



Fachbereich Informatik
AB Natürlichsprachliche Systeme

Studienarbeit

Gewichtung von Aussprachevarianten in der automatischen Erkennung von Spontansprache

Gunnar Evermann

8. März 1998

Inhaltsverzeichnis

1. Einführung	5
2. Aussprachevariabilität	7
2.1. akustische Variation	7
2.2. dialektale Variation	8
2.3. Sprechtempo	9
2.4. gelesene vs. spontane Sprache	9
3. Grundlagen der Spracherkennung	11
3.1. Problemstellung	11
3.2. Merkmalsextraktion	12
3.2.1. MFCC-Merkmale	12
3.2.2. Erweiterungen und Alternativen	15
3.3. Decodierung	15
3.3.1. akustische Modellierung	16
3.3.2. Sprachmodellierung	20
3.3.3. Suchalgorithmen	21
3.4. Daimler Benz Spracherkenner	23
4. Aussprachemodellierung	25
4.1. Einheiten der Modellierung	25
4.1.1. Wortmodelle	25
4.1.2. Wortuntereinheiten	26
4.2. Struktur des Aussprachemodells	28
4.2.1. implizite vs. explizite Modellierung	28
4.2.2. Gewichtung der Varianten	29
4.3. Lexikonerstellung	30
4.4. statisch gewichtete Aussprachevarianten	32
4.4.1. alternative Lexikoneinträge/Aussprachegraphen	33
4.4.2. phonetische Transformationsregeln	34
4.4.3. Entscheidungsbäume	38
4.4.4. andere Repräsentationen	41
4.5. dynamische Gewichtung	42

5. Auswahl der Varianten	45
5.1. verwendetes Korpus	45
5.2. Generierung von Varianten	46
5.2.1. Methoden der Variantengenerierung	46
5.2.2. Auswahl der zu expandierenden Wörter	47
5.3. Struktur der Regeln	48
5.4. Auswahl der Regeln	48
5.4.1. Bewertung eines Regelsatzes	49
5.4.2. Verbesserung des forced alignments	51
5.4.3. MAUS-Regeln	52
5.4.4. Optimierung des Regelsatzes	53
5.5. Ergebnis der Regelauswahl	55
5.5.1. Regelsatz	55
5.5.2. Qualität der Segmentierung	57
5.6. Aufbereitung des Trainingsmaterials	59
6. Erkennungsexperimente	61
6.1. Baseline-System	61
6.2. ungewichtete Varianten	62
6.3. statische Gewichtung	63
6.3.1. wortbasierte Schätzung	64
6.3.2. regelbasierte Schätzung	66
6.4. Zusammenfassung	67
7. Zusammenfassung & Ausblick	69
A. Phonsymbole	71
Literaturverzeichnis	73

1. Einführung

Die Spracherkennungstechnologie hat mittlerweile einen Stand erreicht, der den Einsatz in konkreten Produkten ermöglicht. Allerdings handelt es sich dabei noch um recht eingeschränkte Anwendungen. Die Vision von dem Computer der normale Sprache so gut wie jeder Mensch versteht liegt noch in weiter Ferne.

In den Forschungsprojekten der letzten Jahrzehnte konnte jedoch die Komplexität der Aufgabe in mehreren Dimensionen erhöht werden. Es wurden jeweils spezielle Techniken entwickelt, um auch unter den erschwerten Bedingungen eine sichere Erkennung zu gewährleisten. So ermöglichte die Verwendung statistischer Methoden und der Einsatz von Adaptionstechniken, den Übergang zu sprecherunabhängiger Erkennung. Um die Robustheit gegen Störgeräusche zu erhöhen, wird spektrale Subtraktion oder eine verbesserte Vorverarbeitung verwendet.

Ein Gebiet, auf dem jedoch bislang wenig Fortschritte erzielt wurden, ist die Modellierung von Aussprachevariabilität. So stellen zum Beispiel stark dialektal gefärbte Äußerungen ein erhebliches Problem für die Spracherkennung dar. Im wesentlichen werden in den meisten heutigen Systemen immer noch die Techniken benutzt, die bereits in den 70er Jahren von IBM bzw. im DARPA Speech Understanding Project entwickelt wurden. Lediglich die Modellierung von Koartikulationsphänomenen konnte durch die Einführung kontextabhängiger Wortuntereinheiten verbessert werden. Variationen wie sie etwa in spontaner, schneller Sprache oder in verschiedenen Dialekten auftreten werden jedoch nicht modelliert.

Erst in den letzten Jahren wurden in verschiedenen Forschungsgruppen Versuche unternommen durch eine explizite Aussprachemodellierung die Robustheit zu erhöhen. Diese Entwicklung wurde hauptsächlich durch die Einführung zweier neuer Korpora ausgelöst. Für das *Switchboard*-Korpus wurden Telefongespräche zwischen zwei Personen aufgenommen. Diese Aufnahmen unterscheiden sich stark von den in früheren Projekten verwendeten Aufnahmen von gelesenen Zeitungsartikeln. Phänomene wie Wortabbrüche, Häsitationen und abweichende Aussprachen treten hier sehr häufig auf. In dem *Business-News*-Korpus wurden zum ersten Mal Aufnahmen verwendet, die nicht speziell zur Entwicklung von Spracherkennern aufgenommen wurden. Es handelt sich um Aufnahmen von Nachrichtensendungen aus Radio und Fernsehen. Interessant ist hier insbesondere die Kombination von gelesener, sorgfältig artikulierter Sprache (Nachrichtensprecher) und spontaner Sprache (Interviews, o.ä.).

Es scheint notwendig Methoden zu entwickeln, die es erlauben die Aussprachemo-

dellierung einerseits robuster gegen solche Veränderungen zu machen und andererseits eine Adaption an die konkret beobachteten Aussprachen vorzunehmen. Der Mensch ist in der Lage sich in kürzester Zeit an die Art der Aussprache eines Gesprächspartners anzupassen (extreme Beispiele sind hier Dialekte oder Sprachfehler).

In dieser Arbeit sollen zunächst die verschiedenen zur Zeit diskutierten Herrangehensweisen an die Aussprachemodellierung diskutiert werden. Konkrete Versuche werden mit einer dieser Methoden durchgeführt, um zu untersuchen, wie alternative Aussprachen in einem Spracherkenner modelliert werden können.

2. Aussprachevariabilität

In diesem Kapitel wird diskutiert, welche Arten von Variation im Sprachsignal beobachtet werden, und es wird beschrieben, welche Phänomene in dieser Arbeit untersucht werden.

2.1. akustische Variation

Das entscheidende Problem in der automatischen Spracherkennung ist der hohe Grad an Variabilität, der sich bei sprachlichen Äußerungen zeigt. Sogar Aufnahmen mehrerer Äußerungen derselben Wortfolge stimmen niemals exakt überein. Selbst wenn man von den Veränderungen, die durch Umgebungsgeräusche und den Aufnahmekanal verursacht werden, absieht, so finden sich immer noch erhebliche Unterschiede in dem Sprachsignal.

Als Beispiel sind in Abbildung 2.1 Spektrogramme von zwei Aufnahmen dargestellt. Die beiden Sprecher haben beide die Wortfolge „Montag und Dienstag den“ geäußert. Offensichtliche Unterschiede zeigen sich bezüglich der Länge, der Lautstärke und der Tonhöhe der Äußerungen. Es gibt jedoch auch Unterschiede, die sich nicht so leicht erkennen lassen können. So unterscheiden sich die Äußerungen auch in der *Aussprache*.

Unter *Aussprache* soll in dieser Arbeit die Folge von Lautsymbolen (*Phonen*) verstanden werden, die das Ergebnis einer phonetischen Segmentierung des Signals sind. Jedes Phon wird dabei durch bestimmte Parameter definiert, dies sind z.B. der *Artikulationsort* (labial, apikal, dorsal, etc.), die *Phonation* (oder *Stimmhaftigkeit*) und die *Artikulationsart* (Frikativ, Plosiv, Nasal, etc.). Für eine umfassende Diskussion des Begriffs *Laut* und der Phonetik im allgemeinen sei auf [Kohler 1995] verwiesen. Von supra-segmentalen Faktoren wie der Prosodie wird bei der Festlegung der Aussprache abstrahiert. Ebenso bleiben zahlreiche sprecherspezifische oder situative Faktoren, die sich zum Teil stark auf das akustische Signal auswirken, bei der phonetischen Segmentierung unberücksichtigt.

Eine sehr große Bedeutung spielt in diesem Zusammenhang der Begriff der *Standardaussprache* (auch *kanonische Aussprache* oder *Standardlautung*). Dies ist die für jedes Wort bestimmte Phonfolge, die die gegenwärtige Aussprache eines Wortes durch einen geübten Sprecher des *Hochdeutschen* am besten wiedergibt. Sie orientiert sich stark an der Schriftsprache und soll eine Aussprache darstellen, die von allen Sprechern des Deutschen verstanden und realisiert werden kann. Diese Aussprache stellt jedoch nicht in einem normativen Sinne die *richtige* Aussprache dar, sondern soll rein *deskriptiv* den gegenwärtigen Sprachgebrauch widerspiegeln. Eine detailliertere Diskussion dieses Begriffs kann im Rahmen dieser Arbeit nicht geleistet werden, findet sich aber z.B. in der bereits

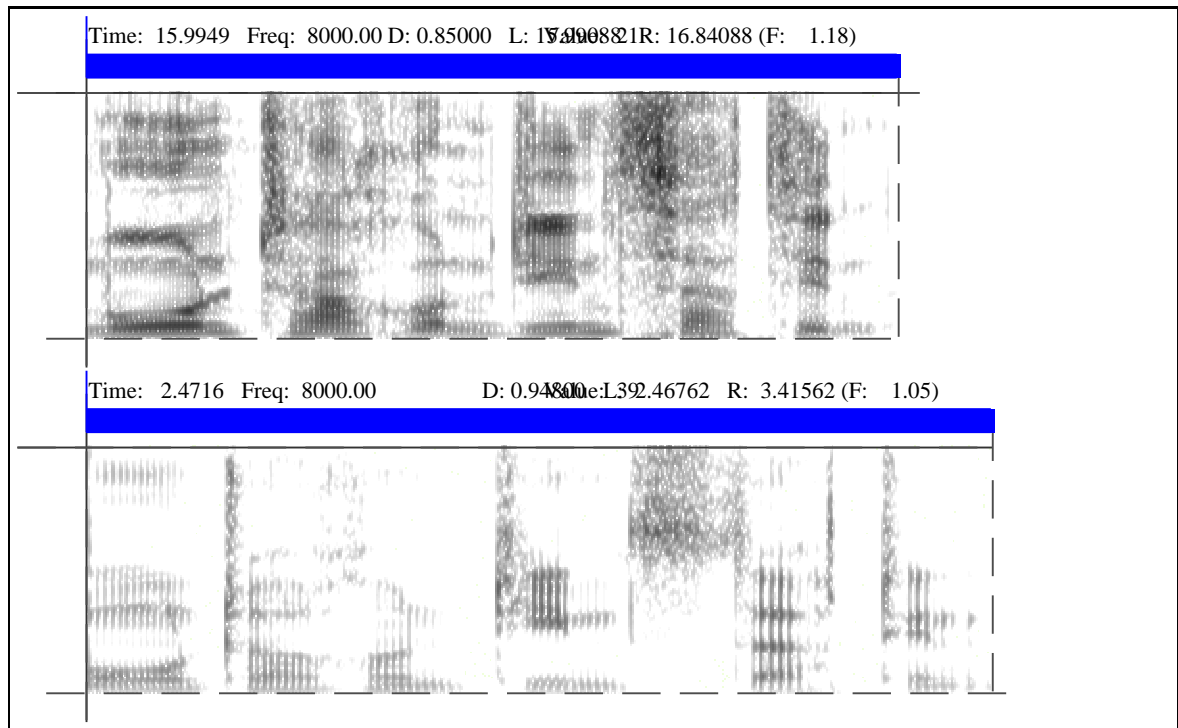


Abbildung 2.1.: Spektrogramme der Äußerung „Montag und Dienstag den“

erwähnten Monographie von Kohler.

In dieser Arbeit wird nur der Aspekt der Aussprachevariabilität untersucht. Die anderen Aspekte der Variabilität sprachlicher Äußerungen (wie alle *klassischen* Sprecheradaptionstechniken, die z.B. eine Adaption an die Stimmcharakteristika vornehmen) werden nicht explizit betrachtet, obwohl Standardtechniken verwendet werden, um die Modellierung robuster zu machen.

In den folgenden Abschnitten werden einige der Phänomene beschrieben, die bewirken, daß die tatsächlich beobachteten Aussprachen von den Standardaussprachen abweichen.

2.2. dialektale Variation

Eine offensichtliche Variation der Aussprachen ist auf die Verwendung verschiedener Dialekte zurückzuführen. Da sich die Standardaussprache am Hochdeutschen bzw. der Schriftsprache orientiert, ergeben sich starke Abweichungen zu den Aussprachen, die in manchen Dialekten als *normal* angesehen werden.

Sieht man einmal von syntaktischen Unterschieden und dialektspezifischen idiomatischen Wendungen ab, so kann ein Dialekt relativ einfach modelliert werden, indem die

Standardaussprache an die dialektspezifischen Aussprachen angepaßt wird. Ein großes Problem ist jedoch, daß in der Regel nicht davon ausgegangen werden kann, daß der Dialekt des Sprechers a-priori bekannt ist. Somit wäre es notwendig, Aussprachevarianten für alle relevanten Dialekte vorzusehen. Außerdem verwenden viele *Dialektsprecher* abhängig von der konkreten Sprechsituation ihren Dialekt mehr oder weniger stark ausgeprägt. Man kann beobachten, daß Sprecher sich in einer Konversation mit Personen aus anderen Dialektregionen bemühen, *hochdeutsch* zu sprechen. Eine relativ plausible Annahme ist somit, daß auch ein Nutzer eines automatischen Dialogsystems sich bemühen wird, möglichst die allgemein als Standard akzeptierte Aussprache zu verwenden.

Da es bislang auch keine Sprachkorpora mit großen Mengen an Sprachaufnahmen in verschiedenen Dialekten des Deutschen gibt, wurde der Einfluß der Dialekte meist ignoriert. In dem Verbmobil-Korpus zum Beispiel wurden Aufnahmen von Sprechern aus vielen verschiedenen Teilen Deutschlands gemacht, dies schlägt sich jedoch kaum in den beobachteten Aussprachen nieder. So kann hier höchstens von einer *dialektalen Färbung* gesprochen werden.

Dialektale Variationen werden in dieser Untersuchung nicht explizit betrachtet, allerdings lassen sich einige der modellierten Varianten wahrscheinlich auf den Dialekt der Sprecher zurückführen.

2.3. Sprechtempo

Bei Erhöhung der Sprechgeschwindigkeit sinkt nicht nur die Länge der einzelnen Segmente, sondern es finden auch andere Prozesse statt. So besteht die Möglichkeit, daß Phone der Standardlautung komplett ausgelassen werden (*Elisionen*). Ein anderer Prozeß, der insbesondere bei schneller Artikulation beobachtet wird, ist die *Assimilation*. Hierbei wird eines der oben beschriebenen phonetischen Merkmale von einem Nachbarphon übernommen. Durch diese Angleichung z.B. des Artikulationsortes wird der Bewegungsablauf der Artikulation insgesamt einfacher und kann damit auch schneller ausgeführt werden.

Insgesamt gilt, daß bei einer schnelleren Artikulation die Exaktheit der Bewegungsabläufe sinkt und insbesondere die eher trägen Artikulationsorgane wie z.B. der Zungenrücken einfachere Bewegungen ausführen. Die aufgrund der geringeren Trägheit beweglichere Zungenspitze jedoch kann meistens den *normalen* Ablauf in der hohen Geschwindigkeit ausführen.

In Kapitel 5 werden in Zusammenhang mit der Optimierung der Modellierung konkrete Beispiele dieser Prozesse diskutiert.

2.4. gelesene vs. spontane Sprache

Die spezifischen Probleme, die beim Übergang von gelesener zu spontaner, ungeplanter Sprache auftreten, sind bisher noch nicht systematisch im Detail untersucht worden. Offensichtlich treten in spontaner Sprache auf syntaktischer Ebene andere Strukturen

auf, aber dies hat nicht notwendigerweise Einfluß auf die Artikulation der einzelnen Wörter. Außerdem finden sich Phänomene wie Wortabbrüche und Restarts wesentlich häufiger als in vorgelesener oder sorgfältig geplanter Sprache.

Es wurde von vielen Forschungsgruppen beobachtet, daß die Erkennung spontaner Äußerungen ein wesentlich größeres Problem darstellt als die Erkennung gelesener Texte. So wurden z.B. auf dem *Switchboard-Korpus* noch vor wenigen Jahren bei Verwendung der normalen, auf gelesener Sprache entwickelten Erkennen Fehllraten von ca. 50% beobachtet.

Insgesamt muß davon ausgegangen werden, daß spontane Sprache einen wesentlich höheren Grad der artikulatorischen Verschleifung aufweist. So werden vor allem die Wortgrenzen noch weniger als in gelesener Sprache beachtet, und insbesondere häufige, idiomatische Phrasen werden zusammengezogen und als eine Einheit artikuliert (z.B. *haben wir* → *hamma*). Außerdem zeigt es sich, daß die kanonische Aussprache zum Teil stark verkürzt wird, so daß die Verwendung der kanonischen Formen in informellen Gesprächen schon fast als Überartikulation empfunden wird.

3. Grundlagen der Spracherkennung

In diesem Kapitel sollen kurz die grundlegenden Probleme der automatischen Spracherkennung und die wichtigsten Techniken zur Lösung dieser Probleme vorgestellt werden. Der Schwerpunkt liegt dabei auf den Techniken, die direkt mit der Modellierung der Aussprache zusammenhängen. Andere Problemkreise können hier nur kurz angerissen werden. Für eine umfassendere Diskussion sei auf die Monographien [Schukat-Talamazzini 1995] und [Lee 1989] verwiesen, auf den die folgende Darstellung auch in weiten Teilen basiert.

3.1. Problemstellung

Das Ziel der automatischen (maschinellen) Spracherkennung ist es, aus dem akustischen Signal einer sprachlichen Äußerung die gesprochene Wortfolge zu rekonstruieren. Die Entwicklung von Verfahren, die diese menschliche Fähigkeit nachbilden, ist nach wie vor ein aktives Forschungsfeld, in dem es zahlreiche Probleme gibt, die noch völlig ungeklärt sind. Es gibt zur Zeit noch kein System, das das menschliche Vorbild in puncto Zuverlässigkeit und Robustheit auch nur annähernd erreicht.

Seit Ende der siebziger Jahre dominieren *statistische Verfahren* die Forschung. Davor wurde lange Zeit ein mehr wissensbasierter (KI-orientierter) Ansatz verfolgt, der versuchte die Funktionsweise der menschlichen Spracherkennung explizit in *Regeln* zu kodieren, die dann zur Konstruktion eines automatischen Spracherkennungssystems benutzt wurden. Jedoch zeigte es sich, daß mit dem statistischen Ansatz, der diese *Regeln* nur implizit aus einer sehr großen Trainingsstichprobe lernt, wesentlich zuverlässigere Erkenner entwickelt werden konnten. Im folgenden werden daher hauptsächlich die vorherrschenden statistischen Verfahren beschrieben.

Bei der Spracherkennung handelt es sich um ein *Mustererkennungsproblem*, in dem es gilt, eine Beobachtung (das akustische Signal) korrekt zu klassifizieren, d.h. es der richtigen Klasse (der gesprochenen Wortfolge) zuzuordnen (siehe Abbildung 3.1. Dieser Prozeß wird üblicherweise in zwei Stufen zerlegt: In der ersten Stufe (der *Merkmalsextraktion*) wird das Signal zunächst in eine maschinell verarbeitbare Form gebracht, wobei versucht wird, eine Repräsentation zu gewinnen, in der die Unterscheidung der verschiedenen Laute möglichst leicht vorzunehmen ist. Hierbei wird gleichzeitig versucht, die Menge der anfallenden Daten so klein wie möglich zu halten, ohne jedoch relevante Information zu verlieren (siehe Abschnitt 3.2). In der zweiten Stufe (der eigentlichen *Klassifikati-*

on) werden diese Daten Wortfolgen zugeordnet. Dieser Vorgang wird in Anlehnung an das informationstheoretische Modell der Spracherkennung *Decodierung* genannt (siehe Abschnitt 3.3).

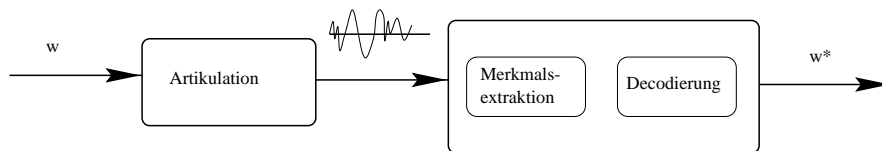


Abbildung 3.1.: Modell der Spracherkennung

3.2. Merkmalsextraktion

Ziel der Merkmalsextraktion (im Kontext der Spracherkennung häufig auch (*akustische Vorverarbeitung* genannt) ist es, eine parametrische Repräsentation des Signals zu generieren, die einerseits so kompakt wie möglich ist, aber andererseits noch alle zur Klassifikation nötigen Informationen enthält. Insbesondere wird versucht, die Variabilität, die auf verschiedene Mikrofoncharakteristiken, Umgebungsgeräusche, verschiedene Sprecher oder ähnliches zurückzuführen ist, zu reduzieren. Im folgenden wird ein kurzer, informeller Überblick über die zur Zeit dominierende Technik der *mel frequency cepstral coefficients* (kurz *MFCC*) gegeben. Erweiterungen und andere Verfahren werden in Abschnitt 3.2.2 beschrieben. Eine ausführlichere Darstellung findet sich z.B. in [Schukat-Talamazzini 1995].

3.2.1. MFCC-Merkmale

Die erste Stufe in der Codierung der Sprachsignale bildet immer die *Digitalisierung*. Hierbei wird das analoge Sprachsignal, das mit einem Mikrofon aufgenommen wurde, mittels eines *A/D-Wandlers* in eine digitale Repräsentation umgewandelt.

Eine in der Verarbeitung von Sprachsignalen übliche (aber nicht korrekte) Annahme ist, daß das Signal in kurzen Zeitintervallen näherungsweise stationär ist. Daher wird die Folge von Abtastwerten in sich überlappende Segmente (*frames*) zerlegt, die mit einer *Fensterfunktion* (z.B. Rechteck oder Hamming-Fenster) gewichtet werden. Die spektralen Eigenschaften dieser frames werden anschließend in jeweils einem *Merkmalsvektor* codiert.

Die frames werden durch eine Bank von Bandpaßfiltern (siehe Abbildung 3.2 gefiltert. Hieraus erhält man eine Schätzung für die Energie pro frame in den verwendeten Frequenzbändern. Diese *Filterbank* besteht dabei meist aus – auf der *mel*-Skala äquidistanten – Dreiecksfiltern. Üblicherweise werden hierbei etwa 20 Filter verwendet.

Aufgrund der Tatsache, daß Menschen in der Lage sind, niedrige Frequenzen feiner aufzulösen als hohe, nehmen die Breite und der Abstand der Filter mit zunehmender Mittelfrequenz exponentiell zu. Die Verwendung der *mel*-Skala ist ein Beispiel dafür, daß die Erforschung der *menschlichen* Spracherkennung Erkenntnisse liefert, die erfolgreich in statistische Erkennungssysteme integriert werden können.

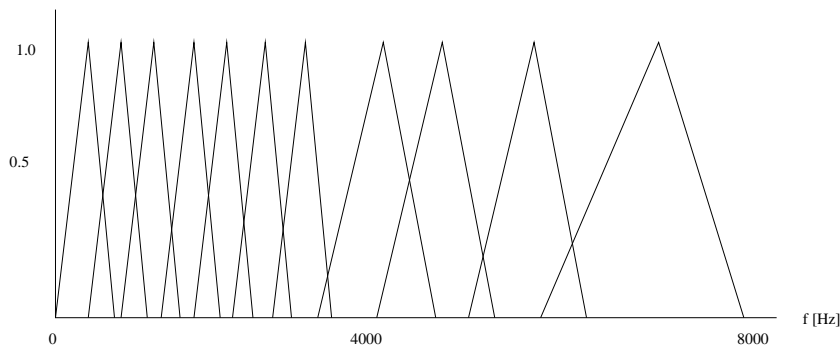


Abbildung 3.2.: Filterbank

Da eine direkte Anwendung der Filter im Zeitbereich zu aufwendig wäre, wird das Signal zunächst mit Hilfe einer Fourier-Transformation transformiert und die Filterbank dann auf dem Betragsspektrum im Frequenzbereich simuliert. Die so errechneten Koeffizienten werden *mel frequency coefficients* (kurz *MFC*) oder *mel-Spektrum* genannt und können bereits als Merkmale für den Klassifikator benutzt werden.

In der Regel wird das *mel-Spektrum* jedoch noch einer Logarithmierung und einer inversen Fourier-Transformation unterworfen. Dieser Verarbeitungsschritt ist aus zwei Gründen sinnvoll: Der erste Grund basiert auf dem *source-filter*-Modell der Spracherzeugung, das von Gunnar Fant entwickelt wurde. In Fants Modell der Spracherzeugung wird ein Anregungssignal (z.B. von der Glottis erzeugt) durch ein lineares System (Vokaltrakt und Lippenabstrahlung) gefiltert. Die beiden Anteile des Signals (Anregungssignal und Impulsantwort des Filters) können durch die Logarithmierung und inverse Fourier-Transformation getrennt werden. Der zweite Grund ist, daß durch diese Transformation die Koeffizienten relativ gut dekorreliert werden, was die statistische Modellierung erheblich vereinfacht.

In der Regel wird die inverse Fourier-Transformation durch eine *Kosinus-Transformation* berechnet, wobei die Koeffizienten höherer Ordnung weggelassen werden (*Lifiting*). Außerdem wird in den meisten Systemen der nullte Cepstralkoeffizient durch ein logarithmisches Energiemaß ersetzt.

Das so transformierte Spektrum wird *Cepstrum* genannt. Wenn es auf dem Ergebnis einer *mel*-Filterbank aufsetzt, erhält man die *mel-frequency cepstral coefficients*.

In Abbildung 3.3 wird noch einmal ein Überblick über alle verwendeten Verarbeitungsschritte gegeben.

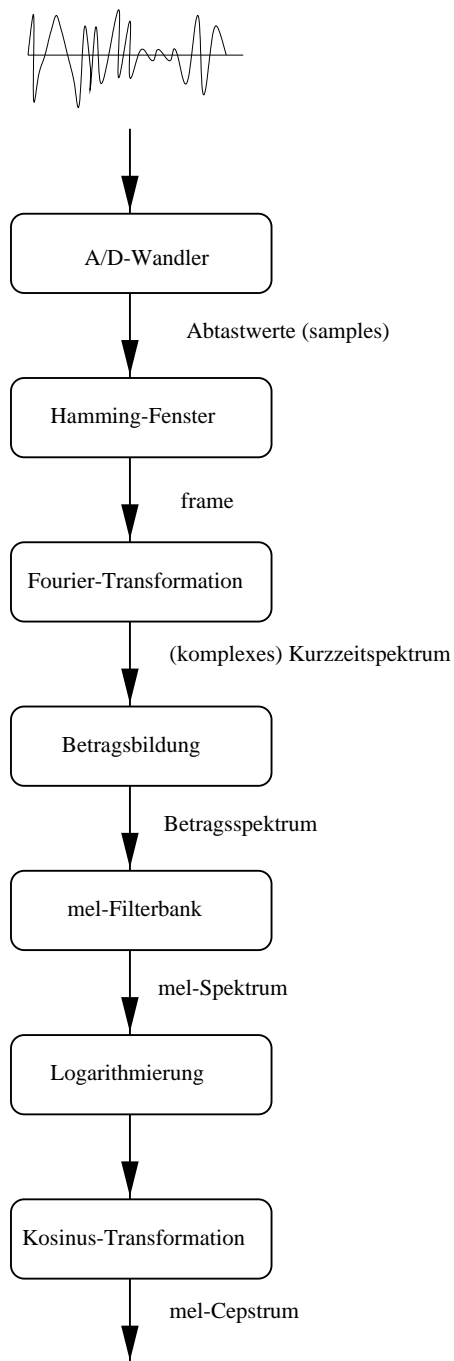


Abbildung 3.3.: MFCC-Merkmalsextraktion

3.2.2. Erweiterungen und Alternativen

Basierend auf *MFCC*-Merkmalen können leistungsfähige Spracherkennungssysteme aufgebaut werden, um jedoch die Erkennungsleistung und die Robustheit des Systems gegen Störeinflüsse weiter zu verbessern, wurden verschiedene Erweiterungen vorgeschlagen. An dieser Stelle sollen nur zwei Techniken beschrieben werden, die heutzutage in den meisten Systemen verwendet werden. Es handelt sich dabei um die *cepstrale Mittelwertsubtraktion* und die Verwendung *dynamischer Merkmale*.

Es hat sich gezeigt, daß Systeme, die direkt cepstrale Merkmale verwenden, relativ empfindlich auf eine Veränderung des akustischen Kanals reagieren, insbesondere die Verwendung eines anderen Mikrofons führt zu einer drastischen Verschlechterung der Erkennungsleistung. Handelt es bei dem akustischen Kanal in guter Näherung um ein lineares, zeitinvariantes System, so verursacht er im Cepstrum nur eine konstante additive Störung, die durch eine Subtraktion des Mittelwerts des Cepstrums über einen längeren Zeitraum beseitigt werden kann. Da hierdurch natürlich auch der mittlere Cepstralwert der „ungefilterten“ Sprache subtrahiert wird, muß der Mittelwert möglichst robust geschätzt werden. Üblicherweise wird der Mittelwert einer kompletten Äußerung subtrahiert, bevor mit der eigentlichen Erkennung begonnen wird.

Ein weiteres Problem ist, daß in der *Decodierung* (siehe Abschnitt 3.3) angenommen wird, daß die Merkmalsvektoren aufeinanderfolgender *frames* stochastisch unabhängig sind. Da die Merkmalsvektoren nur statische Informationen enthalten, wird die zeitliche Veränderung des Sprachsignals nicht erfaßt. Dieses Manko kann durch die Verwendung von *dynamischen Merkmalen* gemildert werden. Der weitverbreitetste Ansatz ist hierbei das Hinzufügen der ersten und zweiten Ableitung der statischen Cepstralparameter zu dem Merkmalsvektor. Die Schätzung der Ableitung kann durch einfache Bildung von Differenzen oder durch lineare Regression erfolgen. Die zusätzlichen Merkmale werden Δ - bzw. $\Delta\Delta$ -Parameter genannt.

3.3. Decodierung

Als Ergebnis der oben beschriebenen Merkmalsextraktion erhält man eine Folge von *Merkmalsvektoren* $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. Diese bilden die Eingabe für die *Decodierung*, in der versucht wird, die tatsächlich gesprochene Wortfolge $\mathbf{w} = w_1, w_2, \dots, w_m$ zu bestimmen. In dem vorherrschenden statistischen Paradigma wird \mathbf{X} als Ergebnis eines stochastischen Prozesses betrachtet, der – gegeben die Wortfolge – die Merkmalsvektoren „produziert“. Daher wird während der Erkennung versucht, die Wortfolge $\hat{\mathbf{w}}$ zu finden, deren *a-posteriori Wahrscheinlichkeit* (gegeben die beobachteten Merkmalsvektoren) am größten ist:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} P(\mathbf{w}|\mathbf{X}) \quad (3.1)$$

Diese Entscheidungsvorschrift nennt man auch *maximum a-posteriori (MAP) Kriterium*. Da jedoch die a-posteriori Wahrscheinlichkeiten nicht bekannt sind, wird Gleichung 3.1 zunächst mit Hilfe der Bayes-Formel in eine geeignetere Form gebracht:¹

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} P(\mathbf{w}|\mathbf{X}) = \operatorname{argmax}_{\mathbf{w}} \frac{p(\mathbf{X}|\mathbf{w})P(\mathbf{w})}{p(\mathbf{X})} \quad (3.2)$$

Da die a-priori Wahrscheinlichkeitsdichte der Merkmalsvektoren $p(\mathbf{X})$ für eine gegebene Äußerung konstant über alle Wortfolgen \mathbf{w} ist, beeinflußt sie die Entscheidung nicht und kann bei der Maximierung weggelassen werden:

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})P(\mathbf{w}) \quad (3.3)$$

Das Problem der Spracherkennung reduziert sich somit auf eine möglichst robuste Schätzung dieser beiden Funktionen. Der erste Faktor $p(\mathbf{X}|\mathbf{w})$ modelliert die akustische Realisierung der Äußerung, während der zweite Faktor $P(\mathbf{w})$ die a-priori Wahrscheinlichkeit einer Wortfolge angibt. Die beiden Verteilungen werden in der Regel getrennt voneinander geschätzt und optimiert. Dies hat zwar den Vorteil, daß das Gesamtproblem in überschaubare Teilprobleme zerlegt wird, aber hinsichtlich der *globalen* Optimalität der Schätzung wäre eine gemeinsame Optimierung vorzuziehen (siehe z.B. [Bourlard et al. 1996]).

Nähere Details zur Modellierung und verwendeten Schätzverfahren finden sich in den folgenden beiden Abschnitten (zur akustischen Modellierung in 3.3.1 und zur Sprachmodellierung in 3.3.2). In Abschnitt 3.3.3 wird schließlich beschrieben, wie die Maximierung in Gleichung 3.3 praktisch ausgeführt werden kann.

3.3.1. akustische Modellierung

Ziel der akustischen Modellierung ist es, die Wahrscheinlichkeitsdichte $p(\mathbf{X}|\mathbf{w})$ für alle Wortfolgen \mathbf{w} anzugeben. Da hierfür keine a-priori besonders plausible Verteilung existiert, muß sie anhand einer möglichst großen *Trainingsstichprobe* geschätzt werden.

Es werden jedoch starke Einschränkungen hinsichtlich der Struktur dieser Verteilung gemacht, um das *Trainingsproblem* auf die Schätzung einer endlichen Zahl von Parametern zu reduzieren. Nicht-parametrische Methoden spielen hingegen heutzutage in der Spracherkennung nur noch eine untergeordnete Rolle. Die früher dominierende Technik

¹An dieser Stelle ist zu beachten, daß es sich bei $p(\mathbf{X}|\mathbf{w})$ und $p(\mathbf{X})$ um kontinuierliche Wahrscheinlichkeitsdichtefunktionen (*probability density functions, pdf*) und nicht Wahrscheinlichkeiten oder Wahrscheinlichkeitsverteilungen handelt. In vielen Arbeiten wird diese Unterscheidung nicht konsequent gemacht, was teilweise für erhebliche Verwirrung sorgt. Hier soll versucht werden, die Notation möglichst exakt zu halten, um Approximationen, die vorgenommen werden, deutlich zu machen, ohne die gesamte Darstellung auf diskrete Symbolfolgen \mathbf{X} zu beschränken. Im folgenden werden Dichten immer mit p bezeichnet, während P für Wahrscheinlichkeiten oder Verteilungen vorbehalten bleibt.

des *dynamic time warping* (kurz: DTW), die die zu klassifizierende Äußerung mit allen Trainingsäußerungen vergleicht, wird heute nur noch in Spezialanwendungen eingesetzt (z.B. in Telefonsystemen mit benutzererweiterbarem Wortschatz). Eine Beschreibung der Anwendung des DTW-Verfahrens in der Einzelworterkennung findet sich z.B. in [Sakoe und Chiba 1978].

Die akustische Wahrscheinlichkeitsdichte $p(\mathbf{X}|\mathbf{w})$ wird in der Regel durch *hidden markov models* (kurz HMM) approximiert. Hierbei wird die Wortfolge \mathbf{w} durch eine Folge von HMMs $\lambda(\mathbf{w})$ modelliert. Die Wortfolge wird dabei in der Regel noch in *Wort-untereinheiten* (z.B. Phone) zerlegt, den dann jeweils ein HMM zugeordnet wird. Die Details dieser Abbildung und ihre Relevanz für die Aussprachemodellierung werden in Abschnitt 4.1 diskutiert.

Da sich die Folge von HMMs $\lambda(\mathbf{w})$ zu einem großen HMM λ verketteten läßt, wird in der folgenden Darstellung in der Regel das Argument \mathbf{w} weggelassen und nur λ verwendet.

Aufgrund der zentralen Bedeutung, die HMMs in der Spracherkennung haben, soll an dieser Stelle eine (relativ informelle) Definition gegeben werden. Anschließend werden die für die Spracherkennung wichtigsten Algorithmen kurz vorgestellt.

HMM Definition

Ein HMM λ ist ein Modell eines zweistufigen stochastischen Prozesses und wird benutzt, um den Wert der Dichte $p(\mathbf{X}|\lambda)$ für eine Folge von Merkmalsvektoren $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_T$, mit $\mathbf{x}_i \in \mathbf{R}^D$, zu berechnen.

Ein HMM besteht aus eine Menge von Zuständen $\mathcal{Q} = \{1, \dots, N\}$ und gewichteten Transitionen zwischen diesen Zuständen. Der zugrundeliegende (stationäre) Markov-Prozeß erster Ordnung kann jedoch nicht direkt beobachtet werden. Aus diesem Grund wird die Folge $\mathbf{q} = q_1, \dots, q_T$ der Zustände, die das System einnimmt, auch *versteckte* Zustandfolge (*hidden state sequence*) genannt. Die Transitionsgewichte a_{ij} geben die Wahrscheinlichkeit an, daß ein Übergang von Zustand i nach Zustand j stattfindet und werden als Matrix $\mathbf{A} = [a_{ij}]$ notiert.

$$a_{ij} = P(q_{t+1} = j | q_t = i) \quad (3.4)$$

In der Sprachverarbeitung werden hauptsächlich HMMs benutzt, in den Transitionen zu Zuständen mit kleinerem Index nicht möglich sind. Für diese *Links-Rechts-Modell* genannten HMMs hat die Transitionsmatrix die Form einer oberen Dreiecksmatrix.

Die initiale Zustandsverteilung ist durch den Vektor $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)$, mit $\pi_i = P(q_1 = i)$, gegeben.

Jedem Zustand i ist eine Ausgabe-Verteilungsdichte b_i zugeordnet, für die gilt:

$$b_i(\mathbf{x}) = p(\mathbf{x} | q_t = i) \quad (3.5)$$

Meist werden hier Normalverteilungen oder Linearkombinationen von Normalverteilungen (sogenannte *Mischverteilungen*, *mixture gaussians*) verwendet, d.h.:

$$b_i(\mathbf{x}) = \sum_{k=1}^K c_{ik} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}) \quad \text{mit} \quad \sum_{k=1}^K c_{ik} = 1 \quad (3.6)$$

Ein HMM $\boldsymbol{\lambda}$ wird also charakterisiert durch Angabe der initialen Zustandsverteilung $\boldsymbol{\pi}$, der Transitionsmatrix \mathbf{A} und der Emissionsdichten \mathbf{B} , d.h. $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$.

Die gesuchte Wahrscheinlichkeit der Merkmalsfolge erhält man durch erneutes Aufteilen und Summation über alle Zustandsfolgen:

$$p(\mathbf{X}|\boldsymbol{\lambda}) = \sum_{\mathbf{q} \in \mathcal{Q}^T} p(\mathbf{X}|\mathbf{q}, \boldsymbol{\lambda}) P(\mathbf{q}|\boldsymbol{\lambda}) \quad (3.7)$$

$$= \sum_{\mathbf{q} \in \mathcal{Q}^T} \pi_{q_1} b_{q_1}(\mathbf{x}_1) \prod_{t=2}^T a_{q_{t-1}q_t} b_{q_t}(\mathbf{x}_t) \quad (3.8)$$

Um diese Berechnung effizient durchführen zu können definiert man zwei Hilfsgrößen: die sogenannte *Vorwärtswahrscheinlichkeit* $\alpha_t(i)$ und die *Rückwärtswahrscheinlichkeit* $\beta_t(i)$:²

$$\alpha_t(i) = p(\mathbf{x}_1, \dots, \mathbf{x}_t, q_t = i | \boldsymbol{\lambda}) \quad (3.9)$$

$$\beta_t(i) = p(\mathbf{x}_{t+1}, \dots, \mathbf{x}_T | q_t = i, \boldsymbol{\lambda}) \quad (3.10)$$

Die beiden Größen lassen sich rekursiv sehr effizient (in $O(TN^2)$) berechnen. Die akustische Wahrscheinlichkeit ergibt sich als Summe des Produkts der beiden Faktoren (für ein beliebiges t) über alle Zustände.

$$p(\mathbf{X}|\boldsymbol{\lambda}) = \sum_{i \in \mathcal{Q}} p(\mathbf{X}, q_t = i | \boldsymbol{\lambda}) = \sum_{i \in \mathcal{Q}} \alpha_t(i) \beta_t(i) \quad (3.11)$$

Training

In der Trainingsphase wird versucht, auf Basis eine Trainingsstichprobe einen Parametersatz $\boldsymbol{\lambda}$ zu finden, der eine möglichst zuverlässige Erkennung ungesehener Äußerungen erlaubt. Da es sehr schwierig ist, direkt die zu erwartende Fehlerrate als Optimierungskriterium zu benutzen, wird üblicherweise ein anderes, suboptimales Kriterium verwendet (einen guten Überblick über verschiedene Kriterien bietet [Valtchev 1995]). In den

²hierbei handelt es sich natürlich auch um Dichtefunktionen

meisten Systemen wird das *Maximum-Likelihood-Kriterium* (ML-Kriterium) verwendet. Hierbei gilt es, die sogenannte *Likelihood-Funktion* \mathcal{L} durch geeignete Wahl des Parameters λ zu maximieren:

$$\mathcal{L}_\lambda = p(\mathbf{X}|\lambda) = \sum_{q \in Q} p(\mathbf{X}, q|\lambda) \quad (3.12)$$

Die Likelihood \mathcal{L}_λ ist ein Maß für die Wahrscheinlichkeit der Trainingsstichprobe bezüglich eines gegebenen Modellsatzes λ .

Da für dieses Optimierungsproblem jedoch keine Lösung in geschlossener Form existiert, wird ein iteratives *Expectation-Maximization*-Verfahren angewandt. Dieses unter dem Namen *Baum-Welch*- (oder *forward-backward*-)Algorithmus bekannte Verfahren schätzt ausgehend von einem Parametersatz λ neue Parameter λ' mit $\mathcal{L}_{\lambda'} > \mathcal{L}_\lambda$. Diese *lokale* Optimierung garantiert aber nicht die *globale* Optimalität der Schätzung und setzt damit eine geeignete Wahl der initialen Parameter λ_0 voraus. Die Schätzformeln für π , \mathbf{A} , μ_{ik} , Σ_{ik} und c_{ik} finden sich z.B. in [Schukat-Talamazzini 1995].

Erkennung

In der Erkennung muß die Wortfolge \mathbf{w} gefunden werden, die die beste Wahrscheinlichkeitsbewertung $p(\mathbf{X}|\lambda(\mathbf{w}))$ besitzt. Da diese Berechnung für alle Wortfolgen extrem aufwendig wäre, wird die folgende Näherung gemacht:

$$p(\mathbf{X}|\lambda) = \sum_{q \in Q^T} p(\mathbf{X}|q, \lambda)P(q|\lambda) \quad (3.13)$$

$$\approx \max_{q \in Q^T} p(\mathbf{X}|q, \lambda)P(q|\lambda) \quad (3.14)$$

Dies bedeutet, daß nur der beste Pfad für jede Wortfolge gefunden werden muß. Dies kann mit geeigneten Suchverfahren sehr effizient implementiert werden (siehe Abschnitt 3.3.3). Für den wahrscheinlichsten Pfad \hat{q} gilt:

$$\hat{q} = \operatorname{argmax}_q P(q|\mathbf{X}, \lambda) = \operatorname{argmax}_q \frac{p(\mathbf{X}, q|\lambda)}{p(\mathbf{X}|\lambda)} \quad (3.15)$$

$$= \operatorname{argmax}_q p(\mathbf{X}, q|\lambda) \quad (3.16)$$

Da $p(\mathbf{X}|\lambda)$ in Gleichung 3.15 unabhängig von q ist, kann es bei der Maximierung vernachlässigt werden (Gleichung 3.16).

Die Bewertung dieses Pfades $\hat{p}(\mathbf{X}|\lambda) = p(\mathbf{X}, \hat{q}|\lambda)$ wird auch Viterbi-Bewertung (*viterbi score*) genannt. Sie wird als Näherung von $p(\mathbf{X}|\lambda)$ benutzt und bildet somit in der Erkennung die Grundlage für die Klassifikation.

Der Pfad \hat{q} kann durch ein rekursives Verfahren sehr effizient berechnet werden, das auf der Hilfsgröße $\vartheta_i(t)$ basiert:

$$\vartheta_i(t) = \max_{\mathbf{q} \in \mathcal{Q}^T, q_t = i} p(\mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{q} | \boldsymbol{\lambda}) \quad (3.17)$$

forced alignment

Eine besonders während der Entwicklung eines Erkenners sehr wichtige Technik ist das sogenannte *forced-alignment*. Im Prinzip handelt es sich um eine Abwandlung des oben beschriebenen Viterbi-Erkennungsalgorithmus. Jedoch wird die tatsächlich gesprochene Wortfolge vorgegeben und nur noch die beste Zustandsfolge \hat{q} berechnet. Hierdurch wird das Sprachsignal mit Hilfe der trainierten Modelle segmentiert. Die so bestimmte Segmentierung kann z.B. zusammen mit einem Spektrogramm dargestellt werden, um so Fehler in der Modellierung oder der Transkription zu finden. Außerdem können aus einem *alignment* des gesamten Trainings globale Statistiken wie zum Beispiel Längenverteilungen einzelner Phone o.ä. generiert werden.

Eine leichte Verallgemeinerung ergibt sich, wenn statt der linearen Kette der Modelle, die der Transkription entsprechen, ein Graph mit alternativen Pfaden benutzt wird. Denkbar sind hier etwa alternative Aussprachen der transkribierten Wörter oder detaillierte Modelle für gefüllte Pausen zwischen den Wörtern. Diese Verallgemeinerung wird manchmal auch als *constrained alignment* bezeichnet.

3.3.2. Sprachmodellierung

Durch die *Sprachmodellierung* werden Wortfolgen \mathbf{w} a-priori Wahrscheinlichkeiten $P(\mathbf{w})$ zugeordnet. Das Sprachmodell stellt damit neben dem akustischen Modell eine weitere Wissensquelle dar, die zur Steigerung der Erkennungsleistung genutzt wird.

Das Wissen, welche Wortfolgen besonders wahrscheinlich sind, kann entweder in einer deterministischen Grammatik oder durch ein stochastisches Modell repräsentiert werden.

Die Benutzung von Grammatiken ist vor allem in einfachen Dialogsystemen sinnvoll, in den nur eine sehr eingeschränkte Menge von Kommandoeingaben zugelassen ist, die z.B. mit einer regulären Grammatik spezifiziert werden kann. Hier kann die Menge der über dem Vokabular möglichen Wortfolgen in grammatische und ungrammatische partitioniert werden, wobei die ungrammatischen eine a-priori Wahrscheinlichkeit von Null und alle grammatischen Folgen der gleichen Länge die gleiche Wahrscheinlichkeit erhalten.

Die stochastischen Sprachmodelle werden in der Regel auf sehr großen Textkorpora trainiert. Um die Schätzung der Verteilung über die (unendlich vielen) Wortfolgen auf die Schätzung eines endlichen Satzes von Parametern zu reduzieren, wird dabei die folgende Näherung gemacht:

$$P(\mathbf{w}) = \prod_{i=1}^m P(w_i | w_1^{i-1}) \quad \text{wobei} \quad w_i^k = w_i, \dots, w_k \quad (3.18)$$

$$\approx \prod_{i=1}^m P(w_i | w_{i-n+1}^{i-1}) \quad (3.19)$$

Die Faktorisierung in bedingte Wahrscheinlichkeiten in Gleichung 3.18 ist noch exakt, aber die Beschränkung der Länge der *Geschichte* \mathbf{h} auf $(n-1)$ Worte stellt eine Näherung dar. Sprachmodelle, die $P(\mathbf{w})$ auf diese Weise approximieren, werden *n-Gramm-Modelle* oder *n-Gramm-Grammatiken* genannt. Die am meisten benutzten Modelle dieser Klasse sind Unigramm- ($n = 1$), Bigramm- ($n = 2$) und Trigramm-Grammatiken ($n = 3$).

Bei der Schätzung der Parameter wird wiederum ein Maximum-Likelihood-Verfahren benutzt, wobei die Likelihood-Funktion die Wahrscheinlichkeit des Trainingstexts – gegeben die Parameter – ist. Es ergibt sich:

$$P_{ML}(w|\mathbf{h}) = \frac{c(\mathbf{hw})}{c(\mathbf{h})} \quad (3.20)$$

hier bezeichnet $c(\mathbf{v})$ die absolute Häufigkeit der Wortfolge \mathbf{v} im Trainingstext.

Bei Benutzung dieser Parameter ergibt sich für Sätze, die im Training ungesene n -Gramme enthalten, eine Wahrscheinlichkeit von Null. Dies führt zu erheblichen Problemen bei der Erkennung, insbesondere wenn das Erkennungsvokabular Wörter enthält, die in der Trainingsstichprobe nicht enthalten waren.

Um die Schätzung robuster zu machen, wird daher ein Teil der „Wahrscheinlichkeitsmasse“ von den gesehenen auf die ungesehenen Ereignisse umverteilt (*discounting*). Zumeist wird eine *Rückfall-Strategie* (*back-off*) angewandt, die die Wahrscheinlichkeit eines ungesehenen n -Gramms anhand der Häufigkeit des $(n-1)$ -Gramms schätzt (siehe [Katz 1987]).

3.3.3. Suchalgorithmen

Die Aufgabe des *Decoders* ist es, die Wortfolge \mathbf{w} mit der besten Bewertung $p(\mathbf{X}|\mathbf{w})$ zu finden (siehe Gleichung 3.3). Um diese Aufgabe zu effizient zu lösen, wird das Problem zunächst aus einer etwas anderen Sichtweise betrachtet: Da die Wörter (bzw. Aussprachen) durch HMMs modelliert werden, können die HMM-Folgen, die allen möglichen Wortfolgen entsprechen, zu einen (unendlich großen) Graphen zusammengefaßt werden. Die Knoten sind die Zustände der HMMs (mit ihren zugehörigen Emissionsdichten). Die Kanten sind innerhalb der HMMs die normalen Transitionen und zwischen den HMMs neu einzufügende Verbindungskanten. Diese Kanten, die aufeinanderfolgende Wörter verbinden, werden mit der entsprechenden Sprachmodellwahrscheinlichkeit gewichtet. Aufgrund der *Viterbi-Annahme* (Gleichung 3.14) reduziert sich das Decodierungsproblem auf die Suche nach dem billigsten Pfad durch diesen Graphen.

Die Graphstruktur kann stark vereinfacht werden, indem man gemeinsame Präfixe der Pfade zusammenfaßt, so daß sich eine Baum-artige Struktur ergibt.

Aufgrund der Größe des Graphen muß dieser dynamisch während der Suche generiert werden, so daß jeweils nur der Teil expandiert wird, in dem aktuell gesucht wird. Hierbei können sogar Knoten, die sich in ihren Fortsetzungen weder hinsichtlich der Modellfolgen noch des Sprachmodells unterscheiden, zusammengefaßt werden (*path merging*).

Prinzipiell gibt es zwei verschiedene Herangehensweisen, um dieses Suchproblem zu lösen: zeitsynchrone und -asynchrone Verfahren. Im folgenden wird jedoch nur auf die synchronen Verfahren eingegangen. Zu den asynchronen zählen insbesondere die auf dem A*-Algorithmus basierenden Verfahren, die mit Hilfe von Kellerspeichern sehr effektiv implementiert werden können.

Im Prinzip ist es möglich, die in der Suche relevanten scores $\vartheta_i(t)$ in einer großen Matrix zu speichern, dies ist jedoch weder besonders anschaulich, noch kann es gut implementiert werden. Eine sehr anschauliche Sicht auf das Suchproblem liefert die *token passing* Technik (siehe [Young et al. 1989]). Die scores werden dabei in sogenannten *token* gespeichert, die durch das Suchnetz propagiert werden. Zu jedem Zeitpunkt t befindet sich in jedem Zustand i des Netzes genau ein token, der den score $\vartheta_i(t)$ trägt. Bei der Verarbeitung eines neuen frames werden alle token in die mit einer Transition erreichbaren Nachfolgezustände propagiert, wobei die aktuelle Bewertung des token mit der Transitionswahrscheinlichkeit multipliziert wird. Nun werden in jedem Zustandsknoten alle token außer dem mit der besten Bewertung verworfen. Dies entspricht der Viterbi-Näherung (Maximierung über die scores in Gleichung 3.17). Anschließend werden die scores aller token noch mit dem Wert der jeweiligen Emissionsdichte multipliziert.

Zur Initialisierung werden token in allen Anfangszuständen plaziert. Danach werden alle frames schrittweise wie beschrieben verarbeitet und unter den token, die sich in Wortendezuständen befinden, derjenige mit dem besten score gesucht. Dieser token repräsentiert den besten Pfad und enthält dessen Viterbi-Bewertung.

Da ein Absuchen des kompletten Netzwerkes wegen der enormen Größe viel zu aufwendig wäre, ist es nötig, die Suche auf den relevanten Teil des Netzes zu beschränken (*pruning*). Da a-priori natürlich nicht bekannt ist, in welchem Teil des Netzes sich der beste Pfad befindet, werden verschiedene Heuristiken eingesetzt, um diese Entscheidung zu treffen. Die erfolgreichste Technik ist die sogenannte *Strahlsuche* (*beam search*, siehe [Lowerre 1976]), die zu jedem Zeitpunkt alle token verwirft, deren (logarithmierte) Bewertung mehr als die konstante Strahlbreite von der Bewertung des besten token abweicht. Dies bewirkt, daß nur die Hypothesen „in der Nähe“ der aktuell besten verfolgt werden.

Eine weitere Technik ist die zusätzliche Begrenzung der Zahl der zu einem Zeitpunkt im Suchnetz existierenden token (*maximum model pruning*, siehe [Odell 1995]), hierdurch kann sichergestellt werden, daß der Speicherbedarf des Erkenners nicht den physikalisch verfügbaren Speicher überschreitet.

Diese beiden Techniken bewirken, daß die Suche um mehr als eine Größenordnung beschleunigt wird, ohne viele Suchfehler zu machen und damit die Erkennungsleistung

zu verschlechtern. Es gibt zahlreiche weitere Methoden, die eine Beschleunigung der Suche bewirken, wie z.B. *Wortende-pruning* oder verschiedene *look-ahead-Techniken* (einen Überblick gibt z.B. [Odell 1995]).

Bei der direkten Anwendung der Decodierungsvorschrift (Gleichung 3.3) zeigen sich jedoch zwei Probleme. Einerseits bewegen sich die akustischen Dichtewerte und die Sprachmodellwahrscheinlichkeiten in verschiedenen Größenordnungen, so daß das Sprachmodell nur einen sehr geringen Einfluß auf die Entscheidung des Erkenners hat. Dies wird üblicherweise durch einen Gewichtungsfaktor γ behoben, der als Exponent der Sprachmodellwahrscheinlichkeit in Gleichung 3.3 verwendet wird. Ein weiteres Problem ist die Tendenz des Erkenners, Wortfolgen mit vielen kurzen Wörtern zu produzieren. Dies läßt sich durch Einführen der sogenannten *Wortstrafe* (*word insertion penalty*) ρ , die an jeder Wortgrenze auf die Bewertung aufmultipliziert wird, beseitigen. Als neue Zielfunktion ergibt sich somit:

$$f(\mathbf{w}|\mathbf{X}) = p(\mathbf{X}|\mathbf{w})P(\mathbf{w})^\gamma \rho^{|\mathbf{w}|} \quad (3.21)$$

Die beiden Faktoren γ und ρ werden in der Regel rein empirisch anhand der Fehlerrate der produzierten Wortfolgen optimiert.

Mit der oben beschriebenen Methode kann lediglich die Wortfolge mit der besten Bewertung gefunden werden. Falls die Ausgabe des Erkenners noch weiterverarbeitet wird (z.B. in einer syntaktischen/semantischen Komponente eines Dialogsystems), so können in den nachfolgenden Verarbeitungsschritten Fehler, die vom Erkener gemacht wurden, nicht oder nur schwer revidiert werden. Es empfiehlt sich daher, nicht nur die *beste Kette*, sondern auch weitere (schlechter bewertete) Hypothesen bereitzustellen. Dies geschieht entweder in Form einer Liste der n besten Satzhypothesen oder aber kompakter in Form eines Worthypothesengraphen (*word lattice*). Die verschiedenen Hypothesen werden dabei mit den vom Erkener bestimmten Bewertungen annotiert, so daß die nachgeschalteten Komponenten diese in ihre Verarbeitung einbeziehen können.

3.4. Daimler Benz Spracherkennung

In diesem Abschnitt soll auf einige Besonderheiten des bei Daimler Benz entwickelten Spracherkennungssystems eingegangen werden.

Das System basiert im wesentlichen auf Standardtechniken, wobei jedoch darauf geachtet wird, daß das System (gegebenenfalls bei entsprechender Einschränkung des Wortschatzes) echtzeitfähig ist. Die Merkmalsextraktion verwendet die in Abschnitt 3.2.1 beschriebenen MFCC-Merkmale, verbunden mit einer dynamischen cepstralen Mittelwertsubtraktion. Eine Normalisierung auf Basis einer kompletten Äußerung liefert zwar bessere Ergebnisse, ist aber wegen der resultierenden Zeitverzögerung nicht akzeptabel. Zusätzlich zu 12 MFCC-Merkmalen wird ein logarithmisches Energiemaß verwendet. Diese 13 Merkmale werden mit den Merkmalsvektoren der vier Nachbarframes zu

beiden Seiten zu einem *Supervektor* mit 117 Elementen kombiniert. Die Dimension wird mit einer linearen Transformation, die durch eine *Lineare Diskriminanz Analyse (LDA)* bestimmt wurde, auf 23 reduziert.

Für die Decodierung werden *semi-kontinuierliche HMMs* verwendet, d.h. daß es einen globalen Pool von Normalverteilungsdichten (das *Codebuch*) gibt, und für die einzelnen Zustände nur die Mischungsgewichte geschätzt werden. Das Codebuch enthält 1024 Normalverteilungen mit vollen Kovarianzen.

Die Auswahl der Wortuntereinheiten erfolgt mit einem automatischen Verfahren, das die Häufigkeiten der *n*-Phone und phonologisches Wissen nutzt, um einen Satz von Triphonen zu finden, die mit dem vorhandenen Trainingsmaterial robust trainierbar sind. Hierbei werden nur wortinterne Kontexte benutzt, obwohl eine Verwendung von Kontexten über Wortgrenzen hinaus sicherlich die Erkennungsleistung steigern würde. Dies hätte aber eine erhebliche Komplexitätssteigerung in der Suche zur Folge. Bei dem Clustering wird von allen im Training vorhandenen Triphonen ausgegangen, die dann *bottom-up* nach phonologischen Kriterien auf Modellebene geclustert werden (für Details zum Verfahren siehe [Fach 1996]).

Bei dem Decoder handelt es sich um einen zeitsynchronen dynamischen Netzwerk-Decoder, der auf einem als Baum strukturierten Lexikon aufsetzt. Der Decoder erzeugt einen Worthypothesengraphen, der in einer zweiten Verarbeitungsstufe unter Anwendung eines Sprachmodells auf die gewünschte Größe reduziert wird (siehe [Kuhn et al. 1996]).

4. Aussprachemodellierung

In diesem Kapitel soll ein Überblick über die in der Spracherkennung verwendeten Techniken zur Aussprachemodellierung gegeben werden.

Wie bereits in Kapitel 3 erwähnt, wird die akustische Modellierung einer Wortfolge w in der Regel durch eine Folge von verketteten HMMs geleistet. Hinsichtlich der Granularität der Modellierung muß eine Entscheidung getroffen werden, die zwischen einer detaillierten, spezifischen Modellierung und einer robusten Schätzung der Parameter abwägt. In Abschnitt 4.1 werden die Alternativen und insbesondere die heute übliche Polyphon-Modellierung vorgestellt.

In Abschnitt 4.2 wird die allgemeine Struktur eines Aussprachemodells definiert und die Integration dieses Modells in die Decodierung diskutiert. Anschließend wird in Abschnitt 4.3 exemplarisch eine Technik zur manuellen Erstellung von (kanonischen) Aussprachelexika beschrieben.

In Abschnitt 4.4 schließlich werden die wichtigsten in der Literatur diskutierten Verfahren zur expliziten Modellierung von Aussprachevarianten und Methoden zur Gewichtung dieser Varianten vorgestellt und verglichen. Eine Verallgemeinerung dieser Techniken auf eine *dynamische* Gewichtung der Varianten wird in 4.5 eingeführt.

4.1. Einheiten der Modellierung

Optimalerweise sollte für jede Wortfolge eine spezielle HMM-Folge verwendet werden, um eine exakte Modellierung der akustischen Realisierung der Wortfolge zu ermöglichen. Da jedoch fast alle möglichen Wortfolgen im Training nie oder zu selten vorkommen, ist es zwingend notwendig, das vorhandene Trainingsmaterial besser auszunutzen. Dazu werden Einheiten definiert, die in verschiedenen Wortfolgen trainiert und benutzt werden können. Dies führt zwar dazu, daß die Modelle nicht mehr so speziell sind und damit Eigenheiten der Realisierung bestimmter Wortfolgen nicht mehr explizit modellieren können, ist aber notwendig, um eine robuste Schätzung der Parameter zu ermöglichen.

4.1.1. Wortmodelle

Die naheliegendste Vereinfachung ist, davon auszugehen, daß die Realisierung eines Wortes nicht von dem Rest der Wortfolge abhängt, d.h.:

$$\lambda(w_1, w_2, \dots, w_n) = \lambda(w_1), \lambda(w_2), \dots, \lambda(w_n) \quad (4.1)$$

Diese Vereinfachung stammt noch aus der Zeit der Isoliertwort-Erkennung und war dort sicherlich auch angemessen, da die Pausen zwischen den Worten sicherstellen, daß die Realisierung eines Wortes nicht von den Nachbarworten beeinflusst wird. Für kontinuierliche Sprache gilt diese Annahme jedoch sicherlich nicht mehr, da die Wortgrenzen sich hier hinsichtlich der Koartikulation nicht besonders von wortinternen Kontexten unterscheiden.

Es bleibt jedoch nach wie vor das Problem der Trainierbarkeit der Modelle, da viele der Wörter (besonders bei den heute üblichen Vokabularen mit bis zu 60.000 Einträgen) im Training zu selten oder gar nicht auftreten.

4.1.2. Wortuntereinheiten

Die nächste Stufe der Vergrößerung der Modellierung bildet die Einführung einer zusätzlichen Repräsentationsebene, in der Phone als elementare Symbole verwendet werden. Somit wird die bedingte Wahrscheinlichkeitsdichte aus Gleichung 3.3 noch weiter aufgespaltet:

$$p(\mathbf{X}|\mathbf{w}) = \sum_{\mathbf{a}} p(\mathbf{X}|\mathbf{a}, \mathbf{w})P(\mathbf{a}|\mathbf{w}) \quad (4.2)$$

wobei \mathbf{a} eine Aussprache (d.h. eine Folge von Phonen) ist. Die Dichte $p(\mathbf{X}|\mathbf{a}) = p(\mathbf{X}|\mathbf{a}, \mathbf{w})$ stellt die eigentliche akustische Modellierung dar, während mit $P(\mathbf{a}|\mathbf{w})$ die Aussprache modelliert wird.

Diese Modellierung läßt sich jedoch nur sehr schwer in die Suche (siehe Abschnitt 3.3.3) integrieren, da sie eine Summation über die Aussprachevarianten der Wortfolge erfordert. Üblicherweise wird daher auch hier gemäß der *Viterbi-Annahme* die Summation durch eine Maximumbildung ersetzt:

$$p(\mathbf{X}|\mathbf{w}) \approx \max_{\mathbf{a}} p(\mathbf{X}|\mathbf{a})P(\mathbf{a}|\mathbf{w}) \quad (4.3)$$

In vielen Systemen wird jedoch noch einen Schritt weiter gegangen und angenommen, daß jede Wortfolge \mathbf{w} adäquat durch eine einzige Phonfolge $\mathbf{a}_{can}(\mathbf{w})$ modelliert werden kann. Dies schließt jedoch eine explizite Modellierung der Aussprachevariabilität auf der phonetischen Ebene aus (siehe Abschnitt 4.2.1).

$$p(\mathbf{X}|\mathbf{w}) = p(\mathbf{X}|\mathbf{a}_{can}(\mathbf{w})) \quad (4.4)$$

Monophone

Die Kombination der Vereinfachung 4.4 mit der Annahme, daß die Modellierung Wörter unabhängig voneinander erfolgt (Gleichung 4.1), bedeutet, daß zu jedem Wort w im Vokabular genau eine phonetische Repräsentation $\mathbf{a}_{can}(w)$ festgelegt wird, die dann direkt als Basis für die Modellauswahl dient. In der Regel umfaßt das Inventar der phonetischen Symbole 40–60 Einheiten (siehe Anhang A für eine Tabelle der in dieser Arbeit verwendeten Symbole). Diese können direkt Modellen zugeordnet werden, so daß insgesamt genau ein HMM pro Phon trainiert wird. Diese Modelle werden auch *Monophone* oder *Kontext-unabhängige Modelle* genannt.

$$\lambda(w_1, w_2, \dots, w_n) = \lambda(\mathbf{a}_{can}(w_1, w_2, w_n)) = \lambda(a_1), \lambda(a_2), \dots, \lambda(a_m) \quad (4.5)$$

Monophone haben den Vorteil, daß sie auf sehr vielen Trainingsbeispielen trainiert werden können. Der Nachteil ist jedoch, daß die Modellierung damit sehr unspezifisch wird. Verschiedene Arten der Artikulation eines Phons können nicht modelliert werden, da über alle Varianten „gemittelt“ wird.

Polyphone

Da der phonetische Kontext eines Phons seine Artikulation stark beeinflusst, liegt es nahe, diesen Kontext bei der Abbildung auf die Modelle zu berücksichtigen. In der Regel wird eine Kontext von einem Phon nach beiden Seiten einbezogen. Die so entstehenden Modelle werden *Triphone* genannt.

$$\lambda(w_1, w_2, \dots, w_n) = \lambda(\#, a_1, a_2), \lambda(a_1, a_2, a_3), \dots, \lambda(a_{m-1}, a_m, \#) \quad (4.6)$$

Wird der Kontext auf mehr als ein Phon erweitert, so spricht man allgemein von *Polyphonen*. Viele der möglichen Polyphone werden jedoch wiederum im Training zu selten beobachtet, so daß man nicht alle Kontexte getrennt modellieren kann. Es wurden verschiedene Verfahren entwickelt, die ausgehend von dem Trainingsmaterial und seiner phonetischen Transkription einen Satz von Modellen auswählen, die robust trainiert werden können. Hierbei handelt es sich meist um Verfahren, die einen Baum generieren, der die verschiedenen speziellen Modelle eines Phons enthält. Jeder Knoten repräsentiert dabei die Vereinigung der Kontexte seiner Tochterknoten. Der Wurzelknoten enthält alle Kontexte des Phons, und die Blätter repräsentieren die speziellsten Kontexte, die trainiert werden können (z.B. Triphone oder Quinphone). Dieser Baum kann entweder bottom-up ausgehend von allen im Training auftretenden Kontexten oder aber top-down mit Hilfe von sogenannten *phonetischen Fragen* generiert werden. Das top-down clustering hat den Vorteil, daß auch im Training ungesehene Kontexte einem Blattknoten zugeordnet werden. Dieses clustering-Verfahren kann entweder mit ganzen Modellen oder aber auf der

Ebene der Modellzustände durchgeführt werden. Ein Überblick über top-down clustering Verfahren bietet [Nock et al. 1997].

Auch bei dem Einsatz von Polyphonen kann die Vereinfachung benutzt werden, daß die Modellierung eines Wortes nicht von den anderen Worten beeinflusst wird. Dies bedeutet, daß nur wortinterne Kontexte berücksichtigt werden. Die zusätzliche Berücksichtigung von wortübergreifenden Kontexten erlaubt eine bessere Modellierung, erfordert aber ein clustering Verfahren, daß auch für ungesehene Kontexte Modelle bereitstellt.

Ein Beispiel für die verschiedenen Modellierungsebenen findet sich in Abbildung 4.1.

Wort	kanonische Aussprache
Grüß	g r y: s
Gott	g 0 t

Modellierungsebene	Modellfolge $\lambda(\mathbf{w})$
Wortmodelle	$\lambda_{\text{Grüß}} \lambda_{\text{Gott}}$
Monophone	g r y: s g 0 t
wortinterne Triphone	g+r g-r+y: r-y:+s y:-s g+0 g-0+t 0-t
x-word Triphone	g+r g-r+y: r-y:+s y:-s+g s-g+0 g-0+t 0-t

Abbildung 4.1.: Modellfolgen für verschiedene Granularitäten

Neben der phonbasierten Modellierung gibt es noch verschiedene andere Verfahren, die z.B. den Übergangsbereich zwischen Phonen (*Diphone*) oder ganze Silben modellieren. Diese alternativen Modellierungsansätze konnten bislang jedoch nicht mit dem Erfolg der Polyphonmodellierung mithalten und werden daher im Rahmen dieser Arbeit nicht weiter diskutiert. In [Schukat-Talamazzini 1995] werden zahlreiche Verweise auf relevante Arbeiten gegeben.

4.2. Struktur des Aussprachemodells

4.2.1. implizite vs. explizite Modellierung

Im Prinzip gibt es im Rahmen der Spracherkennung mit Wortuntereinheiten-basierten Modellen zwei Möglichkeiten, Aussprachevarianten zu modellieren. Entweder man gibt nur eine *kanonische* Aussprache im Lexikon an und geht davon aus, daß die Modelle die verschiedenen auftretenden Varianten *implizit* „lernen“, oder man sieht *explizit* verschiedene Aussprachen pro Wort vor.

Die implizite Modellierung kann die Varianten gut modellieren, die eindeutig durch den (Polyphon-)Kontext bestimmt werden. Dies gilt jedoch nur für die Fälle, in den die Phone nur anders artikuliert werden als in der kanonischen Aussprache. Handelt es sich um Auslassungen oder Einfügungen von Phonen, so ergeben sich erhebliche Probleme. Solche Phänomene können durch eine explizite Modellierung wesentlich besser modelliert

werden, da dort eine entsprechend kürzere bzw. längere Modellfolge vorgesehen werden kann. Ein Beispiel ist die Verkürzung des Wortes *haben* ($ha:b\textcircled{n}$) zu $ha:m$. Hierbei müßte die Modellfolge $b\textcircled{n}$ implizit die Aussprache m modellieren. Dies kann jedoch selbst mit Polyphonen nicht geleistet werden.

Insgesamt gilt, daß durch die implizite Modellierung „verschmutzte“ Modelle trainiert werden, d.h. daß die Modelle auf Bereichen des Signals trainiert werden, in den eigentlich ein anderes Phon artikuliert wurde. Zu einem gewissen Grad sind diese „breiten“ Modelle durchaus erwünscht, da Koartikulationsphänomene und leicht verschiedene Sprechweisen im Training erfaßt werden und damit die Modelle robuster gegen solche Variationen werden.

Grundlage für eine explizite Modellierung ist jedoch fast immer auch ein initiales kanonisches Aussprachelexikon. Zu diesem Lexikon werden dann Varianten hinzugefügt, um die Modellierung zu verbessern. Die Ergänzung der Varianten kann dabei auf verschiedene Weisen geschehen. Die Extrema sind hierbei einerseits vollautomatische datengetriebene Verfahren, die die Varianten selbständig „lernen“ und eine rein manuelle Auswahl von Varianten andererseits. In Abschnitt 4.4 werden diese Methoden im Detail diskutiert.

4.2.2. Gewichtung der Varianten

Durch die Ergänzung neuer Varianten steigt die akustische Verwechselbarkeit der Wörter unter Umständen erheblich. Insbesondere steigt die Zahl der *Homophone*, d.h. es finden sich Wörter, die gleiche Aussprachevarianten besitzen. Um in solchen Situationen entscheiden zu können, welches Wort gesprochen wurde, ist es hilfreich, die Varianten mit Gewichten zu versehen. Diese Gewichtung wird durch den Faktor $P(\mathbf{a}|\mathbf{w})$ in Gleichung 4.3 geleistet. Bei der Verwendung des normalen Viterbi-Decoders ergibt sich das schon in Abschnitt 3.3.3 im Zusammenhang mit den Sprachmodellwahrscheinlichkeiten erwähnte Problem der verschiedenen Größenordnungen der beteiligten Faktoren. Dieses Problem wird durch die Einführung eines *Aussprachegewichtsfaktors* ϕ gelöst.

Bei der Gewichtung der Aussprachevarianten ergibt sich zusätzlich die Schwierigkeit, daß Wörter, die besonders viele Varianten besitzen, stärker bestraft werden als Wörter, die z.B. nur eine Variante besitzen, da die Wahrscheinlichkeitsmasse auf mehr Ereignisse verteilt werden muß. Ein häufig gemachter Lösungsvorschlag ist, die Gewichte so zu skalieren, daß die wahrscheinlichste Variante das Gewicht 1 erhält:

$$g(\mathbf{a}, w) = \frac{P(\mathbf{a}|w)}{\max_{\mathbf{a}} P(\mathbf{a}|w)} \quad (4.7)$$

Durch diese Skalierung bleibt die relative Gewichtung der Varianten eines Wortes untereinander erhalten, aber Worte mit vielen Varianten werden nicht „benachteiligt“. Insbesondere im Zusammenhang mit der Viterbi-Annahme erscheint diese Skalierung sinnvoll, da hier nur die beste Variante berücksichtigt wird. Würde man tatsächlich

über alle Varianten summieren, so wäre diese Skalierung sicherlich nicht angebracht. In Experimenten verschiedener Forschungsgruppen gab es bis jetzt jedoch kein schlüssiges Bild, bzgl. des Nutzens dieser Skalierung (siehe [Riley et al. 1997]).

Als neuer Ausdruck für die vom Decodierer zu optimierende Zielfunktion ergibt sich (vergleiche Gleichung 3.21):

$$f(\mathbf{w}, \mathbf{a} | \mathbf{X}) = p(\mathbf{X} | \mathbf{a}) g(\mathbf{a}, \mathbf{w})^\phi P(\mathbf{w})^\gamma \rho^{|\mathbf{w}|} \quad (4.8)$$

Diese Zielfunktion kann mit dem üblichen Viterbi-Verfahren maximiert werden, da eine Zustandsfolge \mathbf{q} eindeutig einer Aussprache \mathbf{a} einer Wortfolge \mathbf{w} entspricht. Das Gewicht ϕ kann ähnlich dem Sprachmodellgewicht γ empirisch optimiert werden.

Im folgenden wird zwischen *statischer* und *dynamischer* Gewichtung der Varianten unterschieden. Bei der statischen Gewichtung ist die Funktion $g(\mathbf{a}, \mathbf{w})$ für alle Äußerungen konstant, während sie bei der dynamischen Gewichtung sogar innerhalb einer Äußerung verändert werden kann. Es ist z.B. vorstellbar, daß man eine Möglichkeit findet, die Gewichtsfunktion dynamisch an die Sprechweise des Sprechers anzupassen. Ein möglicher Ansatz hierzu findet sich in [Ostendorf et al. 1996] und wird in Abschnitt 4.5 näher beschrieben.

4.3. Lexikonerstellung

In diesem Abschnitt werden die Erstellung und Verbesserung eines initialen Aussprachelexikons diskutiert. Die Darstellung orientiert sich hauptsächlich an dem am *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur* (LIMSI) praktizierten Verfahren, da dieses Verfahren durch zahlreiche Veröffentlichungen (z.B. [Gauvin et al. 1994, Lamel und Adda 1996, Adda-Decker und Lamel 1997]) sehr gut dokumentiert ist. Außerdem wurde das am LIMSI für den *Wall Street Journal Korpus* entwickelte Aussprachelexikon von verschiedenen Forschungsgruppen erfolgreich in ihre Erkennen integriert, was die Qualität des Lexikons eindrucksvoll unter Beweis stellt.

Die Erstellung eines Aussprachelexikons kann grob in zwei Stufen gegliedert werden. In der ersten Stufe wird zunächst der Wortschatz festgelegt, der in dem Erkennen benutzt werden soll. Dies ist eine sehr kritische Entscheidung, da alle Wörter, die nicht im Vokabular sind, in der realen Erkennung zwangsläufig niemals erkannt werden können. Andererseits hat eine Vergrößerung des Erkennervokabulars aber auch eine stärkere akustische Verwechselbarkeit der Wörter zur Folge, und die Decodierung wird unter Umständen erheblich aufwendiger. Das Erkennungsvokabular wird – sofern es sich nicht gerade um ein Dialogsystem in einer sehr stark eingeschränkten Domäne handelt – meist anhand eines extrem großen Textkorpus bestimmt. Beliebt sind hierbei vor allem Zeitungstexte, die in großen Mengen verfügbar sind (in der Regel mehr als 100 Millionen Wörter). Einige der bei dieser Art der Wortlistenstellung auftretenden Probleme werden z.B. in [Adda et al. 1997] diskutiert. [Rosenfeld 1995] versucht, eine Abschätzung für

die optimale Größe des Erkennungsvokabulars zu finden, um den Anteil der unbekanntem Wörter zu minimieren.

Ein weiteres Problem bei der Lexikonerstellung ist, daß wortübergreifende Verschleifungen in einem wortbasierten Lexikon natürlich nicht erfaßt werden können. Solche Verschleifungen treten insbesondere bei häufigen Phrasen auf (im Englischen z.B.: *haben wir*). Um diese Varianten im Lexikon modellieren zu können, werden diese Phrasen als zusätzliche Einträge ins Lexikon aufgenommen (*what_did_you*). Diese Einträge werden *compound words* oder *multi-words* genannt. Die Auswahl erfolgt in der Regel von Hand oder aufgrund ihrer Häufigkeit und der *mutual information* zwischen den Wörtern (siehe [Gauvain et al. 1997, Finke und Waibel 1997]).

In der zweiten Stufe der Lexikonerstellung müssen für alle Wörter im Erkennungsvokabular und zusätzlich für weitere nur im akustischen Training vorkommende Wörter Aussprachen festgelegt werden. Ausgangsbasis bilden in der Regel mehrere große Aussprachelexika, die aus verschiedenen Quellen stammen (z.B. Lexika von anderen Forschungsgruppen oder sogar klassische Wörterbücher). Eine weitere Möglichkeit, Aussprachen zu finden, ist die Nutzung von Graphem-nach-Phonem Regeln, die meist im Rahmen von Sprachsynthesystemen entwickelt wurden. Bei der Kombination dieser verschiedenen Aussprachelexika ist die Hauptschwierigkeit, daß sie in der Regel verschiedene Phonsymbole und Konventionen benutzen. Meist kann man jedoch eine relativ gute Abbildung in den vom Erkennen verwendeten Symbolsatz finden.

Da es sich bei dem Aussprachelexikon in der Regel um ein Vollformenlexikon handelt, tauchen gerade bei stark flektierenden Sprachen sehr viele Flektionsformen bzw. abgeleitete Formen eines Wortes auf. Die Aussprachen solcher Wörter können relativ zuverlässig durch einfache Regeln aus der Grundform abgeleitet werden.

Eine weitere Klasse von Wörtern, deren Aussprache relativ sicher generiert werden können, sind die Komposita. Wenn für die Bestandteile Aussprachen vorhanden sind, so kommt ihre Verkettung meist der echten Aussprache zumindest sehr nahe.

Um die Aussprachen aus den verschiedenen Quellen zu nutzen und die Arbeit zu beschleunigen, wurde am LIMSI ein interaktives System entwickelt, das für unbekannte Wörter automatisch die oben skizzierten Methoden anwendet (Komposita müssen natürlich von Hand zerlegt werden). Die Ergebnisse werden dem Nutzer präsentiert, der sich für eine der Aussprachen entscheidet oder selbst eine Verschriftung eingibt.

Diese manuelle Kontrolle ist unbedingt nötig, um Fehler zu finden und die Konsistenz der Verschriftung sicherzustellen. Die Konsistenz kann auch durch automatische Prüfung und Angleichung der Aussprache bestimmter Wortteile (Silben o.ä.) erhöht werden.

Wurden für den gesamten Wortschatz Aussprachen festgelegt, so wird ein forced-alignment des Trainingsmaterials mit bereits trainierten Modellen durchgeführt. Hierbei wird in der Regel aufgrund des *beam prunings* für einige Sätze kein Pfad gefunden. Diese Sätze werden nochmals von Hand untersucht, um etwaige Fehler in der Worttranskription zu korrigieren. Hierbei ist es manchmal notwendig, zu einem Wort eine weitere Aussprache hinzuzufügen. Meist handelt es sich um stark dialektale Varianten oder besonders verschliffen artikulierte Wörter. Diese Aussprachevarianten erhalten zu diesem

Zeitpunkt in der Regel kein Gewicht, da es keine Möglichkeit gibt, diese a-priori robust zu schätzen. Somit gibt es zu einigen Wörtern mehrere Aussprachen \mathbf{a}_i mit einem Gewicht $g(\mathbf{a}_i, w) = 1$.

Durch ein forced-alignment mit bereits trainierten Modellen wird entschieden, welche Variante jeweils artikuliert wurde, die dann im Training als Referenz verwendet wird (siehe Abschnitt 5.4.2).

In [Markey und Ward 1997] wird ein Verfahren vorgestellt, das noch einen Schritt weiter geht und die im forced-alignment beobachteten scores nutzt, um Fehler im Aussprachelexikon oder fehlende Varianten zu finden. Weicht der mittlere score eines Triphons in einem bestimmten Wort stark von dem globalen Mittelwert des scores dieses Triphons ab, so sollte die Aussprache korrigiert werden.

Das Ergebnis der beschriebenen Verfahren ist ein Aussprachelexikon, das ungewichtete Aussprachevarianten enthält. Typischerweise werden im Schnitt ca. 1.2 Aussprachen pro Wort angegeben.

4.4. statisch gewichtete Aussprachevarianten

Es gibt zahlreiche Vorschläge, wie Aussprachevarianten modelliert werden können, hier soll jedoch nur ein Überblick über die in der statistischen Spracherkennung direkt einsetzbaren Verfahren gegeben werden. Außerdem werden in diesem Kapitel nur die „klassischen“ Ansätze beschrieben, die entweder nur ungewichtete Varianten generieren oder aber nur eine konstante, statische Gewichtungsfunktion $g(\mathbf{a}, \mathbf{w})$ bereitstellen. Möglichkeiten dynamischer Modellierung werden in Abschnitt 4.5 diskutiert.

Um die Darstellung zu strukturieren, werden die Arbeiten nach der Art der Repräsentation der Funktion $g(\mathbf{a}, \mathbf{w})$ klassifiziert. An alternativen Repräsentationen finden sich:

explizite Lexikoneinträge Die alternativen Aussprachen können explizit als neue Lexikoneinträge für die relevanten Wörter eingetragen werden. Hierbei besteht kein Zusammenhang zwischen den Varianten der verschiedenen Wörter.

phonetische Transformationsregeln Die auftretenden phonologischen Prozesse können in Transformationsregeln codiert werden, die dann auf die kanonischen Aussprachen aller Wörter angewendet werden. Eine Gewichtung der Wortvarianten kann über a-priori Wahrscheinlichkeiten der Regeln vorgenommen werden.

Entscheidungsbäume Die tatsächliche Realisierung eines Phons der kanonischen Aussprache kann probabilistisch durch eine bedingte Wahrscheinlichkeitsverteilung modelliert werden (als Bedingungen bieten sich z.B. die Nachbarphone an). Diese Verteilung kann sehr gut mit Entscheidungsbäumen gelernt und repräsentiert werden.

Man muß beachten, daß einige der Repräsentationen in andere konvertiert werden können. So lassen es natürlich alle Ansätze zu, die Varianten „auszumultiplizieren“ und

damit ein Variantenlexikon zu generieren (siehe 4.4.1), was üblicherweise auch getan wird, um die Varianten in der Erkennung zu verwenden. Die Einordnung erfolgt hier jedoch nach der ursprünglichen Intention der jeweiligen Autoren und ordnet die Ansätze in die „natürlichste“ Kategorie ein.

Ein weiteres Unterscheidungsmerkmal ist, ob das Modell lediglich alternative Aussprachen generiert (mit $g(\mathbf{a}, \mathbf{w})=1$) oder aber ob es auch eine wirkliche Gewichtung der Varianten leistet.

4.4.1. alternative Lexikoneinträge/Aussprachegraphen

Wie schon in Abschnitt 4.3 erwähnt, ist die einfachste Methode, Aussprachevarianten zu definieren, sie bei Bedarf von Hand zu der kanonischen Aussprache hinzuzufügen. Hierbei ergibt sich aber das Problem, daß es kaum möglich ist zu entscheiden, welche Varianten tatsächlich für die Erkennung relevant sind. Um diese Entscheidung zu treffen, können die oben erwähnten Methoden, die auf den scores eines forced alignments basieren, verwendet werden ([Adda-Decker und Lamel 1997] und [Markey und Ward 1997]). Trotzdem müssen mögliche Varianten von Hand erstellt werden, was sehr zeitaufwendig und fehleranfällig ist.

Einen Schritt weiter geht Tilo Sloboda, der ein Verfahren entwickelt hat, das aus dem Trainingsmaterial automatisch Aussprachen generiert (siehe [Sloboda 1995, Sloboda und Waibel 1996]). Mittels eines forced-alignments werden alle Signalabschnitte, die Äußerungen eines bestimmten Wortes enthalten, extrahiert. Auf diesen Signalabschnitt wird dann eine Phonererkennung (mit einem Phonbigramm) durchgeführt. So erhält man für jedes Auftreten eines Wortes eine Phontranskription. Hieraus können zum Beispiel die besonders häufigen Transkriptionen ermittelt werden und als neue Varianten zum Lexikon hinzugefügt werden. Dieses neue Aussprachelexikon kann auch für ein erneutes Training der Modelle verwendet werden. Ein großer Vorteil dieses Verfahrens ist, daß der Erkenner direkt zur Generierung der neuen Varianten benutzt wird, da auf diese Weise die Aussprachen wesentlich konsistenter sind als die von Hand erstellten. Sloboda erreichte auf dem VM-Korpus eine relative Fehlerreduktion von ca. 6%.

Ein etwas allgemeineres Verfahren wurde an der TU-Dresden entworfen (siehe [Westendorf und Jelitto 1996]). Ein Phonerkenner wird benutzt, um für jede Äußerung einen Phonhypothesengraphen zu erzeugen. Mit dem aktuellen Aussprachelexikon wird aus der Worttranskription ebenfalls ein Phongraph erzeugt. Durch ein Suchverfahren werden zwei Pfade in den beiden Graphen bestimmt, so daß die jeweils korrespondierenden Phone möglichst „ähnlich“ sind (hierbei wird z.B. eine Verwechslungsmatrix benutzt und auch die Einfügung und Löschung von Phonem zugelassen). Der beste Pfad durch den Phonhypothesengraphen wird nun wieder in die Worte segmentiert, und die jeweiligen Phonfolgen werden bei Bedarf als neue Aussprachen ins Lexikon aufgenommen.

In [Wooters 1993, Wooters und Stockle 1994] wird ein Algorithmus diskutiert, der auf Aussprachegraphen beruht und auch in der Lage ist, die beobachteten Varianten zu generalisieren und somit im Training nicht gesehenen Varianten ein Gewicht größer Null

zuweisen kann. Zunächst wird für jedes Wort ein Aussprachegraph konstruiert, der alle verfügbaren Aussprachen (z.B. aus verschiedenen Wörterbüchern, Graphem nach Phonem Regeln, o.ä.) als parallele Pfade enthält. Mit diesen Graphen wird ein alignment der Trainingsäußerungen durchgeführt. Aus den nun beobachteten Aussprachen wird ein neuer Graph konstruiert, wobei durch eine a-priori Verteilung über die Modellstrukturen der Grad der Verallgemeinerung gesteuert werden kann. Die Gewichte in dem Aussprachegraphen können mit einem normalen ML-Algorithmus geschätzt werden. Die Verallgemeinerung, die bei der Graphkonstruktion stattfindet, führte allerdings nicht zu einer Steigerung der Erkennungsleistung, während die Verwendung der Graphen mit beobachteten Aussprachen eine Verbesserung von etwa 10% gegenüber dem System mit nur einer Aussprache pro Wort einbrachte.

4.4.2. phonetische Transformationsregeln

Die in 4.4.1 beschriebenen Methoden haben den Nachteil, daß sie alle wortbasiert arbeiten und daher nur für die Wörter angewandt werden können, die (relativ häufig) im Trainingskorpus vorkommen. Für im Training nicht gesehene Wörter können höchstens vollständig manuell Varianten definiert werden, die z.B. einem Phonetiker besonders relevant erscheinen. Dabei leidet aber wiederum die Konsistenz der Verschriftung, und eine Gewichtung der Varianten ist nicht möglich.

Eine Möglichkeit, diese Probleme zu umgehen, ist, die Varianten ausgehend von einem initialen Aussprachelexikon durch die Anwendung eines Satzes von (optionalen) Transformationsregeln zu generieren. Diese können auf beliebige (auch ungesehene) Wörter angewendet werden. Aus den Regeln zugeordneten Wahrscheinlichkeiten kann eine Gewichtung der Varianten erreicht werden.

Die Transformationsregeln haben im allgemeinen die Form:

$$C_l/A/C_r \rightarrow B \tag{4.9}$$

wobei A, B, C_l, C_r (evtl. leere) Folgen von Phonsymbolen sind. Die Regel wird auf eine kanonische Aussprache angewendet und ersetzt die Folge A im Kontext C_lC_r durch B . Beispiele typischer Regeln finden sich in Tabelle 4.1.

/ə/→	Schwa-Elision
b/ən/→m	Assimilation
/rə/→ʁ	r-Reduktion

Tabelle 4.1.: typische Transformationsregeln

Man muß zwischen *obligatorischen* und *optionalen* Regeln unterscheiden. Obligatorische Regeln werden hauptsächlich verwendet, um eine eher phonemische Lexikonrepräsentation in die entsprechende phonetische Realisierung umzuwandeln oder um die

an Wortgrenzen zwangsläufig stattfindenden Prozesse (z.B. Tilgung von Doppelkonsonanten) zu modellieren. Durch Anwendung der optionalen Regeln entstehen alternative Aussprachen.

In den meisten Arbeiten wird jeder optionalen Regel r eine a-priori Wahrscheinlichkeit $P(r)$ zugeordnet, die benutzt wird, um die Wahrscheinlichkeit einer Wortvariante zu berechnen. Hierzu werden zu jeder Variante $\mathbf{a} \in \mathcal{V}(w)$ eines Wortes w die folgenden Mengen definiert:

$$\mathcal{R}(w) = \{r | r \text{ ist auf } w \text{ anwendbar}\} \quad (4.10)$$

$$\mathcal{R}^+(w, \mathbf{a}) = \{r \in \mathcal{R}(w) | r \text{ wurde zur Generierung von } \mathbf{a} \text{ angewandt}\} \quad (4.11)$$

$$\mathcal{R}^-(w, \mathbf{a}) = \mathcal{R}(w) \setminus \mathcal{R}^+(w, \mathbf{a}) \quad (4.12)$$

In Gleichung 4.10 können die Regeln, die in dem Wort w *mehrfach* anwendbar sind, mit unterschiedlichen Indizes versehen werden. Hierdurch ist es möglich, daß eine anwendbare Regel an einer Stelle in dem Wort angewendet wird und an einer anderen Stelle jedoch nicht. Da auf diese Weise die Zahl der möglichen Varianten nochmals steigt, wird in einigen Arbeiten (z.B. [Kemp 1995]) die Einschränkung gemacht, daß eine Regel in einem Wort entweder an allen möglichen Stellen oder gar nicht angewendet wird.

Die Wahrscheinlichkeit $P(\mathbf{a}|w)$ einer Variante \mathbf{a} wird nun folgendermaßen berechnet (siehe [Tajchman et al. 1995a] oder [Finke und Waibel 1997]):

$$P(\mathbf{a}|w) = \frac{\prod_{r \in \mathcal{R}^+(\mathbf{a}, w)} P(r) \prod_{r \in \mathcal{R}^-(\mathbf{a}, w)} (1 - P(r))}{Z} \quad (4.13)$$

wobei die Normalisierungskonstante Z so gewählt wird, daß sich die Wahrscheinlichkeiten zu eins addieren:

$$Z = \sum_{\mathbf{a} \in \mathcal{V}(w)} \left(\prod_{r \in \mathcal{R}^+(\mathbf{a}, w)} P(r) \prod_{r \in \mathcal{R}^-(\mathbf{a}, w)} (1 - P(r)) \right) \quad (4.14)$$

Die Notwendigkeit dieser Normalisierung ergibt sich nur, wenn es Regeln gibt, deren Anwendung sich gegenseitig ausschließt (ein Beispiel ist in Abbildung 4.2 angegeben). Falls sich die anwendbaren Regeln frei kombinieren lassen, so ergibt sich von selbst $Z = 1$. Ein anderes Verfahren zur Normalisierung der Variantenwahrscheinlichkeiten wird in [Kipp et al. 1997] beschrieben.

In den beschriebenen Arbeiten werden verschieden starke Einschränkungen bzgl. der Länge der Kontexte und des zu ersetzenden Bereichs gemacht. Im folgenden wird zwischen den Verfahren, die eine Erstellung der Regeln von Hand voraussetzen, und den automatischen Verfahren unterschieden.

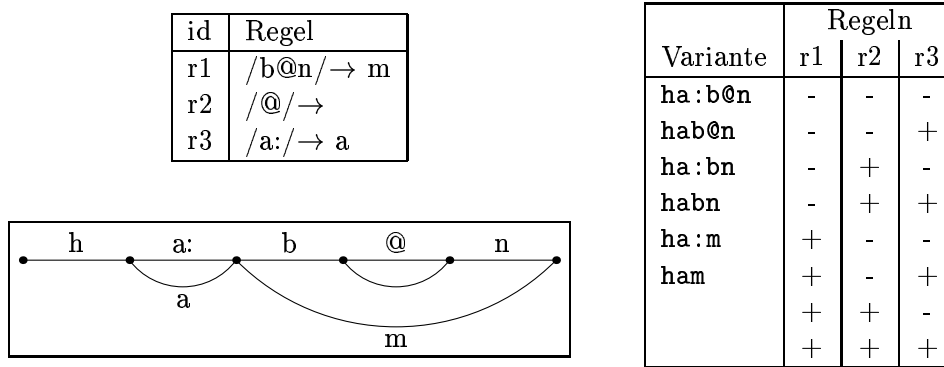


Abbildung 4.2.: Beispiel für Regelanwendung

manuelle Regelerstellung

Eine der ersten Arbeiten, die systematisch die Anwendung von optionalen Transformationsregeln in der statistischen Spracherkennung diskutiert, ist [Cohen 1989]. Die Auswahl und Verwendung der Regeln wird auf Basis einer Stichprobe aus dem TIMIT-Korpus, der sowohl auf Wort- als auch auf Phonebene transkribiert wurde, untersucht. Cohen definiert als die beiden zentralen Maße für die Beurteilung eines Regelsatzes die *coverage* und die *overgeneration*. Für jede Äußerung werden die Transformationsregeln auf die kanonische Aussprache der Worttranskription angewendet. In dem so entstandenen Phonographen wird mit einem DP-Algorithmus nach dem Pfad gesucht, der der Phontranskription am besten entspricht. Das Abstandsmaß ist hierbei die Anzahl der Fehler (Einfügungen, Löschungen und Ersetzungen). Die Fehlerrate f ist die Anzahl der Fehler dividiert durch die Zahl der Phone in der Transkription. Cohen definiert die *coverage* als $c = 1 - f$.

Durch Verwendung sehr allgemeiner Regeln ließe sich natürlich leicht eine coverage von 1 erreichen, dies wäre jedoch in der Erkennung nicht hilfreich, da so im Prinzip für jedes Wort jede Aussprache möglich wäre. Es ist also notwendig, die Übergenerierung der Regeln zu begrenzen. Cohen definiert als Maß für die *overgeneration* die mittlere Anzahl von Kanten in dem Phonographen, die nicht auf dem besten Pfad liegen.

Eine Gewichtung der Regeln gewinnt Cohen durch die Anwendung der Schätzformel für die Transitionswahrscheinlichkeit aus dem Baum-Welch-Algorithmus auf seine Phonographen. Die Wahrscheinlichkeit einer Regel wird hier als Verhältnis der Zahl der tatsächlichen Anwendungen der Regel zu der Zahl der Kontexte, in den sie anwendbar war, berechnet. Die Verwendung der so gewichteten Aussprachegraphen brachte eine relative Verbesserung von ca. 30% bei Erkennungsexperimenten mit dem RM-Task.

Der Nutzen von regelbasiert generierten Aussprachen wird auch in [Tajchman et al. 1995a, Tajchman et al. 1995b] demonstriert. Ein Satz von 10 Regeln wird zur Expandierung eines Lexikons des Wall Street Journal Tasks benutzt. Hiermit werden Regelwahrscheinlichkeiten über die Häufigkeiten im forced-alignment geschätzt und zu Gewichten

der Wortvarianten kombiniert. Besonders interessant ist, daß die so auf den WSJ-Daten geschätzten Regelanwendungswahrscheinlichkeiten sehr gut mit Werten übereinstimmen, die aus den Handtranskriptionen des TIMIT-Korpus berechnet wurden.

Regelsätze für die Deutsche Sprache werden in [Mühlenfeld 1986] und [Wesenick 1996] vorgestellt. Insbesondere der von Wesenick entwickelte Regelsatz versucht mit großem Aufwand (ca. 1500 Regeln) alle relevanten Aussprachephänomene abzudecken, aber gerade die große Zahl der Regeln hat bislang die Integration in ein Erkennungssystem verhindert. Die Regeln wurden jedoch erfolgreich benutzt, um auf Basis einer Worttranskription eine detaillierte Phontranskription zu generieren [Kipp et al. 1996], die beinahe an die Qualität einer von Hand erstellten Transkription heranreicht.

Ein ähnliches Verfahren wurde von [Downey und Wiseman 1997] implementiert, um Aussprachephänomene in kontinuierlicher Englischer Sprache zu untersuchen. Downey et al. beobachteten eine Phonemfehlerrate von 20-25% bei dem Vergleich der kanonischen Aussprache mit von Hand erstellten Transkriptionen. Sie entwickelten einen Satz von 7 Regeln, mit denen neue Aussprachen generiert werden.

Ein Satz von Regeln, der speziell zur Modellierung deutscher Spontansprache entwickelt wurde, wird in [Flach 1995] auf einem extrem kleinen Testset (194 Wörter im Lexikon) getestet. Auf diesem Testset ergaben sich erhebliche Verbesserungen (ca. 20% relativ). Diese guten Ergebnisse konnten jedoch nicht auf ein größeres Erkennungssystem übertragen werden. In [Kemp 1995] wird der Regelsatz in leicht abgewandelter Form verwendet, bewirkt aber keine signifikante Veränderung der Erkennungsleistung trotz eines Neutrainings mit dem Variantenlexikon. Kemp konnte jedoch eine sehr interessante Abhängigkeit der Häufigkeit der Regelanwendung von der Dialektregion der Sprecher und der Sprechgeschwindigkeit feststellen. Diese Abhängigkeit wurde jedoch nicht in der Erkennung genutzt.

Speziell mit der Modellierung der Aussprachephänomene, die an Wortgrenzen auftreten, beschäftigt sich [Giachin et al. 1991]. Ein Satz von 11 (obligatorischen oder optionalen) Regeln wird in einen auf dem RM-Task trainierten Monophon-Erkenner integriert. Um diese Regeln anzuwenden, mußte der Decodierungsalgorithmus dahingehend verändert werden, daß nur noch Wortübergänge zwischen bestimmten Wortvarianten möglich sind. Die Fehlerrate sinkt um knapp 10% relativ.

In [Aubert und Dugast 1995] wird diese Technik zur Behandlung von Liaison-Effekten im Französischen eingesetzt. Im Französischen kann optional nach einem Wort, dessen finaler Konsonant normalerweise nicht artikuliert wird, ein Phon eingefügt werden, falls das folgende Wort mit einem Vokal beginnt. Dieses Phänomen wird in dem Philips-Erkenner modelliert, indem neue Aussprachevarianten aufgenommen werden und dynamisch in der Decodierung die Liaison nur bei den entsprechenden Wortpaaren zugelassen wird. Hierbei werden sogar Crossword-Triphone eingesetzt. Experimentelle Untersuchungen wurden im Rahmen des SQALE-Projekts auf dem BREF-Korpus durchgeführt und ergaben eine Verbesserung von ca. 10% relativ.

automatische Regelgenerierung

In den letzten Jahren wurden verschiedene Verfahren publiziert, die Transformationsregeln automatisch aus dem Vergleich einer phonetischen Transkription mit der kanonischen Aussprache der Worttranskription erzeugen. Die phonetische Transkription wird bei diesen Verfahren entweder tatsächlich von Hand erstellt oder automatisch mit einem Phonerkenner erzeugt.

In [Cremelie und Martens 1995] wird ein DP-alignment der kanonischen Aussprache mit von Hand erstellten Labels erzeugt, und aus den abweichenden Regionen werden Regeln generiert. Hierbei kann durch Berücksichtigung eines sehr großen Kontexts eine sehr spezielle Regel generiert werden. Wird der Kontext eingeschränkt oder sogar ganz weggelassen, so generalisieren die Regeln. Anhand einer Statistik über die Häufigkeit des Auftretens der Kontexte und der Anwendung der Regeln kann entschieden werden, welche Regeln besonders relevant für die Modellierung sind. Die relative Anwendungshäufigkeit wird als Schätzung für die Regelwahrscheinlichkeit benutzt. Cremelie et al. konzentrieren sich in ihrer Studie besonders auf Regeln, die an Wortgrenzen angewendet werden (*inter-word rules*). Die Wahrscheinlichkeit der Regelanwendung auf Wortpaare wird als zusätzlicher Faktor zum Sprachmodell in die Decodierung integriert. In [Cremelie und Martens 1997] wird das Verfahren auf die Verwendung automatisch generierter Phonlabel übertragen. Es ergibt sich auf verschiedenen Tasks eine Fehlerraten-Reduktion von 36%–13%. Interessanterweise brachte die Hinzunahme von *intra-word* Regeln keine weitere Verbesserung, was auf die besondere Bedeutung von *inter-word* Regeln in Spontansprache hinweist.

Andreas Kipp verwendet ebenfalls ein Verfahren, das automatisch aus dem Vergleich der kanonischen Form mit von Experten erstellten Transkriptionen, Transformationsregeln generiert (siehe [Kipp et al. 1997, Kipp 1998]). Diese Regeln werden anhand ihrer Anwendungshäufigkeit gewichtet und verwendet, um große Mengen von Sprachaufnahmen automatisch zu segmentieren. Hierbei wird mit den Regeln aus der bekannten Worttranskription ein (gewichteter) Phonograph erstellt und für ein forced-alignment verwendet. In einem Vergleich mit den von drei menschlichen Segmentierern erstellten Transkriptionen zeigt sich, daß die Qualität der automatisch generierten Transkription durchaus mit den von Hand erstellten mithalten kann. Die Integration der Regeln in ein Erkennungssystem ist insbesondere aufgrund der großen Zahl der Regeln (ca. 1200) noch nicht gelungen.

4.4.3. Entscheidungsbäume

Eine alternative Repräsentation der Transformationsregeln bilden Entscheidungsbäume (*decision trees*, siehe [Breiman et al. 1984, Gelfand et al. 1991]).

Entscheidungsbäume sind hierarchische Klassifikatoren, die benutzt werden können, um bedingte Wahrscheinlichkeitsverteilungen zu repräsentieren. In der Aussprachemodellierung werden sie z.B. benutzt, um die allophonische Realisierung β eines Phonems α

zu modellieren ([Randolph 1990]). Die benachbarten Phoneme γ_l, γ_r werden dabei als zusätzliche Bedingungen verwendet, d.h. mit dem Entscheidungsbaum wird die Verteilung $P(\beta|\alpha, \gamma_l, \gamma_r)$ geschätzt. Üblicherweise wird für jedes Phonem ein eigener Baum erzeugt, in dem an jedem (inneren) Knoten jeweils eine Teilmenge der Merkmalsmenge (γ_l, γ_r) angegeben ist. Um für ein konkretes Phonem die Verteilung zu finden, wird der Baum von der Wurzel zu einem Blatt traversiert, wobei jeweils der linke Tochterbaum gewählt wird, falls der konkrete Kontext in der jeweiligen Menge enthalten ist, ansonsten wird im rechten Tochterbaum fortgefahren. Für jeden Blattknoten existiert eine spezielle Verteilung. Üblicherweise werden die Mengen an den Knoten nicht direkt, sondern über phonetische Fragen (siehe Abschnitt 4.1.2) angegeben. Hierbei können z.B. Fragen bzgl. der phonologischen Kategorie (z.B. Nasal, Frikativ, etc.) der Nachbarphoneme benutzt werden. Abbildung 4.3 zeigt einen binären Entscheidungsbaum als Beispiel.

[Chen 1990] schlägt als Alternative eine Clustering-Technik vor, die für jede Zerlegung die optimale Partition findet. Hierbei wird die Einschränkung auf *binäre* Bäume aufgegeben.

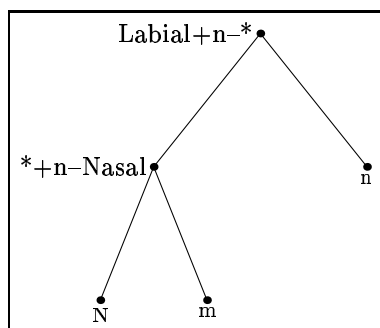


Abbildung 4.3.: Entscheidungsbaum für die Realisierung von /n/

Bei dem Training eines Entscheidungsbaums werden zunächst alle Realisierungen des Phonems in dem Wurzelknoten gesammelt. Anhand einer Frage wird diese Menge in zwei disjunkte Teilmengen zerlegt (die beiden Tochterknoten). Über ein geeignetes Kriterium wird lokal die Frage ausgewählt, die den „besten“ Teilbaum erzeugt. Ein häufig verwendetes Kriterium ist die Entropiereduktion der beiden Tochterknoten im Vergleich zum Wurzelknoten. Diese Aufspaltung wiederholt man rekursiv, bis z.B. die Entropiereduktion unter eine gewisse Schwelle fällt oder sich die Fehlerrate auf einer Kreuzvalidierungsmenge nicht mehr verbessert. Eine Alternative ist es, zuerst den Baum so lange zu erweitern, bis alle Blätter nur noch Allophone jeweils einer Klasse enthalten und dann den Baum mit Hilfe der Kreuzvalidierungsmenge zu *prunen*, d.h. ihn von unten so lange zu verkleinern bis der Teilbaum mit der minimalen Fehlerrate gefunden ist (eine detailliertere Diskussion der verschiedenen Techniken findet sich in [Breiman et al. 1984] und in [Gelfand et al. 1991]). Aus den Häufigkeiten der Allophone in den Blättern werden die entsprechenden Verteilungen geschätzt. Jedes Blatt kann auch in einen äquivalenten

Satz von Transformationsregeln umgesetzt werden (siehe Abbildung 4.3).

Der große Vorteil von Entscheidungsbäumen ist, daß automatisch die „optimalen“ Regeln gelernt werden. Der Nachteil ist, daß es relativ schwierig ist, variable Längen der Phonemkette (nicht des Kontexts) zu verwenden.

In [Riley 1991] wird beschrieben, wie durch die Verwendung von Entscheidungsbäumen die Vorhersagesicherheit der allophonischen Realisierung der Phonemfolge erheblich gesteigert werden kann. Auf dem TIMIT-Korpus gibt Riley die mittlere bedingte Entropie $H(\text{Phonem}|\text{Phon})$ mit ca. 1.5 Bits an. Die trainierten Bäume benutzen die drei benachbarten Phoneme zu beiden Seiten und reduzieren die bedingte Entropie auf 0.8 Bits. Sie prädictieren in 84% der Fälle das richtige Allophon (gegenüber 69% ohne Einbeziehung des Kontexts). Bei dieser Art der Vorhersage wird angenommen, daß die phonetische Realisierung der Phoneme unabhängig voneinander ist. Diese Einschränkung umgeht Riley durch die Berücksichtigung des Allophons, das das vorhergehende Phonem realisiert. Dies bewirkt eine Steigerung der Vorhersagegenauigkeit auf 85%.

Ein generelles Problem bei dieser Modellierung ist, daß Löschungen und Einfügungen nicht direkt modelliert werden können. Eine mögliche Lösung ist die Verwendung neuer spezieller Symbole, die als Folge von null (Löschung) bzw. zwei Allophonen (Einfügung) interpretiert werden.

Die Verwendung von Entscheidungsbäumen zur Vorhersage von Phonemrealisierungen hat jedoch stark an Bedeutung verloren, da in den meisten Systemen heutzutage entweder eher phonbasierte Lexika verwendet werden oder davon ausgegangen wird, daß die Triphonmodelle die entsprechende Realisierung des Phonems in den konkreten (Triphon-)Kontext automatisch lernen. Die Entscheidungsbäume wurden jedoch in den letzten Jahren wiederentdeckt, um die Aussprachevariabilität in Spontansprache oder verschiedene Dialekte zu modellieren.

In [Humphries et al. 1996, Humphries und Woodland 1997] wird das Aussprachelexikon eines auf Britisch-Englischen Äußerungen (WSJCAM0) trainierten Systems zur Erkennung von Amerikanisch-Englisch (WSJ) adaptiert. Hierzu wird zunächst eine Phonerkennung der amerikanischen Trainingsäußerungen durchgeführt, die dann durch einen DP-Algorithmus mit der kanonischen, britischen Aussprache ausgerichtet wird. Für diese Zuordnung (britisches Phon nach amerikanisches Phon) werden Entscheidungsbäume trainiert. Aus diesen Bäumen werden gewichtete Regeln extrahiert, mit denen das Lexikon expandiert wird. Die Verwendung dieses neuen Lexikons reduziert die Fehlerrate von ca. 31% Prozent auf knapp 25% Prozent.

In dem AT&T Erkennungssystem (siehe [Riley et al. 1995]) werden Entscheidungsbäume eingesetzt, um Aussprachevarianten zu modellieren. Die Bäume werden zunächst auf dem (von Hand transkribierten) TIMIT-Material trainiert und benutzt, um die Worttranskriptionen des WSJ-Trainingsmaterials zu Phonographen zu expandieren. Mit diesen Phonographen wird ein forced-alignment durchgeführt, mit dem dann die Verteilungen an den Blättern des Baums neu geschätzt werden. Diese Bäume wurde zur Expansion des WSJ-Lexikons eingesetzt und brachten eine absolute Verbesserung der Wortfehlerrate von 2.7%.

Eine Anwendung der Aussprachemodellierung auf ein Spontansprachkorpus fand im Rahmen des DoD LVSR-Workshop'97 statt (siehe [Riley et al. 1997]). Die Experimente wurden dabei auf dem Switchboard-Korpus (siehe [Godfrey et al. 1992]) durchgeführt, für den mittlerweile auch große Mengen von Hand erstellte Phontranskriptionen vorliegen. Zum Training wurden die TIMIT- und Switchboard-Phontranskriptionen benutzt. Mit den Bäumen wurden Phonlabel für das komplette Switchboard-Trainingsmaterial erstellt, die zum Neuschätzen der Verteilungen verwendet wurden. Mit diesem neuen Aussprachemodell wurden wiederum Phontranskriptionen erstellt, die zu Neutraining der akustischen Modelle verwendet wurden. Dieses System zeigte allerdings wider Erwarten eine schlechtere Erkennungsleistung (WER 45.48%) als das Baseline System (44.66%). Die Entscheidungsbäume verbessern aber die bedingte Entropie von 0.714 bits auf 0.485 bits.

Besonders interessant ist die Technik, die zur Integration des Aussprachemodells in die Decodierung vorgeschlagen wird. Die Entscheidungsbäume werden verwendet, um einen Wortgraphen zu einem Phongraphen zu expandieren, in dem der Decoder die beste Phonkette sucht. Die Wortkette, die dieser Phonkette entspricht, wird dann als Erkennungsergebnis angegeben. Bemerkenswert ist, daß hierbei auch wortübergreifende Aussprachevarianten berücksichtigt werden können.

Insgesamt kann man sagen, daß die Aussprachemodellierung mit Entscheidungsbäumen ein sehr vielversprechender Ansatz ist, der Vorteile der verschiedenen anderen Ansätze kombiniert. Faszinierend ist insbesondere die Möglichkeit, das Expertenwissen über phonologische Phänomene in Form von phonetischen Fragen anhand der konkreten Sprachdaten zu evaluieren und *optimal* in die Modellierung zu integrieren.

4.4.4. andere Repräsentationen

Es wurden noch verschiedene andere Repräsentationen des Aussprachemodells vorgeschlagen. An dieser Stelle sollen jedoch nur zwei weitere erwähnt werden:

In [Fukada und Sagisaka 1997] wird ein neuronales Netz trainiert, um alternative Aussprachen vorherzusagen. Das Vorgehen entspricht der oben beschriebenen Modellierung durch Entscheidungsbäume (Phonererkennung, DP-alignment, Training des Modells, Variantengenerierung). Es ist nicht klar, warum ein neuronales Netz einem Entscheidungsbaum überlegen sein sollte. Insbesondere kann man bei dem Netz keine Einsichten über die besonders relevanten Kontextfaktoren gewinnen. In den trainierten Bäumen kann man jedoch anhand der Struktur sehr gut die entscheidenden Faktoren erkennen. Fukada et al. berichten von einer Verbesserung der Worterkennungsrate von 27.4% auf 32.6%.

In [Jost et al. 1997] wird versucht, auch die Aussprachemodellierung durch HMMs vorzunehmen. Auf einer mit einem Phonerkenner generierten Phonkette wird für jede Silbe ein diskretes HMM trainiert (ein emittierender Zustand pro Phon in der kanonischen Aussprache). In der Erkennung wird auch diese zweistufige Technik angewandt, d.h. die Silben-HMMs benutzen eine Phonkette als Eingabe. Durch die frühe Entscheidung für eine Phonkette geht natürlich die globale Optimalität der Lösung bzgl. des Modells verloren. Es wäre besser, die beiden Stufen zu integrieren oder einen Phongraphen an der

Schnittstelle zu verwenden. Ebenfalls sinnvoll wäre es, mit den Silben-HMMs Aussprachen zu generieren und diese als Alternativen in das Lexikon einzutragen. Ähnlich wie bei den neuronalen Netzen ist jedoch auch hier nicht klar, welchen Vorteil diese Modellierung gegenüber den Entscheidungsbäumen hat, da hier gewisse Annahmen (Abhängigkeit von der Silbe, tying von Zuständen) hart in die Struktur der Modelle codiert werden. Diese Annahmen brauchen bei der Verwendung von Entscheidungsbäumen nicht a-priori gemacht werden, sondern können datengetrieben gelernt werden.

4.5. dynamische Gewichtung

Eine Einschränkung aller oben beschriebenen Arbeiten ist die Annahme, daß die Gewichtungsfunktion $g(\mathbf{a}, w)$ konstant ist. Verschiedene Untersuchungen deuten darauf hin, daß es hilfreich wäre, die Gewichtung dynamisch innerhalb einer Äußerung zu verändern.

In [Weintraub et al. 1996] werden Erkennungsexperimente durchgeführt, in den versucht wird, nur den Sprechstil (*speaking style*) zu variieren, während alle anderen Parameter (Sprecher, Text, etc.) konstant gehalten werden. Die Fehlerrate für spontane Äußerungen liegt mit ca. 53% erheblich über der für gelesene Diktiersprache (29%). Es wird darauf hingewiesen, daß in Spontansprache ein relativ großer Anteil der Phone der kanonischen Aussprache ausgelassen werden.

Der Zusammenhang zwischen der Sprechgeschwindigkeit und der Fehlerrate wird in [Mirghafori et al. 1995] untersucht. Anhand von Äußerungen aus dem WSJ- und dem TIMIT-Korpus wird festgestellt, daß es eine starke Korrelation zwischen der Sprechgeschwindigkeit (gemessen in Phonen pro Sekunde) und der Wahrscheinlichkeit der extremen Verkürzung oder Auslassung von Phonen gibt. Durch verschiedene Adaptionstechniken gelingt es Mirghafori et al., die Fehlerrate auf den schnellen Sätzen zu reduzieren, allerdings steigt dadurch die Fehlerrate auf den langsamen Sätzen an. Es wird angedeutet, daß mit einem zuverlässigen Maß für die Sprechgeschwindigkeit die Adaption gezielter vorgenommen werden könnte.

Diese beiden Studien legen nahe, daß eine Aussprachemodellierung, die den Sprechstil und die aktuelle Sprechgeschwindigkeit mit einbezieht, erheblichen Einfluß auf die Leistungsfähigkeit des Erkenners haben könnte.

Eine sehr interessante Verallgemeinerung der in Abschnitt 4.4 vorgestellten Aussprachemodellierung wird in [Ostendorf et al. 1996] vorgeschlagen. Es wird die Existenz eines *hidden speaking modes* postuliert, der als Indikator für die Wahrscheinlichkeit verschiedener Aussprachen verwendet wird. Der *speaking mode* soll den aktuellen Sprachstil widerspiegeln und von Faktoren wie z.B. der Sprechgeschwindigkeit oder auch der syntaktischen Struktur der Äußerung beeinflusst werden.

Es wird vorgeschlagen, den *speaking mode* auf Wortebene zu modellieren. Der aktuelle mode m_i eines Wortes soll ausgehend von akustischen Merkmalen \mathbf{Y} und von der Wortfolge abhängigen Merkmalen $f(\mathbf{w})$ bestimmt werden. Es muß also eine Möglichkeit gefunden werden, die Verteilung $P(\mathbf{m}|\mathbf{Y}, f(\mathbf{w}))$ zu schätzen. Dies kann z.B. mit Hilfe

von Entscheidungsbäumen getan werden. Diese Verwendung von Entscheidungsbäumen sollte allerdings nicht mit der in Abschnitt 4.4.3 vorgestellten verwechselt werden.

Ostendorf et al. schlagen zwei Techniken vor, um den *speaking mode* in die Erkennung zu integrieren:

akustische Modellierung Die Abhängigkeit vom *speaking mode* kann bereits beim Training der akustischen Modelle berücksichtigt werden. Sie könnte z.B. als zusätzliche Frage beim clustering der Polyphonmodelle (bzw. der Zustände) eingesetzt werden. Dies entspricht einer Erweiterung der akustischen Dichte $p(\mathbf{X}|\mathbf{a})$ auf $p(\mathbf{X}|\mathbf{a}, \mathbf{m})$.

Aussprachemodellierung Die Aussprachewahrscheinlichkeit $P(\mathbf{a}|\mathbf{w})$ kann ebenfalls um eine Abhängigkeit von dem *speaking mode* erweitert werden: $P(\mathbf{a}|\mathbf{w}, \mathbf{m})$

Für die Kombination der beiden Techniken erhält man als Verallgemeinerung von Gleichung 4.3:

$$p(\mathbf{X}|\mathbf{w}) \approx \max_{\mathbf{a}} \sum_{\mathbf{m}} p(\mathbf{X}|\mathbf{a}, \mathbf{m}) P(\mathbf{a}|\mathbf{w}, \mathbf{m}) P(\mathbf{m}|\mathbf{Y}, f(\mathbf{w})) \quad (4.15)$$

In [Finke und Waibel 1997] wird eine Implementierung dieses Ansatzes beschrieben. Finke konzentriert sich auf eine mode-abhängige Modellierung der Aussprachegewichte und ließ die akustische Modellierung unverändert.

Um die Wahrscheinlichkeit $P(\mathbf{a}|\mathbf{w}, \mathbf{m})$ zu modellieren, wurden 21 Transformationsregeln entwickelt, die die häufigsten Aussprachephänomene in Englischer Spontansprache erfassen sollen. Jeder Regel r wird eine a-priori Wahrscheinlichkeit $P(r)$ zugeordnet. Mit den Regeln werden für alle Wörter Varianten erzeugt, deren Gewichte aus den Regelwahrscheinlichkeiten berechnet werden (siehe Gleichung 4.13).

Die Regelwahrscheinlichkeiten wurden zunächst einfach aus den Häufigkeiten in einem forced-alignment des Trainingsmaterials geschätzt. Diese (globale) Schätzung konnte durch Einbeziehung des phonetischen Kontextes des zu transformierenden Bereichs verbessert werden. Prinzipiell könnte man einfach die vorhandenen Regeln von Hand verfeinern, indem man die Kontexte vergrößert und somit speziellere Regeln erhält. Hierbei würden jedoch Regeln entstehen, die nur noch sehr selten anwendbar sind und deren Gewichtung damit nur schlecht trainiert werden kann. Finke verwendet Entscheidungsbäume mit phonetischen Fragen nach den benachbarten Phonen, um automatisch die relevanten Kontextfaktoren zu bestimmen. Für jede Regel wird ein spezieller Baum erzeugt, der $P(r|\mathbf{w})$ approximiert. Für drei der 21 Regeln konnte die Vorhersagesicherheit durch Berücksichtigung des phonetischen Kontextes verbessert werden.

Eine weitere Verbesserung der Vorhersagegüte konnte durch die Einbeziehung von weiteren Faktoren in das Training der Entscheidungsbäume erreicht werden. Finke berichtet in Übereinstimmung mit den in [Ostendorf et al. 1996] geäußerten Beobachtungen, daß insbesondere Maße für die Sprechgeschwindigkeit gute Indikatoren sind. Finke verzichtet auf eine explizite Modellierung des *speaking modes* und schätzt stattdessen direkt

die Verteilung $P(r|w, \mathbf{Y})$ mit Entscheidungsbäumen. Durch diese Modellierung kann die Vorhersagesicherheit der Regelanwendung zum Teil erheblich verbessert werden.

Der Nutzen der gewichteten Aussprachevarianten für die Erkennung wurde auf den englischsprachigen Switchboard- und CallHome-Korpora evaluiert. Durch ein forced-alignment wurden die jeweils verwendeten Aussprachen bestimmt und die Modelle auf diesen Transkriptionen neu trainiert. Bei einer Verwendung der Varianten ohne Gewichtung im Test und im Training ergab sich eine leichte Verschlechterung gegenüber der Verwendung des initialen Lexikons in Test und Training. Dies wird auf die erhöhte akustische Verwechselbarkeit der Wörter zurückgeführt.

Durch Verwendung der wortabhängig gewichteten Regeln ($P(r|w)$) konnte die Fehlerrate um 5.5% bzw. 2.3% relativ (Switchboard bzw. CallHome) gegenüber dem System mit ungewichteten Varianten verbessert werden. Die mode-abhängige Gewichtung ($P(r|w, \mathbf{Y})$) brachte 7.0% bzw. 6.5% relative Reduktion der Fehlerrate.

5. Auswahl der Varianten

In diesem Kapitel soll eine Technik zur Generierung und Gewichtung von Aussprachevarianten vorgestellt werden, die auf dem Verbmobil-Korpus entwickelt und evaluiert wurde. Zuerst wird in Abschnitt 5.1 ein Überblick über die Besonderheiten der verwendeten Stichproben gegeben. In Abschnitt 5.2 werden allgemeine Kriterien für die Auswahl eines Aussprachemodells angegeben, und die Entscheidung für einen regelbasierten Ansatz wird begründet. Die Struktur der verwendeten Regeln wird in 5.3 beschrieben.

Anschließend werden in 5.4 verschiedene Gütemaße und Methoden zur Auswahl der Transformationsregeln diskutiert und die erzielten Ergebnisse diskutiert. Der so gefundene *optimale* Regelsatz wird in Abschnitt 5.5 vorgestellt und an Beispielen illustriert.

In Abschnitt 5.6 wird die Aufbereitung der Trainingstranskriptionen und das Neutraining der Modelle beschrieben.

5.1. verwendetes Korpus

Die experimentellen Untersuchungen wurden auf dem Sprachkorpus, das im Rahmen des Verbmobil-Projekts (siehe [Wahlster 1993]) erstellt wurde, durchgeführt. Es handelt sich um Terminverhandlungsdialoge, die jeweils zwei Sprecher miteinander führen. Die Aufnahmen wurden mit einem Nahbesprechungsmikrofon in einer Büroumgebung aufgenommen. Zur Zeit stehen etwa 33 Stunden an Dialogen auf Deutsch zur Verfügung (789 Dialoge mit insgesamt 568 Sprechern und 13892 turns). Um die Vergleichbarkeit mit alten Ergebnissen und Untersuchungen anderer Gruppen sicherzustellen, wurden die Bedingungen der Akustik-Evaluation 96 eingehalten (siehe [Reinecke 1996]). Das Gesamtkorpus wurde in die folgenden *disjunkten* Stichproben eingeteilt:

training Die zum Training der akustischen Modelle und des Aussprachemodells verwendete Stichprobe (11404 turns, ca. 27 Stunden)

xval96 Die Kreuzvalidierungsstichprobe, auf der die Parametereinstellungen aller Modelle optimiert wurden (243 turns, ca. 26 Minuten)

eval96 Die offizielle Evaluierungsstichprobe, auf der nicht optimiert wurde, stellt völlig ungesehenes Material dar und wurde nur für die Bestimmung endgültigen Ergebnisse benutzt. (305 turns, ca. 35 Minuten)

Zusätzlich wurden bei der Auswahl der Regeln (siehe Abschnitt 5.4) zwei weitere Teilmengen des Trainingskorpus verwendet. In der Entwicklungsphase wurden viele Experimente zur Messung der Qualität der Regeln aus Zeitgründen nur auf den Dialogen der VM-CD2 (1537 turns, ca. 3.5 Stunden) durchgeführt. Außerdem standen für ein kleines Teilkorpus von 240 turns (ca. 22 Minuten) von Hand erstellte Transkriptionen auf Phonebene zur Verfügung, mit denen die Qualität der Aussprachemodellierung direkt gemessen werden konnte.

5.2. Generierung von Varianten

5.2.1. Methoden der Variantengenerierung

Da in dieser Arbeit insbesondere die Gewichtung von großen Aussprachelexika untersucht werden soll, muß eine Methode gewählt werden, die auch für seltene und im Training ungesene Wörter Varianten generiert. Somit kommen die in Abschnitt 4.4.1 vorgestellten wortbasierten Verfahren nicht in Frage, da sie nur auf häufige Wörter anwendbar sind, bzw. einen unakzeptabel großen Aufwand an manueller Arbeit erfordern. Somit bleiben von den in Kapitel 4 diskutierten Verfahren nur noch die Verwendung von phonetischen Transformationsregeln oder phonetischen Entscheidungsbäumen übrig.

Die Verfahren, die die Varianten datengetrieben automatisch lernen, sind sicherlich interessanter und auch vielversprechender, da hierbei das Expertenwissen (z.B. in Form phonetischer Fragen) auf dem Korpus hinsichtlich seiner Relevanz untersucht und bewertet wird. Der Nachteil dieser Techniken ist, daß sie große Mengen von auf Phonebene transkribierten Daten erfordern. Für das verwendete Verbmobil-Korpus stehen solche Daten nicht zur Verfügung bzw. hätten mit einem Phonerkenner erzeugt werden müssen, was einerseits eine weitere Fehlerquelle eingeführt und einen erheblichen zusätzlichen Entwicklungs- und Optimierungsaufwand bedeutet hätte. Aus diesen Gründen wurde ein Satz von optionalen, phonetischen Transformationsregeln manuell erstellt. Die Relevanz dieser Regeln wurde jedoch anhand der vorliegenden Daten verifiziert, so daß sich ein iteratives Vorgehen zur Auswahl und Verbesserung der Regeln ergab, das in Abschnitt 5.4 beschrieben wird.

Ein komplett von Hand erstellter Regelsatz ohne Beschränkung bei der Auswahl der Regeln wie z.B. in [Wesenick 1996] hat zwar den Vorteil, daß er wirklich alle denkbaren Varianten abdecken kann, aber der Nachteil ist, daß dies meist mit sehr vielen Regeln und einer starken Übergenerierung verbunden ist.

Das Problem der Übergenerierung wird man bei der Verwendung von phonetischen Regeln wahrscheinlich immer haben, man sollte sie aber so klein wie möglich halten, um nicht Varianten zu modellieren, die nicht artikulierbar sind oder niemals verwendet werden. Eine begrenzte Übergenerierung ist natürlich in der Anwendung in einem Segmentationsystem wie MAUS (*Münchener Automatisches Segmentationsystem*, siehe Abschnitt 5.4.3) wesentlich weniger störend als in einem echten Erkennen, da bei der Segmentation die Wortfolge vorgegeben wird und somit das Problem der akustischen

Verwechselbarkeit verschiedener Wörter keine Rolle spielt.

Da die Regeln verwendet werden sollen, um den Aussprachevarianten wie in Abschnitt 4.4.2 beschrieben Gewichte zuzuweisen, ist es notwendig, die Anzahl der Regeln zu begrenzen, damit die a-priori Wahrscheinlichkeiten der Regeln robust geschätzt werden können. Dies ist nur möglich, wenn die Regeln in dem Trainingskorpus häufig anwendbar sind und somit die relative Anwendungshäufigkeit eine gute Schätzung für die Anwendungswahrscheinlichkeit darstellt. Neben dieser Art die Gewichtung zu bestimmen, können die Gewichte auch direkt wortbasiert für die häufigen Wörter bestimmt werden. Somit würden die Regeln nur zur Generierung aber nicht zur Gewichtung der Varianten verwendet werden. Eine genauere Beschreibung der untersuchten Gewichtungsverfahren wird in Kapitel 6 gegeben.

5.2.2. Auswahl der zu expandierenden Wörter

Da in der Erkennung der vorhandene Viterbi-Decoder verwendet wurde, konnten die Regeln nur *innerhalb* von Wörtern angewendet werden. Um trotzdem die wichtigsten wortübergreifenden Phänomene zu erfassen, wurden einige Wortpaare zu *multi-words* zusammengefaßt und explizit in das Lexikon eingetragen. Für die Auswahl dieser Wortpaare standen die folgenden Informationen zur Verfügung:

- Auftretenshäufigkeit im Trainingskorpus
- Extreme Abweichungen von der Standardaussprache werden in den Transliterationen in den sogenannten *Aussprachekommentaren* vermerkt. Hierbei handelt es sich um eine orthographische Umschrift der abweichenden Aussprache. Da dies auch für Wortpaare getan wird, konnten die Wortpaare bestimmt werden, die besonders häufig abweichend artikuliert wurden.
- Anhand einer Segmentation des gesamten Trainingskorpus mit dem MAUS-System wurden die Wortpaare gesucht, bei deren Auftreten am häufigsten eine wortübergreifende Regel angewendet wurde.

Unabhängig von der Modellierung von Aussprachevarianten hat die Verwendung von Wortpaaren den Vorteil, daß so auch in einem System, das nur wortinterne Polyphontexte verwendet, bei allen Wortgrenzen innerhalb der Wortpaare der volle Kontext berücksichtigt werden kann. Daher bringt die Benutzung der häufigsten Wortpaare auch ohne die Generierung von Varianten schon eine Verbesserung der Erkennungsleistung. Man sollte also bei der Auswahl der Wortpaare insbesondere ihre Auftretenshäufigkeit berücksichtigen.

Für diese Untersuchung wurden die 500 häufigsten Wortpaare noch um besonders häufig in Aussprachekommentaren vorkommende Paare und Paare ergänzt, an deren Grenze in der MAUS-Segmentation häufig Transformationsregeln angewandt wurden. Auffallend war, daß die meisten Paare, die gemäß der letzten beiden Kriterien häufig

nicht-kanonisch artikuliert werden, bereits aufgrund ihrer Häufigkeit gute Kandidaten zur Auswahl sind. Dies deutet darauf hin, daß insbesondere immer wiederkehrende Wortfolgen, die meist schon Phrasencharakter haben, anders ausgesprochen werden als es das kanonische Lexikon vorhersagt. Insgesamt wurde 547 Wortpaare verwendet.

5.3. Struktur der Regeln

Hier soll kurz eine Definition der verwendeten Transformationsregeln gegeben und ihre Anwendung an einem Beispiel illustriert werden.

Eine phonetische Transformationsregel hat die Form:

$$C_l/A/C_r \rightarrow B \tag{5.1}$$

wobei A, B, C_l, C_r (evtl. leere) Folgen von Phonsymbolen sind.

In dieser Arbeit wurden bei der Entwicklung des Regelsatzes a-priori keine Einschränkung bezüglich der Länge der Phonfolgen und der Kontexte gemacht. Allerdings wurde bei der Regelanwendung sichergestellt, daß keine Regel auf Phone angewendet wurde, die nur aufgrund einer anderen Regel entstanden sind. Diese Beschränkung auf *nicht-rekursive* Regeln hat zur Folge, daß einige Aussprachephänomene, die sich als Folge mehrerer phonologischer Prozesse erklären lassen, in eine Regel *ausmultipliziert* werden müssen. Ein Beispiel hierfür ist eine Assimilation, die häufig in Verbindung mit einer Elision stattfindet: $ha : b@n \rightarrow ha : bn \rightarrow ha : bm$

Diese Abfolge muß explizit in der Regel $b/@n/\rightarrow m$ kodiert werden und kann nicht als Folge einer allgemeinen @-Elisionsregel und einer Assimilationsregel modelliert werden. Der Verzicht auf *rekursive* Regeln erfordert daher eine größere Zahl von Regeln, vereinfacht aber die Verwendung der Regeln erheblich.

Praktisch implementiert wurde die Regelanwendung als Expansion der kanonischen Aussprache zu einem Aussprachegraphen. Die Phonsymbole in der Regel beziehen sich dabei aber immer auf die kanonische Phonfolge, wodurch die rekursive Anwendung verhindert wird. Jede Anwendung einer Regel führt eine neue Kante in den Graphen ein. Jeder Pfad durch den Graphen entspricht einer Aussprachevariante. Abbildung 5.1 zeigt ein Beispiel für einen kleinen Regelsatz.

5.4. Auswahl der Regeln

In diesem Abschnitt wird zunächst erläutert nach welchen Kriterien die Qualität eines Regelsatzes beurteilt wird und welche Maße hierbei verwendet werden. Um die Erstellung des Regelsatzes zu beschleunigen, wurde als Ausgangsbasis ein großer Regelsatz verwendet, der von Andreas Kipp automatisch erzeugt wurde (siehe [Kipp 1998]) und zur automatischen Segmentation großer Korpora eingesetzt wurde. Die Regeln wurden in dem *Münchener Automatischen Segmentationssystem* verwendet und werden daher im

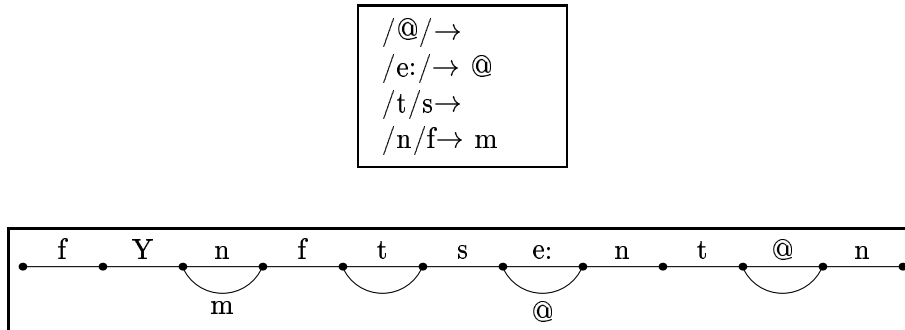


Abbildung 5.1.: Aussprachevariantengraph

folgenden kurz *MAUS-Regeln* genannt. In Abschnitt 5.4.3 wird ein kurzer Überblick über die Generierung dieses Regelsatzes gegeben. In Abschnitt 5.4.4 schließlich wird das in dieser Arbeit verwendete Verfahren zur Optimierung dieses Regelsatzes dargestellt und versucht, seine Qualität zu bestimmen.

5.4.1. Bewertung eines Regelsatzes

Das letztlich entscheidende Maß zur Bewertung der Qualität eines Regelsatzes ist die Reduktion der Wortfehlerrate in einem Erkennen, der diese Regeln verwendet. Dieses Maß ist jedoch praktisch kaum zu verwenden, da es ein komplettes, sehr zeitaufwendiges Neutrainning der Modelle erfordert. Außerdem müssen insbesondere bei der Gewichtung der erzeugten Varianten viele Entscheidungen getroffen werden, die mit dem Regelsatz nicht direkt zusammenhängen, aber die Fehlerrate massiv beeinflussen. Es ist also notwendig einen Weg zu finden, einen Regelsatz direkt, d.h. ohne das komplette neu trainierte System zu bewerten.

Eine sehr sinnvolle Forderung ist, daß alle durch die Regeln erzeugten Varianten tatsächlich als Aussprachen des jeweiligen Wortes möglich sind und überhaupt artikuliert werden können. Diese Forderung kann offensichtlich nur von Hand überprüft werden, was bei der Zahl der verschiedenen Varianten nur noch stichprobenhaft möglich ist.

Die Qualität des generierten Aussprachelexikons kann direkt auf dem Korpus geprüft werden, indem durch ein forced alignment mit der bekannten Worttranskription entschieden wird, welche Aussprache jeweils von dem Sprecher benutzt wurde. Hierzu wird ein Phonograph erstellt, der alle möglichen Aussprachen der bekannten Wortfolge als alternative Pfade enthält. Zwischen den einzelnen Wörtern sowie zu Beginn und Ende der Äußerungen werden optionale Pausemodelle eingefügt. Abbildung 5.2 zeigt einen Beispielgraphen.

Mit diesen Graphen wird dann das forced alignment durchgeführt. Die so erhaltene Segmentation der Äußerungen kann wiederum von Hand untersucht werden und anhand

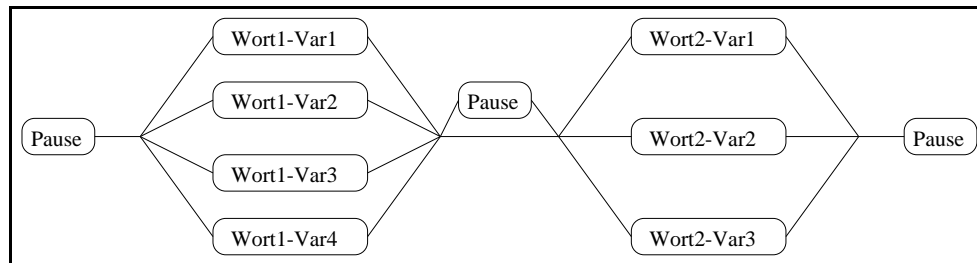


Abbildung 5.2.: Phongraph für das forced alignment

des Signals oder eines Spektrogramms überprüft werden. Diese Überprüfung kann jedoch wiederum nur für einen kleinen Teil der Äußerungen durchgeführt werden und ist auch rein qualitativ. Außerdem ist die Entscheidung, welche Aussprache die konkrete Äußerung eines Wortes besser beschreibt, häufig kaum reproduzierbar objektiv zu treffen.

Falls für ein Teilkorpus Phonlabel, die von einem Experten von Hand erstellt wurden, vorliegen, so kann die Übereinstimmung direkt gemessen werden. Diese von Hand erstellten Transkriptionen leiden natürlich auch unter der oben erwähnten subjektiven Einschätzung der Experten. Mit diesen manuell erstellten Phontranskriptionen lassen sich drei relevante Maße bestimmen:

LPER die Phonfehlerrate des wie oben beschrieben konstruierten Phongraphen

PER die Phonfehlerrate der durch das forced alignment bestimmten Phonkette

CENT falls eine Gewichtung $P(\mathbf{a}|\mathbf{w})$ vorliegt, die bedingte Entropie der richtigen Phonfolge

Die Phonfehlerrate im Graphen entspricht 100%–*coverage* in [Cohen 1989] und ist eine obere Schranke für die Fehlerrate der Kette aus dem forced alignment. Im Graphen läßt sich, wie bereits in 4.4.2 erwähnt, die Fehlerrate LPER bis auf Null reduzieren. Dies senkt aber nicht notwendigerweise auch PER, da hierbei mit den akustischen Modellen die beste Kette bestimmt wird und die erhöhte Verwechselbarkeit zu neuen Fehlern führt. Dies läßt sich gut daran erkennen, daß eine *freie Phonererkennung* typischerweise eine höhere Fehlerrate aufweist als das alignment mit einem kleineren Phongraphen. Die freie Phonererkennung entspricht einem Graphen, der alle möglichen Phonfolgen enthält und weist damit eine Graphfehlerrate von Null auf.

Die Berechnung der bedingten Entropie ist nur möglich, wenn eine Gewichtung vorliegt und zusätzlich die *richtige* Phonfolge immer eine von Null verschiedene Wahrscheinlichkeit erhält. Dies ist normalerweise nicht gegeben, so daß man ein Teil der Wahrscheinlichkeitsmasse auf die durch die Regeln nicht vorhergesagten Folgen verteilen müßte (dies entspricht dem *discounting* bei der Sprachmodellsschätzung, siehe 3.3.2).

5.4.2. Verbesserung des forced alignments

Da durch das forced alignment entschieden wird, welche Variante eines Wortes jeweils geäußert wurde und diese neue Transkription für das Training der neuen akustischen Modelle verwendet wird, ist die Qualität des alignments von sehr großer Bedeutung.

Eine problematische Entscheidung hierbei ist die Auswahl der akustischen Modelle, mit den das alignment vorgenommen wird. Optimalerweise sollte man hier Modelle verwenden, die nur auf manuell phonetisch transkribierten Daten trainiert wurden. Solche Daten stehen jedoch üblicherweise nicht in ausreichender Menge zur Verfügung, so daß man hier einen Kompromiß eingehen muß.

Die naheliegendste Lösung ist die Verwendung der Polyphon-Modelle, die für den normalen Worterkenner trainiert wurden. Diese Modelle werden üblicherweise auf der kanonischen Aussprache trainiert und haben somit die Aussprachevariationen *implizit* gelernt. Aus diesem Grund wird im forced alignment mit diesen Modellen sehr häufig die kanonische Aussprache gewählt, obwohl in Wirklichkeit eine andere Variante artikuliert wurde.

Eine Alternative zu den Polyphonen ist die Verwendung kontextunabhängiger Monophone, die zwar im Prinzip unter der gleichen *Verschmutzung* leiden, aber insgesamt robuster trainiert wurden.

Bei der Entwicklung des Regelsatzes wurde z.B. beobachtet, daß das Triphon $Y-n+f$ häufig als m artikuliert wird (fünfundzwanzig: $fYnfUnttsvantsIC$ vs. $fYmfUnt \dots$ oder sogar $fYmUnt \dots$). Da das n -Polyphon nun hauptsächlich auf m -Segmenten trainiert wurde, wird es im alignment häufig auch auf gerade diesen m -Segmenten einen sehr guten score haben. Bei den Monophonen wirkt sich dieses Phänomen nicht so stark aus, da das n -Modell auf wesentlich mehr Segmenten trainiert wurde, von den sicherlich nur ein kleiner Anteil in Wirklichkeit als m artikuliert wurde. Ein ähnliches Phänomen tritt bei der Elision von Phonen auf, falls sie in bestimmten Kontexten stattfindet – so wird z.B. das \emptyset im Kontext $n-\emptyset+n$ sehr häufig elidiert und das entsprechende Triphon somit auf einem Teil des n -Segments trainiert.

Ein weiteres Problem bei der Verwendung von Polyphonen liegt darin, daß die Polyphon-Kontexte, die in kanonischen Aussprachen vorkommen, im Training viel häufiger auftreten, als die Kontexte der neu erzeugten Varianten. Dies hat zur Folge, daß im forced alignment die kanonischen Aussprachen zu einem großen Teil durch *echte* Polyphone repräsentiert werden, während für die Varianten häufig auf die Monophone *zurückgefallen* werden muß. Da die Monophone aber viel *breiter* als die speziellen Polyphone trainiert sind, verursachen sie meist schlechtere scores. Auf diese Weise werden wieder die kanonischen Aussprachen (oder zumindest Kontexte) bevorzugt.

All diese Argumente sprechen für die Verwendung der Monophone, die üblicherweise in dem Trainingsablauf sowieso erzeugt werden und daher auch zur Verfügung stehen. Eine denkbare Erweiterung wäre ein iteratives Verfahren, in dem zunächst die kanonisch trainierten Monophone für das alignment verwendet werden. Mit den neuen Transkriptionen können dann neue Modelle trainiert werden, mit den wieder ein alignment

durchgeführt wird. Dieser Zyklus kann mehrmals wiederholt werden, um die Modelle immer weiter zu verbessern. Dies Experiment konnte aber leider aus Zeitgründen nicht durchgeführt werden.

Eine weitere Möglichkeit, die Qualität des alignments zu verbessern, liegt in der Nutzung der Sprecherinformation. Da über die Worttranskriptionen hinaus bekannt ist, welche Sätze von dem gleichen Sprecher geäußert wurden, können diese zusammen für eine Adaption der akustischen Modelle an diesen Sprecher benutzt werden. In dieser Arbeit wurde das *Maximum Likelihood Linear Regression*-Adaptionsverfahren benutzt, das eine lineare Abbildung zur Transformation der Mittelwerte der Normalverteilungen nach dem ML-Kriterium schätzt (siehe [Legetter 1995, Haiber 1998]). Um diese Transformation zu schätzen, müssen zunächst alle Äußerungen eines Sprechers mit der Referenztranskription segmentiert werden. Aus dieser Zuordnung einzelner frames auf Modellzustände wird dann die Transformation geschätzt, die die Likelihood-Funktion (siehe Abschnitt 3.3.1) maximiert. Die Adaption birgt natürlich die Gefahr, daß die Modelle noch stärker in Richtung der kanonischen Form adaptiert werden. Um dies zu verhindern, wurden bereits bei dem Alignment für die Adaption die Phonographen mit den Aussprachevarianten benutzt. Nach Anwendung der Transformation wurde das Alignment noch einmal durchgeführt. Auch dieses Verfahren könnte prinzipiell iteriert werden.

Es wurde nur eine globale Transformation pro Sprecher geschätzt, um zu verhindern, daß spezielle (z.B. phonspezifische) Transformationen eine implizite Adaption der *kanonischen* Modelle an die abweichenden Aussprachen ermöglichen. Mit der globalen Transformation wird versucht, nur die Charakteristik des jeweiligen Sprechers zu erfassen.

Die Ergebnisse der beschriebenen Verbesserungsversuche werden in Abschnitt 5.5.2 in Zusammenhang mit der Untersuchung der Qualität des Regelsatzes diskutiert.

5.4.3. MAUS-Regeln

Das *Münchener Automatische Segmentationssystem* (kurz MAUS, siehe [Kipp 1998]) wurde im Rahmen des Verbmobil-Projekts entwickelt, um für große Mengen von Sprachdaten qualitativ hochwertige Segmentierungen auf phonetischer Ebene zu erzeugen. Üblicherweise werden heutzutage Teile eines neuen Sprachkorpus von erfahrenen Phonetikern von Hand mit Phonsymbolen annotiert. Diese Arbeit ist jedoch extrem zeitaufwendig (im amerikanischen Switchboard-Projekt wird hierbei von 400-facher Echtzeit ausgegangen) und damit nicht für das gesamte Korpus durchführbar. In dem MAUS-Projekt werden die vorhandenen von Hand erstellten Label verwendet, um phonetische Regeln zu lernen. Diese werden zusammen mit der Worttranskription benutzt, um das gesamte Korpus zu segmentieren.

An der Universität Kiel wurde für ca. 1200 Äußerungen des VM-Korpus von Hand zu jedem Phon der kanonischen Aussprache notiert, wie es tatsächlich artikuliert wurde, bzw. ob es ausgelassen wurde. Hieraus erhält man für jede Äußerung zwei Phonfolgen die gegeneinander ausgerichtet sind. Diese Folgen bestehen aus Fehlerregionen und Regionen, in den die Phone übereinstimmen. Aus jeder Fehlerregion (beliebiger Länge) wird eine

Transformationsregel erzeugt, wobei zu beiden Seiten jeweils ein Phon als konstanter Kontext benutzt wird. Die so gefundenen Regeln können nach ihrer Häufigkeit sortiert werden und mit a-priori Wahrscheinlichkeiten versehen werden. Hierbei werden sehr selten beobachtete Regeln verworfen.

In dem MAUS-System ist das Ergebnis dieser *Lernphase* ein Satz von 1213 Regeln. Diese werden verwendet, um die übrigen Äußerungen des Korpus zu segmentieren. Aus dieser Segmentierung können aufgrund der Anwendungshäufigkeiten neue a-priori Wahrscheinlichkeiten geschätzt werden.

Die bei der Segmentierung verwendeten akustischen Modelle wurden ebenfalls mit manuell erstellten Phontranskriptionen trainiert, um eine *Verschmutzung* durch die *falschen*, kanonischen Label zu vermeiden.

In [Kipp et al. 1997] wird die Übereinstimmung zwischen den automatisch und manuell erstellten Labeln (mehrerer Phonetiker) mit 75–80% angegeben. Zwischen den verschiedenen Experten wurde eine Übereinstimmung von 80–83% beobachtet.

Ein Problem dieses Ansatzes ist, daß nur die Phänomene in Regeln erfaßt werden können, die in dem relativ kleinen Korpus, das von hand transkribiert wurde, häufig auftreten. Anhand der Segmentierung des gesamten VM-Trainingsmaterials können zwar die Wahrscheinlichkeiten robuster geschätzt werden, aber neue Regeln werden hierbei nicht generiert. Auch die Einschränkung auf den konstanten Kontext von einem Phon scheint problematisch, denn einige phonologische Prozesse sind unabhängig von dem Kontext (zumindest auf einer Seite) anwendbar. Die *Begrenzung* der Kontextlänge auf ein Phon hingegen ist gut vertretbar.

5.4.4. Optimierung des Regelsatzes

Der oben beschriebene MAUS-Regelsatz bildete die Ausgangsbasis für die Erstellung der hier verwendeten Regeln. Der MAUS-Regelsatz umfaßt 1213 Regeln, ein Ausschnitt ist in Abbildung 5.3 zu sehen.

n/t/s→	u:/t@/n→d
k/@/#→	v/E:r@/n→E:6
t/E:/g→e:	z/i:/#→
l/@/n→	#/b/aI→m
f/a:r@/n→a:6	#/de:/n→
C/@/n→	6/l/I→
m/b/6→m	I/t@/n→?
6/@/n→	a/x@/n→h
t/@/#→	a:/b@/#→p
#/?/O→	aI/g@ntl/I→Nd

Abbildung 5.3.: Ausschnitt aus dem MAUS-Regelsatz

Ausgehend von der Anwendungsstatistik bei der Segmentierung des VM-Trainingskorpus wurden zunächst die Regeln gestrichen, die nur selten angewandt wurden. Dies geschah sowohl nach der absoluten als auch der relativen Häufigkeit (Verhältnis zu der Zahl der Stellen, an den die Regel anwendbar war). Unter den verbleibenden Regeln unterschieden sich viele nur durch ihre Kontexte. So gab es z.B. 60 Regeln, die eine @-Elision in verschiedenen Kontexten vorhersagten. Solche Regeln wurde zusammengefaßt und durch Streichung der Kontextphone verallgemeinert – in diesem Beispiel wurde eine Regel ohne Kontext verwendet.

Bei vielen Regeln konnte man deutlich erkennen, daß nur der Kontext auf einer der Seiten für das Aussprachephänomen relevant war. Die Kontexte auf der anderen Seite ergaben sich meist einfach aus besonders häufigen Worten (bzw. Wortfolgen). Diese Art von Kontexten wurden auch gestrichen.

In einigen der Regeln ließen sich die modellierten phonologischen Prozesse recht deutlich erkennen (z.B. Assimilation: $m/b/6 \rightarrow m$). Sofern es sich hierbei um allgemeinere Phänomene handelte, die für eine ganze Klasse von Lauten möglich sind, wurden die entsprechenden zusätzlichen Regeln ergänzt.

Mit dem so entwickelten Regelsatz wurde das kanonische Aussprachelexikon expandiert und ein forced alignment von ca. 1500 Äußerungen durchgeführt, aus dessen Ergebnis wiederum eine Anwendungsstatistik erstellt wurde. Zum Vergleich wurden diese Äußerungen sowohl mit dem kanonischen als auch dem MAUS-Lexikon segmentiert. Das MAUS-Lexikon enthielt dabei für jedes Wort alle Aussprachevarianten, die bei der ursprünglichen MAUS-Segmentierung dieser Äußerungen beobachtet wurden. Dies stellt noch eine Verallgemeinerung der MAUS-Regeln dar, denn hierbei wurden Wortvarianten, die eigentlich nur in bestimmten phonetischen Kontexten (des Nachbarwortes) zugelassen waren, generell als Aussprache des Worts erlaubt.

Die Segmentierung dieser drei Systeme wurde nun von Hand untersucht, wobei sich zeigte, daß das MAUS-Lexikon tatsächlich teilweise zu starke Variationen, insbesondere extreme Verkürzungen, zuließ, die so nicht artikulierbar sind. Die Untersuchung fand mit einem interaktiven tool statt, das sowohl das Signal mit einem Spektrogramm, als auch die drei Segmentierungen anzeigt und das selektive Anhören einzelner Signalabschnitte erlaubt. Hierdurch konnten zahlreiche Fehler und offensichtlich fehlende Regeln gefunden werden.

Da natürlich nicht alle Äußerungen untersucht wurden, mußte eine geeignete Auswahl getroffen werden. Hierbei wurden die mittleren akustischen scores des alignment benutzt, um besonders interessante Äußerungen zu finden. Da sowohl das regelbasierte Variantenlexikon als auch das MAUS-Lexikon alle kanonischen Aussprachen enthielten, war der score des kanonischen Lexikons immer der schlechteste. Der score des MAUS-Lexikons war fast immer besser als der des Regel-Lexikons, da noch extremere Varianten vorhergesagt wurden. Hierbei handelte es sich teilweise um starke Übergenerierung, was beweist, daß der score alleine kein geeignetes Kriterium für die Auswahl des Regelsatzes darstellt.

Die Abweichung der drei scores einer Äußerung war jedoch sehr hilfreich, um proble-

matische Stellen zu finden. Stimmt z.B. die scores des kanonischen und des Varianten-Lexikons überein, aber der des MAUS-Lexikons lag deutlich darunter, so war dies ein Indiz für eine fehlende Variante. Mit dem Varianten-Lexikon wurde offensichtlich auch die kanonische Form gewählt, aber das MAUS-Lexikon enthielt eine akustisch besser passende Aussprache.

Ein weiteres Kriterium bei der Optimierung war die Phonfehlerrate im Vergleich zu manuell erstellten Phontranskriptionen. Hierfür wurde allerdings nur ein sehr kleiner Korpus von 240 Äußerungen verwendet, das nur einen sehr begrenzten Ausschnitt des Trainingsmaterials darstellt und damit nur einen Teil der tatsächlich auftretenden Phänomene enthält. Aus diesem Grund wurde für die Auswahl *neuer* Regeln vor allem die oben beschriebene Technik verwendet.

Dieser Zyklus (Regelmodifikation, Alignment, manuelle Untersuchung) wurde mehrfach wiederholt, um einen geeigneten Regelsatz zu finden. Hierbei wurde auch eine leichte Übergenerierung der Regeln in Kauf genommen, um sie nicht in viele spezielle Regeln aufspalten zu müssen. Anschließend wurden noch einmal alle besonders selten angewandten Regeln gestrichen, um die Größe des Lexikons zu reduzieren.

5.5. Ergebnis der Regelauswahl

In diesem Abschnitt sollen kurz die Regeln, die verwendet wurden, vorgestellt und die Qualität der Segmentierung diskutiert werden.

5.5.1. Regelsatz

In vielen der 35 Regeln, die in Tabelle 5.5.1 aufgeführt sind, kann man die zugrundeliegenden phonologischen Prozesse gut erkennen. Es handelt sich vor allem um Elisionen, Assimilationen und Abschwächungen von Segmenten.

Ein Beispiel für ein Elision ist die Regel, die eine Auslassung des reduzierten Vokals @ zuläßt. Vermutlich ist dies der Prozeß, der am stärksten zu der Verkürzung der tatsächlichen Aussprache gegenüber der kanonischen Form beiträgt. Typische Beispiele aus [Kohler 1995] sind:

- *lebend*: $le:b@nt \rightarrow le:bnt$
- *ebene*: $e:b@ne \rightarrow e:bn@$

An diesen Beispielen wird deutlich, daß die kanonischen Form insbesondere in einer spontanen, nicht-formellen Sprechsituation *ungewöhnlich* oder sogar *überartikuliert* klingt. Als anderes Beispiel führt Kohler die Kombination von Hilfsverben in der 1. Person Singular mit dem Pronomen an:

- *hatte ich*: $hat@IC \rightarrow hatIC$

Regel	Beispiel
/?/→	ähnlich: ?E:nIIC→E:nIIC
/@/→	reden: re:d@n→re:dn
b/@n/→ m	Abend: ?a:b@nt→?a:bmt
/b@n/→ m	abends: ?a:b@nts→?a:mts
g/@n/→ N	sagen: za:g@n→za:gN
/g@n/→ N	irgendwie: ?I6g@ntvi:→?I6Ntvi:
x/@n/→ N	einfachen: aInfax@n→aInfaxN
m/@n/→	bestimmen: b@StIm@n→b@StIm
N/@n/→	Erfahrungen: ?E6fa:rUN@n→?E6fa:rUN
/n@/m→	einem: ?aIn@m→?aIm
/r@/→ 6	wäre: vE:r@→vE:6
/e:/→ @	Detail: de:taI1→d@taI1
v/i:6/→ @	wir: vi:6→v@
/E:/→ E	Vorschläge: fo:6S1E:g@→fo:6S1Eg@
/a:/→ a	Montag: mo:nta:k→mo:ntak
/E:/→ e:	spät: SpE:t→SpE:t
/t/t→	Mitteilung: mIttaI1UN→mItaI1UN
/t/d→	gut_dann: gu:tdan→gu:dan
/t/z→	entsetzlich: ?EntzEtsIIC→?EntzEtsIIC
/t/s→	zwanzig: tsvantsIC→tsvansIC
/C/s→ k	zwanzigste: tsvantsICst@→tsvantsIkst@
a:/k/→ x	Montag: mo:nta:k→mo:nta:x
/n/f→ m	fünf: fYnf→fYmf
/nf/U→ m	fünfundzwanzig: fYnfUnttsvantsIC→fYmUnttsvantsIC
f/I/l→	vielleicht: fIlaICt→flaICt
n/d/6→	hundert: hUnd6t→hUn6t
m/b/6→	September: zEptEmb6→zEptEm6
/n/m→	einmal: ?aInma:l→?aIma:l
/Unt/→ n	zweiundzwanzig: tsvalUnttsvantsIC→ tsvalIntsvantsIC
v/i:6/→ 6	wir: vi:6→v6
?Is/t/→	ist: ?Ist→?Is
t/?E/s→	ist_es: ?Ist?Es→?Ists
/g@nt/l→ N	eigentlich: aIg@ntIIC→aINIIC
/n@/n→	einen: ?aIn@n→?aIn
/?aI/n@→	einen: ?aIn@n→n@n

Tabelle 5.1.: endgültiger Regelsatz

- *wurde ich*: vUrd@IC→vUrdIC

Bei den Assimilationsprozessen wird ein phonetisches Merkmal eines Segments an das entsprechende Merkmal eines benachbarten Lauts angepaßt. Mögliche Merkmale sind z.B. Artikulationsort, Nasalität und Stimmhaftigkeit. Weiterhin werden die Assimilationen hinsichtlich ihrer Wirkungsrichtung unterschieden. Bei der *progressiven Assimilation* wird ein Merkmal eines Lauts auf das nachfolgende Segment übertragen. Findet die Wirkung in die andere Richtung statt, so spricht man von einer *regressiven Assimilation*.

Assimilationen treten häufig in Verbindung mit einer Elision auf. So wird in vielen Fällen erst durch die Elision eines @, das zwei Konsonanten trennt, die Assimilation möglich. Wie bereits erwähnt, müssen solche Effekte aufgrund des Verzichts auf rekursive Regelanwendung in einer Regel kombiniert werden.

Beispiele für Regeln, die Assimilationsprozesse codieren, sind:

- /n/f→ m (fünf: fYnf→fYmf) regressive Assimilation des Artikulationsortes (apikaler Nasal wird labialer Nasal)
- g/@n/→ N (sagen: za:g@n→za:gN) Elision gefolgt von progressiver Assimilation des Artikulationsortes (apikaler Nasal wird dorsaler Nasal)

Ein weiterer wichtiger Prozeß ist die Reduktion von Geminaten (das sind lange Konsonanten, die auf zwei angrenzende Silben verteilt sind). Ein Beispiel ist:

- /t/t→ (Mitteilung: mIttaIlUN→mItaIlUN)

Eine ausführlichere Beschreibung der verschiedenen Prozesse, die im Deutschen beobachtet werden, findet sich in [Kohler 1995].

Mit den 35 Regeln, die das Ergebnis des Optimierungsverfahrens bildeten, wurden die kanonischen Aussprachen der 5867 Lexikoneinträge expandiert. Das Ergebnis waren 34428 Aussprachevarianten. Zu einigen kanonischen Formen gab es keine Varianten, während besonders lange Wörter extrem viele verschiedene Varianten besaßen (für das Wort *siebenundzwanzigsten* wurden 192 Varianten generiert!). Hierbei zeigte sich wieder die für den regelbasierten Ansatz typische Übergenerierung. Im Schnitt wurden ca. 5.9 Varianten pro Wort erzeugt.

5.5.2. Qualität der Segmentierung

Um die verschiedenen Methoden des forced alignments zu vergleichen und die Qualität des regelbasierten Variantenlexikons zu beurteilen, wurde ein forced alignment des Teilkorpus vorgenommen, für den manuell erstellte Phonlabel vorlagen. Es handelt sich dabei um 240 Äußerungen, die zusammen ca. 13.000 Phone umfassen.

Zunächst wurden die Worte der Transkription durch ihre kanonischen Aussprachen ersetzt und die Fehlerrate dieser Phonfolge gemessen. Die Phonfehlerrate betrug 29.0%,

wobei mehr als die Hälfte der Fehler Einfügungen von Phonen sind. Dies bestätigt die Vermutung, daß die kanonische Aussprache in der Regel mehr Phone enthält als tatsächlich artikuliert werden.

Zur Bewertung des Regelsatzes wurden die Worttranskriptionen, wie oben beschrieben, mit dem generierten Variantenlexikon zu Phonographen expandiert. In diesen Phonographen wurde die Kette mit der geringsten Fehlerrate gesucht, um eine untere Schranke für die Fehlerrate des forced alignments zu bestimmen. Die Fehlerrate von 13.9% liegt immernoch relativ hoch, stellt aber eine erheblich Verbesserung gegenüber der kanonischen Form dar. Insbesondere die Zahl der Einfügungen konnte erheblich gesenkt werden.

Anschließend wurde mit diesen Phonographen und den verschiedenen Modellsätzen das forced alignment ausgeführt. Die Fehlerraten der verschiedenen Systeme sind in Tabelle 5.2 zusammengefaßt.

Lexikon	PER (sub/del/ins)
kanonisch	29.0% (1520/81/2204)
Graph	13.9% (1216/135/468)
var-mono	19.3% (1501/226/807)
var-tri	23.4% (1512/122/1432)
var-mono-MLLR	19.1% (1502/217/787)
var-mono-retrain	18.8% (1497/253/713)
var-tri-retrain	19.0% (1499/212/784)

Tabelle 5.2.: Fehlerraten der verschiedenen Systeme

Wie erwartet zeigt sich, daß das alignment mit den Monophonen wesentlich bessere Ergebnisse liefert als mit den Triphonen. Interessant ist, daß diese Verbesserung vor allem auf eine Reduktion der Einfügungen von Phonen zurückzuführen ist, was sich aber, wie in Abschnitt 5.4.2 beschrieben, relativ plausibel erklären läßt.

Die Anwendung der MLLR-Adaption brachte nur eine enttäuschend kleine Verbesserung, wurde aber trotzdem verwendet, da sich diese leichte Verbesserung konsistent bei allen Tests zeigte. Das Monophon System mit einer Iteration MLLR wurde zur Erstellung der neuen Transkriptionen für das Neutraining der akustischen Modelle verwendet (siehe Abschnitt 5.6).

Mit den neu trainierten Modellen wurde das forced alignment noch einmal durchgeführt, um abzuschätzen, ob eine iterative Verbesserung der Modelle möglich ist. Es zeigte sich bei dem Monophon-System nur eine leichte Verbesserung von 0.5% absolut. Sehr erfreulich ist hingegen, daß das Triphon-System beinahe die Qualität des Monophon-alignments erreicht. Dies deutet darauf hin, daß durch das Training die *Verschmutzung* der akustischen Modelle stark reduziert werden konnte.

Der immer noch erhebliche Abstand zwischen der minimalen Fehlerrate im Graphen und der Fehlerrate des alignments (ca. 5% absolut) deutet auf Ungenauigkeiten in der

akustischen Modellierung hin. Er kann aber sicher zumindest zum Teil auf die Inkonsistenzen der manuellen Phonlabel zurückgeführt werden. Leider gibt es keine Möglichkeit, die Anteile dieser beiden Fehlerquellen abzuschätzen. Es bleibt jedoch zu hoffen, daß die automatisch generierten Transkriptionen eine höhere Konsistenz aufweisen als es bei den manuell erstellten der Fall ist.

Die Fehlerrate von ca. 19% kann durchaus mit den Fehlerraten aufwendigerer Systeme (MAUS oder dem Hamburger HMM-Ansatz) konkurrieren. Ein Unterschied zu den beiden anderen Ansätzen ist, daß in dieser Arbeit im forced alignment keinerlei Gewichtung der Varianten vorgenommen wurde. Zunächst stehen natürlich auch keine a-priori Wahrscheinlichkeiten für die manuell erstellten Regeln zu Verfügung, jedoch könnten diese aus dem ersten forced alignment (ohne Gewichte) gewonnen werden. Dieses Vorgehen könnte wieder iteriert werden, z.B. in Verbindung mit dem Neutraining. Es ist anzunehmen, daß hierdurch die Fehlerrate noch einmal reduziert werden könnte, allerdings ist fraglich, ob diese weitere Verbesserung einen großen Einfluß auf die Fehlerrate in der eigentlichen Worterkennung hat.

5.6. **Aufbereitung des Trainingsmaterials**

Mit den in Abschnitt 5.5.1 beschriebenen Variantenregeln und den akustischen Monophonmodellen wurde ein forced alignment des gesamten Trainingsmaterials vorgenommen. Die so erzeugte neue Transkription wurde verwendet, um die sprecherspezifischen MLLR-Transformationen zu schätzen. Unter Anwendung dieser Transformationen wurde noch einmal das gesamte Korpus segmentiert. Diese Transkription diente dann als Ausgangsbasis für ein vollständiges Neutraining der akustischen Modelle.

In den neuen Transkriptionen tauchten nur 10807 verschiedene Varianten der 5811 unterschiedlichen Wörter auf, obwohl das Lexikon 36153 Varianten enthielt. Dies deutet darauf hin, daß der Regelsatz relativ stark übergeneriert (im Sinne von [Cohen 1989]), und es eventuell sinnvoll gewesen wäre noch weitere Regeln zu streichen. Hierdurch hätte der Suchraum erheblich reduziert werden können. Tabelle 5.3 zeigt die Anwendungshäufigkeiten der einzelnen Regeln.

Man erkennt, daß eine weitere Reduktion der Regeln relativ leicht möglich gewesen wäre. Die selten angewandten Regeln hätten ohne großen Verlust gestrichen werden können. Alle Regeln, die in mehr als 50% der Fälle angewendet wurden und gestrichen werden sollen, könnten in obligatorische Regeln umgewandelt werden. Dies würde die Korrektheit der Modellierung verbessern, ohne die Variantenzahl zu erhöhen. Aus Zeitgründen konnte dies jedoch nicht durchgeführt werden, da ein neues forced alignment des gesamten Trainings nötig gewesen wäre.

Anhand der neuen Phontranskription wurde ein neues clustering der Triphonmodelle vorgenommen, das 1780 Modelle generierte (im Vergleich zu 1658 Modellen im ursprünglichen System). Diese Modelle wurden nach dem üblichen Trainingsverfahren trainiert.

Regel	absolut	relativ	Regel	absolut	relativ
/?/→	36135	0.552912	m/@n/→	709	0.667608
/@/→	32114	0.498703	x/@n/→N	703	0.316809
/a:/→a	7923	0.234881	/E:/→E	664	0.164153
/n@/n→	4044	0.798894	t/?E/s→	659	0.885753
/t/s→	3016	0.131347	f/l/l→	600	0.575264
/r@/→6	2952	0.644120	/C/s→k	465	0.124132
?Is/t/→	2852	0.809537	/n/m→	455	0.953878
/e:/→@	2510	0.121040	/E:/→e:	388	0.095921
v/i:6/→@	2480	0.446605	n/d/6→	337	0.346351
/Unt/→n	1928	0.281871	/g@nt/l→N	318	0.685345
g/@n/→N	1618	0.350976	b/@n/→m	312	0.158376
v/i:6/→6	1541	0.277508	/n/f→m	309	0.108918
/t/t→	1451	0.552973	/n@/m→	266	0.585903
/t/d→	1331	0.956178	/nf/U→m	259	0.806854
/g@n/→N	1250	0.271150	m/b/6→	253	0.291811
/b@n/→m	1168	0.592893	N/@n/→	217	0.643917
a:/k/→x	930	0.128064	/t/z→	132	0.687500
/?aI/n@→	815	0.299963			

Tabelle 5.3.: Regelanwendungshäufigkeiten (absolut/relativ)

6. Erkennungsexperimente

In diesem Kapitel werden die Ergebnisse der Erkennungsexperimente mit den expandierten Lexika und den neu trainierten Modellen diskutiert. Alle Experimente wurden gemäß den Bedingungen der Verbmobil-Akustikevaluation 96 ausgeführt (siehe [Reinecke 1996]). Es wurden die offizielle 5k Wortliste und das Philips-Bigramm-Sprachmodell verwendet. Für die Entwicklung und Optimierung wurde die xval96-Stichprobe verwendet. Die endgültigen Fehlerraten wurden auf der eval96-Stichprobe bestimmt. Bei beiden Stichproben wurden wie in der Akustikevaluation die Äußerungen, die Buchstabiereinheiten enthalten, nicht ausgewertet.

In Abschnitt 6.1 wird zunächst das Referenzsystem vorgestellt, das Ausgangsbasis der Entwicklung war und dessen Fehlerrate unterboten werden sollte.

In Abschnitt 6.2 wird die Verwendung von Aussprachevarianten ohne eine explizite Gewichtung diskutiert. Anschließend werden in Abschnitt 6.3 die verschiedenen Arten der statischen Gewichtung der Varianten untersucht.

6.1. Baseline-System

Üblicherweise wird in Forschungsarbeiten auf dem Gebiet der Spracherkennung zu Beginn der experimentellen Untersuchung ein System als Ausgangsbasis definiert, das dann durch den Einsatz neuer Methoden verbessert wird (das sogenannte *baseline system*). Dies ist vor allem auch deshalb wichtig, weil nur so exakt eingeschätzt werden kann, welchen Beitrag zur Verbesserung eine neue Methode leistet. Somit muß also im Prinzip jeder einzelne Verbesserungsschritt getrennt evaluiert werden. Dies ist jedoch praktisch nur schwer möglich, da viele Veränderungen einen erheblichen Optimierungsaufwand erfordern.

Auch im Rahmen dieser Arbeit stellte sich das Problem, daß nur ein komplettes Training durchgeführt werden konnte. Dies hatte zur Folge, daß kein wirklich geeignetes Baseline-System zur Verfügung stand. Optimal wäre ein System gewesen, in dem nur die kanonischen Aussprachen und keinerlei multi-words verwendet werden. So ein System stand jedoch leider nicht zur Verfügung, so daß auf das aktuell beste System, das auch für die VM-Evaluation 96 eingesetzt wurde, zurückgegriffen werden mußte.

Beim Training dieses Modellsatzes wurden neben den Standardaussprachen ca. 100 Aussprachevarianten verwendet, die manuell erstellt wurden. Außerdem wurden 571 Wortpaare benutzt, die auf Basis ihrer absoluten Häufigkeiten ausgewählt wurden. Diese

Modelle wurden auch für das forced alignment, das in Kapitel 5 beschrieben wurde, eingesetzt.

Um den Vergleich der verschiedenen Systeme zu ermöglichen, wurden bei allen Experimenten die gleichen pruning-Schwellen verwendet. Die Einstellung der pruning-Schwellen (und damit die Beschränkung des Suchraums) stellt immer einen Kompromiß zwischen der Geschwindigkeit und der Genauigkeit des Erkenners dar. Die verwendete Einstellung erlaubte eine Erkennung in etwa 4-facher Echtzeit, verursachte jedoch noch einige Suchfehler.

Mit diesem System ergab sich auf der xval-Stichprobe eine Wortfehlerrate von 21.3%, auf der eval-Stichprobe 22.7%.

6.2. ungewichtete Varianten

Als erstes Experiment wurden verschiedene Aussprachelexika mit Varianten verwendet, in den auf eine Gewichtung verzichtet wurde. Es wurde jeweils eine Erkennung mit den baseline-Modellen und den neu trainierten Modellen durchgeführt. Die Fehlerraten der einzelnen Konfigurationen auf der xval-Stichprobe sind in Tabelle 6.1 zusammengefaßt.

Lexikon	Anzahl Varianten	Modellsatz	
		baseline	retrain
baseline	5965	21.3	28.0
kanon	5332	22.8	30.1
kanon-mw	5876	21.8	29.2
mf-pron	5876	24.6	21.8
min20-fa	7645		21.1
all-fa	10870	22.4	21.6

Tabelle 6.1.: Fehlerraten bei ungewichteten Varianten (xval)

Zunächst wurden die Aussprachen und multi-words benutzt, die auch bei dem Training des baseline-Modellsatzes verwendet wurden (Lexikon **baseline** in der Tabelle). Bemerkenswert ist, daß die *saubereren* Modelle eine viel höhere Fehlerrate aufweisen als die baseline-Modelle. Dies läßt sich auf die *implizite* Modellierung der Varianten durch die baseline-Modelle zurückführen. Die neu trainierten HMMs setzen eine *explizite* Aussprachemodellierung voraus, da im Training mit den *richtigen* Transkriptionen (gemäß dem alignment) trainiert wurde. Fehlen im Erkennungsvokabular diese Varianten, so äußert sich das durch die erhöhte Fehlerzahl. Diese Beobachtung stützt die Forderung, in Test und Training die gleichen Varianten zu verwenden.

Als nächstes Experiment wurden für alle Wörter der offiziellen Wortliste (*ohne* multi-words) nur die kanonischen Aussprachen verwendet (**kanon**). Einige Fehler in den baseline-Aussprachen wurden hierbei korrigiert. Die Verschlechterung beider Systeme

läßt sich, wie oben beschrieben, durch die nochmals reduzierte Zahl der Varianten und den damit noch größeren Gegensatz zwischen Test- und Trainingsmodellierung erklären. Außerdem fehlen natürlich die multi-words, und somit können die speziell trainierten Triphone nicht eingesetzt werden.

In einem weiteren Versuch wurden die neu ausgewählten multi-words, die auch im Neutraining verwendet wurden, hinzugefügt (**kanon-mw**). Als Aussprachen wurde dabei jeweils die Verkettung der kanonischen Aussprachen ihrer Bestandteile eingetragen. Mit beiden Modellsätzen gab es eine leichte Verbesserung, jedoch wurden die Ergebnisse des baseline-Lexikons nicht erreicht. Dies läßt sich einerseits auf die geringere Zahl der Varianten und im Falle der baseline-Modelle auf die Verwendung anderer multi-words zurückführen. Im Training dieser Modelle war die Hauptfunktion der multi-words, für häufige Wortfolgen das Trainieren von wortübergreifenden Triphonen zu erlauben. Durch den Einsatz anderer multi-words im Test und Training kamen diese Triphone jedoch nur teilweise zum Einsatz. In allen folgenden Untersuchungen wurden die multi-words mit ihren kanonischen Aussprachen genauso wie *normale* Worte behandelt.

Für die folgenden drei Experimente wurden die Häufigkeiten der einzelnen Varianten benutzt, die im forced alignment mit dem 35k-Lexikon bestimmt wurden. Zuerst wurden die kanonischen Aussprachen jeweils durch die im Training am häufigsten beobachtete Variante ersetzt (**mfpron**). Für die Wörter, die nicht im Training vorkamen, wurde weiterhin die kanonische Form verwendet. Die Fehlerraten zeigen, daß eine Übereinstimmung der in Test und Training verwendeten Aussprachen von entscheidender Bedeutung für die Qualität der Modelle ist.

Bei Verwendung aller in der Trainingsstichprobe beobachteten Varianten (**all-fa**) ergab sich eine weitere Verbesserung. Hier wirkt sich jedoch auch die akustische Verwechselbarkeit der vielen Varianten aus, denn die minimale Fehlerrate wurde erreicht, indem die häufigste Variante nur um diejenigen Varianten ergänzt wurde, die in mindestens 20% der Vorkommen eines Wortes beobachtet wurden (**min20-fa**).

Insgesamt sind die Ergebnisse der Experimente mit den ungewichteten Varianten sehr ermutigend. Insbesondere ist es sehr vielversprechend, daß bereits durch das einfache Auswahlverfahren anhand der relativen Häufigkeiten die Erkennungsrate des baseline Systems übertroffen wurde.

6.3. statische Gewichtung

Durch die Einführung einer Gewichtung der Varianten soll den verschiedenen a-priori Wahrscheinlichkeiten der Varianten Rechnung getragen werden. Diese Wahrscheinlichkeiten stellen neben der akustischen Modellierung und dem Sprachmodell eine weitere Wissensquelle dar, die in der Decodierung genutzt werden kann. Die Aussprachegewichte werden für jede Variante eines Wortes als zusätzlicher Faktor in die Zielfunktion des Decoders integriert:

$$f(\mathbf{w}, \mathbf{a} | \mathbf{X}) = p(\mathbf{X} | \mathbf{a}) P(\mathbf{w})^\phi \rho^{|\mathbf{w}|} \prod_{i=1}^n g(v_i, w_i)^\phi \quad (6.1)$$

Hierbei ist $p(\mathbf{X} | \mathbf{a})$ die akustische Wahrscheinlichkeit und $P(\mathbf{w})^\phi \rho^{|\mathbf{w}|}$ die Sprachmodellwahrscheinlichkeit. Die Gewichte $g(v_i, w_i)$ der verwendeten Varianten v_i werden multipliziert, wobei der Aussprachegewichtsfaktor ϕ als Exponent verwendet wird. Der Faktor ϕ wurde bei allen der folgenden Experimente jeweils empirisch optimiert.

Für die Schätzung der Variantengewichte $g(v_i, w_i)$ wurden zwei verschiedene Techniken untersucht. Zunächst wurde ein Verfahren, das allein auf den Häufigkeiten der Wortvarianten im alignment basiert, evaluiert. Als Alternative wurde für jede Variantenregel eine Wahrscheinlichkeit geschätzt, die dann zu den Gewichten der Wortvarianten kombiniert wurden.

6.3.1. wortbasierte Schätzung

Als Schätzung der a-priori Wahrscheinlichkeit $P(\mathbf{a} | w)$ einer Aussprache \mathbf{a} des Wortes w wurde die relative Häufigkeit dieser Variante, die im Trainingsalignment beobachtet wurde, benutzt. Somit hatten nur die tatsächlich beobachteten Varianten eine Wahrscheinlichkeit größer Null. Für ein Experiment mit dem reduzierten Variantenlexikon (**min20-fa**, 7645 Varianten, siehe oben) wurden die Varianten, die in weniger als 20% der Vorkommen beobachtet wurden, verworfen. Die Häufigkeiten der anderen Varianten wurden anschließend renormalisiert.

Die tatsächlich im Erkenner benutzten Gewichte $g(v_i, w_i)$ wurden auf zwei verschiedene Arten bestimmt. Zunächst wurden direkt die a-priori Wahrscheinlichkeiten verwendet, d.h. die Gewichte der Varianten eines Wortes addieren sich zu eins (in Tabellen 6.2 Eintrag **sum** als Skalierung):

$$\sum_{v \in \mathcal{V}(w)} g(v, w) = 1 \quad (6.2)$$

Alternativ wurde die bereits in Abschnitt 4.2.2 vorgestellte Renormierung der besten Variante auf eins vorgenommen (**best**):

$$g(v_i, w) = \frac{P(v_i | w)}{\max_{v \in \mathcal{V}(w)} P(v | w)} \quad (6.3)$$

Die Erkennung wurde mit zwei der oben beschriebenen Lexika vorgenommen (**min20-fa**: 7645 Varianten, **all-fa**: 10870 Varianten). Die im Training nicht gesehenen Wörter wurden nur mit ihrer kanonischen Aussprache und Gewicht eins eingetragen. Für jede

Konfiguration wurde separat der Faktor ϕ optimiert. Wie erwartet liegt er für die best-Skalierung jeweils höher als in der entsprechenden anderen Konfiguration, da durch die Renormierung in Gleichung 6.3 die Gewichte aller Varianten größer werden und daher stärker gewichtet werden müssen.

Die Wirksamkeit der Gewichtung kann man daran erkennen, daß alle im Training gesehenen Varianten benutzt werden können, ohne daß die Erkennungsrate gegenüber dem kleineren Lexikon fällt, wie es ohne Gewichte beobachtet wurde. Außerdem ist die Fehlerrate um 1% absolut besser als in dem besten System mit ungewichteten Varianten.

Um die Fehlerrate weiter zu senken, wurde versucht, die Häufigkeiten aus einem neuen forced alignment des Trainings mit den neu trainierten Triphonen zu verwenden. Die Überlegung hierbei ist, daß es für das Erkennungsergebnis irrelevant ist, welche Variante eines Wortes gewählt wird. Insbesondere muß nicht die *richtige* Variante gewählt werden, die ein Experte bestimmt. Insofern sollte es besser sein, für die Bestimmung der Gewichte im Training die gleichen Modelle wie in der Erkennung zu benutzen. Hierdurch leidet zwar vielleicht die Übereinstimmung mit den manuell erstellten Labels, aber die Gewichtung sollte besser an den Erkennenner angepaßt sein. Um diese Vermutung zu verifizieren, wurde mit den neu trainierten Triphonen noch einmal ein forced alignment vorgenommen, hierbei wurde allerdings aus Zeitgründen auf die MLLR-Sprecheradaption verzichtet. Aus diesem alignment wurden wiederum die Gewichte bestimmt und in der Erkennung benutzt (**all-fa-tri**). Allerdings verschlechtert sich hierbei die Fehlerrate gegenüber dem entsprechenden System, das das ursprüngliche Monophon-alignment verwendet. Diese Verschlechterung kann leider nicht plausibel erklärt werden.

Als abschließendes Experiment wurde die Gewichtung aus einem forced alignment der *Teststichprobe* bestimmt (**orakel-tri**). Dies bietet eine Abschätzung der Fehlerrate, die mit einer statischen Gewichtung bestenfalls erreicht werden kann. Die Gewichte spiegeln nun genau die Häufigkeiten in der Teststichprobe wieder. In der normalen Erkennung kann solches Wissen natürlich nicht angewandt werden, da die *korrekte* Worttranskription nicht zur Verfügung steht.

Die Ergebnisse der Versuche mit der wortbasiert geschätzten Gewichtung sind in Tabelle 6.2 zusammengefaßt.

Lexikon	Skalierung	ϕ_{opt}	WER
min20-fa	sum	6	20.5
min20-fa	best	24	20.5
all-fa	sum	15	20.4
all-fa	best	20	20.1
all-fa-tri	best	23	20.3
orakel-tri	best	25	19.2

Tabelle 6.2.: Fehlerraten bei statischer, wortbasierter Gewichtung (xval)

Insgesamt kann man erkennen, daß die wortbasierte statische Gewichtung noch einmal eine Verbesserung gegenüber den ungewichteten Variantenlexika bietet. Es ist erstaunlich, wie dicht man mit den im Training geschätzten Gewichten an die optimale Gewichtung des Orakelexperimentes herankommt. Andererseits ist es enttäuschend, daß man selbst mit dem Orakelwissen nur eine Fehlerrate von 19.2% erreicht. Es wäre denkbar, daß sich hier insbesondere die Beschränkung auf die im Training gesehenen Wörter und Varianten negativ auswirkt.

6.3.2. regelbasierte Schätzung

In den bisher beschriebenen Experimenten wurden die Regeln nur benutzt, um die möglichen Varianten zu generieren. Bezüglich der Gewichtung wurden die Regeln jedoch nicht verwendet. In den folgenden Versuchen werden die Gewichte aus a-priori Wahrscheinlichkeiten der einzelnen Regeln berechnet. Dies hat den Vorteil, daß auch für ungesehene Varianten und sogar ungesehene Wörter Gewichte zur Verfügung stehen.

Im forced alignment wurden die Stellen bestimmt, an den eine der Regeln anwendbar war. Für jede Regel r wurde nun die relative Häufigkeit der tatsächlichen Regelanwendung als Schätzung der Regelwahrscheinlichkeit $P(r)$ verwendet.

Zur Schätzung der Wahrscheinlichkeiten der einzelnen Aussprachevarianten eines Wortes wurde das bereits in Abschnitt 4.4.2 vorgestellte Verfahren angewandt. Hierbei wurden wie im Training auch Varianten zugelassen, in den eine Regel nur an einigen der möglichen Stellen angewandt wurde. Zur Berechnung der Variantenwahrscheinlichkeit wurde Gleichung 4.13 verwendet, die hier noch einmal angegeben wird:

$$P(\mathbf{a}|w) = \frac{\prod_{r \in \mathcal{R}^+(\mathbf{a},w)} P(r) \prod_{r \in \mathcal{R}^-(\mathbf{a},w)} (1 - P(r))}{Z} \quad (6.4)$$

Die Normierungskonstante Z wird so berechnet, daß sich eine echte Wahrscheinlichkeitsverteilung ergibt:

$$Z = \sum_{\mathbf{a} \in \mathcal{V}(w)} \left(\prod_{r \in \mathcal{R}^+(\mathbf{a},w)} P(r) \prod_{r \in \mathcal{R}^-(\mathbf{a},w)} (1 - P(r)) \right) \quad (6.5)$$

Die Experimente wurden mit dem Lexikon durchgeführt, das durch Anwendung der Regeln auf die kanonischen Formen generiert wurde. Dieses Lexikon wurde auch beim forced alignment des Trainingsmaterials verwendet und enthält knapp 35000 Aussprachen. Wie bei den wortbasierten Experimenten wurden wieder beide Arten der Skalierung der Gewichte ausprobiert. Außerdem wurden auch die Regelwahrscheinlichkeiten aus dem Triphonalignment benutzt. Hier konnte auch die erwartete Verbesserung gegenüber dem Monophonalignment beobachtet werden. Allerdings gelang es nicht, mit dem regelbasierten Ansatz die Fehlerraten der wortbasierten Gewichtung zu unterbieten. Die Ergebnisse der drei verschiedenen Konfigurationen sind in Tabelle 6.3 aufgeführt.

Lexikon	Skalierung	ϕ_{opt}	WER
all	sum	3	20.8
all	best	5	20.7
all-tri	best	7	20.5

Tabelle 6.3.: Fehlerraten bei statischer, regelbasierter Gewichtung (xval)

Eine mögliche Ursache für die Unterlegenheit der regelbasierten Gewichtung ist die große Zahl der verwendeten Varianten und die damit verbundenen hohen akustischen Verwechselbarkeit.

6.4. Zusammenfassung

Um die Effektivität der verschiedenen Modellierungen objektiv einzuschätzen, wurden sie auf einer ungesehenen Stichprobe getestet. Dies ist notwendig, da anhand der xval-Stichprobe verschiedene Parameter (z.B. ϕ) optimiert wurden. Eine unvoreingenommene Schätzung der Wortfehlerrate kann daher nur durch ein Experiment auf ungesehenen Daten ermittelt werden.

Die Fehlerraten der verschiedenen Systeme jeweils auf der xval- und der eval-Stichprobe sind in Tabelle 6.4 zusammengefaßt.

System	Anzahl Varianten	Stichprobe	
		xval	eval
baseline	5965	21.3	22.7
no-weight	7645	21.1	
word-based	10870	20.1	22.5
rule-based	34470	20.5	22.0

Tabelle 6.4.: Fehlerraten der verschiedenen Modellierungen

Erwartungsgemäß liegen die Fehlerraten auf der ungesehenen Stichprobe etwas höher. Überraschend ist allerdings, daß auf der eval-Stichprobe die Erkennung mit den regelbasiert gewichteten Varianten besser funktioniert als mit den wortbasiert gewichteten. Dies deutet darauf hin, daß der regelbasierte Ansatz eine robustere Gewichtung leistet als der wortbasierte. Vermutlich sollte auch die Optimierung von ϕ auf einer größeren Stichprobe vorgenommen werden, damit sich die Ergebnisse der Entwicklungsstichprobe besser auf ungesehene Daten übertragen lassen.

Insgesamt muß jedoch festgestellt werden, daß die Verbesserung durch die statische Gewichtung eher enttäuschend ist. Insbesondere liegt die Fehlerrate von 22.0% gerade an der Grenze des 95%-Konfidenzintervalls des baseline-Systems, insofern wurde keine statistisch signifikante Verbesserung der Erkennungsleistung erreicht.

7. Zusammenfassung & Ausblick

Das zentrale Problem der automatischen Spracherkennung ist die große Variabilität sprachlicher Äußerungen. In dieser Arbeit wurde speziell der Aspekt der Aussprachvariabilität untersucht. Es wurde versucht, das üblicherweise verwendete kanonische Aussprachelexikon durch ein besseres Modell zu ersetzen.

Zunächst wurde diskutiert, wie eine Aussprachemodellierung in einen statistischen Spracherkennung integriert werden kann. Die Verwendung eines kanonischen Lexikons ergibt sich hierbei ganz natürlich als Spezialfall.

Die verschiedenen in der Literatur vorgeschlagenen Verfahren der Aussprachemodellierung wurden diskutiert. Ein Hauptunterschied zwischen verschiedenen Verfahren ist die Erstellung des Modells. Prinzipiell kann das Modell manuell erstellt werden oder automatisch aus einer Trainingsstichprobe gelernt werden.

Für die Experimente in dieser Arbeit wurde eine Kombination dieser beiden Verfahren gewählt. Die zulässigen Aussprachen wurden mit phonetischen Transformationsregeln, die manuell ausgewählt wurden, generiert. Die Gewichtung der so erzeugten Varianten wurde dann automatisch anhand der Trainingsstichprobe gelernt.

Der Vorteil dieses Vorgehens ist, daß durch die manuelle Regelerstellung ein zu starke Übergenerierung von Varianten verhindert werden kann. Die Gewichtung der Varianten wird dann anhand einer Statistik, die auf der Trainingsstichprobe erstellt wurde, vorgenommen. Auf diese Weise kann Expertenwissen mit statistischen Informationen kombiniert werden, um die Robustheit des Modells zu erhöhen.

Die Regeln wurden aufbauend auf die MAUS-Regeln (siehe [Kipp 1998]) durch Untersuchung des forced alignments entwickelt. Das Ergebnis war ein Regelsatz, der 35 Regeln enthält.

Mit den durch die Regeln definierten Varianten wurde ein forced alignment des Trainingsmaterials durchgeführt, um neue, bessere Transkriptionen für das Training der akustischen Modelle zu erhalten. Die Phonfehlerrate dieser neuen Transkription im Vergleich zu manuell erstellten Phonlabeln konnte so um mehr als 30% relativ reduziert werden.

Bei der Verwendung der Varianten mit den neuen akustischen Modellen ergab sich auf einer unabhängigen Stichprobe eine Reduktion der Fehlerrate, die jedoch statistisch nicht signifikant ist (auf dem 95%-Niveau). Die verschiedenen Experimente auf der Krossvalidierungsstichprobe belegen allerdings eindeutig die Notwendigkeit einer Gewichtung der Varianten insbesondere bei der Verwendung großer Lexika.

Eine Technik, die eventuell eine weitere Verbesserung der Erkennungsrate bewirken

würde, ist die dynamische Gewichtung der Varianten. Diese Verallgemeinerung konnte jedoch im Rahmen dieser Arbeit nicht experimentell untersucht werden.

Die Ergebnisse sprechen dafür, daß die Zahl der Varianten noch weiter reduziert werden sollten. Hierdurch würde die akustische Verwechselbarkeit der Wörter reduziert werden, was möglicherweise zu einer besseren Erkennung beitragen würde. Auf diese Weise würde man einen Kompromiß zwischen einer expliziten und einer impliziten Variantenmodellierung erreichen.

Es ist anzunehmen, daß durch die Einbeziehung von weiteren Kontextfaktoren in die Regelanwendung die Vorhersagegenauigkeit noch weiter verbessert werden kann. Hierbei sind besonders die Silbenstruktur und die Position der Betonung im Wort plausible Indikatoren.

Sehr interessant wäre auch ein Vergleich des regelbasierten Ansatzes mit der Verwendung von Entscheidungsbäumen, die die Phonrealisierung vorhersagen. Möglich wäre auch eine Kombination der beiden Ansätze, indem von Hand erstellte Regeln verwendet werden, aber ihre Anwendungswahrscheinlichkeit durch Entscheidungsbäume vorhergesagt wird. Dies würde auch eine dynamische Veränderung der Variantengewichte ermöglichen.

A. Phonsymbole

Tabelle A.1 zeigt den in dieser Arbeit verwendeten Symbolsatz mit Beispielen zu den einzelnen Phonen. Der Symbole basieren im wesentlichen auf dem SAMPA-Inventar (siehe [Wells 1987]).

Plosive			Liquide		
p	Pein	paIn	l	Leim	laIm
b	Bein	baIn	r	Reim	raIm
t	Teich	taIC	Vokale		
d	Deich	daIC	I	Sitz	zIts
k	Kunst	kUnst	E	Gesetz	g@zEts
g	Gunst	gUnst	a	Satz	zats
Glottalverschluß			O	Trotz	tr0ts
?	Verein	fE6?aIn	U	Schutz	SUts
Frikative			Y	hübsch	hYpS
f	fast	fast	9	plötzlich	p19tslIC
v	was	vas	i:	Lied	li:t
s	Tasse	tas@	e:	Beet	be:t
z	Hase	ha:z@	E:	spät	SpE:t
S	waschen	vaS@n	a:	Tat	ta:t
Z	Etage	?e:ta:Z@	o:	rot	ro:t
C	sicher	zIC6	u:	Blut	blu:t
j	Jahr	ja:6	y:	süß	zy:s
x	Buch	bu:x	2:	blöd	bl2:t
h	Hand	hant	aI	Eis	aIs
Nasale			aU	Haus	haUs
m	mein	maIn	OY	Kreuz	kr0yts
n	nein	naIn	@	bitte	bit@
N	Ding	dIN	6	besser	bEs6

Tabelle A.1.: Symbolsatz

Literaturverzeichnis

- [Adda-Decker und Lamel 1997] Adda-Decker, Martine und L. Lamel (1997). *The use of lexica in automatic speech recognition*. course material for the fifth ELSNET Summer School on Language and Speech Communication.
- [Adda et al. 1997] Adda, Gilles, M. Adda-Decker, J.-L. Gauvain und L. Lamel (1997). *Text Normalization and Speech Recognition in French*. In: *Proc. Eurospeech*.
- [Aubert und Dugast 1995] Aubert, Xavier und C. Dugast (1995). *Improved Acoustic-Phonetic Modeling in Philips' Dictations System by Handling Liasons and Multiple Pronunciations*. In: *Proc. Eurospeech'95*, S. 767–770.
- [Bourlard et al. 1996] Bourlard, Hervè, H. Hermansky und N. Morgan (1996). *Towards increasing speech recognition error rates*. *Speech Communication*, 18(3):205–231.
- [Breiman et al. 1984] Breiman, Leo, J. H. Friedman, R. A. Olsen und C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software.
- [Chen 1990] Chen, Francine R. (1990). *Identification of contextual factors for pronunciations networks*. In: *Proc. ICASSP'90*, S. 753–756.
- [Cohen 1989] Cohen, Michael Harris (1989). *Phonological Structures for Speech Recognition*. Doktorarbeit, Computer Science Division, University of California, Berkley.
- [Cremelie und Martens 1995] Cremelie, Nick und J.-P. Martens (1995). *On The Use Of Pronunciation Rules For Improved Word Recognition*. In: *Proc. Eurospeech'95*, S. 1747–1750.
- [Cremelie und Martens 1997] Cremelie, Nick und J.-P. Martens (1997). *Automatic Rule-based Generation of Word Pronunciation Networks*. In: *Proc. Eurospeech*.
- [Downey und Wiseman 1997] Downey, Simon und R. Wiseman (1997). *Dynamic and Static Improvements to Lexical Baseforms*. In: *Proc. Eurospeech*.
- [Fach 1996] Fach, Marcus (1996). *Optimierung von Wortuntereinheiten bei der Erkennung gesprochener Sprache*. Diplomarbeit, Universität Stuttgart. in German.

- [Finke und Waibel 1997] Finke, Michael und A. Waibel (1997). *Speaking Mode dependent Pronunciation Modeling in Large Vocabulary Speech Recognition*. In: *Proc. Eurospeech*.
- [Flach 1995] Flach, Gudrun (1995). *Modelling Pronunciation Variability for Special Domains*. In: *Proc. Eurospeech'95*, S. 1743–1746.
- [Fukada und Sagisaka 1997] Fukada, Toshiaki und Y. Sagisaka (1997). *Automatic Generation of a Pronunciation Dictionary Based on a Pronunciation Network*. In: *Proc. Eurospeech*.
- [Gauvain et al. 1997] Gauvain, J.L., G. Adda, L. Lamel und M. Adda-Decker (1997). *Transcribing Broadcast News: The LIMSI Nov96 Hub4 System*. In: *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia.
- [Gauvain et al. 1994] Gauvain, J.L., L. Lamel, G. Add und M. Adda-Decker (1994). *Speaker-Independent continuous speech dictation*. *Speech Communication*, 15(1-2):21–37.
- [Gelfand et al. 1991] Gelfand, Saul B., C. Ravishankar und E. J. Delp (1991). *An iterative growing and pruning algorithm for classification*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):163–174.
- [Giachin et al. 1991] Giachin, Egidio P., A. E. Rosenberg und C.-H. Lee (1991). *Word Junction modeling using phonological rules for HMM-based continuous speech recognition*. *Computer Speech & Language*, 5:155–168.
- [Godfrey et al. 1992] Godfrey, J. J., E. C. Holliman und J. McDaniel (1992). *SWITCHBOARD: Telephone speech corpus for research and development*. In: *Proc. ICASSP'92*, S. 517–520.
- [Haiber 1998] Haiber, Udo (1998). *Sprecheradaptation in einem statistischen Spracherkennungssystem*. Doktorarbeit, Universität Ulm.
- [Humphries und Woodland 1997] Humphries, J.J. und P. Woodland (1997). *Using Accent-specific Pronunciation Modelling for improved Large Vocabulary Continuous Speech Recognition*. In: *Proc. Eurospeech*.
- [Humphries et al. 1996] Humphries, J.J., P. Woodland und D. Pearce (1996). *Using Accent-specific Pronunciation Modelling for Robust Speech Recognition*. In: *Proc. ICSLP 96*, Philadelphia, USA.
- [Jost et al. 1997] Jost, Uwe, H. Heine und G. Evermann (1997). *What's wrong with the lexicon – An attempt to model pronunciations probabilistically*. In: *Proc. Eurospeech*.
- [Katz 1987] Katz, Slava M. (1987). *Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer*. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35:400–401.

-
- [Kemp 1995] Kemp, Thomas (1995). *Regelbasiert generierte Aussprachevarianten für Spontansprache*. In: *Natural Language Processing and Speech Technology*. Dafyd Gibbon (Ed.) Mouton de Gruyter, Berlin 1996, ISBN 3-11-015449-8.
- [Kipp 1998] Kipp, Andreas (1998). Doktorarbeit, Ludwig-Maximilian Universität München.
- [Kipp et al. 1996] Kipp, Andreas, M.-B. Wesenick und F. Schiel (1996). *Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora*. In: *Proc. ICSLP 96*, Philadelphia, USA.
- [Kipp et al. 1997] Kipp, Andreas, M.-B. Wesenick und F. Schiel (1997). *Pronunciation Modeling applied to Automatic Segmentation of Spontaneous Speech*. In: *Proc. Eurospeech*.
- [Kohler 1995] Kohler, Klaus (1995). *Einführung in die Phonetik des Deutschen*. Erich Schmidt Verlag, Berlin, 2. Aufl.
- [Kuhn et al. 1996] Kuhn, Thomas, P. Fetter, A. Kaltenmeier und P. Regel-Brietzmann (1996). *DP-Based Wordgraph Pruning*. In: *Proc. ICASSP'96*, Bd. 2, S. 861, Atlanta, USA.
- [Lamel und Adda 1996] Lamel, Lori und G. Adda (1996). *On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition*. In: *Proc. ICSLP 96*, Philadelphia, USA.
- [Lee 1989] Lee, Kai-Fu (1989). *Automatic Speech Recognition — The Development of the SPHINX-System*. Kluwer Academic Publishers.
- [Legetter 1995] Legetter, C. J. (1995). *Improved Acoustic Modelling for HMMs Linear Transformations*. Doktorarbeit, Cambridge University.
- [Lowerre 1976] Lowerre, B. (1976). *The Harpy Speech Recognition System*. Technischer Bericht Carnegie Mellon University.
- [Markey und Ward 1997] Markey, K.L. und W. Ward (1997). *Lexical Tuning Based on Triphone Confidence Estimation*. In: *Proc. Eurospeech*.
- [Mirghafori et al. 1995] Mirghafori, Nikki, E. Fosler und N. Morgan (1995). *Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes*. In: *Proc. Eurospeech'95*.
- [Mühlenfeld 1986] Mühlenfeld, Reinhard (1986). *Verifikation von Worthypothesen*. Doktorarbeit, Uni Erlangen.
- [Nock et al. 1997] Nock, H.J., M. Gales und S. Young (1997). *A Comparative Study of Methods for Phonetic Decision-Tree State Clustering*. In: *Proc. Eurospeech*.

- [Odell 1995] Odell, Julian (1995). *The Use of Context in Large Vocabulary Speech Recognition*. Doktorarbeit, Cambridge University.
- [Ostendorf et al. 1996] Ostendorf, M., B. Byrne, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shiber, D. Talkin, A. Waibel, B. Wheatley und T. Zeppenfeld (1996). *Modeling Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode*. In: *Proc. ICSLP96*.
- [Randolph 1990] Randolph, Mark A. (1990). *A data-driven method for discovering and predicting allophonic variation*. In: *Proc. ICASSP'90*, S. 1177–1180.
- [Reinecke 1996] Reinecke, Jörg (1996). *Evaluierung der signalnahen Spracherkennung im Verbundprojekt VERBMOBIL (Herbst 1996)*. Memo 113, Verbmobil, TU Braunschweig.
- [Riley et al. 1997] Riley, M., W. Byrne, S. Khudanpur, J. McDonough, H. Nock, M. Saraclar, C. Wooters und G. Zavaliagkos (1997). *Pronunciation Modelling using Decision Trees*. slides of the final presentation at the Summer Research Workshops on Speech Recognition.
- [Riley et al. 1995] Riley, Michael, A. Ljolje, D. Hindle und F. Pereira (1995). *The AT&T 60,000 Word Speech-to-Text System*. In: *Proc. Eurospeech'95*, S. 207–210.
- [Riley 1991] Riley, Michael D. (1991). *A statistical model for generating pronunciation networks*. In: *Proc. ICASSP'91*, Bd. 2, S. 737–740.
- [Rosenfeld 1995] Rosenfeld, Ronald (1995). *Optimizing lexical and N-gram coverage via judicious use of linguistic data*. In: *Proc. Eurospeech'95*, S. 1763–1766.
- [Sakoe und Chiba 1978] Sakoe, Hiroaki und S. Chiba (1978). *Dynamic programming Algorithm optimization for Spoken Word Recognition*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26:43–49.
- [Schukat-Talamazzini 1995] Schukat-Talamazzini, Ernst Günther (1995). *Automatische Spracherkennung*. Vieweg.
- [Sloboda 1995] Sloboda, Tilo (1995). *Dictionary Learning: Performance through Consistency*. In: *Proc. ICASSP'95*, Bd. 1, S. 453, Detroit, USA.
- [Sloboda und Waibel 1996] Sloboda, Tilo und A. Waibel (1996). *Dictionary Learning for Spontaneous Speech Recognition*. In: *Proc. ICSLP 96*, Philadelphia, USA.
- [Tajchman et al. 1995a] Tajchman, Gary, E. Fosler und D. Jurafsky (1995a). *Building Multiple Pronunciation Models for Novel Words using Explanatory Computational Phonology*. In: *Proc. Eurospeech'95*, S. 2247–2250.

-
- [Tajchman et al. 1995b] Tajchman, Gary, D. Jurafsky und E. Fosler (1995b). *Learning Phonological Rule Probabilities from Speech Corpora with Explanatory Computational Phonology*. In: *Proc. ACL'95*.
- [Valtchev 1995] Valtchev, Valtcho (1995). *Discriminative Methods in HMM-based Speech Recognition*. Doktorarbeit, Cambridge University.
- [Wahlster 1993] Wahlster, W. (1993). *Verbmobil—Translation of Face-to-Face Dialogs*. In: *Proc. Eurospeech'93*, S. 29–38, Berlin, Germany.
- [Weintraub et al. 1996] Weintraub, Mitch, K. Taussig, K. Hunicks-Smith und A. Snodgrass (1996). *Effect of Speaking Style on LVCSR Performance*. In: *unknown Proc.*
- [Wells 1987] Wells, J.C. (1987). *Computer-coded phonetic transcription*. *Journal of the International Phonetic Association*, 17(2):94–114.
- [Wesenick 1996] Wesenick, Maria-Barbara (1996). *Automatic Generation of German Pronunciation Variants*. In: *Proc. ICSLP 96*, Philadelphia, USA.
- [Westendorf und Jelitto 1996] Westendorf, Christian-Michael und J. Jelitto (1996). *Learning Pronunciation Dictionary from Speech Data*. In: *Proc. ICSLP 96*, Philadelphia, USA.
- [Wooters 1993] Wooters, Charles Clayton (1993). *Lexical Modeling in a Speaker Independent Speech Understanding System*. Technischer Bericht TR-93-068, International Computer Science Institute, Berkeley, CA.
- [Wooters und Stockle 1994] Wooters, Chuck und A. Stockle (1994). *Multiple-Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System*. In: *Proc. ICSLP'94*.
- [Young et al. 1989] Young, S.J., N. Russell und J. Thornton (1989). *Token Passing: a Simple Conceptual model for Connected Speech Recognition Systems*. Technischer Bericht TR38, Cambridge University Engineering Department.