlicht. Der Zuordnungs- und Extraktionsprozeß wird über ein Trainingsmaterial mit möglichst vielen sprachlichen Äußerungen durchgeführt.

Alignment von Referenzaussprache und Realisierung

$$\boldsymbol{c} = \underbrace{\#\ \text{j}\ \text{a:}\ \#\ \text{g}\ \text{u:}}_{\boldsymbol{s}_0}\ \overbrace{\text{t}\ @}^{\boldsymbol{a}_0}\ \underbrace{\text{n}\ \#\ \text{t}}_{\boldsymbol{s}_1}\ \overbrace{\text{a:}\ \text{k}}^{\boldsymbol{a}_1}\ \underbrace{\#}_{\boldsymbol{s}_2}$$

$$\boldsymbol{r} = \overbrace{\#\ \text{j}\ \text{a:}\ \#\ \text{g}\ \text{u:}}^{\boldsymbol{s}_0}\ \underbrace{\quad}_{\boldsymbol{b}_0}\ \overbrace{\text{n}\ \#\ \text{t}}^{\boldsymbol{s}_1}\ \underbrace{\text{a}\ \text{x}}_{\boldsymbol{b}_1}\ \overbrace{\#}^{\boldsymbol{s}_2}$$

↓

Extrahierte Mikro-Aussprachevarianten:

$(\mathbf{0}, [\text{u:}], [\text{t@}], [\text{n}])$

$([\text{ax}], [\text{t}], [\text{ak}], [\#])$

↓

Zu inkrementierende Häufigkeiten:

$n(\mathbf{0}, [\text{u:}], [\text{t@}], [\text{n}])$

$n([\text{ax}], [\text{t}], [\text{ak}], [\#])$

Abbildung A.12: Extraktion der Mikro-Aussprachevarianten aus einer Äußerung im Trainingsmaterial. Als Kontext wird im gezeigten Beispiel jeweils ein Symbol Präund Postkontext verwendet.

Wie aus den Gleichungen A.38 und A.39 ersichtlich, ist bei der Berechnung der Häufigkeiten aus den alignierten Phonemstrings immer $\boldsymbol{a}_i \neq \boldsymbol{b}_i$. Damit wäre immer

$n(\boldsymbol{a},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) = 0$. Das Ereignis $(\boldsymbol{a},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ bedeutet aber, daß keine Variante angewendet wurde und kann daher durchaus auftreten. Die Häufigkeiten $n(\boldsymbol{a},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ können aus den absoluten Häufigkeiten der möglichen Anwendungen von $n(\mathbf{m}) = n(\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$, d.h. dem Auftreten des Strings ***xay*** im Trainingsmaterial, berechnet werden:

$$\begin{aligned} n(\boldsymbol{a},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) &= n(\mathbf{m}) - \sum_{\boldsymbol{b}\in\Sigma\setminus\boldsymbol{a}} n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \\ &= n(\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) - \sum_{\boldsymbol{b}\in\Sigma\setminus\boldsymbol{a}} n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \end{aligned} \tag{A.40}$$

Die Auftretenshäufigkeiten $n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ und $n(\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ können also aus der Trainingsstichprobe geschätzt werden und als Basis für die Schätzung der Mikro-Aussprachevarianten-Wahrscheinlichkeiten $P(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ verwendet werden.

Maximum Likelihood Schätzung der Mikro-Aussprachevarianten-Gewichte

Die einfachste (und intuitivste) Möglichkeit, Wahrscheinlichkeiten empirisch zu bestimmen, ist, sie durch die relativen Häufigkeiten abzuschätzen. Gegeben sei folgende Situation:

- eine Ereignismenge (Ereignisalgebra) $\mathbf{U} = \{u_0, u_1 \ldots u_{K-1}\}$, die K verschiedene Ereignisse enthält.
- eine Stichprobe von N Ereignissen $x_t \in \mathbf{U}$, $t = 0 \ldots N-1$
- absolute Ereignishäufigkeiten $n(u_i)$ der Ereignisse in der Stichprobe

In diesem Fall können die Ereigniswahrscheinlichkeiten $P(u_i)$ anhand der Ereignishäufigkeiten in der Stichprobe als $\hat{p}(u_i) = \frac{n(u_i)}{N}$geschätzt werden. Empirische Schätzungen werden im folgenden immer mit einem kleinen $\hat{p}$ notiert. Da $\sum_{i=0}^{K-1} n(u_i) = N$, ist $\sum_{i=0}^{K-1} \hat{p}(u_i) = 1$. Damit ist das Ereigniswahrscheinlichkeitensystem zumindest formal richtig.

Daß eine Abschätzung der Ereigniswahrscheinlichkeiten mit relativen Häufigkeiten auch sinnvoll ist, zeigt die folgende Überlegung. Sieht man die Stichprobe als N voneinander unabhängige Zufallsexperimente, bei denen ein System jeweils eines der

Ereignisse u_i mit der Wahrscheinlichkeit $P(u_i)$ produziert, so ist die Gesamtwahrscheinlichkeit oder die *Likelihood*[7] der Stichprobe $L = \prod_{t=0}^{N-1} P(x_t)$. Dieser Wert ist ein sinnvolles Maximierungskriterium, denn die Stichprobe soll natürlich mit möglichst großer Wahrscheinlichkeit von dem System produziert worden sein. Werden die $P(u_i)$ aus einer gegebenen Sichprobe durch Maximieren der Gesamtwahrscheinlichkeit berechnet, so spricht man von einer Schätzung mit der Maximum-Likelihood-Methode.

Nimmt man den negativen Logarithmus von L und normiert ihn mit der Größe der Stichprobe, so ergibt sich die Perplexität, also der Grad an Überraschtheit, den ein Ereignis aus der Stichprobe durchschnittlich beim Beobachter hervorruft. Da der negative Logarithmus eine streng monoton fallende Funktion für L ist, muß $L^* = -\frac{\mathrm{ld}(L)}{N}$ minimiert werden, um eine maximale Likelihood zu erreichen. Dies ist auch sofort einsichtig, da die Stichprobe beim Beobachter – mit den geschätzten $P(u_i)$ – natürlich eine minimale Überraschung verursachen soll.

Die Minimierung von L^* ist aufgrund der Eigenschaften des Logarithmus häufig einfacher zu bewerkstelligen. Für L^* ergibt sich dann, mit einer beliebigen Schätzung der $\hat{p}(u_i)$:

$$\begin{aligned} L^* = -\frac{1}{N}\sum_{t=0}^{T-1} \mathrm{ld}(\hat{p}(x_t)) &= -\frac{1}{N}\sum_{i=0}^{K-1} n(u_i)\,\mathrm{ld}(\hat{p}(u_i)) \\ &= -\sum_{i=0}^{K-1} \frac{n(u_i)}{N}\,\mathrm{ld}(\hat{p}(u_i)) \end{aligned} \tag{A.41}$$

Es kann gezeigt werden, daß für zwei Mengen reeller Zahlen $\{q_0, q_1 \ldots q_{K-1}\}$ und $\{p_0, p_1 \ldots p_{K-1}\}$ mit

$$p_i \geq 0, \quad q_i \geq 0 \quad \text{für alle } 0 \leq i < K \tag{A.42}$$

und

$$\sum_{i=0}^{K-1} p_i = \sum_{i=0}^{K-1} q_i \tag{A.43}$$

[7] von engl. "likely": wahrscheinlich. Der Begriff hat sich auch in der deutschsprachigen Literatur eingebürgert

gilt:

$$-\sum_{i=0}^{K-1} p_i \,\mathrm{ld}(p_i) \leq -\sum_{i=0}^{K-1} p_i \,\mathrm{ld}(q_i) \tag{A.44}$$

wobei das Gleichheitszeichen gilt, wenn

$$p_i = q_i \quad \text{für alle } 0 \leq i < K \tag{A.45}$$

Die Gln. A.44 und A.45 sind unter dem Namen "Gibb's Theorem"[8] bekannt. Bezogen auf Gl. A.41 bedeutet dies nichts anderes, als daß der Term L^* seinen minimalen Wert für $\hat{p}(u_i) = \frac{n(u_i)}{N}$ annimmt. Damit ist gezeigt, daß die relativen Häufigkeiten die Perplexität minimieren und folglich die Likelihood maximieren. Deshalb sind die relativen Häufigkeiten eine Schätzung der Wahrscheinlichkeiten nach dem Maximum-Likelihood-Kriterium.

Damit ergibt sich als Schätzung der $P(\boldsymbol{b}|\boldsymbol{a},\boldsymbol{x},\boldsymbol{y})$ mit den entsprechenden $n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ nach diesem Kriterium:

$$\hat{p}(\boldsymbol{b}|\boldsymbol{a},\boldsymbol{x},\boldsymbol{y}) = \frac{n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})}{n(\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})} \tag{A.46}$$

Es werden bei dieser Schätzung jeweils nur Wahrscheinlichkeiten größer Null für tatsächlich in der Trainingsstichprobe beobachtete Ereignisse angegeben. Damit erhält man eine Menge von Mikro-Aussprachevarianten $\hat{\mathbf{U}}_0$, die sich darstellen läßt als:

$$\hat{\mathbf{U}}_0 = \left\{ \; (\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \;\middle|\; n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) > 0 \; \right\} \tag{A.47}$$

Ein Problem besteht jedoch darin, daß die klassische Maximum-Likelihood Methode für Ereignisse mit niedrigen absoluten Häufigkeiten wenig verlässliche und für nicht beobachtete Ereignisse gar keine Schätzungen liefert. Ersteres tritt aber in realen Stichproben in diesem Fall häufig auf. Letzteres führt dazu, daß aus einer geschätzten Ereigniswahrscheinlichkeit von Null nicht geschlossen werden kann, ob es sich um ein unmögliches Ereignis oder ein mögliches, nur zufällig in der Trainingsstichpobe nicht beobachtetes Ereignis handelt.

Die nächsten beiden Abschnitt wenden sich deshalb diesem Problemkomplex zu.

[8] Für einen Beweis siehe [52].

Robuste Schätzung der Mikro-Aussprachevarianten-Gewichte

Um sich über die Notwendigkeit von Schätzmethoden klar zu werden, die eine robustere Schätzung der Mikro-Aussprachevarianten-Gewichte zulassen, als dies mit der Maximum-Likelihood Methode möglich ist, betrachte man zunächst die Beziehung zwischen Ereignisraum $\mathbf{U}$ und der Trainingsstichprobe, die zur Schätzung der Wahrscheinlichkeiten verwendet wird.

Prinzipiell ist die Menge $\mathbf{U}^{(\infty)} = \mathbf{\Sigma} \times \mathbf{\Sigma} \times \mathbf{\Sigma} \times \mathbf{\Sigma}$ der möglichen Mikro-Aussprachevarianten unendlich groß. In einem idealen Trainingsmaterial würde genau die Mikro-Aussprachevariantenmenge $\mathbf{U}^{(L)}$ beobachtet werden, die zur Aussprachemodellierung einer Sprache nötig ist. Durch das Hinzunehmen von weiterem Trainingsmaterial kämen dann auch keine neuen Elemente zu $\mathbf{U}^{(L)}$ hinzu. Ideales Trainingsmaterial, das diesen Anforderungen genügt, wäre in vielerlei Hinsicht für die Theoriebildung von großem Interesse, ist praktisch aber nicht verfügbar.

In realem Trainingsmaterial, das vor allem nicht umfangreich genug ist, um den Idealanforderungen zu genügen, wird nur eine relativ kleine Untermenge von $\mathbf{U}^{(L)}$ überhaupt beobachtbar sein. Desweiteren sind die absoluten Auftretenshäufigkeiten für manche Mikro-Aussprachevarianten sehr klein. Man spricht in diesem Fall von spärlichen Trainingsdaten (*sparse data*): Viele mögliche Ereignisse werden selten oder überhaupt nicht beobachtet, und die Wahrscheinlichkeiten dieser Ereignisse können nicht mit befriedigender Verläßlichkeit geschätzt werden.

Trotzdem soll aber eine Menge $\hat{\mathbf{U}}$ anhand des vorhandenen Trainingsmaterials angegeben werden, die dem theoretischen $\mathbf{U}^{(L)}$ möglichst nahe kommt, also eine möglichst große Untermenge von $\mathbf{U}^{(L)}$ ist. Weiterhin sollen für alle Mikro-Aussprachevarianten $(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) \in \hat{\mathbf{U}}$ die Wahrscheinlichkeiten $P(\boldsymbol{b}|\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y})$ so geschätzt werden, daß sich die zu erwartenden Fehler für die beiden Größen in Grenzen halten und beim späteren Einsatz in einem Segmentierungs- oder Spracherkennungssystem möglichst nicht zu Performanzeinbußen führen. Die Schätzung der Wahrscheinlichkeiten soll also robust gegen die Unvollständigkeit der Trainingsdaten sein.

Mit der in Abschnitt A.1.3 vorgeschlagenen Methode erhält man ein $\hat{\mathbf{U}}_0$, das alle in einem Trainingsmaterial beobachteten Mikro-Aussprachevarianten enthält, und Maximum-Likelihood Schätzungen für die entsprechenden Wahrscheinlichkeiten. Diese werden aber die Robustheitsforderung aufgrund des geringen Umfangs, der reales Trainingsmaterial stets kennzeichnet, wohl nicht optimal erfüllen.

Im Rahmen dieses statistischen Aussprachemodells werden zwei speziellere Metho-

den verwendet, um trotz der eigentlich zu geringen Menge an Trainingsmaterial robuste Schätzungen für die Wahrscheinlichkeiten $P(\boldsymbol{b}|\boldsymbol{a},\boldsymbol{x},\boldsymbol{y})$ angeben zu können. Diese Methoden sind in den nächsten beiden Unterabschnitten beschrieben.

Bei beiden Verfahren wird die Menge $\mathbf{U}^{(\infty)}$ der möglichen Aussprachevarianten etwas beschränkt und zwar in der Form $\mathbf{U}^* = \boldsymbol{\Theta}_{\boldsymbol{b}} \times \boldsymbol{\Theta}_{\boldsymbol{x}} \times \boldsymbol{\Theta}_{\boldsymbol{a}} \times \boldsymbol{\Theta}_{\boldsymbol{y}}$. Für die geschätzten Mengen $\hat{\mathbf{U}}$ gilt dann stets $\hat{\mathbf{U}} \subseteq \mathbf{U}^*$.

Die Einschränkung der möglichen Aussprachevarianten betrifft zunächst den betrachteten Prä- und Postkontext: Er wird auf jeweils ein Symbol reduziert, d.h $\boldsymbol{\Theta}_{\boldsymbol{x}} = \mathbf{S}$ und $\boldsymbol{\Theta}_{\boldsymbol{y}} = \mathbf{S}$. Dadurch werden Probleme umgangen, die durch die zwangsläufig bei der Schätzung auftretenden statistischen Bindungen zwischen Mikro-Varianten entstehen, die sich nur im Prä- oder Postkontext unterscheiden, und zwar so, daß jeweils der Präkontext der einen Variante ein Suffix des Präkontextes der anderen ist, bzw. der Postkontext der einen ein Präfix des Postkontextes der anderen ist. Zusätzlich ist es unmittelbar einsichtig, daß die Reduktion des Kontextes auf nur ein Symbol im Hinblick auf die spätere Verwendung in einem Modell erster Ordnung von Vorteil ist.

Weiterhin sind selbst in einem theoretischen $\mathbf{U}^{(L)}$ die Stringmengen, aus denen $\boldsymbol{a}$ und $\boldsymbol{b}$ entnommen sind, endlich. Diese Beschränkung ergibt sich schon allein aus der Tatsache, daß einzelne sprachliche Äußerungen in irgendeiner Form zeitlich begrenzt sind. Für $\boldsymbol{\Theta}_{\boldsymbol{a}}$ und $\boldsymbol{\Theta}_{\boldsymbol{b}}$ werden deshalb Stringmengen angenommen deren Elemente eine maximale Stringlänge l haben, also $\Theta_a = \{\boldsymbol{a}|\boldsymbol{a} \in \boldsymbol{\Sigma} \wedge |\boldsymbol{a}| \leq l_a\}$ und $\boldsymbol{\Theta}_{\boldsymbol{b}} = \{\boldsymbol{b}|\boldsymbol{b} \in \boldsymbol{\Sigma} \wedge |\boldsymbol{b}| \leq l_b\}$.[9]

Mit dieser Einschränkungen ist $\mathbf{U}^* = \boldsymbol{\Theta}_b \times \boldsymbol{\Theta}_{\boldsymbol{x}} \times \boldsymbol{\Theta}_{\boldsymbol{a}} \times \boldsymbol{\Theta}_y$ eine Menge mit endlich vielen Elementen. Damit ist die Zahl der möglichen Mikro-Aussprachevarianten, und damit die der Elementarereignisse bei der Schätzung der Wahrscheinlichkeiten ebenfalls endlich. Dies ist zwar keine notwendige Voraussetzung, erleichtert aber die folgenden Betrachtungen.

Annahme statistischer Unabhängigkeit zwischen Prä- und Postkontext Zunächst wird statistische Unabhängigkeit zwischen dem Prä- und dem Postkontext angenommen, d.h es wird davon ausgegangen, daß die statistischen Bindungen der Aussprache nicht so weitreichend sind, daß der Laut vor der variierenden Lautfolge den Laut nach dieser Folge beeinflußt. Da in der Lautmodellierung nur ein Modell erster Ordnung verwendet wird, liegt diese Vereinfachung nahe.

[9]Die Schreibweise $|\boldsymbol{a}|$ bezeichnet die Länge des Strings $\boldsymbol{a}$.

Formt man $P(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ unter Berücksichtigung von $P(\boldsymbol{y}|\boldsymbol{x}) = P(\boldsymbol{y})$ um, so erhält man:

$$\begin{aligned} P(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) &= \frac{P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a})P(\boldsymbol{b}|\boldsymbol{a})}{P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{a})} \\ &= \frac{P(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a})P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{a})P(\boldsymbol{b}|\boldsymbol{a})}{P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{a})} \end{aligned} \tag{A.48}$$

Es werden lediglich Schätzungen für $P(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a})$, $P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{a})$, $P(\boldsymbol{b}|\boldsymbol{a})$ und $P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{a})$ benötigt. Diese Wahrscheinlichkeiten können robuster geschätzt werden, da die zugrundeliegenden Ereignisse größere Teilmengen der Elementarereignismenge sind. So ist z.B. das Ereignis, die Vertauschung von $\boldsymbol{b}$ durch $\boldsymbol{a}$ mit Präkontext $\boldsymbol{x}$ zu beobachten, gegeben durch $\bigcup_{\boldsymbol{y}\in\boldsymbol{\Theta}_{\boldsymbol{y}}}(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ und zur Berechnung der Ereignishäufigkeiten werden alle Häufigkeiten der Einzelereignisse in dieser Menge addiert. Die Schätzung der Teilwahrscheinlichkeiten im linken und rechten Kontext beruht dann auf größeren absoluten Häufigkeiten, als dies ohne der Annahme der statistischen Unabhängigkeit zwischen Prä- und Postkontext der Fall gewesen wäre und ist deshalb robuster.

Mit der Maximum-Likelihood Methode anhand relativer Häufigkeiten können $P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{a})$, $P(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a})$ und $P(\boldsymbol{b}|\boldsymbol{a})$ aus den $n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ wie folgt geschätzt werden.

$$\hat{p}(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{a}) = \frac{n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a})}{\sum_{\boldsymbol{x}'\in\Theta_x} n(\boldsymbol{b},\boldsymbol{x}',\boldsymbol{a})} \tag{A.49}$$

$$\hat{p}(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a}) = \frac{n(\boldsymbol{b},\boldsymbol{a},\boldsymbol{y})}{\sum_{\boldsymbol{y}'\in\Theta_y} n(\boldsymbol{b},\boldsymbol{a},\boldsymbol{y}')} \tag{A.50}$$

mit:

$$n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a}) = \sum_{\boldsymbol{y}\in\boldsymbol{\Theta}_{\boldsymbol{y}}} n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \tag{A.51}$$

$$n(\boldsymbol{b},\boldsymbol{a},\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\boldsymbol{\Theta}_{\boldsymbol{x}}} n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \tag{A.52}$$

$$\tag{A.53}$$

Die Wahrscheinlichkeit $P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{a})$ kann aus den Häufigkeiten $n(\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ bestimmt werden:

$$\hat{p}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{a}) = \frac{n(\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})}{\sum_{\boldsymbol{y}'\in\Theta_y}\sum_{\boldsymbol{x}'\in\Theta_x} n(\boldsymbol{x}',\boldsymbol{a},\boldsymbol{y}')} \tag{A.54}$$

Durch die Schätzung der Wahrscheinlichkeiten mit dieser Methode werden für eine größere Anzahl von Mikro-Aussprachevarianten, als im Trainingsmaterial beobachtet, von Null verschiedene Auftretenswahrscheinlichkeiten angeben. Die Menge $\hat{\mathbf{U}}_1$ der mit der Methode generierten Mikro-Aussprachevarianten ist:

$$\hat{\hat{U}}_1 = \{\ (\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \mid \sum_{\hat{\boldsymbol{y}}\in\Theta_y} n(\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\hat{\boldsymbol{y}}) > 0 \quad \wedge \quad \sum_{\hat{x}\in\Theta_x} n(\boldsymbol{b},\hat{\boldsymbol{x}},\boldsymbol{a},\boldsymbol{y}) > 0\ \} \tag{A.55}$$

Damit enthält $\hat{\mathbf{U}}_1$ mehr Elemente als $\hat{\mathbf{U}}_0$ und es wird auch im Trainingsmaterial unbeobachteten Ereignissen eine Wahrscheinlichkeit größer Null zugeordnet. Bedingung dafür ist, daß jeweils der linke und der rechte Mikro-Aussprachevariantenteil beobachtet wurden, aber nicht notwendigerweise in Verbindung miteinander.

Schätzung der Ereigniswahrscheinlichkeiten mit der *absolute-discounting*-Technik Bei dieser Methode werden Wahrscheinlichkeiten für Ereignisse nicht mit der traditionellen Maximum-Likelihood Methode sondern mit Hilfe von Varianten der *Touring-Good* [31],[19] Formeln bzw. mit Kreuzvalidierungstechniken geschätzt. Kerngedanke hierbei ist, einen Teil der "Wahrscheinlichkeitsmasse" umzuverteilen: die Maximum-Likelihood-Wahrscheinlichkeiten von Ereignissen, die in einem Trainingskorpus tatsächlich beobachtet wurden, werden dekrementiert und die Summe der Dekremente auf nicht beobachtete, aber mögliche Ereignisse verteilt. Das geschieht dadurch, daß ein *discounting*[10] der absoluten Auftretenshäufigkeiten stattfindet, d.h. diese Häufigkeiten selbst dekrementiert werden.

Die discounting-Technik hat sich bei Sprachmodellen, die die Wahrscheinlichkeit für das Aufeinanderfolgen zweier Wörter in gesprochener Sprache angeben, als erfolgreich erwiesen. Bei der Schätzung der Wahrscheinlichkeiten für Sprachmodelle sind spärliche Trainingsdaten ebenfalls ein Kernproblem.

Die Überlegungen, die dazu führen, von den absoluten Häufigkeiten einen Teil abzuziehen und ungesehenen Ereignissen zuzuschlagen, sollen nun kurz dargelegt werden.

Zunächst führt man sich vor Augen, daß Ereignisse, die in einer Trainingsstichprobe gleich oft vorkommen, auch gleiche Ereigniswahrscheinlichkeiten erhalten sollten.

[10] von engl. "to discount": abziehen (vom Nominalwert).

Angenommen, zur Schätzung der Wahrscheinlichkeiten $P(u_i)$ für K verschiedene Ereignisse einer Menge $\mathbf{U} = \{u_0 \dots u_{K-1}\}$ steht ein Trainingsmaterial mit N Ereignissen $x_t \in \mathbf{U}$, $t = 0 \dots N-1$ zur Verfügung. Die absoluten Auftretenshäufigkeiten sind wiederum $n(u_i)$. Alle Ereignisse mit gleichen $n(u_i) = r$ bilden eine Äquivalenzklasse: Alle $P(u_i)$ aus dieser Klasse sollen gleich sein und den Wert p_r haben.[11] Die Mächtigkeit der Äquivalenzklasse, d.h. die Zahl der Ereignisse die genau r mal in der Stichprobe auftraten sei n_r.

$$p_r = P(u_i) \quad \text{wenn} \quad n(u_i) = r \tag{A.56}$$

$$n_r = |\{u_i | n(u_i) = r\}| \tag{A.57}$$

Um zu testen, ob die $P(u_i)$ vernünftig geschätzt sind, wäre es denkbar, eine bestimmte Anzahl von Ereignissen aus dem gesamten Material auszublenden, also zwei Indexmengen $\mathbf{X}$ und $\mathbf{Y}$ mit $\mathbf{X} \cup \mathbf{Y} = \{0, 1 \dots N-1\}$ und $\mathbf{X} \cap \mathbf{Y} = \emptyset$ zu bilden.

Die Menge $\mathbf{X}$ könnte zur Schätzung der Parameter verwendet werden, die Menge $\mathbf{Y}$ zum Testen und zwar mit der früher bereits eingeführten Perplexität – oder bildlicher, dem Grad an Überraschtheit – in folgender Weise: Die in der Testmenge auftretenden Ereignisse sollen mit den aus der Trainingsmenge geschätzten Wahrscheinlichkeiten $\hat{p}(u_i)$ insgesamt möglichst wenig Überraschung hervorrufen, d.h. Ereignisse mit niedrigem $\hat{p}(u_i)$ sollen selten vorkommen und solche mit hohem $\hat{p}(u_i)$ eher häufiger. Die Überraschung, die ein einzelnes Ereignis bei geschätzten $\hat{p}(u_i)$ verursacht, ist $-\operatorname{ld}(\hat{p}(u_i))$. Um die Testsetperplexität zu bestimmen, werden die einzelnen Perplexitätsbeiträge der Ereignisse über das Testset summiert und auf seine Größe normiert.

Nimmt man die Aufteilung in Test- und Trainingsset so vor, daß $\mathbf{X}$ genau $N-1$ Elemente enthält und $\mathbf{Y}$ genau eines, so ist die Testsetperplexität natürlich gleich dem Perplexitätsbeitrages des einen Elements in $\mathbf{Y}$. Es sind N verschiedene Aufteilungen in Testset und Trainingsset in der beschriebenen Art und Weise möglich. Die gesamte Perplexität aller möglichen Testsets stellt ein sinnvolles Minimierungskriterium dar, da jedes Ereignis einmal zu Testzwecken verwendet wird, sonst aber immer im Trainingsset ist (*leave-one-out*-Verfahren).

Mit dem weiter oben eingeführten Konzept der Äquivalenzklasseneinteilung können die p_r global optimiert werden. Die Trainingssets sind jeweils mit $N-1$ Elementen

[11] Der Index r ist hier kursiv gesetzt, da es sich – im Gegensatz zu z.B. n_e – um eine Variable handelt.

gleich groß, so daß in jedem Set Ereignisse mit gleicher Auftretenshäufigkeit auch dieselben Wahrscheinlichkeiten zugeordnet bekommen sollen. Die $\hat{p}(u_i)$ werden aus den optimierten p_r berechnet.

Man betrachte folgende Aufteilung:

- Der Testkorpus sei ein Ereignis, das im Gesamtkorpus die Auftretenshäufigkeit $n(u_i) = r + 1$ hat.
- Im Traningskorpus hat das Ereignis nur noch die Auftretenshäufigkeit r. Seine Äquivalenzklasse ist also r und die zugeordnete Wahrscheinlichkeit ist p_r.

Da das Ereignis im Gesamtkorpus $r + 1$ mal aufgetreten ist, gibt es $r + 1$ derartige Aufteilungen für dieses eine Ereignis. Die Äquivalenzklasse umfaßt n_{r+1} Ereignisse. Der gesamte Perplexitätsbeitrag aller n_{r+1} Ereignisse der Äquivalenzklasse $r + 1$ ist folglich $(r + 1)n_{r+1} \operatorname{ld}(p_r)$.

Wenn es R Äquivalenzklassen gibt, d.h. das maximale $n(u_j) = R$ war, ergibt sich die Gesamtperplexität über alle Äquivalenzklassen zu:

$$L_{\mathrm{d}}^{*} = \sum_{r=0}^{R-1} (r + 1) n_{r+1} \operatorname{ld}(p_r) \tag{A.58}$$

Die Größe L_{d}^{*} muß nun unter Berücksichtigung der Nebenbedingung, daß die Summe über alle Ereigniswahrscheinlichkeiten Eins sein soll, minimiert werden. Die Nebenbedingungen sind:

$$\sum_{i=0}^{K-1} n(u_i) = \sum_{r=1}^{R} r n_r = N \tag{A.59}$$

$$\sum_{i=0}^{K-1} \hat{p}(u_i) = \sum_{r=0}^{R} n_r p_r = 1 \tag{A.60}$$

Das Problem der Suche nach Extrema mit Nebenbedingung kann mit Lagrange-Multiplikatoren gelöst werden. Formuliert man G. A.58 mit Gl. A.60 als

$$L_{\mathrm{d}}^{*} = \sum_{r=0}^{R-1} (r + 1) n_{r+1} \operatorname{ld}(p_r) - \lambda \left(\sum_{r=0}^{R} n_r p_r - 1 \right) \tag{A.61}$$

wobei λ der Lagrange-Multiplikator ist, ergibt partielles Differenzieren für die einzelnen p_r den Ausdruck:

$$p_r = \frac{1}{\lambda}\frac{(r+1)n_{r+1}}{n_r} \quad \text{für } r = 0, 1 \ldots R-1 \tag{A.62}$$

Der Wert des Lagrange-Multiplikators kann mit den Gln. A.59 und A.60 bestimmt werden:

$$\sum_{r=0}^{R} n_r p_r = \frac{1}{\lambda} \underbrace{\sum_{r=0}^{R-1}(r+1)n_{r+1}}_{\sum_{r=1}^{R} r n_r = N} + p_R n_R = 1$$

damit ergibt sich:

$$\lambda = \frac{1 - p_R n_R}{N} \tag{A.63}$$

und die p_r werden geschätzt als:

$$p_r = \frac{1 - p_R n_R}{N}\frac{(r+1)n_{r+1}}{n_r} \tag{A.64}$$

Die Gesamtwahrscheinlichkeit für die in der Trainingsstichprobe beobachteten Ereignisse ist dann:

$$\sum_{r=1}^{R} p_r = 1 - \frac{n_1}{N}(1 - n_R p_R) \tag{A.65}$$

Das bedeutet, daß die "Wahrscheinlichkeitsmasse" $\frac{n_1}{N}(1 - n_R p_R)$ noch frei ist und über unbeobachtete, aber mögliche Ereignisse verteilt werden kann.

Nimmt man an, daß $p_R n_R \ll 1$ gilt, was für reale Trainingsstichproben oft gerechtfertigt ist, ergibt sich die Touring-Good-Schätzung:

$$p_r = \frac{1}{N}\frac{(r+1)n_{r+1}}{n_r} \tag{A.66}$$

Die in diesem Fall freie Wahrscheinlichkeitsmasse ist dann $\frac{n_1}{N}$. Die so geschätzten Wahrscheinlichkeiten p_r sind also durchschnittlich niedriger, als die relativen Häufigkeiten $\frac{r}{N}$. Die geschätzte Wahrscheinlichkeit kann also in der Form

$$p_r = \frac{r - d(r)}{N} \tag{A.67}$$

geschrieben werden. Die Funktion $d(r)$ wird als *Discountingfunktion* bezeichnet. Sie ist eine Funktion der Auftretenshäufigkeit r mit $d : r \to \mathbb{R}$. Für eine Schätzung nach Gl. A.64 ergibt sich:

$$d(r) = r - \frac{(1 - n_R p_R) n_{r+1}}{n_r}(r + 1) \tag{A.68}$$

Diese Discountingfunktion hat jedoch den Nachteil, daß sie keine monoton steigenden p_r für wachsende r fordert, was für eine sinnvolle Abschätzung der Ereigniswahrscheinlichkeiten jedoch wünschenswert ist, denn natürlich sollten Ereignisse mit größeren Auftretenshäufigkeiten auch größere Wahrscheinlichkeiten erhalten.

Im hier beschriebenen Ansatz wird deshalb eine wesentlich einfachere Discountingfunktion verwendet, nämlich $d(r) = b_0 = \text{const}$. Dieses Vorgehen wird als *absolute discounting* [32] bezeichnet. Hierbei wird zwar Gl. A.61 nicht exakt minimiert, aber die Grundidee der Umverteilung der Wahrscheinlichkeitsmasse verwirklicht, und zwar so, daß Ereignisse mit hohen Auftretenshäufigkeiten relativ gesehen weniger stark betroffen sind als solche mit geringen Häufigkeiten.

Das ist durchaus sinnvoll, denn ob ein Ereignis nur einmal oder überhaupt nicht in der Stichprobe aufgetreten ist, kann als zufällig angesehen werden, und die Schätzung anhand der relativen Häufigkeit ist eher unsicher. Sie sollte verhältnismäßig stärker nach unten korrigiert werden als bei häufigeren Ereignissen.

Wendet man die absolute Discountingfunktion mit konstantem b_0 zur Schätzung der $P(\boldsymbol{b}|\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y})$ an, so ergibt sich:

$$\hat{p}(\boldsymbol{b}|\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \dfrac{n(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) - b_0}{n(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})} & \text{für } n(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) > 0 \\ b_0 \dfrac{K - n_0}{N n_0} & \text{sonst} \end{cases} \tag{A.69}$$

mit

$$n_0 = |\{(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) | (\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) \in \mathbf{U}^* \ \wedge \ n(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) = 0\}| \tag{A.70}$$

$$K = |\mathbf{U}^*| \tag{A.71}$$

$$N = \sum_{\substack{(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) \\ \in \mathbf{U}^*}} n(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) \tag{A.72}$$

Es werden also *allen* Ereignissen aus $\mathbf{U}^*$ von Null verschiedene Wahrscheinlichkeiten zugeordnet, d.h. die von dieser Methode geschätzte Mikro-Aussprachevariantenmenge $\hat{\mathbf{U}}_3 = \mathbf{U}^*$ hat die größte Mächtigkeit im Vergleich zu den beiden bisherigen Schätzungen nach dem Maximum-Likelihood-Kriterium und mit Annahme der Unabhängigkeit zwischen linken und rechtem Kontext.

Die große Mächtigkeit der Menge führt bei deren Anwendung zu einer starken Übergenerierung von Aussprachehypothesen. Es sind sicher vielen – im phonetischen Sinn – unmöglichen Varianten kleine, aber doch von Null verschiedene Wahrscheinlichkeiten zugeordnet. Dies führt mit der real vorhandenen Rechenkapazität und der nicht optimalen akustischen Modellierung zu Problemen.

Eine sinnvollere Basis-Aussprachenvariantenmenge ist die des regelbasierten Aussprachemodells. Man kann dann in den Gln. A.70 - A.72 die Menge $\mathbf{U}^*$ ersetzen durch:

$$\mathbf{U}^{**} = \{(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) | n(\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y}) > 0\} \quad \cup \quad \mathbf{U}_{\text{man}} \tag{A.73}$$

Formulierung des Modells

Kontextuelle Einflüsse Ein Problem bei der Berechnung der Makro-Aussprachevariantenwahrscheinlichkeiten sind $\mathbf{q}' = (i', \boldsymbol{b}', \mathbf{m})')$ und $\mathbf{q}'' = (i'', \boldsymbol{b}'', \mathbf{m})'')$ (mit verschiedenen Anwendbarkeitsbedingungen $\mathbf{m}'$ und $\mathbf{m}''$), die sich überlappen. In Abb. A.13 ist dies der Fall für $\mathbf{q}_0$ und $\mathbf{q}_1$ sowie für $\mathbf{q}_5$ und $\mathbf{q}_0$.

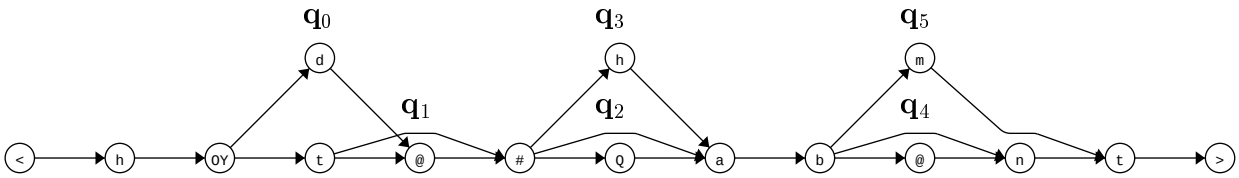


Abbildung A.13: Überlappung von Mikro-Aussprachevariantenanwendungen.

Wenn Mikro-Aussprachevarianten die gleiche Anwendbarkeitsbedingung haben – wie $\mathbf{q}_2$ und $\mathbf{q}_3$ im Beispiel – tritt kein Problem auf, da sich die Schätzung der Ereigniswahrscheinlichkeiten immer auf den gleichen Kontext bezieht. Es ist aber nicht gesichert, daß alle möglichen Überlappungen beobachtet wurden. Darum kann keine

Aussage darüber gemacht werden, ob nun Varianten mit der Anwendbarkeitsbedingung $\mathbf{m}'$ wahrscheinlicher sind als solche mit $\mathbf{m}''$ oder umgekehrt. Die Wahrscheinlichkeitsmasse soll deshalb auf solche Ereignisse gleichverteilt werden. Dazu wird eine *Kontextwahrscheinlichkeit* $p_{\mathrm{e}}(\mathbf{q})$ eingeführt. Um sie zu berechnen, wird die gesamte Menge der anwendbaren Mikro-Aussprachevarianten in K Untermengen $\mathbf{D}_l^{(\boldsymbol{c})} \subseteq \mathbf{Q}^{(\boldsymbol{c})}, l = 0 \ldots K-1$ aufgeteilt, für die gilt:

- die Mikro-Aussprachevarianten $\mathbf{q}'$ und $\mathbf{q}''$ in $\mathbf{D}_l^{(\boldsymbol{c})}$ überlappen sich nicht oder sind im exakt gleichen Kontext anwendbar.
- Es gibt keine Mikro-Aussprachevariante in $\mathbf{Q}^{(\boldsymbol{c})} \setminus \mathbf{D}_l^{(\boldsymbol{c})}$, die sich nicht mit mindestens einer in $\mathbf{D}_l^{(\boldsymbol{c})}$ überlappt.
- Es gibt keine Mikro-Aussprachevariante in $\mathbf{Q}^{(\boldsymbol{c})} \setminus \mathbf{D}_l^{(\boldsymbol{c})}$, die die exakt gleiche Anwendbarkeitsbedingung wie eine innerhalb von $\mathbf{D}_l^{(\boldsymbol{c})}$ hat.

Jede dieser Untermengen stellt ein System dar, innerhalb dessen die geschätzten Mikro-Aussprachevariantenwahrscheinlichkeiten gelten. Die Kontextwahrscheinlichkeiten $p_{\mathrm{e}}(\mathbf{q})$, mit denen sie beaufschlagt werden, um eine Gleichverteilung der Wahrscheinlichkeitsmasse zu erreichen, sind nun die relativen Auftretenshäufigkeiten der jeweiligen $\mathbf{q}$ in den Mengen[12] $\mathbf{D}_l^{(\boldsymbol{c})}$:

$$p_{\mathrm{e}}(\mathbf{q}) = \frac{\left| \{l \mid \mathbf{q} \in \mathbf{D}_l^{(\boldsymbol{c})}\} \right|}{K} \tag{A.74}$$

Diese Kontextwahrscheinlichkeiten geben eine Abschätzung für die statistischen Abhängigkeiten der Mikro-Aussprachevarianten im Kontext einer *bestimmten* Referenzaussprache an.

Die Berechnung der Größe $p_{\mathrm{e}}(\mathbf{q})$ erfolgt rekursiv, da viele der Mengen $\mathbf{D}_l^{(\boldsymbol{c})}$ gleiche Untermengen enthalten. Eine Darstellung, mit der sich dieser Sachverhalt kompakt ausdrücken läßt, sind gerichtete Graphen. Dazu wird jedes Element $\mathbf{q} \in \mathbf{Q}^{(\boldsymbol{c})}$ als Knoten angesehen. Zwischen zwei Knoten $\mathbf{q}'$ und $\mathbf{q}''$ wird eine gerichtete Kante eingefügt, wenn sich $\mathbf{q}''$ unmittelbar nach $\mathbf{q}'$ anwenden läßt, ohne daß dabei ein anderes Element $\breve{\mathbf{q}}$ übersprungen werden muß. Knoten mit exakt gleichem Anwendungskontext werden zu Superknoten zusammengefaßt.

Jeder Weg durch den Graphen von einem Anfangs- zu einem Endknoten repräsentiert eine Menge $\mathbf{D}_l^{(\boldsymbol{c})}$, da Kanten entlang des Weges sich nicht überschneidende und

[12] Die Mächtigkeit einer Menge wird hier mit $|\mathbf{D}|$ notiert.

direkt aufeinander folgende Mikro-Aussprachenvariantenanwendungen $\mathbf{q}$ darstellen. Insgesamt sind folglich K Wege im Graphen enthalten. Den Kanten werden nun Wahrscheinlichkeiten zugeordnet. Das Produkt über die Wahrscheinlichkeiten aller Kanten, die auf einem Weg liegen, der ein $\mathbf{D}_l^{(\boldsymbol{c})}$ repräsentiert, soll immer $P(\mathbf{D}_l^{(\boldsymbol{c})}) = \frac{1}{K}$ sein.

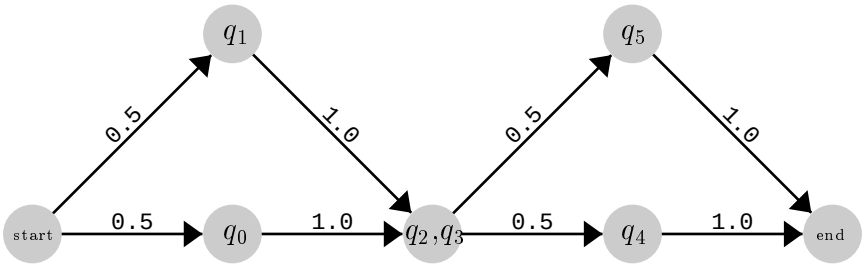


Abbildung A.14: Graph der unmittelbar hintereinander anwendbaren $\mathbf{q}$ für das Beispiel aus Abb. A.13. Man verifiziert leicht, daß hier $p_\mathrm{e}(\mathbf{q}_0) = p_\mathrm{e}(\mathbf{q}_1) = 0.5$, $p_\mathrm{e}(\mathbf{q}_2) = p_\mathrm{e}(\mathbf{q}_3) = 1$ und $p_\mathrm{e}(\mathbf{q}_2) = p_\mathrm{e}(\mathbf{q}_3) = 0.5$ gilt.

Die Berechnung dieser Wahrscheinlichkeiten kann mit dem in Abschnitt A.1.2 vorgestellten Algorithmus rekursiv und effizient erfolgen. Die resultierenden Knotenwahrscheinlichkeiten sind dann die $p_\mathrm{e}(\mathbf{q})$, da sie genau die relativen Häufigkeiten, mit denen der Knoten $\mathbf{q}$ in einem Weg durch den Graphen enthalten war, angeben (siehe Gl. A.74).

Abb. A.14 zeigt den Graphen der unmittelbar hintereinander anwendbaren $\mathbf{q}$ für das Beispiel aus Abb. A.13.

Konstruktion der Makro-Aussprachevarianten-Wahrscheinlichkeit Zur Darstellung aller Makro-Aussprachevarianten zu einer Äußerung und der jeweiligen Wahrscheinlichkeiten der Varianten wird nun wiederum ein endlicher Zustandsautomat konstruiert. Die Struktur des beschreibenden Variantengraphen ist durch die Menge $\mathbf{Q}^{(\boldsymbol{c})}$, die alle auf die Referenzaussprache $\boldsymbol{c}$ anwendbaren Mikro-Aussprachevarianten und die Lokation der möglichen Anwendung umfaßt, gegeben.

Den Kanten des Graphen müssen noch Wahrscheinlichkeiten gemäß den Überlegungen aus den Abschnitten A.1.3 bis A.1.3 zugeordnet werden. Diese sollen anhand der Schätzungen für $P(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ und der Kontextwahrscheinlichkeiten $p_{\mathrm{e}}(\mathbf{q})$ bestimmt werden.

Die Knotenmenge $\mathbf{D}$ des Graphen setzt sich – wie in Abschnitt A.1.1 eingeführt – zusammen aus Knoten h_i, die die Referenztranskription emittieren, und Knoten $q_{k,l}$, die jeweils den Ersetzungsphonemstring b von $\mathbf{q}_k = (i, \boldsymbol{b}, \mathbf{m})$ emittieren. Um eine Fallunterscheidung zu vermeiden, werden Elisionen, also Fälle, in denen $\boldsymbol{b} = \mathbf{0}$ ist, durch die Einfügung eines nichtemittierenden Knotens $q_{k,0}$ gelöst. Dann kann Für jede $\mathbf{q}_k$ ein Eintrittsknoten $s(\mathbf{q})$ und ein Austrittsknoten $a(\mathbf{q})$ definiert werden:

$$s(\mathbf{q}_k) = q_{k,0} \tag{A.75}$$

$$a(\mathbf{q}_k) = \begin{cases} q_{k,|\boldsymbol{b}|-1} & \text{für } \boldsymbol{b} \neq \mathbf{0} \\ q_{k,0} & \text{sonst} \end{cases} \tag{A.76}$$

$$\text{mit } \mathbf{q}_k = (i, \boldsymbol{b}, \mathbf{m})$$

Eine Kante von einem Knoten aus der Menge $\{h_0 \ldots h_{N-1}\}$ zu einem Knoten $s(\mathbf{q}_k)$ repräsentiert einen Zustandswechsel, der die Anwendung der entsprechenden Mikro-Aussprachevariante bedeutet. Kanten, die zwei Knoten aus der Menge der h_i verbinden repräsentieren Zustandswechsel, bei denen keine Anwendung eventuell möglicher Mikro-Aussprachevarianten stattfindet.

Elemente aus der Menge der $q_{k,l}$ haben immer genau einen Vorgänger- und genau einen Nachfolgerknoten (siehe Fall 1 aus Abschnitt A.1.1). Berücksichtigt man dies, so sind folgende bedingte Wahrscheinlichkeiten durch die Anwendung von $\mathbf{q}_k = (i, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ bestimmt.

$$P(h_{i-1}|s(\mathbf{q}_k)) = 1 \tag{A.77}$$

$$P(h_{i+|\boldsymbol{a}|}|a(\mathbf{q}_k)) = 1 \tag{A.78}$$

$$P(q_{k,l}|q_{k,l-1}) = 1 \quad 1 < l < |\boldsymbol{b}| \tag{A.79}$$

$$\begin{aligned} P(s(\mathbf{q}_k)|h_{i-1}) &= \frac{P(s(\mathbf{q}_k))P(h_{i-1}|s(\mathbf{q}_k))}{P(h_{i-1})} \\ &= \frac{P(s(\mathbf{q}_k))}{P(h_{i-1})} \qquad \text{(mit Gl. A.77)} \end{aligned} \tag{A.80}$$

$$\begin{aligned} P(h_i|h_{i-1}) &= 1 - \sum_{\substack{\{k'|s(\mathbf{q}_{k'} \in \\ \Gamma^+(h_{i-1})\}}} P(s(\mathbf{q}_{k'})|h_{i-1}) \\ &= 1 - \frac{1}{P(h_{i-1})} \sum_{\substack{\{k'|s(\mathbf{q}_{k'}) \in \\ \Gamma^+(h_{i-1})\}}} P(s(\mathbf{q}_{k'}) \end{aligned} \tag{A.81}$$

In Abb. A.15 ist ein Ausschnitt aus einem möglichen Graphen dargestellt und die entsprechenden Wahrscheinlichkeiten eingezeichnet.

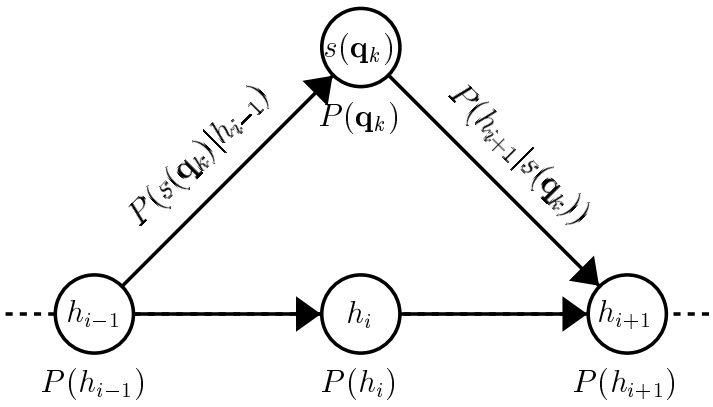


Abbildung A.15: Ausschnitt aus einem möglichen Graphen mit Knotenbezeichnungen und Wahrscheinlichkeiten.

Die Wahrscheinlichkeiten $P(h_i)$ und $P(s(\mathbf{q}_k))$ aus den Gln. A.80 und A.81 sind die Gesamtwahrscheinlichkeiten, daß sich der jeweilige Knoten in einer in einer emittierten Zustandsfolge befindet. Die Wahrscheinlichkeit $P(\mathbf{q}_k)$, daß die Mikro-Aussprachevariante $\mathbf{q}_k$ angewendet wird, ist folglich gleich der Knotenwahrscheinlichkeit von $s(\mathbf{q}_k)$. Mit Gl. A.79 ergibt sich:

$$P(\mathbf{q}_k) = P(s(\mathbf{q}_k)) = P(\mathbf{q}_{k,0}) = P(\mathbf{q}_{k,1}) = \cdots = P(a(\mathbf{q}_k)) \tag{A.82}$$

$P(s(\mathbf{q}_k))$ soll aus den geschätzten Wahrscheinlichkeiten $P(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{y})$ und der Kontextwahrscheinlichkeit $p_{\mathrm{e}}(\mathbf{q})$ berechnet werden:

$$P(\mathbf{q}_k) = P(i,\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) = p_{\mathrm{e}}(\mathbf{q}_k)P(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{y}) \tag{A.83}$$

Die Wahrscheinlichkeiten $P(h_i)$ können dann aus den lokalen Normierungsbedingungen berechnet werden. Mit den Gln. A.80 und A.78 ergibt sich:

$$\begin{aligned}
P(h_i) &= \sum_{d_j \in \Gamma^-(h_i)} P(d_j)P(h_i|d_j) \\
&= P(h_i|h_{i-1})P(h_{i-1}) + \sum_{\substack{\{k'|a(\mathbf{q}_{k'})\in \\ \Gamma^-(h_i)\}}} P(a(\mathbf{q}_{k'}))\underbrace{P(h_i|a(\mathbf{q}_{k'}))}_{=1} \\
&= P(h_{i-1})\Big(1 - \frac{1}{P(h_{i-1})} \sum_{\substack{\{k'|s(\mathbf{q}_{k'})\in \\ \Gamma^+(h_{i-1})\}}} P(s(\mathbf{q}_{k'}))\Big) + \sum_{\substack{\{k''|a(\mathbf{q}_{k''})\in \\ \Gamma^-(h_i)\}}} P(a(\mathbf{q}_{k''})) \\
&= P(h_{i-1}) - \sum_{\substack{\{k'|s(\mathbf{q}_{k'})\in \\ \Gamma^+(h_{i-1})\}}} P(\mathbf{q}_{k'}) + \sum_{\substack{\{k''|a(\mathbf{q}_{k''})\in \\ \Gamma^-(h_i)\}}} P(\mathbf{q}_{k''})
\end{aligned} \tag{A.84}$$

Auch hier kann also die Rekursivität zur Berechnung ausgenützt werden. Führt man die Rekursion bis zum Anfangsknoten (Wahrscheinlichkeit Eins) durch, so heben sich alle Terme mit negativem Vorzeichen auf, mit Ausnahme der Wahrscheinlichkeiten von Varianten, die den Knoten h_i aussparen. Es ergibt sich:

$$\begin{aligned}
P(h_i) &= 1 - \sum_{k' \in \mathbf{R}_i} P(\mathbf{q}_{k'}) \\
&\text{wobei } R_i = \{k'|n_{\mathrm{a}}(\mathbf{q}_{k'}) < i \wedge n_{\mathrm{r}}(\mathbf{q}_{k'}) > i\}
\end{aligned} \tag{A.85}$$

Die Wahrscheinlichkeit, daß lokal die Referenztranskription realisiert wird, ist also – wie von der Anschauung her klar – das Komplement der Wahrscheinlichkeit, daß eine mögliche Variante angewendet wird.

Damit ist der Zustandsautomat vollständig beschrieben. Es kann zu jeder Referenzaussprache $\boldsymbol{c}$ ein Variantengraph konstruiert werden und die Zustandsübergänge mit

den Gln. A.77 – A.81, A.83 und A.85 gewichtet werden. Die expliziten Wahrscheinlichkeiten $P(\boldsymbol{r}|\boldsymbol{c})$ können gemäß der Standard-Formeln für Markov-Ketten berechnet werden, da jedem $\boldsymbol{r}$ genau eine Zustandsfolge zugeordnet ist.

A.2 Aufbau des MAUS Systems

In diesem Kapitel ist der Aufbau des gesamten Systems zur Segmentierung und Etikettierung von Sprache – kurz MAUS[13] (Münchner AUtomatisches Segmentationssystem) – dargestellt. Die Funktionsweise der beiden Hauptkomponenten, nämlich der Sprachlautmodellierung[14] und der Aussprachemodellierung, wurde in den vorhergehenden beiden Kapiteln detailliert beschrieben. Hier wird auf die weiteren Komponenten und ihre Integration in das Gesamtsystem eingegangen. Alle Komponenten sind modular und in ihrer Funktionsweise voneinander unabhängig aufgebaut. Sie werden in einfacher Weise seriell verschaltet, d.h. das Ergebnis, das eine Komponente ausgibt, wird von der nächsten Komponente unmittelbar als Eingabe verwendet.

Konkret wird nach der Aufbereitung des Datenmaterials zunächst eine Referenztranskription aus der vorliegenden orthographischen Repräsentation der Äußerung erzeugt. Danach wird anhand dieser Referenztranskription und dem verwendeten Aussprachemodell ein Variantengraph erzeugt, der in kompakter Form alle hypothetisierten Aussprachevarianten der Äußerung enthält.

Daran schließt sich der Aufbau eines HMM-Zustandsraumes anhand des Variantengraphen an. In diesem Zustandsraum wird eine Viterbi-Suche nach der optimalen Zustands- bzw. Laut-HMM-Folge durchgeführt. Die optimale Zustands- bzw. Laut-HMM-Folge ist dadurch charakterisiert, daß sie mit der größten Wahrscheinlichkeit die Merkmalsvektorfolge erzeugt hat, die aus der akustischen Repräsentation der Äußerung extrahiert wurde.

Aus der optimalen Folge der Lautmodelle ergibt sich eine Zuordnung von Merkmalsvektor-Teilfolgen zu den Lautmodellen und damit eine Segmentierung: Jede Merkmalsvektor-Teilfolge stellt ein zeitliches Segment dar und ist durch den Laut, den das zugeordnete HMM repräsentiert, klassifiziert. Da der Viterbi-Algorithmus

[13] Der Name wurde am Institut für Phonetik und Sprachliche Kommunikation vom Autor und anderen gewählt.

[14] In dieser Arbeit aus Platzgründen nicht enthalten.

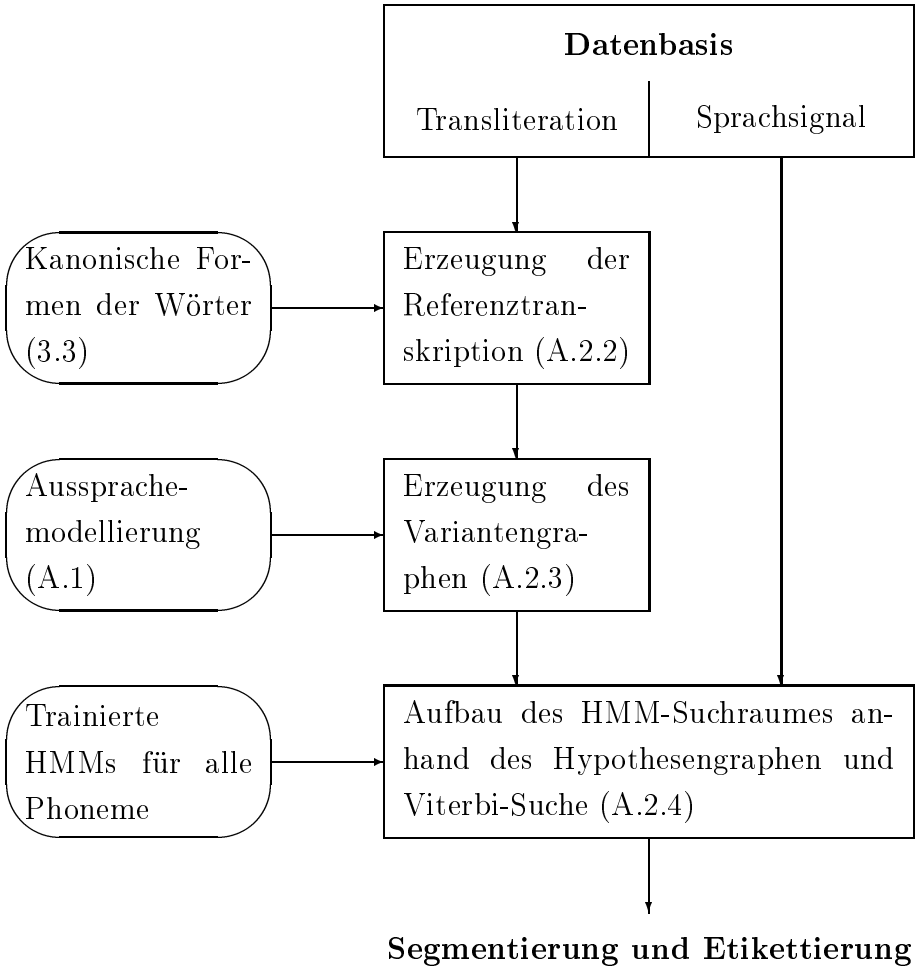


Abbildung A.16: Schematischer Aufbau von MAUS. In Klammern die Abschnitte, in denen die einzelnen Module bzw. Wissensquellen besprochen werden.

zur Suche verwendet wird, ist die Zuordnung der Merkmalsvektor-Teilfolgen zu den Modellen eindeutig, und die Segmente sind nichtüberlappend.

Die eben kurz angerissenen Schritte werden in den nächsten Abschnitten detaillierter beschrieben. Abb. A.16 gibt einen Überblick über das Gesamtsystem.

A.2.1 Aufbereitung des Datenmaterials

Wie im Abschnitt 3 bereits angesprochen, kann eine phonetisch befriedigende automatische Segmentierung und Etikettierung nur mit zusätzlichem Wissen erzeugt werden, das über die bloße akustische Information im Sprachsignal hinausgeht. Eine orthographische Repräsentation des Äußerungsinhaltes, eine Transliteration, stellt, wie erwähnt, eine mit vertretbarem Aufwand manuell generierbare und relativ zuverlässige Wissensquelle dar. Die Transliteration einer Äußerung in dem im Rahmen dieser Arbeit verwendetem Format (vgl. [9],[7]) kann beispielsweise wie folgt aussehen:

```
UTB008: <A> wie <:<#Klopfen> sieht:> es denn direkt <Schmatzen> nach
        Ihrem Kongre"s aus <A> ? <"ahm> vielleicht Mittwoch ,
        Donnerstag , f"unfzehnten , sechzehnten Dezember ?
```

Die Transliteration enthält Markierungen für artikulatorische Phänomene wie Häsitationen (`<äh>,<ähm>,<hm>`), Atmen und merklichen Pausen während der Äußerung (`<A>,<P>`) und Zögern während der Äußerung eines Wortes (`<Z>`). Weiterhin sind artikulatorische und nicht-artikulatorische Geräusche (z.B. `<Schmatzen>,<#Klopfen>`) markiert. Das Format sieht noch eine Fülle von weitern Markierungen für Äußerungsteile vor, die – wie in Spontansprache häufig zu beobachten – von der normalen Grammatik abweichen. Diese finden hauptsächlich in der linguistischen Analyse von Spontansprache Anwendung und wurden hier nicht berücksichtigt.

Spontansprachliche Äußerungen, von denen eine akustische Repräsentation in Form eines gespeicherten Sprachsignals einerseits und eine orthographische Repräsentation in Form einer Transliteration andererseits vorliegen, sind der Ausgangspunkt des Gesamtsystems zur Segmentierung und Etikettierung.

A.2.2 Erzeugung der Referenztranskription

Die Aufgabe des ersten Moduls am Eingang des Systems ist die Umsetzung der Transliteration in eine Referenztranskription. Dazu wird ein kanonisches Aussprachelexikon verwendet, das für jedes Wort, das in den Transliterationen auftritt, eine eindeutige Abbildung seiner orthographischen Repräsentation in eine phonetische Transkription enthält. Die Anforderungen an eine derartige Abbildung wurden bereits in Abschnitt 3.3 spezifiziert.

Jedes Element der Transliteration wird seriell in einen String von Transkriptionssymbolen transformiert, und zwar die einzelnen verschiedenen Elemente in folgender Weise:

Vollständige Wörter werden im Aussprachelexikon gesucht und der String von Transkriptionssymbolen an den Referenzaussprachestring angehängt.

Unvollständige, abgebrochene und unterbrochene Wörter: Eine Transkription, die das Wortfragment lautlich wiedergibt, wird mit Hilfe des Aussprachelexikons erzeugt und an den Referenzaussprachestring angehängt.

Artikulatorische und nichtartikulatorische Geräusche: Ein einziges Symbol für einen nichtsprachlichen Bereich wird an den Referenzaussprachestring angehängt.

Leerzeichen, die Elemente trennen werden durch ein Elementtrennungssymbol (#) im Referenzaussprachestring repräsentiert.

Als Referenzaussprachestring für das oben in Abschnitt A.2.1 angegebene Beispiel ergibt sich dann:

```
<nib> v i: # <nib> z i: t # Q E s # d E n # d I r E k t # <nib> n a: x #
Q i: r @ m # k O N g r E s # Q aU s # <nib> Q E: m # f I l aI C t # m I
t v O x # d O n 6 s t a: k # f Y n f t s e: n t @ n # z E C t s e: n t @
n # d e t s E m b 6
```

A.2.3 Generierung eines Variantengraphen

Die Aufgabe des nächsten Moduls ist es, einen Variantengraphen zur Aussprachemodellierung zu erzeugen. Der Variantengraph beschreibt einen Zustandsautomaten der einen Markov-Prozeß implementiert (vgl. A.1). In den Zuständen des Automaten werden Transkriptionssymbole emittiert. Alle hier verwendeten Zustandsautomaten haben einen eindeutigen, nichtemittierenden Anfangs- und Endzustand und jede Zustandsfolge vom Anfangs- zum Endzustand emittiert genau eine hypothetisierte Transkriptionssymbolfolge.

Die Generierung des Variantengraphen erfolgt durch Anwendung der Ausspracheregeln im Falle des regelbasierten Aussprachemodells bzw. der Mikro-Aussprachevarianten im Falle der statistischen Lautmodellierung auf die Referenztranskription. Der Aufbau der Struktur für beide Modelle ist dabei identisch, die

Gewichtung der Zustandsübergänge ist jedoch modellspezifisch (siehe hierzu Kapitel A.1).

Mit dem Konzept des Variantengraphen können aber auch andere Möglichkeiten der Aussprachemodellierung als die in Kapitel A.1 diskutierten verwendet werden. Die dabei gewonnen Ergebnisse können zur Bewertung der statistischen und der regelbasierten Aussprachemodellierung herangezogen werden.

Ein sinnvoller Vergleichspunkt ist immer die Referenztranskription selbst. In diesem Fall ist der Variantengraph entartet und enthält nur eine einzige Hypothese. Die Etikettierung ist in diesem Fall bereits an dieser Stelle festgelegt und die nachfolgende HMM-Stufe ist gezwungen, für alle Transkriptionssymbole ein Segment im Zeitsignal zu finden, ob es nun tatsächlich vorhanden ist oder nicht. In Kapitel 3 wurde dafür der Terminus *forced alignment* eingeführt.

Um die Leistungsfähigkeit der Lautmodellierung zu testen, wird als Vergleich auch eine freie Phonemerkennung duchgeführt. Der entsprechende Zustandsautomat besitzt für jedes Element aus dem Transkriptionssymbolinventar **S** einen Zustand, der das Symbol emittiert, und Zustandsübergänge von jedem Zustand zu jedem anderen.

A.2.4 Segmentierung und Etikettierung

Die HMM-Stufe bestimmt eine Lösung für die Maximierungsaufgabe aus Kapitel 3 und findet zu einem gegebenen Signal die von der Aussprachemodellierung *und* der Sprachlautmodellierung her wahrscheinlichste Segmentierung. Dazu müssen alle von der Aussprachemodellierung gelieferten Hypothesen akustisch, also mit dem Zeitsignal anhand der Sprachlautmodellierung, bewertet werden.

Die verwendete Darstellung der Aussprachevarianten als Variantengraph ist strukturkompatibel mit den für die Aussprachemodellierung verwendeten HMMs. Es kann für jede Äußerung ein HMM-System durch die Vernetzung der Laut-HMMs entsprechend des Variantengraphen aufgebaut und die optimale Zustandsfolge durch das System mit dem zugrundeliegenden Zeitsignal bestimmt werden. Diese Zustandsfolge verläuft im allgemeinen durch mehrere einzelne Laut-HMMs, und deren Abfolge bestimmt die Etikettierung. Die Segmentierung ist durch die zeitliche Zuordnung des Sprachsignals zu den Laut-HMMs gegeben.

Aufbau eines HMM-Zustandsraumes anhand des Variantengraphen

Der HMM-Stufe steht für jedes Element aus dem Transkriptionssymbolinventar **S** ein HMM zur Verfügung, welches das entsprechende Phonem[15] modelliert. Die Parameter dieser Laut-HMMs können mit den dafür üblichen Methoden geschätzt werden (vgl. dazu z.B. [39]).

Um die Hypothesen, die die Aussprachemodellierung in Form des Variantengraphen liefert, akustisch zu modellieren, wird anhand dieses Graphen ein HMM-System für die zu segmentierende Äußerung aufgebaut. Jedem Knoten des Graphen ist ein Transkriptionssymbol eindeutig zugeordnet, nämlich dasjenige, das im entsprechenden Zustand des aussprachemodellierenden Markovprozesses emittiert wird.

Jeder Knoten im Graphen wird durch eine Instanz (Kopie) des Laut-HMMs, welches das betreffende Transkriptionssymbol akustisch modelliert, ersetzt. Auf diese Weise werden zwei abstrakte Modellierungsebenen, jede auf der Basis endlicher Zustandsautomaten, miteinander kombiniert.[16] Es sei an dieser Stelle darauf hingewiesen, daß ein Laut-HMM in einem derartigen System mehrmals instantiiert[17] werden muß, wenn verschiedene Knoten des Graphen das gleiche Symbol mit sich assoziiert haben. Ein Variantengraph mit K Knoten, die durch HMM-Instanzen $\lambda_0, \lambda_1 \ldots \lambda_{K-1}$ ersetzt wurden, spannt dann ein HMM-System mit einer Zustandsmenge $\mathbf{D} = \{d_{i,j} | 0 \leq i < K \ \wedge \ 0 \leq j < M(\lambda_i\}$ auf, wobei $i = 0$ den Knoten des Variantengraphen indexiert und j den Zustand des Laut-HMMs. Im HMM-System existieren nun zwei Arten von Übergängen.

- Übergänge innerhalb von Laut-HMMs $(d_{i,j}, d_{i,k})$. Die Wahrscheinlichkeiten dieser Übergänge sind durch die akustische Parameterschätzung bestimmt.
- Übergänge zwischen Laut-HMMs $(d_{i,M(\lambda_i)-1}, d_{j,0})$. Für jede Kante des Variantengraphen wird ein solcher Übergang eingeführt und mit der entsprechenden Zustandsübergangswahrscheinlichkeit des aussprachemodellierenden Markovprozesses gewichtet.

---

[15] da ein phonemisches Inventar verwendet wird, haben die Laute, die durch die Symbole repräsentiert werden, Phonemstatus.

[16] Eine gute und weitergehende Betrachtung dieses Konzeptes findet sich in [38]

[17] Instantiieren bedeutet, eine Instanz erzeugen. In der Literatur wird auch oft davon gesprochen, daß ein Modell "geklont" wird.

Durch das HMM-System wird auf diese Weise die akustische Modellierung *und* die der Aussprache gemeinsam repräsentiert.

Im Variantengraph wird die Existenz eines eindeutigen, nichtemittierenden Anfangszustandes und die eines eindeutigen, nichtemittierenden Endzustandes gefordert. Diese beiden Zustände werden zum globalen Anfangs- bzw. Endzustand des HMM-Systems.

Jede Zustandsfolge $\boldsymbol{S}$ vom globalen Anfangs- zum globalen Endzustand gibt eine Laut-HMM-Folge wieder und entspricht damit einer Transkriptionssymbolfolge $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{S})$. Es können mit den aus Kapitel 3 bekannten Formeln Wahrscheinlichkeiten $P(\boldsymbol{O}|\boldsymbol{S})$ dafür angegeben werden, daß das HMM-System ein bestimmtes Sprachmuster $\boldsymbol{O}$ generiert und daraus wiederum die Zustandsfolge – und damit die Segmentierung und Etikettierung – bestimmt werden, die mit der größten Wahrscheinlichkeit ein gegebenes Sprachmuster erzeugt hat.

Viterbi Suche im HMM-System

Um mit Hilfe des aufgebauten HMM-Systems die wahrscheinlichste Segmentierung und Etikettierung zu einer gegebenen Merkmalsvektorfolge $\boldsymbol{O}$ zu finden, in die das Zeitsignal durch die Vorverarbeitung (vgl. B) transformiert wurde, werden alle Zustandsfolgen entsprechender Länge durch das HMM-System betrachtet. Jeder Zustandsfolge $\boldsymbol{S}$ ist eindeutig eine Segmentierung $\boldsymbol{K}$ zugeordnet

$$\boldsymbol{S} = d_{0,0}\,\underbrace{d_{i,0}\ldots d_{i,M(\lambda_i)}}_{\substack{\tau_0 \text{ Zustände}\\ \text{im Modell } \lambda_i\\ k_0=(0,\tau_0,\rho\lambda_i)}}\underbrace{d_{j,0}\ldots d_{j,M(\lambda_j)}}_{\substack{\tau_1 \text{ Zustände}\\ \text{im Modell } \lambda_i\\ k_1=(\tau_0,\tau_1,\rho\lambda_j)}}\ldots d_{K-1,0} \tag{A.86}$$

$$\boldsymbol{K}(\boldsymbol{S}) = k_0 k_1 \ldots k_N$$

Die endliche Menge der Zustandsfolgen $\boldsymbol{S}_j$, $j = 0 \ldots N_\mathrm{S}-1$ die ein bestimmtes Muster $\boldsymbol{O}$ erzeugt haben können, kann also auf eine ebenfalls endliche Menge von Segmentierungen $\boldsymbol{K}_i$, $i = 0 \ldots N_\mathrm{K} - 1$ abgebildet werden. Die Abbildung der Zustandsfolge auf die Segmentierungen ist nicht bijektiv, denn die gleiche Segmentierung kann durch mehrere Zustandsfolgen erzeugt werden. Um die Wahrscheinlichkeit für eine Segmentierung $\boldsymbol{K}_i$ zu finden, muß über alle diese Zustandsfolgen summiert werden.

$$P(\boldsymbol{K}_i, \boldsymbol{O}) = \sum_{\{\boldsymbol{S}_j|\boldsymbol{K}(\boldsymbol{S}_j)=\boldsymbol{K}_i\}} P(\boldsymbol{O}, \boldsymbol{S}_j) \tag{A.87}$$

Um die wahrscheinlichsten Segmentierung zu finden, müßte die Maximierungsaufgabe

$$P(\hat{\boldsymbol{K}}, \boldsymbol{O}) = \max_{\{\boldsymbol{K}_i | 0 \leq i < N_\mathrm{K}\}} P(\boldsymbol{K}_i, \boldsymbol{O}) \tag{A.88}$$

$$\hat{\boldsymbol{K}} = \operatorname*{argmax}_{\{\boldsymbol{K}_i | 0 \leq i < N_\mathrm{K}\}} P(\boldsymbol{K}_i, \boldsymbol{O}) \tag{A.89}$$

gelöst werden. Die Komplexität der Maximierungsaufgabe in dieser Form ist allerdings sehr hoch; zu hoch, um mit heute verfügbaren Rechnern in endlicher Zeit zu einer Lösung zu kommen.

Nähert man die Wahrscheinlichkeit für eine bestimmte Segmentierung durch die Wahrscheinlichkeit der *optimalen* Zustandsfolge an, die diese Segmentierung generiert hat, so vereinfacht sich die Maximierungsaufgabe wesentlich:

$$P(\boldsymbol{O}, \hat{\boldsymbol{K}}) = \max_{\{\boldsymbol{K}_i | 0 \leq i < N_\mathrm{K}\}} \left[\max_{\{\boldsymbol{S}_j | \boldsymbol{K}(\boldsymbol{S}_j) = \boldsymbol{K}_i\}} P(\boldsymbol{O}, \boldsymbol{S}_j) \right] \tag{A.90}$$

$$= \max_{\{\boldsymbol{S}_j | 0 \leq j < N_\mathrm{S}\}} P(\boldsymbol{O}, \boldsymbol{S}_j) \tag{A.91}$$

$$\hat{\boldsymbol{S}} = \operatorname*{argmax}_{\{\boldsymbol{S}_j | 0 \leq j < N_\mathrm{S}\}} P(\boldsymbol{O}, \boldsymbol{S}_j) \tag{A.92}$$

$$\hat{\boldsymbol{K}} = \boldsymbol{K}(\hat{\boldsymbol{S}}) \tag{A.93}$$

Die wahrscheinlichste Zustandsfolge kann effizient mit der Viterbi-Suche bestimmt werden. Die – mit der Viterbi-Näherung – optimale Segmentierung ergibt sich aus der eindeutigen Abbildung der Zustandsfolgen auf die Segmentierungen.

Anhang B

Merkmalsextraktion MFCC

Der folgende Anhang enthält eine kurzgefaßte Beschreibung der *Merkmalsextraktion* bzw. *Vorverarbeitung* des Schalldrucksignals in spektrale Parameter, wie er im MAUS-System Verwendung findet. Es sollte hier betont werden, daß es sich bei den *Mel-Frequency-Cepstral-Coefficiants* (MFCC) um eine etablierte Standard-Technik handelt, die mittlerweile im Bereich der Spracherkennung routinemäßig zum Einsatz kommt.

Bei der Vorverarbeitung wird formal aus N aufeinanderfolgende Abtastwerten $x_i \ldots x_{i+N-1}$ des Sprachsignals ein Vektor $\underline{o}_t$ der Dimension M berechnet. Dabei ist $M < N$, d.h. es findet eine Datenreduktion statt. Die Bereiche, aus denen aufeinanderfolgende Merkmalsvektoren berechnet werden, überlappen sich. In Abb. B.1 ist das Vorgehen schematisch dargestellt.

Die Begrenzung des zeitlichen Bereiches, der einer Spektraltransformation unterworfen wird, entspricht systemtheoretisch einer Fensterung. Das Signal wird im Zeitbereich mit einer Fensterfunktion multipliziert. Diese Fensterfunktion hat die Koeffizienten:

$$\begin{aligned} w_j &\neq 0 \quad \text{für } j = 0 \ldots N-1 \\ w_j &= 0 \quad \text{sonst} \end{aligned} \tag{B.1}$$

Für die Berechnung des Vektors $\underline{o}_t$ werden dann Werte $x'_i = w_{i-tM} x_i$ verwendet. Wie man sieht, sind alle Werte außerhalb des interessierenden Bereiches $tM \leq i < tM + n$ gleich Null. Im einfachsten Fall hat die Fensterfunktion die Koeffizienten $w = 1,\ j = 0 \ldots N-1$. Man spricht dabei von einem Rechteckfenster.

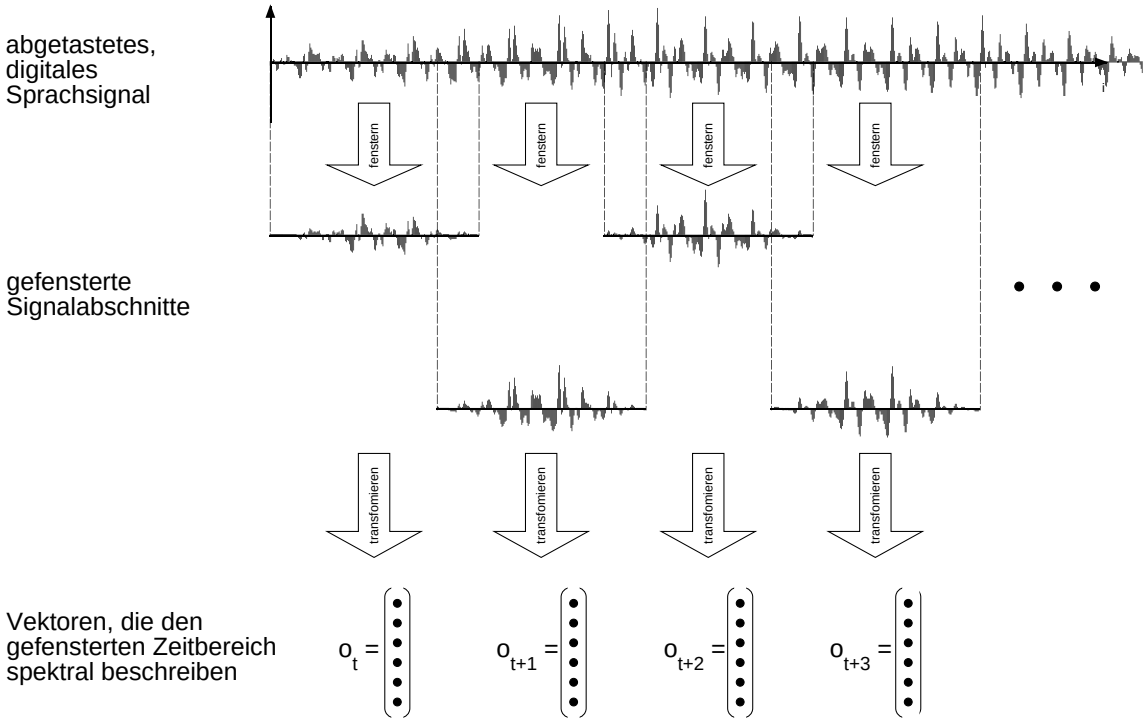


Abbildung B.1: Schematische Darstellung der Vorverarbeitung

Für den Spektralbereich bedeutet eine Fensterung, daß das Spektrum des Signals mit der Spektraltransformation der Fensterfunktion gefaltet wird und somit die spektralen Eigenschaften des Fensters mit in die Berechnung des Datenvektors eingehen. Ein Rechteckfenster hat eher ungünstige spektrale Eigenschaften. Deshalb wird meist – so auch hier – ein sogenanntes *Hamming-Fenster* verwendet, das in dieser Hinsicht besser geeignet ist. Die Koeffizienten des Hamming-Fensters werden wie folgt bestimmt:

$$w_j = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi j}{N-1}\right) & \text{für } j = 0 \ldots N-1 \\ 0 & \text{sonst} \end{cases} \tag{B.2}$$

Die N von Null verschiedenen Werte werden nun einer Spektraltransformation unterworfen werden. In MAUS werden *Mel Frequency Cepstral Coefficients* (MFCC) verwendet.

Das gefensterte Signal wird hierbei zunächst einer Filterbank mit P parallelen Filtern zugeführt. Die Filter haben Bandpaßcharakteristik (Dreieck) und sind äquidistant auf einer Mel-Frequenz-Achse angeordnet. Die Mel-Frequenzachse f_{mel} ist eine Koordinatentransformation der linearen Frequenzachse f und ist definiert durch ([57]):

$$f_{\text{mel}} = 2595 \log_{10}(1 + \frac{f}{700}) \tag{B.3}$$

Die Filterbank wird dadurch implementiert, daß das gefensterte Zeitsignal mittels einer *Fast Fourier Transformation* in den Spektralbereich transformiert wird. Das Spektrum ist dann eine Folge von N komplexen Werten, von denen hier aber nur die logarithmierten Beträge $y_j,\ j = 0 \ldots N-1$ interessieren[1]. Diese werden nun im Spektralbereich mit den P Dreiecksfiltern korreliert. Das Ergebnis ist ein Folge von P Mel-Frequenz-Koeffizienten $m_0 \ldots m_{P-1}$.

$$m_p = \sum_{j=0}^{N-1} g(a_p j - f_p) y_j \tag{B.4}$$

Die Dreiecks-Filterfunktion g ist dabei gegeben durch

[1] Phasenbeziehungen sind für das Sprachverstehen irrelevant

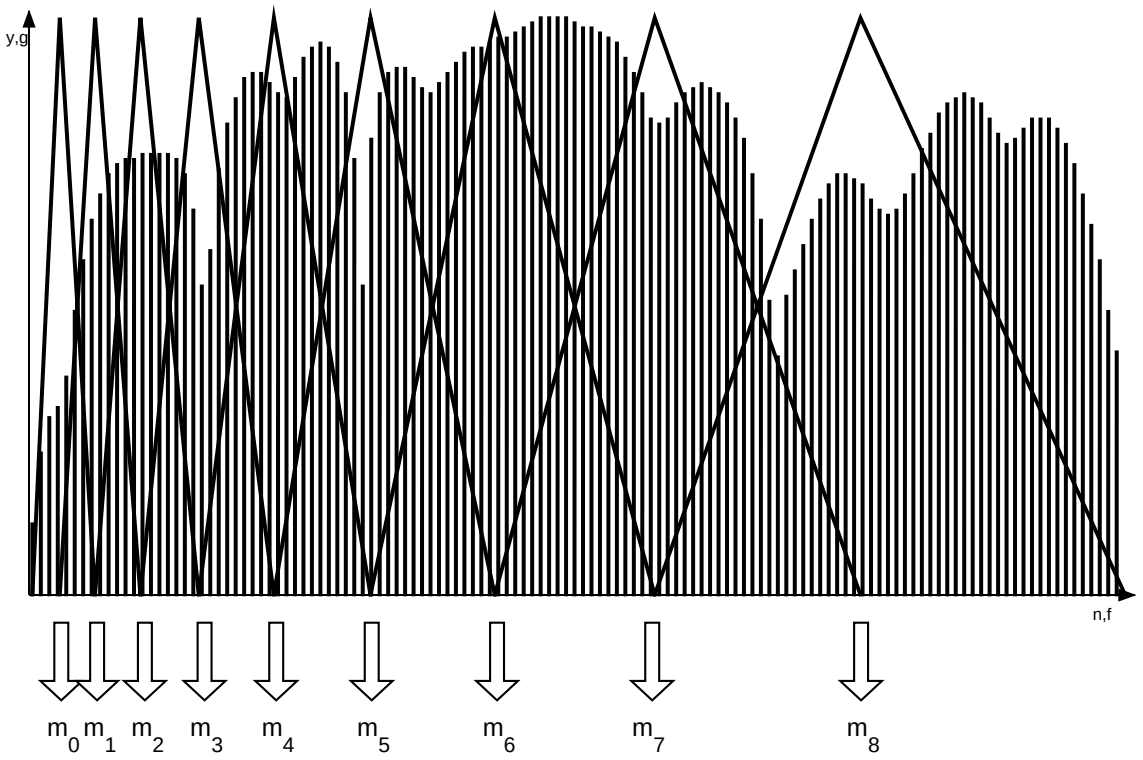


Abbildung B.2: Mel-Filterbank Analyse

$$g(x) = \begin{cases} x+1 & \text{für } -1 \leq x < 0 \\ -x+1 & \text{für } 0 \leq x < 1 \\ 0 & \text{sonst} \end{cases} \tag{B.5}$$

Die Mittelpunkte f_p der Filter werden (im Spektralbereich) so verschoben und die Form des Filters mit a_p so gestreckt, daß sich die in in Abb. B.2 dargestellte Situation ergibt.

Um die MFCCs c_i zu erhalten werden die Koeffizienten m_j mit einer Diskreten Cosinus Transformation in den Cepstralbereich überführt. Es ergeben sich damit dann die Komponenten o_i des Merkmalsvektors $\underline{o}$:

$$o_i = \sqrt{\frac{2}{N}} \sum_{j=0}^{N-1} m_j \cos\left(\frac{\pi i}{N}\left(j+\frac{1}{2}\right)\right) \tag{B.6}$$

Anhang C

Untersuchtes Sprachmaterial

Dieser Anhang enthält eine Kurzbeschreibung der für die empirischen Untersuchungen bzw. Evaluierungen verwendeten Sprachdaten.

C.1 Gelesene Sprache – Phondat 2

Dieser Korpus wurde im Rahmen des vom BMB+F geförderten Projektes *Architectures for Speech and Language* (ASL) in den Jahren 1989 bis 1992 an drei Universitäten – Kiel, Bonn und München – nach einheitlichen Richtlinien aufgenommen und bearbeitet[1]. Der Korpus umfaßt die Sprache von 16 Sprechern, welche jeweils 200 Äußerungen aus der Domäne *Zugauskunft* vom Bildschirm ablesen. Die Sprecher wurden angewiesen, möglichst natürlich, flüssig und nicht im typischen Vorlesestil zu sprechen. Die Inhalte des Korpus sind typische Anfragen, wie sie an ein automatisches Auskunftsystem gestellt werden, z.B.:
"Guten Morgen, ich möchte heute zwischen acht und elf Uhr abends in Hamburg sein."
Fehler wie Versprecher, Hesitationen, Geräusche, etc. waren nicht zugelassen. Die Aufnahmebedingungen waren

- Studio-Umgebung (echoarm)
- Unterschiedliche Mikrophone der Fa. Sennheiser, z.B. MKH 20 P48

[1]Da es sich um eine Weiterführung des vorangegangenen *PhonDat* Projektes handelte, wurde des Name Phondat 2 gewählt

- Sampling-Frequenz 48 kHz
- Digitale Filterung auf 8 kHz und Reduktion der Sampling-Frequenz auf 16 kHz.

Zu allen im Korpus enthaltenen Wörtern existiert eine eindeutige kanosche Zitierform in Form eines Aussprache-Lexikons (SAM-PA).
Zu jeweils einen Subkorpus der gesprochenen Äußerungen wurden manuelle *Wortsegmentierungen, Phonem-Segmentierungen und prosodische Etikettierungen* angefertigt. Die für diese Untersuchung relevanten Phonem-Segmentierungen unfassen jeweils 64 Äußerungen pro Sprecher.
Insgesamt 200 Äußerungen wurden von unabhängigen Bearbeitern mehrfach etikettiert und segmentiert.

C.2 Spontansprache – Verbmobil 1

Der Verbmobil 1 Korpus wurde in den Jahren 1993 bis 1996 als emprische Grundlage für die Entwicklung einer tragbaren Übersetzungshilfe Deutsch - Englisch und Deutsch - Japanisch produziert. An der Erstellung des entgültigen Korpus waren die Universitäten Bonn, Braunschweig, Hildesheim, Karlsruhe, Kiel und München beteiligt. Der Korpus besteht aus Aufnahmen zweier Dialogpartner, die – in Form einer Problemlösungsaufgabe – einen Termin für ein geschäftliches Treffen anhand ihrer beider Terminkalender finden sollen. Mit Hilfe einer *push-to-talk* Einrichtung konnte immer nur einer der beiden Partner sprechen; die Partner waren akustisch getrennt, konnten sich aber sehen. Abgesehen von amerikanischen und japanischen Sprechern enthält Verbmobil 1 Aufnahmen von 779 deutschen Sprechern. Die Sprecher wurden nicht kontrolliert, d.h. alle typischen Effekte, wie sie in spontaner Sprache auftreten, z.B. Versprecher, *false starts*, *repairs*, artikulatorische Geräusche (Husten, Räuspern,...), Hintergrundgeräusche, Hesitationen, etc. wurden mit aufgezeichnet. Die Aufnahmebedingungen waren

- normale, ruhige Büroumgebung (“trockene“ Akustik)
- Head-Set-Mikrophone HD 410 der Fa. Sennheiser
- Sampling-Frequenz 48 kHz
- Digitale Filterung auf 8 kHz und Reduktion der Sampling-Frequenz auf 16 kHz.

Die Nachbearbeitung des Daten umfaßte

- Transliteration nach Verbmobil Konvention
- Manuelle Etikettierung und Segmentierung in SAM-PA ("Kiel Korpus of spontaneous Speech" und eigene Arbeiten)
- Prosodische Etikettierung
- Wort-Segmentierung
- Dialogakt-Etikettierung
- Übersetzungen

In der vorliegenden Untersuchung wurde das manuell etikettierte und segmentierte Material der Universität Kiel ("Kiel Korpus of spontaneous Speech") als *Trainingsmaterial*, und die an der Universität München angefertigten, zum Teil mehrfach bearbeiteten Dialoge als *Entwicklungs- und Testmaterial* verwendet.

Beide Korpora sind über das *Bayerische Archiv für Sprachsignale* (BAS) beziehbar (*www.bas.uni-muenchen.de/Bas*).