Regelbasiert generierte Aussprachevarianten für Spontansprache

Thomas Kemp
Interactive Systems Labs, ILKD
Universität Karlsruhe, 76128 Karlsruhe

Abstract

We investigate the occurrence of six classes of pronounciation variants in a very large corpus of spontaneous german speech. A high correlation between the dialect region of the speaker, the speaking rate and the word position within the utterance is observed. The integration of the pronounciation variants into a speech recognition system yields moderate improvement of the word error rate.

In einem sehr großen Korpus spontaner deutscher Sprache wird das Auftreten von sechs Klassen von phonetischen Aussprachevarianten untersucht. Dabei ergibt sich eine deutliche Abhängigkeit der Realisierungshäufigkeit von der Sprechgeschwindigkeit, der Dialektregion des Sprechers und der Stellung des Wortes in der Äußerung. Durch die Aufnahme der Aussprachevarianten in die Modellparameter eines Spracherkenners konnte eine Reduktion der Fehlerrate des Spracherkenners erzielt werden.

1 Einleitung

Zu den wichtigsten Wissensquellen in einem maschinellen Spracherkennungssystem gehört das phonetische Wörterbuch, eine Datenbank über die phonetische Realisierung der zu erkennenden Worte. Dieses Wörterbuch ist meist so aufgebaut, daß für jedes Wort eine *kanonische* phonetische Ausspracheform enthalten ist. Setzt man ein Spracherkennungssystem in einer Applikation ein, sollte es außer den kanonischen Aussprachen der Worte auch über Wissen über die möglichen Aussprachevarianten und deren Wahrscheinlichkeitsverteilung verfügen. Man benötigt daher Methoden, um ein gegebenes phonetisches Wörterbuch mit Aussprachevarianten anzureichern.
Die möglichen Aussprachevarianten von Hand zu ermitteln ist ermüdend und fehlerträchtig. Daher sind verschiedene Algorithmen zum automatischen Bestimmen von Aussprachevarianten vorgeschlagen worden. Grob lassen sie sich in zwei Kategorien klassifizieren:

1. die regelbasierte Generierung von phonetischen Aussprachevarianten aufgrund von linguistisch-phonetischem Wissen, und
2. die automatische Transkription von unterschiedlichen Realisierungen eines Wortes durch maschinelle phonetische Analyse in einem Korpus (z.B. mittels eines Phonemerkenners)

Mit den Verfahren der zweiten Kategorie (vgl. [1], [3] oder [6]) kann die tatsächliche a-priori-Häufigkeit jeder gefundenen Aussprachevariante abgeschätzt werden. Dadurch kann erreicht werden, daß das Wörterbuch nur um relevante Varianten erweitert wird und damit das Problem der Zunahme der Verwechselbarkeit weniger stark entsteht. Allerdings müssen zur robusten statistischen Bestimmung einer Aussprachevariante mehrere Instanzen des zu modellierenden Wortes in der Datenbasis vorhanden sein. Für selten auftretende Worte ist eine Variantengenerierung nach dieser Methode daher i.allg. nicht möglich.

Ein solches Problem entsteht nicht bei der regelbasierten Generierung von Aussprachevarianten. Hier können auch für selten belegte Worte des Vokabulars Varianten generiert werden. Allerdings ist die Gefahr der Übergenerierung gegeben, wenn mehrere Regeln z.T. mehrfach auf eine kanonische Wortform angewendet werden können.
Ziel dieser Untersuchung ist es, den Nutzen der Verwendung von regelbasiert generierten Aussprachevarianten zu analysieren. Konkret soll geklärt werden

1. inwieweit die Realisierung von bestimmten Varianten vom Kontext abhängt und von daher vorhersagbar ist,
2. wie hoch der Einfluß der inadäquaten Wortmodellierung in der Trainingsphase eines Spracherkenners auf die Erkennungsleistung ist,
3. wie hoch der Einfluß der inadäquaten Modellierung der Wörter während der eigentlichen Spracherkennung ist.

Dieses Papier ist wie folgt aufgebaut. Zunächst werden der verwendete Korpus und die eingesetzten Regeln zur Generierung von Aussprachevarianten vorgestellt. Dann werden das Auftreten von Aussprachevarianten im Korpus sowohl global als auch in Abhängigkeit des Kontextes untersucht. Dabei werden zur Untersuchung Entscheidungsbäume herangezogen. Zum Schluß werden Experimente mit dem Ziel der Verbesserung der Qualität des Spracherkenners durch explizite Variantenmodellierung vorgestellt.

2 Verwendeter Korpus

Ein Teil des zur Analyse verwendeten Korpus wurde im Rahmen des BMBF-Verbundprojekts VERBMOBIL an insgesamt vier Datensammelorten (Bonn, Karlsruhe, Kiel und München) gesammelt und transliteriert. Ein weiterer Teil wurde im Rahmen des JANUS-Projekts an der Universität Karlsruhe gesammelt. In beiden Fällen handelt es sich um Aufnahmen von spontanen Mensch-zu-Mensch Dialogen aus der

Terminvereinbarungs-Domäne.
Insgesamt standen zur Analyse 10735 deutsche Äußerungen mit zusammen knapp 2310000 gesprochenen Worten zur Verfügung. Die dialektale Herkunft der Sprecher wurde nicht für alle Aufnahmen erfaßt. Für eine Stichprobe von 106 Sprechern wurde überprüft, inwieweit der Aufnahmeort als Näherung für die Herkunft des Sprechers geeignet ist. Die Annahme erwies sich dabei in 83 Fällen (78%) als korrekt.

3 Regeln zur Generierung von Aussprachevarianten

Um allgemeingültige Regeln zur Generierung von Aussprachevarianten in einer Sprache aufzustellen, ist eine genaue Korpusanalyse erforderlich. Zwei solche Korpusanalysen wurden unabhängig voneinander von [4] und [5] durchgeführt. Dabei wurden die folgenden Aussprachevariationen beobachtet.

1	keine
2	Ausfall des Glottalverschlußlautes
3	'schwa'-Elision
4	Nasalassimilation nach 'schwa'-Elision
5	Reduktion des 'r' bei Vokal-r-Verbindungen
6	Änderungen der Vokaldauer
7	Änderungen der Vokalqualität
8	stimmhaft-stimmlos-Änderung [v - f]
9	Lautverschmelzung bei gleichem Silbenaus- und Anlaut
10	Monophthongierung von Diphthongen
11	Glottalisierung eines Plosivs in einem Kontext von Nasalkonsonanten
12	Nasalisierung der Endsilbe '-nden'

Table 1: Beobachtete Typen von Aussprachevarianten

Die ersten sechs Regeln stellten nach [4] die große Mehrzahl der Abweichungen und wurden daher in die Untersuchung aufgenommen.

Um eine Übergenerierung von Varianten zu vermeiden, wurden für ein gegebenes Wort alle anwendbaren Regeln an allen Stellen angewendet. Die so entstandenen Varianten wurden jedoch nicht weiter durch nochmalige Regelanwendungen modifiziert. Dieses Verfahren führte zu einer Generierung von durchschnittlich 1,58 zusätzlichen Varianten pro Wort, was einer Vergrößerung des Vokabulars um den Faktor 2,58 entspricht. Beispiele für die Anwendung der Regeln und die genaue Anzahl der erzeugten Varianten werden in Tabelle 2 zusammengefaßt.

Die erlaubten Vokalqualitätänderungen sind 'e' nach '@', 'E:' nach 'e:', und 'i' nach '@'.

Nr.	Beschreibung	Beispiel	kanonische Wortform	variierte Wortform	Anzahl
1	keine	Äther	QE:t6	QE:t6	9149
2	Glottislautelision	Äther	QE:t6	E:t6	2138
3	Schwa-Elision	haben	hab@n	habn	2578
4	Nasalassimilation	haben	hab@n	habm	2321
5	'r'-Reduktion	Jörg	j2rk	j26k	442
6	Vokalverkürzungen	guten	gu:t@n	gut@n	3224
7	Vokalqualitätänderungen	Messe	mes@	m@s@	3807

Table 2: Verwendete Regeln und Anwendungshäufigkeiten auf einem Wörterbuch von 9149 Worten

4 Analyse der Vorkommenshäufigkeit der einzelnen Varianten

4.1 Häufigkeit des Auftretens im Korpus

Für den gesamten Analysekorpus wurden mittels *forced alignment* die besten Wortketten bestimmt. Dabei konnte für jedes Wort jede seiner Aussprachevarianten gleichwahrscheinlich eingesetzt werden. Der Sprach-erkenner bestimmte dabei für jede Instanz eines Wortes, welche der angebotenen Aussprachevarianten lokal vorlag. Auf diese Weise ergab sich die in Tabelle 3 angegebene Verteilung der Aussprachevarianten.

Nr.	Beschreibung	Häufigkeit (absolut)	Häufigkeit (relativ)
1	keine	205464	88.9
2	Glottislautelision	10597	4.6
3	Schwa-Elision	5320	2.3
4	Nasalassimilation	2915	1.3
5	'r'-Reduktion	385	0.2
6	Vokalverkürzungen	4323	1.9
7	Vokalqualitätänderungen	1956	0.9
-	alle	230960	100.0

Table 3: Häufigkeit der Realisierung der verschiedenen Aussprachevarianten

Vergleicht man die tatsächlichen Realisierungen mit der Anzahl der vorhandenen Varianten im Wörterbuch, so fällt die relativ geringe Häufigkeit der Realisierung von Vokalqualitätänderungen und die große Wahrscheinlichkeit der Elision von Glottalverschlußlaut und 'schwa' auf. Abweichend von den Ergebnissen von Flach [4] war die Zahl der Vokalverkürzungen deutlich größer als die der Vokalqualitätänderungen.

Für die Abweichung von der kanonischen Aussprache sind zahlreiche Gründe denkbar. In den nächsten Abschnitten soll versucht werden, einige dieser Gründe aus dem Datenmaterial abzuleiten.

4.2 Abhängigkeit der Vorkommenshäufigkeit von anderen Faktoren

Weitere Ursachen für das Auftreten von Aussprachevarianten sind Variationen der Sprechgeschwindigkeit und dialektale Färbungen. Auch eine Abhängigkeit von vorangehenden Häsitationen oder Pausen, bzw. eine Abhängigkeit von der Stellung im Satz (erstes Wort, letztes Wort) ist denkbar. Um diese Abhängigkeiten zu ermitteln, wurden in mehreren Versuchen Entscheidungsbäume (*decision trees*, [6]) berechnet. Als Optimierungskriterium wurde die Entropie der Variantenverteilungsfunktion gewählt:

$$H_s = -\sum_{RegelX=1}^{RegelX=7} p(RegelX)\log(p(RegelX)) \quad (1)$$

mit

$$p(RegelX) = \frac{\text{nach Regel X generierte Worte in Gruppe s}}{\text{alle Worte in Gruppe s}} = \frac{N_{s,X}}{N_s} \quad (2)$$

H_s ist ein Maß für die Unsicherheit innerhalb einer Menge s von Worten. Je kleiner H_s ist, desto weniger zusätzliche Information ist erforderlich, um für ein Wort der Menge dessen Variante vorhersagen zu können. Wählt man als Optimierungskriterium das Minimum von $\sum N_s * H_s$, ist die optimale Gruppeneinteilung diejenige, bei der alle Nicht-Varianten in Gruppe 1, alle nach Regel 2 veränderten Worte in Gruppe 2 usw. eingeteilt werden. Für diesen Fall nimmt die Entropie den kleinsten möglichen Wert an.
Die Möglichkeiten einer Aufteilung der Gesamtmenge in Gruppen sind durch die möglichen Fragen begrenzt. Der komplette Fragensatz ist in Tabelle 4 aufgeführt.

Der Algorithmus zum Konstruieren eines *decision trees* geht von einem einzigen Knoten, dem Wurzelknoten, aus. Dieser enthält alle Worte des Korpus. Der Algorithmus wählt nun aus einem vordefinierten Fragensatz diejenige Frage aus, die zur minimalen Entropiesumme der zwei resultierenden Tochterknoten führt. Jeder Tochterknoten besteht dabei aus einer Menge von Worten; bei dem einen Tochterknoten ist die Frage mit 'ja' beantwortet worden, bei dem anderen mit 'nein'. Diese Prozedur wird für die Tochterknoten rekursiv durchgeführt, bis der resultierende Entropieverlust unter eine vordefinierte Schranke fällt. Dabei können die Fragen sowohl auf das Wort selbst (ist *dieses* Wort schnell gesprochen worden?) als auch auf die Nachbarn (ist *das nachfolgende/vorhergehende* Wort schnell gesprochen worden?) gerichtet werden.
Die Blätter des so entstandenen Baumes repräsentieren Äquivalenzklassen. Innerhalb dieser ist im Schnitt eine Klassifizierung einfacher als in der Grundmenge. Je näher am Wurzelknoten eine Frage verwendet wird, desto größer ist ihr Einfluß auf die Entropie. Man kann daher an einem fertig konstruierten *decision tree* zweierlei ablesen:

1. welche Fragen wichtig sind, d.h. welche der abgefragten Merkmale einen großen Einfluß auf das Entstehen von Aussprachevarianten

Nr.	Name der Frage	Bedeutung
1	first_word	erstes Wort der Äußerung?
2	last_word	letztes Wort der Äußerung?
3	pause	Pause?
4	hesitation	Häsitation?
5	Variant0	kanonische Variante?
6	Variant2	Variante nach Regel 2?
7	Variant3	Variante nach Regel 3?
8	Variant4	Variante nach Regel 4?
9	Variant5	Variante nach Regel 5?
10	Variant6	Variante nach Regel 6?
11	Variant7	Variante nach Regel 7?
12	breathing	Atemgeräusch?
13	very_slow	Sprechdauer T grösser als $\overline{T} + 2\sigma$?
14	slow	$\overline{T} + \sigma < T < \overline{T} + 2\sigma$?
15	average	$\overline{T} - \sigma < T < \overline{T} + \sigma$?
16	fast	$\overline{T} - \sigma > T > \overline{T} - 2\sigma$?
17	very_fast	$T < \overline{T} - 2\sigma$?
18	SpeedSlow	$T > \overline{T} + \sigma$?
19	SpeedHigh	$T < \overline{T} - \sigma$?
20	Kiel	Sprecher aus der Region Kiel?
21	Bonn	Sprecher aus der Region Bonn?
22	Karlsruhe	Sprecher aus der Region Karlsruhe?
23	Munchen	Sprecher aus der Region München?

Table 4: Verwendeter Fragensatz

haben

2. *welchen* Einfluß die Fragen haben, d.h. welche Merkmale zu welcher Verteilung von Aussprachevarianten führen.

4.3 Ergebnisse der Anwendung der Entscheidungsbäume

Für die oben bereits definierten Fragen ergab sich im Experiment mit 230.960 annotierten Worten aus dem VERBMOBIL-Korpus der in Bild 1 und in Tabelle 5 dargestellte Baum.

Zusammenfassend läßt sich sagen:

- Zu Beginn einer Äußerung ist die Wahrscheinlichkeit einer Aussprachevariante drastisch kleiner als an anderen Stellen. Die Wahrscheinlichkeit einer initialen Glottalverschlußelision ist sogar nahe Null. Allerdings werden initial 40% mehr Vokalverkürzungen beobachtet als im Mittel.
- Es ist ein deutlicher Einfluß der Dialektregion festzustellen. Die dialektale Färbung weist dabei ein Nord-Süd-Gefälle auf: während bei

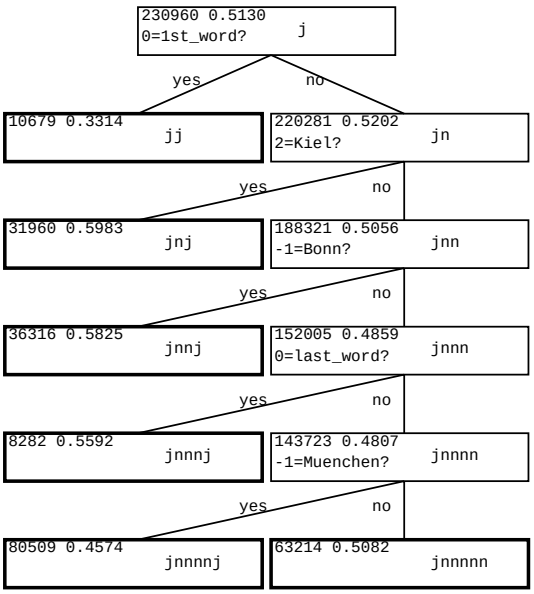

Figure 1: Baum bei Verwendung der Fragen 1-12 und 20-23

Sprechern der Region 'Kiel' die schwa-Elision 50% und die Nasalassimilation sogar 70% häufiger auftritt als im Mittel, ist in der Region 'Bonn' die schwa-Elision nur noch 35% und die Nasalassimilation 50% häufiger. In der Münchner Region schließlich liegt die Häufigkeit von schwa-Elision und Nasalassimilation 22% und 30% unter dem Durchschnitt. Bei den anderen Variantentypen ergibt sich kein einheitliches Bild. Lediglich die um 30% erhöhte Häufigkeit der Glottalverschlußelision in der Region 'Karlsruhe' fällt leicht aus dem Rahmen.

- Am Ende einer Äußerung zeichnet sich ein außergewöhnliches Muster ab. Die Wahrscheinlichkeit der - meist initialen - Glottalverschlußelision liegt 35% unter dem Durchschnitt. Dafür liegt die Häufigkeit der Vokalverkürzung um 40%, der Vokalqualitätänderung um 118% und die der r-Variantenbildung sogar um 300% über dem Durchschnitt.

Für den Fall der Erkennung von Sprache eines unbekannten Sprechers ist die Dialektregion oft nicht bekannt. Die entsprechende Frage läßt sich dann nicht beantworten. Es ist daher vorteilhaft, sich auf Fragen zu beschränken, die kein *a-priori*-Wissen über den Sprecher voraussetzen. Die Fragen 1 bis 20 aus Tabelle 4 erfüllen diese Forderung. Dabei ist zu beachten, daß die Fragen 5 bis 11 nicht auf das Wort selbst, sondern

Knoten	Anzahl Worte	Entropie	Anwendung von Regel Nr. (in %)						
			1	2	3	4	5	6	7
Wurzel	230960	0.5130	88.96	4.59	2.30	1.26	0.17	1.87	0.85
jj	10679	0.3314	93.64	0.62	1.66	0.80	0.09	2.58	0.61
jn	220281	0.5202	88.73	4.78	2.33	1.28	0.17	1.84	0.86
jnj	31960	0.5983	86.81	4.04	3.40	2.17	0.15	2.18	1.23
jnn	188321	0.5056	89.06	4.91	2.15	1.13	0.17	1.78	0.79
jnnj	36316	0.5825	87.05	4.84	3.12	1.84	0.14	2.04	0.97
jnnn	152005	0.4859	89.54	4.92	1.92	0.97	0.18	1.72	0.75
jnnnj	8282	0.5592	88.34	3.03	2.11	1.34	0.71	2.61	1.86
jnnnn	143723	0.4807	89.61	5.03	1.91	0.94	0.15	1.66	0.69
jnnnnj	80509	0.4574	90.42	4.29	1.81	0.88	0.20	1.68	0.73
jnnnnn	63214	0.5082	88.58	5.98	2.05	1.02	0.08	1.65	0.64

Table 5: Baum bei Verwendung der Fragen 1-12 und 20-23

nur auf die benachbarten Wörter angewendet werden dürfen, sonst führen sie trivialerweise zu einer optimalen Aufspaltung. Bei Verwendung dieses Fragensatzes erhält man den in Bild 2 und Tabelle 6 dargestellten Baum.

Knoten	Anzahl Worte	Entropie	Anwendung von Regel Nr. (in %)						
			1	2	3	4	5	6	7
Wurzel	230960	0.5130	88.96	4.59	2.30	1.26	0.17	1.87	0.85
jj	42512	0.7460	82.16	6.05	4.78	2.93	0.29	2.40	1.40
jn	188448	0.4530	90.50	4.26	1.75	0.89	0.14	1.75	0.72
jjj	41155	0.7315	82.60	6.11	4.51	2.73	0.29	2.44	1.32
jjn	1357	1.0609	68.75	4.35	12.9	8.77	0.15	1.25	3.83
jnj	148512	0.4930	89.36	4.80	2.04	1.03	0.13	1.87	0.75
jnn	39936	0.2868	94.70	2.23	0.64	0.34	0.17	1.29	0.63
jnjj	7133	0.2982	94.39	0.42	1.36	0.67	0.10	2.45	0.60
jnjn	141379	0.5008	89.11	5.02	2.08	1.05	0.17	1.85	0.76
jnjnj	6138	0.6164	86.74	3.36	2.22	1.65	0.65	3.08	2.31
jnjnn	135241	0.4944	89.22	5.10	2.07	1.02	0.11	1.79	0.69

Table 6: Baum bei Verwendung der Fragen 1-20

Zusammenfassend läßt sich hier sagen:

- Die Sprechgeschwindigkeit hat einen entscheidenden Einfluß auf die Häufigkeit des Auftretens von Aussprachevarianten. Bei hoher Sprechgeschwindigkeit ist die Gesamthäufigkeit von Varianten 50% höher als der Durchschnitt, bei niedriger Sprechgeschwindigkeit 50% kleiner.
- Bei hoher Sprechgeschwindigkeit verdoppelt sich die Wahrscheinlichkeit der schwa-Elision; die Häufigkeit der Nasalassimilation steigt um 130%. Alle anderen Variantentypen treten um ca. 50% vermehrt auf.

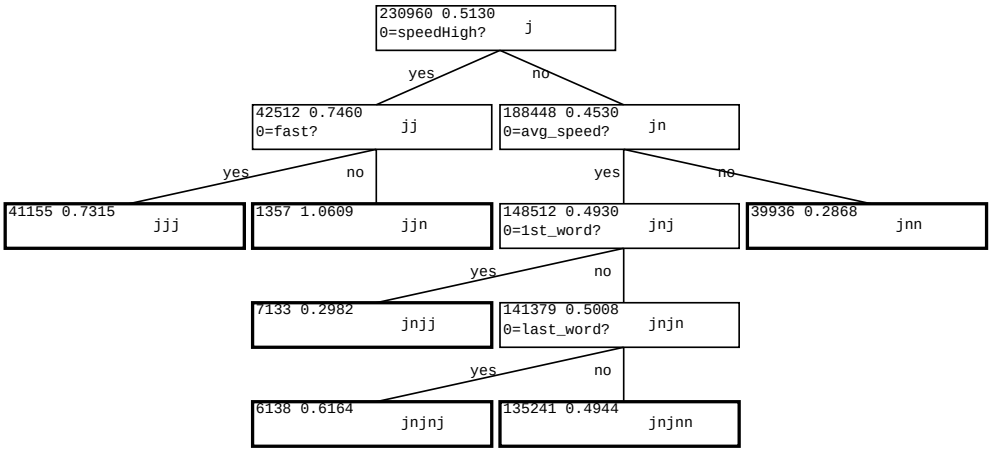

Figure 2: Baum bei Verwendung der Fragen 1-20

- Bei höchster Sprechgeschwindigkeit verfünffacht sich die Wahrscheinlichkeit der schwa-Elision, Nasalassimilationen treten sogar siebenfach häufiger auf als im Durchschnitt aller Worte. Auch Vokalqualitätverschiebungen treten 350% häufiger auf.
- Bei niedriger Sprechgeschwindigkeit sinken dementsprechend die Wahrscheinlichkeiten von schwa-Elision und Nasalassimilation auf unter 30% des Mittelwertes aller Worte. Glottalverschlußelisionen treten um 50% vermindert auf. Hingegen ist kein Einfluß auf die Häufigkeit von r-Varianten zu beobachten.

5 Einsatz im Spracherkenner

5.1 Einsatz beim Training des Spracherkenners

In der sog. Trainingsphase werden dem Spracherkenner Bandaufnahmen realer, gesprochener Sätze und die dazugehörigen Transliterationen zur Verfügung gestellt. Aus der Transliteration berechnet der Spracherkenner mit Hilfe des Wörterbuchs die Phonemsequenz. Diese wird zeitlich mit der Aufnahme ausgerichtet. Aus den Abschnitten des Sprachsignals 'lernt' das System die charakteristischen Eigenschaften der einzelnen Phoneme.
Stimmt nun die phonetische Umschrift im Wörterbuch nicht mit der realisierten Instanz eines Wortes überein, ist die zeitliche Ausrichtung mit der Äußerung zwangsläufig falsch. Ein fehlerhaftes Lernen der Phoneme ist die Folge. Daraus resultiert eine Minderung der Trennschärfe der Phonemmodelle untereinander. Dieser Effekt kann durchaus posi-

tive Folgen haben, nämlich dann, wenn auch bei der eigentlichen Spracherkennung dieselbe Art der Aussprachevariante realisiert wird. In diesem Fall 'paßt' das falsch trainierte Phonem besser zum Sprachsignal, als es ein korrekt trainiertes Phonem würde. Erst wenn bei der Erkennung die Aussprachevarianten explizit im Wörterbuch modelliert sind, können die Vorzüge schärferer Phonemmodellierung genutzt werden.

Um den Einfluß der Aussprachevarianten im Wörterbuch zu bestimmen, wurde ein Spracherkenner zum einen nur mit den kanonischen Varianten und zum anderen mit allen generierten Aussprachevarianten im Wörterbuch trainiert. Dabei kann der Erkenner bei der Berechnung der besten Zeitzuordnungen (*forced alignment*) die am besten passende Variante wählen. Im Schnitt lagen die akustischen Scores bei der Verwendung von Varianten um 0.15% besser. Der Rechenaufwand und damit die Trainingszeit werden jedoch verlängert.

5.2 Einsatz beim Test des Spracherkenners

In der Testphase, der eigentlichen Anwendung eines Spracherkenners, muß das System Sprachaufnahmen unbekannten Inhalts von unbekannten Sprechern analysieren und erkennen.

Als Testsuite wurden die 265 Sätze des 'kurzen' Teilsets der Verbmobil-Akustik-Evaluation 1995 verwendet [2]. Die Wortakkuratheit des verwendeten Systems, das nur auf einer Teilmenge des verfügbaren Materials trainiert war, lag initial bei 67.8%. Davon ausgehend wurden zwei neue Systeme auf dem gesamten Material trainiert, wovon das eine nur die kanonischen plus knapp 200 von Hand hinzugefügte Wortformen, das andere sämtliche automatisch generierte Wortformen zur Verfügung hatte. Beide Systeme wurden dann jeweils mit beiden Wörterbüchern getestet. Die erzielten Erkennungsleistungen zeigt Tabelle 7.

Varianten im Training?	Varianten im Test ?	% korrekt	Wortakkuratheit
nein	nein	71.4%	68.1%
nein	ja	72.3%	68.5%
ja	nein	70.8%	67.7%
ja	ja	72.5%	68.9%

Table 7: Erkennungsergebnisse

Man erkennt eine leichte Steigerung der Wortakkuratheit bei Verwendung automatisch generierter Aussprachevarianten im Testwörterbuch. Diese Steigerung tritt auch dann ein, wenn im Training keine Aussprachevarianten verwendet wurden.
Verwendet man im Training Aussprachevarianten, verzichtet jedoch beim Test darauf, führt die erhöhte Trennschärfe der trainierten Phonemmodelle und die schlechte Modellierung der phonetischen Realisierung im Test zu einer Reduktion der Erkennungsleistung.

6 Zusammenfassung

In dieser Arbeit wurde das Auftreten von sechs Klassen von Aussprachevarianten in einem sehr großen Korpus spontaner Sprache untersucht. Dabei ergab sich eine starke Abhängigkeit der Realisierungshäufigkeit von der Sprechgeschwindigkeit, der Dialektregion des Sprechers und von der Position des Wortes in der Äußerung.
Die Aussprachevarianten wurden in die Modellparameter eines Spracherkenners aufgenommen. Dabei konnte gezeigt werden, daß dadurch im Training eine Verbesserung der Trennschärfe der Phonemmodelle untereinander erreicht werden konnte. Die Verwendung der Varianten in der Erkennung verringerte die Fehlerrate des Spracherkenners um 1,5% relativ.

7 Danksagung

Diese Untersuchungen wurden vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des VERBMOBIL-Projekts gefördert.
Das JANUS-Projekt wurde teilweise durch die Advanced Research Project Agency und das amerikanische Department of Defense gefördert.
Mein Dank gilt insbesondere Frau Gudrun Flach für Hilfe bei der Zusammenstellung der Regeln und Herrn Prof. Alexander Waibel für hilfreiche Hinweise.

References

[1] T. Sloboda, *Dictionary learning: performance through consistency*, Proc. ICASSP 95, pp 453 ff.

[2] E. Paulus, M. Lehning, *Die Evaluierung von Spracherkennungssystemen in Deutschland*, Verbmobil Report 70/95, Juli 1995; auch im Tagungsband *Fortschritte der Akustik*, DAGA 1994, Dresden 1994, S. 147 ff.

[3] R. Haeb-Umbach, P. Beyerlein, E. Thelen, *Automatic transcription of unknown words in a speech recognition system*, in Proc. ICASSP 95, pp. 840 ff.

[4] Gudrun Flach, *Beschreibung von Aussprachevarianten*, Tagungsband 'Elektronische Sprachsignalverarbeitung', Berlin, 1994, ISSN 0940-6832 (Heft 11)

[5] K. Kohler, M. Pätzold, A. Simpson, *Handbuch zur Segmentation und Etikettierung von Spontansprache 2.3*, IPDS Kiel, Verbmobil Technisches Dokument Nr. 16, Dezember 1994

[6] Mei-Yuh Hwang, *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, Ph.D. thesis, CMU-CS-93-230, Carnegie Mellon University, Pittsburgh, PA 15213, December 1993